Presentation and Poster Abstracts

Day 1: Tuesday, 8 December 2015
Science Drivers Project Requirements and Feedback

Title and Presenter	Abstract
WGCM Infrastructure Panel *Karl Taylor (DOE/LLNL, taylor13@llnl.gov)* *V. Balaji (NOAA/GFDL, balaji@princeton.edu)*	The Working Group on Coupled Modeling (WGCM) Infrastructure Panel (WIP) was formed in response to the WGCM's (2013) expressed need to provide scientific guidance and requirements for the global data infrastructure underpinning global climate science and modeling. This infrastructure includes ESGF software, and other tools such as: ES-DOC, CoG, CMOR, CF Conventions, and others. Chaired by V. Balaji (Princeton/GFDL) and K. Taylor (DOE/LLNL/PCMDI), it outlined in 2014 a strategy to develop a series of "position papers" on global data infrastructure and its interaction with the scientific design of experiments; and to present them to the WGCM annual meeting for endorsement by the WGCM, the CMIP Panel, and the modeling groups. A series of position papers were unveiled at the WGCM-19 meeting (2015) in Dubrovnik. The 11 position papers currently in draft, and others in progress, will be available on the WIP website, https://goo.gl/eJxDvL.
DOE Accelerated Climate Modeling for Energy *David Bader (DOE/LLNL, bader2@llnl.gov)* *Dean N. Williams (DOE/LLNL, Williams13@llnl.gov)* *Valentine Anantharaj (DOE/ORNL, anantharajvg@ornl.gov)*	Sponsored by the U.S. Department of Energy's (DOE's) Office of Biological and Environmental Research (BER), the Accelerated Climate Modeling for Energy Project is an ongoing, state-of-the-science Earth system modeling, simulation, and prediction project that optimizes the use of DOE laboratory resources to meet the science needs of the nation and the mission needs of DOE. "A DOE Model on DOE Machines for the DOE Mission" ACME's initial scientific goals address three areas of importance to both climate research and society: 1. Water cycle: How do the hydrological cycle and water resources interact with the climate system on local to global scales? 2. Biogeochemistry: How do biogeochemical cycles interact with global climate change? 3. Cryosphere-ocean system: How do rapid changes in cryosphere-ocean systems interact with the climate system? The high-resolution version of the ACME model simulates the fully coupled climate system at high-resolution (15-25km), and further development is needed to optimize performance on current and future DOE Leadership Class computers. New scalable and extensible solutions for data archival, search, retrieval and analysis are needed for the size and complexity of the ACME output.
Obs4MIPs *Peter Gleckler (DOE/LLNL, gleckler1@llnl.gov)*	
IS-ENES *Sébastien (ENES/IPSL, sebastien.denvil@ipsl.jussieu.fr)*	European climate modelling groups joined together in 2001 to create the European Network for Earth System Modelling (ENES) with the objectives of helping the development and evaluation of climate models of the Earth system, encouraging the exchange of software and model results and promoting the development of high-performance computing facilities. The EU funded project IS-ENES, Infrastructure for ENES (1rst phase 2009-2013, 2nd phase 2013-2017), aims to promote the development of a common distributed climate 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 issue (http://is.enes.org). IS-ENES supports the integration of the European climate modeling community and recently issued the "*Infrastructure strategy for the European Earth System modeling Community: 2012-2022*". It promotes the dissemination of European climate model results from the international WCRP CMIP5 and CORDEX experiments developed in preparation of the IPCC 5th Assessment Report by supporting ESGF developments and operations. IS-ENES also aims at enhancing model development and software sharing and supports the preparation of high-end simulations and the use of high-performance computing.
CREATE-IP *Jerry Potter (NASA/GSFC, jpotter@ucdavis.edu)*	The Climate Model Data Services (CDS) at NASA's GSFC is collaborating with the world's major reanalysis projects to collect reanalysis data and present it through the Distribution, Visualization, Analytics, and Knowledge Services, resulting in the Collaborative REAnalysis Technical Environment-Intercomparison Project (CREATE-IP). CDS has converted monthly mean data from the five major reanalysis projects, including MERRA-2, to the standard Earth System Grid Federation (ESGF) format of one variable per file and published the data in ESGF. The agreement or disagreement among reanalyses enables us to judge the scientific validity of using reanalysis data to evaluate climate models. The differences in the reanalyses may have a variety of causes, including, but not limited to, differences in the input observations, changes in observation instrumentation, or differences in the model physical parameterizations. The data is prepared to the CMIP5 and obs4MIPs specifications using CMOR and is distributed through the ESGF COG interface. So far, we have prepared monthly averages of the primary variables produced by the CMIP5 simulations and 6h frequency of an initial selected set of variables. Additional 6h variables preparation is in process. These initial variables were

	select because they are useful in evaluating weather events in the past for intercomparison among the different reanalyses. Each reanalysis center produces different data structure and organization posing difficulties in preparing the reanalyses for inclusion into ESGF. As a result, each data set requires custom processing. Reanalyses are produced at high horizontal and vertical resolution, and the 6 hour data conversion processing has proven to be particularly challenging requiring several days of dedicated uninterrupted computing to complete one variable. To assist with testing of the processed data, UV-CDAT has proven to be particularly useful for data quality control because of its inherent ease of use and flexibility. In addition to distributing reanalysis data through ESGF, we have implemented a visualization tool, CREATE-V, based on code from the National Center for Atmospheric Research's Climate Inspector. This tool utilizes TDS and OpenLayers to support access by interdisciplinary and reanalysis scientists for the exploration and side-by-side comparison of variables by reanalysis, date, and level.

Day 1: Tuesday, 8 December 2015
Required Data Center and Interoperable Services

Title and Presenter	Abstract
ANU/NCI ***Ben Evan (ANU/NCI, Ben.Evans@anu.edu.au)***	
DDC/DKRZ ***Stephan Kindermann (ENES/DKRZ, kindermann@dkrz.de)***	The DKRZ supports the complete data life cycle of climate data products – model data generation, postprocessing, data ingest, quality assurance, ESGF publication, long term archival and assignment of PIDs, DOIs as well as early citation information. To support end users DKRZ acts also as a replication and long term archival center of external ESGF data products, hosting the world data center for climate (WDCC) and acting as an IPCC data distribution center. Recently the end user requirements got stronger to provide a platform to analyse the huge ESGF data volume hosted at DKRZ. As the (after shutdown) DKRZ ESGF services will be hosted by the same computing platform as the HPC part, such a data near processing service can be efficiently developed (ESGF related data products are hosted as part of the overall ~50PByte Lustre file system). In parallel to this the DKRZ established a data cloud service to support end users in hosting project data collections and to support DKRZ data in-and export. A virtual machine environment allows for the flexible deployment of e.g. project specific data servers and services. Also first investigations to provide docker based compute services in the future were performed.
DOE LLNL/PCMDI ***Dean N. Williams (DOE/LLNL, williams13@llnl.gov)***	LLNL researchers benefit from an institutional IT infrastructure that provides desktop support and experts in server technologies. The latter includes virtualization expertise that has been applied to provide multiple operating systems on shared resources and to create a wide variety of virtual machines to leverage resources across LLNL. An enterprise team also provides networking service to implement the Science demilitarized zone (DMZ) and Data Transfer Nodes (DTNs). Connections into LLNL include ESnet, the dynamic Science Data Network, the ALICE grid system, The Open Science Grid, and a wide variety of programmatic networks. Cisco telepresence nodes are also available to facilitate remote collaboration over these networks. In addition to local group resources, LLNL computational scientists deliver a balanced High Performance Computing (HPC) environment with constantly evolving hardware resources and a wealth of HPC expertise in porting, running, and tuning real-world, large-scale applications, and data management systems, such as ESGF. Currently, LLNL delivers multiple petaflops of compute power, massive shared parallel file systems, powerful data analysis/cluster platforms, and archival storage capable of storing many petabytes of data. This balanced hardware environment supports key collaborations between LLNL applications developers and community experts on the creation, debugging, production use, and performance monitoring of large-scale parallel applications, as well as data analysis in a wide variety of scientific climate application, such as CMIP and ACME. All members of the LLNL Climate projects (i.e., PCMDI, ACME) have desktop workstations available to them. In addition, the LLNL Climate Science Program maintains 12 shared computer servers and two internal file servers with over 250 TBytes of aggregate storage to support its internal research activities. Data management software (i.e., ESGF) and Analysis software (i.e., UV-CDAT) are maintained on these shared systems. The file servers enable seamless integration among the workstation, computer server, and data resources using NFS remote mounting capabilities. The Green Data Oasis (GDO) and Climate Central Systems host data (e.g., hosting the CMIP3 and CMIP5 archives) served to the external community. Additional archival capacity is required for CMIP6.
IS-ENES/IPSL ***Sébastien Denvil (ENES/IPSL, sebastien.denvil@ipsl.jussieu.fr)***	The Institut Pierre Simon Laplace (IPSL) climate modeling group gathers together climate modeling teams from the CNRS (Centre Nationale de la Recherche Scientifique), the CEA (Commissariat à l'Énergie Atomique et aux Énergies Alternatives), and from various university research disciplines: meteorology, oceanography, bio-geochemistry ...etc. The group's objective is the study of natural and anthropogenic variability in the global climate system. IPSL is also studying climate change impacts and usage of climate projections for adaptation to climate change related to industry. IPSL is one of the climate modeling centre of international repute contributing to the IPCC

	(Intergovernmental Panel on Climate Change). The IPSL draws upon a team of 50 engineers and informatics experts. Their collaborations within France and internationally, the diversity of the technologies they exploit, the size and variety of the projects that they handle, are a reflection of the IPSL's desire to be at the cutting edge of climate modeling. During this talk we will present our strategies and propositions on required data center and interoperable services.
IS-ENES/CEDA ***Phil Kershaw (ENES/CEDA, philip.kershaw@stfc.ac.uk)***	CEDA, the Centre for Environmental Data Analysis, hosts data centres managing climate and earth observation data on behalf for the UK environmental science community to facilitate access and support its work in collaborations with international partners. CEDA is underpinned by JASMIN, a petascale storage and cloud computing facility. This, besides hosting the CEDA data archive, provides communities of users with a collaborative environment for analysis of data including group workspaces and hosted processing capability. There are a number of challenges moving forward driven in part by the success of JASMIN to date, and by the increasing data volumes for both model data and observations. Technically these can summarised in terms of the ability to scale computing resources and the effective integration of new and existing technologies to provide services needed to best meet the needs of the user community. Currently the archive holds around 3PB maintained on spinning disk with a full tape back up. This holds overs 250 datasets and in access of 200 million files. Two key programmes in particular, CMIP6 and the data stream from the new generation of ESA earth observation satellites, the Sentinels present challenges both in terms of the data volumes (~10PB reserved for CEDA CMIP6 archive) and velocity (expected 10 TB/day rate for Sentinel datasets). Data from these sources will exceed the disk capacity available to the archive in the near future and will necessitate the development of an integrated solution for disk and nearline tape storage. Work is underway to pilot new technologies associated with the cloud service including the use of containers, orchestration tools and object stores. These will enable the scale of resources to meet demand and provides to federate with other cloud providers be they public or from the research community. For ESGF, there is a need for a robust and stable core service, for the projects we host through the infrastructure. These include SPECS, CCMI, CLIPC and ESA CCI. For CCI, the ESA Climate Change Initiative, CEDA has started a project for ESA over the past year to build an Open Data Portal to serve data products from the programme. This is re-using and building upon technology from ESGF including the Index and Data nodes and will include innovations including support for ISO 19115 search services and the use of Semantic Web technology to develop a machine readable DRS vocabulary and govern its use with client applications.

Day 2: Wednesday, 9 December 2015
Advanced Computational Environments and Data Analytics

Title and Presenter	Abstract
Compute Working Team Overview ***Daniel Duffy (NASA/GSFC, daniel.q.duffy@nasa.gov)***	
WPS Overview and Demo ***Charles Doutriaux (DOE/LLNL, doutriaux1@llnl.gov)***	
The Climate Data Analytic Services (CDAS) Framework **Thomas Maxwell (NASA/GSFC, thomas.maxwell@nasa.gov)** **Mark McInerney (NASA/GSFC, mark.mcinerney@nasa.gov)** ***Daniel Duffy (NASA/GSFC, daniel.q.duffy@nasa.gov)*** ***Jerry Potter (NASA/GSFC, jpotter@ucdavis.edu)*** ***Charles Doutriaux (DOE/LLNL, doutriaux1@llnl.gov)***	Faced with unprecedented growth in the Big Data domain of climate science, NASA has developed the Climate Data Analytic Services (CDAS) framework. This framework enables scientists to execute trusted and tested analysis operations in a high performance environment close to the massive data stores at NASA. The data is accessed in standard (NetCDF, HDF, etc.) formats in a POSIX file system and processed using trusted climate data analysis tools (ESMF, CDAT, NCO, etc.). The framework is structured as a set of interacting modules allowing maximal flexibility in deployment choices. CDAS services are accessed via a WPS API being developed in collaboration with the ESGF Compute Working Team to support server-side analytics for ESGF. The API can be executed using either direct web service calls, a python script or application, or a javascript-based web application. Client packages in python or javascript contain everything needed to make CDAS requests. The CDAS architecture brings together the tools, data storage, and high-performance computing required for timely analysis of large-scale data sets, where the data resides, to ultimately produce societal benefits. It is is currently deployed at NASA in support of the Collaborative REAnalysis Technical Environment (CREATE) project, which centralizes numerous global reanalysis datasets onto a single advanced data analytics platform. This service permits decision makers to investigate climate changes around the globe, inspect model trends, compare multiple reanalysis datasets, and variability.
Ophedia	The Ophidia project is a research effort on big data analytics facing scientific data analysis challenges in multiple domains (e.g. climate change). Ophidia provides declarative, server-side, and parallel data analysis,

Sandro Fiore (ENES/CMCC, sandro.fiore@unisalento.it)	jointly with an internal storage model able to efficiently deal with multidimensional data and a hierarchical data organization to manage large data volumes ("*datacubes*"). The project relies on a strong background on high performance database management and OLAP systems to manage large scientific datasets. The Ophidia analytics platform provides several *data operators* to manipulate *datacubes*, and *array-based primitives* to perform data analysis on large scientific data arrays (e.g. statistical analysis, FFT, DWT, subsetting, compression). Metadata management support (CRUD-like operators) is also provided. The server front-end exposes several interfaces to address interoperability requirements: WS-I$^+$, GSI/VOMS and OGC-WPS (through PyWPS). From a programmatic point of view a Python module (PyOphidia) makes straightforward the integration of Ophidia into Python-based environments and applications (e.g. IPython).The system offers a Command Line Interface (e.g. bash-like) for end-users, with a complete set of commands, as well as integrated help and manuals. A key point of the talk will be the workflow capabilities offered by Ophidia. In this regard, the framework stack includes an internal workflow management system, which coordinates, orchestrates, and optimizes the execution of multiple scientific data analytics & visualization tasks (e.g. statistical analysis, metadata management, virtual file system tasks, maps generation, import/export of datasets in NetCDF format). Specific macros are also available to implement loops, or to parallelize them in case of data independence. Real-time workflow monitoring execution is also supported through a graphical user interface. Some real workflows implemented at CMCC and related to different projects will be also presented: climate indicators in the FP7 EU CLIPC and EUBRAZILCC, fire danger prevention analysis in the INTERREG OFIDIA, and finally, large scale climate model intercomparison data analysis (e.g. precipitation trend analysis, climate change signal analysis, anomalies analysis) in the H2020 INDIGO-DataCloud.
WPS Service and Back-end **Maarten Plieger (ENES/KNMI, maarten.plieger@knmi.nl)**	

Day 2: Wednesday, 9 December 2015
ESGF Development for Data Centers and Interoperable Services

Title and Presenter	Abstract
CoG User Interface Working Team ***Sylvia Murphy (NOAA/ERSL, sylvia.murphy@noaa.gov)***	Throughout 2015, the ESGF User Interface Working Team (UIWT) has worked on upgrading and expanding the Earth System CoG Collaboration Environment to replace the old ESGF web front-end. Major new features include: a) integration of CoG into the ESGF software stack; b) federation of distributed CoGs; c) support for downloading data via Globus; and d) general improvements to the site, infrastructure upgrades, and security fixes. CoG is now ready to be deployed as the ESGF front-end at each node. For the next 6 months, the priority of the ESGF UIWT will be be supporting ESGF administrators and end-users, while at the same time collecting and prioritizing requirements for additional needed functionality.
Dashboard Working Team ***Sandro Fiore (ENES/CMCC, sandro.fiore@unisalento.it) Paola Nassisi (ENES/CMCC, paola.nassisi@cmcc.it) Giovanni Aloisio (ENES/CMCC, giovanni.aloisio@unisalento.it)***	Monitoring the Earth System Grid Federation is a challenging topic. From an infrastructural standpoint the dashboard & desktop components provide the proper environment for capturing usage metrics, as well as system status information at local (node) and global (institution and/or federation) level. The Dashboard and the Desktop are strongly coupled and integrated into the ESGF stack and represent the back- and the front-end of the ESGF monitoring system. The Dashboard acts as information provider, collecting and storing a high volume of heterogeneous metrics, covering machine performance, network topology, host/service mapping and registered users as well as download statistics. The Desktop is a web-based environment and provides an effective, transparent, robust and easy access to all the metrics and statistics provided by the Dashboard. It is written in Java and JavaScript programming languages and presents enhanced views with several gadgets (enriched with charts, tables and maps) for a simple and user-friendly visualization of aggregated and geo-localized information. All the metrics collected by the ESGF monitoring infrastructure are stored in a system catalog that has been extended to support multiple information about the data usage statistics. More specifically, in addition to information like the number of downloads, downloaded datasets, users that have downloaded some data, the amount of data downloaded etc., new metrics are being provided. Some examples are statistics about data downloads grouped by model, variable or experiment, by country or over time, top ten list of the most downloaded datasets or clients distribution maps. To this end, specific data marts have been created to allow a fast access to this information. Finally, to grant a programmatic access to the metrics managed by the Dashboard, a set of RESTful APIs has been defined (based on a JSON data interchange format) allowing the user to design and implement his/her own client applications.
Data Transfer Working Team ***Luckasz Lacinski (DOE/ANL, lukasz@uchicago.edu)***	The past years the focuses have been on updating to latest software and use the new user interface. The key work completed include a) supporting Globus download option with the latest ESGF user interface COG b) updating components on the data node to a recent and supported version of servers and (c) simplifying the installation process and script. The new release includes all of these, and this talk will present details on this work done and impact for ESGF users and administrators.
Identity Entitlement Access Team	The remit of the IdEA team is to maintain and develop ESGF's system for access control to resources hosted within the federation. Over the past few years the scope of this work has grown from the original requirement for securing of access to CMIP5 data, to that for other projects in the federation. In addition, with the

Philip Kershaw (ENES/BADC, philip.kershaw@stfc.ac.uk) *Rachana Ananthakrishnan (DOE/ANL, ranantha@uchicago.edu)*	development of compute capability for ESGF, securing of access to computing resources is an increasingly important aspect for consideration in the evolution of the system. Activities over the last year have been dominated by the security incident with the federation. Nevertheless, some promising work has been undertaken piloting new capability for user delegation using the OAuth 2.0 protocol. User delegation is an important capability to support remote computation of secured resources. Initial integration work has been undertaken between CEDA's OAuth 2.0 service and IS-ENES partners KNMI and DKRZ. This will soon be extended to work with Globus transfer. These activities are providing confirmation of the potential for OAuth to simplify access and to provide a common baseline to a number of different access control use cases. We outline the roadmap and resources needed to take this work into a production service for the federation. Besides the technical development of the system, there are considerations with respect to policy and operation of Identity Providers in the federation. We will set out recommendations for the future to simplify access for users, enhance security and reduce the operational burden.
Installation Working Team *Prashanth Dwarakanath (ENES/Liu, pchengi@nsc.liu.se)* *Nicolas Carenton (ENES/IPSL, ncarenton@ipsl.jussieu.fr)*	ESGF installation working team was created in March 2014. Its main responsibilities are ESGF releases management, installation tool maintenance as well as node administrators' support. 2015 saw many key deliverables being met, which included providing an automated installation mechanism for ESGF, and switching to Apache web server as the frontend, providing support for non-java server-side components. The security incident announced in June of 2015 was the biggest challenge that was encountered by the IWT, and was successfully handled; it however meant additional coordination with developers and extra develop-test-deploy cycles. We will present here the work done since the last F2F and also highlight features of the major releases till date, and upcoming work on the installer.
International Climate Network Working Group *Eli Dart (DOE/ESnet, dart@es.net)* *Mary Hester (DOE/ESnet, mchester@es.net)*	This talk will describe the ICNWG work in 2015, and progress made to date. Next steps for the working group will also be discussed.
Metadata and search Working Team *Luca Cinquini (NASA/JPL, Luca.Cinquini@jpl.nasa.gov)*	For the ESGF Metadata and Search Working Team (ESGF-MSWT), the year 2015 was largely dominated by the general ESGF security incident, which prompted the whole federation to be brought offline. The ESGF-MSWT took advantage of this unfortunate situation to execute a much needed upgrade of the ESGF search services infrastructure, which would have been much more difficult as a backward-compatible upgrade. As a consequence, the upcoming ESGF 2.0 software stack will utilize Solr 5, deployed as a standalone engine embedded within Jetty, which includes many important new features such as atomic updates. The general master/slave/replica architecture hasn't changed, but the Solr slave shard will be exposed through the standard HTTP port 80 to avoid pesky firewall issues. Additional, support for publishing data to a new "local shard" has been introduced. From the User Interface perspective, many improvements have been added to the search pages, the administrator configuration utilities, and the data cart. In the next year, the ESGF-MSWT main focus will be to support the upcoming CMIP6 distributed data archive, and related observational data. Major areas of development will include metadata validation, partition of the global search space into virtual organizations, scalability and performance.
Node Manager Working Team *Sasha Ames (DOE/LLNL, ames4@llnl.gov)* *Prashanth Dwarakanath (ENES/Liu, pchengi@nsc.liu.se)* *Sandro Fiore (ENES/CMCC, sandro.fiore@unisalento.it)*	ESGF nodes of all flavors require a Node Manager component to coordinate automated configuration and federation-wide monitoring activities. To improve scalability over the prior P2P-based node manager, we are implementing a two-tier system that combines aspects of P2P to coordinate the "super-nodes" with client-server to handle the secondary tier of member nodes. We are transitioning to use a Python based implementation that can run under Apache. Development of this component is approaching readiness to test in a test-federation environment as we shore up more of the functionality. Additionally, this talk will incorporate plans for a "Tracking and Feedback" effort for ESGF.
Persistent Identifier Services *Tobias Weigel (ENES/DKRZ, weigel@dkrz.de)* *Stephan Kindermann (ENES/DKRZ, kindermann@dkrz.de)* *Katharina Berger (berger@dkrz.de)*	Persistent Identifier (PID) services for ESGF are concerned with the automated assignment and curation of persistent identifiers for CMIP6 data managed in ESGF at several levels of granularity. PIDs will be assigned to all CMIP6 files as well as several higher levels of aggregation, covering datasets, simulations and models. Identifier names are generated by CMOR and registered as part of the overall publishing workflow. An exemplary application based on the PID service is a smart user workspace tool that can pull additional information on given files from the federation, tell whether a new dataset version is available and ultimately provide access to it. The presentation will give an overview on the service design as also described in the corresponding WIP paper and provide an update on the current development status. The service architecture is based on a distributed message queue to achieve high availability and throughput. The PID services interact with other ESGF components, including versioning, replication and citation services.
Provenance Working Team *Bibi Raju (DOE/PNNL bibi.raju@pnnl.gov)*	Provenance team aims to focus on the development of provenance solutions in support of reproducibility and performance investigations to accomplish the Accelerated Climate Modeling for Energy (ACME) computational goals. This includes development of a provenance format that can capture sufficient information to enable scientists to reproduce their previous calculations correctly as well as capture and link to performance information for specific workflows and model runs to enable in-depth performance analysis. The first step is to investigate methods for the capture, representation and storage, evaluation, access and use of provenance information. Over the last year, we have been developing a comprehensive workflow performance data model called Open Provenance Model-based Workflow Performance Provenance (OPM-WFPP). It enables the

	structured analysis of workflow performance characteristics and variability. It also links provenance information and performance metrics ontology. The provenance capture ontology and system enables the capture of provenance from the high-level workflow through all relevant system levels in one integrated environment. A provenance production and collection framework is in place called Provenance Environment (ProvEn). It provides components supporting the production and collection of provenance information for distributed application environments. Semantic Web technologies and ontologies, including the Open Provenance Model – Workflow Performance Provenance (OPM-WFPP) ontology, are used by ProvEn for the representation, storage, and reporting of provenance. We are currently in the process of developing a provenance capture mechanism that can handle the high-velocity provenance information.
Publication Working Team ***Sasha Ames (DOE/LLNL, ames4@llnl.gov)*** ***Rachana Ananthakrishnan (DOE/ANL, ranantha@uchicago.edu)***	The Publication Working team has concerns within the esg-publisher software, the development of a publications service, and the management of overarching workflows for publication, which include all steps required in preparation for ESGF publication. Accomplishments within the 2015 ESGF year included an initial release of a GUI-based publication service running for ACME. ESGF 2.0 contains a handful of changes to the publisher software including support of upgraded components, improved versioning support, and optional facets support for published data sets. Future work will have a strong focus on workflow and software requirements for CMIP6, and in addition the release of a publication service API that incorporates Globus Transfer of data.
Quality Control Working Team ***Martina Stockhause (ENES/DKRZ, stockhause@dkrz.de)*** ***Guillaume Levavasseur (ENES/IPSL, glipsl@ipsl.jussieu.fr)*** ***Katharina Berger (ENES/DKRZ, berger@dkrz.de)***	The ESGF-QCWT aims to improve the quality of ESGF user services by integration of additional external documentations. The team coordinates the implementation of the errata service (IPSL) and the data citation service (DKRZ). We will present the team's progress over the last 12 months and give a roadmap for the next year with special emphasis on requirements, collaboration and risk aspects.
Replication and Versioning Working Team ***Stephan Kindermann (ENES/DKRZ, kindermann@dkrz.de)*** ***Tobias Weigel (ENES/DKRZ, weigel@dkrz.de)***	Ensuring ESGF CMIP6 data consistency across sites strongly depends on stable and agreed versioning and replication procedures. On one hand this requires common software components (versioning support as part of publication procudure and replication software like synda) - yet on the other hand operational agreements and the adherence to "versioning, replication and publication best practices" is necessary. The presentation will describe the currents status of the software as well as agreements aspects. As part of this also a short summary of the "replication and versioning" WIP paper is given. A roadmap for 2016 will be discussed highlighting open issues to be resoved. The collaboration aspects with the icnwg team and the publication team are summarized as well as future versioning and replication support aspects of the proposed persistent identifier ESGF
Software Security Working Team ***Prashanth Dwarakanath (ENES/LIU, pchengi@nsc.liu.se)***	
Support Working Team ***Matthew Harris (DOE/LLNL, harris112@llnl.gov)***	Last years part one presentation we covered all the misdoings of the technically features of the support working Team. This year, with joy, we will cover the new, replaced, and even removed tools for giving our users a better experience. Topics will cover from FAQ, wikis, sites, and again mail archiving. There have been mass improvements to the support process, though there is continued room for improvement with everyone's help.
User Working Team ***Torsten Rathmann (ENES/DKRZ, rathmann@dkrz.de)***	Besides operational support Torsten has been working on a statistics of user questions via esgf-user@lists.llnl.gov and the former Askbot. December 2013 - September 2015 we got 1133 requests (spam not included, ~3 Askbot questions lost). From the statistics a list of issues shall be distilled, for example concerning registration+login, search, Globus Connect and malfunctioning servers.

Day 3: Thursday, 10 December 2015
Coordinated Efforts with Community Software Projects

Title and Presenter	Abstract
THREDDS Data Server (TDS) ***John Caron (Independent, jcaron1129@gmail.com)***	The THREDDS Data Server (TDS) has had significant development since first adoption by the ESGF. The latest version (5.0) can now scale to the tens of thousands of catalogs and millions of datasets used by ESGF nodes. This talk will cover these improvements and others of interest to the ESGF community.
Science DMZ for ESGF Super Nodes	This talk will describe the Science DMZ model, and its application to large (Super Node) ESGF deployments. In addition, the talk will discuss next-generation portal architectures, and their potential applications in the ESGF.

Eli Dart (DOE/ESnet, dart@es.net)	
Named Data Networking (NDN) ***Christos Papadopoulos (Colorado State, christos@colostate.edu)***	The current Internet names the hosts, leaving it to the application to locate the host with the desired data. However, with the emergence of technologies such as Content Delivery Networks (CDN) and the cloud, and trends such as mobility and Internet of Things (IoT), the need to associate data with an Internet Protocol (IP) address has become a hindrance. This misalignment requires enormous corrective effort at the expense of application complexity and robust security. In this talk will present NDN and some illustrative applications of Named Data Networking (NDN) in the Climate and High Energy Physics (HEP) communities.
Designing Climate Model Output Rewriter version 3 (CMOR3) for Coupled Model Intercomparison Project, Phase 6 (CMIP6) ***Denis Nadeau (DOE/LLNL, nadeau1@llnl.gov)***	A lot of lessons have been learned during the Couple Model Intercomparison Project Phase 5 and a more flexible version of CMOR has became necessary to handle state-of-the-art Model Intercomparison Projects (MIPs). In order to accomplish value delivery in the rapid model developments and growth in climate science, flexibility, adaptability, scalability and robustness are necessary to keep pace with these rapid changes. CMOR is currently being enhanced to line up with continuously growing CMIP6 requirements. Delineating new input tables structure, which empower each model to maintain value delivery into CMIP6, is doing adapting to those requirements. As well, customized global attributes are being designed to accommodate growth in capability needed by new MIPs. Finally, possibility of parallelization of CMOR is also being regarded as improvement since model outputs continuously grow in spatial and temporal resolution.
Synda ***Sébastien Denvil (ENES/IPSL, sebastien.denvil@ipsl.jussieu.fr)***	Synda is a command-line alternative to the ESGF web front-end. Current main features are listed below. • Simple data installation using an apt-get like command • Support every ESGF project (CMIP5, CORDEX, SPECS, ...) • Parallel downloads, incremental process (download only what's new) • Transfer priority, download management and scheduling, history stored in a database • GridFTP enabled • Hook available for automatic publication upon datasets download completion • Install using a doker container and or RPM Synda can download files from the ESGF archive in an easy way, based on a list of facets (variables, experiments, ensemble members, etc..). The program evolves together with the ESGF archive backend functionalities. This talk will walk through Synda main features from a replication and replica publication perspectives. Also, ESGF currently only supports an “offline, on demand” replication procedure, where dedicated replication sites pull replica-sets from ESGF sites, then reorganize them to fit into their internal ESGF data organization structure and finally publish them as “replicas” into the ESGF data federation. No automatic replica synchronization or notification mechanisms are supported. Thereby in general original data can be unpublished or modified without effects on replica sites. We will discuss the current work plan and will expose existing possibilities towards an automatic replication workflow.
Globus and ESGF ***Rachana Ananthakrishnan (DOE/ANL, ranantha@uchicago.edu)***	Globus provides a hosted research data management service, and is widely used for moving and sharing research data on a variety of HPC and campus computing resources. With the recent release of data publication and discovery capabilities, Globus now provides useful tools for managing data across the research lifecycle. This talk will present an overview of Globus capabilities including recently released features, and provide a quick look at some of features that will be released soon. The presentation will discuss how ESGF uses Globus today, and opportunities for future work for ESGF leveraging Globus further.
On-demand Streaming of Massive Climate Simulation Ensembles ***Cameron Christensen (University of Utah, cam@sci.utah.edu)***	The increasing size of climate datasets is a burden that impedes analysis and visualization tasks due to limited storage space, computing power, and network bandwidth available to clients. Our work addresses this issue by providing a framework for interactive visualization and user-directed analysis of massive remote climate simulation ensembles. This framework enables visualization parameters, ensemble members, and analysis scripts to all be modified on the fly. It can be used to experiment with various combinations of analyses for later use in a comprehensive global computation, or directly for out-of-core visualization and analysis tasks at any resolution. The framework supports server-side data blending and regridding to minimize client-side storage, computation, and network bandwidth requirements. Scripting integration is provided by python or java wrappers of the entire framework, and a javascript-based syntax is utilized for an in-application dynamic scripting. The framework is built on the IDX multiresolution data format, so we also provide server side on-demand data reordering for requested fields of an ensemble in order to seamlessly utilize our analysis and visualization tools. This data is cached for later access by other users. On-demand data reordering enables streaming multiresolution access and processing of remote datasets that can be too large to download directly. Even devices that could not store a single timestep of data could utilized for visualization and analysis of climate data ensembles. This system has been deployed at LLNL and is currently being integrated with the ESGF front end. The client application is currently available for download from the University of Utah. No modifications to existing data format or infrastructure need to be made in order for this technology to be utilized by end users. We will present our streaming analysis and visualization framework and demonstrate on-demand data conversion using both disparately located and massive datasets.

Day 3: Thursday, 10 December 2015
Poster Session

Title and Presenter	Abstract
Climate4Impact Portal ***Maarten Pileger (ENES/KNMI, maarten.plieger@knmi.nl)***	The aim of climate4impact (C4I) is to enhance the use of Climate Research Data and to enhance the interaction with climate effect/impact communities. The portal is based on impact use cases from 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 6 years, and its development currently continues within IS-ENES2. As the climate impact community is very broad, the focus is currently mainly on the scientific impact community. This work has resulted in the ENES portal interface for climate impact communities and can be visited at http://climate4impact.eu/. C4I is connected to the Earth System Grid Federation (ESGF). A challenge was to describe the available model data and how it can be used. The portal warns users about possible pitfalls when using climate models. All impact use cases are described in the documentation section, using highlighted keywords pointing to detailed information in the glossary. The main goal for C4I can be summarized by two objectives: The first, to work on a web interface, which generates a graphical user interface on WPS endpoints. These endpoints calculate climate indices and subset data using OpenClimateGIS/icclim on data stored in ESGF data nodes. Data is transmitted from ESGF nodes over secured OpenDAP and becomes available in a new, per user, secured OpenDAP server. The results are visualized using ADAGUC. Dedicated wizards for processing of climate indices are developed in close collaboration with users. The second, to expose C4I services to offer standardized services, which can be used by other portals, like the EU FP7 CLIPC portal. This has the advantage to add interoperability between several portals, as well as to enable the design of specific portals aimed at different impact communities, either thematic or national.
ACME Dashboard ***Matthew Harris (DOE/LLNL, harris112@llnl.gov)***	Supporting the ACME community in model development, testing and usage requires the utilization of many complex and ever-changing components from model modules and script version to computer systems and diagnostics. In particular, in collaborative development efforts it is often difficult to keep track of the latest version of specific model and scripts, which set up parameters where used by collaborators or which runs still need to be completed. The ACME Dashboard is an integrated development environment that aims to support the required 'book keeping' and coordination effort by integrating secure resources access (storage, computing), component registers (data, models, diagnostics, workflows, etc.), provenance (usage information) and work execution (e.g., run workflow, use diagnostics) in one graphical environment.
HPSS Connection to ESGF ***Sam Fries (DOE/LLNL, fries2@llnl.gov)*** ***Alex Sim (DOE/ LBNL, asim@lbl.gov)***	Accessing data stored on tape archives is difficult, time consuming, and prone to error. The ACME project plans to create 100s TB-PBs of data, all of which is not feasible to store on disk-based archives. To address this, we are bridging HPSS and ESGF, allowing data sets stored on tape to be accessed through the same methods that climate scientists are already familiar with. LBNL's Berkeley Archival Storage Encapsulation (BASE) library provides a simple API for retrieving metadata as well as actual data from HPSS and other storage systems. We are creating a Python Web application that uses BASE to access and retrieve data, and allow that data to be published to ESGF. Our initial platform will test HPSS at NERSC with ESGF nodes at LLNL, with plans to deploy at other ACME sites such as OLCF and ALCF.
Distributed Resource for the ESGF Advanced Management (DREAM) ***Dean N. Williams (DOE/LLNL, williams13@llnl.gov)*** ***Luca Cinquini (NASA/JPL, Luca.Cinquini@jpl.nasa.gov)***	We envision that the Distributed Resource for the ESGF Advanced Management (DREAM) project will accelerate discovery by enabling climate researchers, among other types of researchers, to manage, analyze, and visualize data from earth-scale measurements and simulations. DREAM's success will be built on proven components that leverage existing services and resources. A key building block for DREAM will be the ESGF. Expanding on the existing ESGF, the project will ensure that the access, storage, movement, and analysis of the large quantities of data that are processed and produced by diverse science projects can be dynamically distributed with proper resource management. Much of the DOE Office of Science data is currently generated by multiple stand-alone facilities. DREAM can collect data accumulated from these facilities and incorporate it into a fully integrated network accessible from anywhere in the world. The result is a completely new paradigm shift for data management, analysis, and visualization enabling researchers to: 1. Manage their calculations, data, tools, and research results; 2. Ensure that all data are sharable, reproducible and (re)usable—accompanied by appropriate metadata describing its provenance, syntax, and semantics at creation; 3. Advance application performance by selectively adapting APIs and services in response to scientific requirements and architectural complexities; and 4. Provide scalable interactive resource management—navigate data and metadata at multiple levels, provide architecture-aware data integration, analysis and visualization tools. We will engage closely with DOE, NASA, and NOAA science groups working at leading edge compute facilities. These engagements—in domains such as biology, climate, and hydrology—will allow us to advance disciplinary science goals and inform our development of technologies that can accelerate discovery across DOE and other U.S. agencies more broadly.
Observation Data Publication into the ESGF ***Misha B. Krassovski (DOE/ORNL, krassovskimb@ornl.gov)*** ***Tom Boden (DOE/ORNL, bodenta@ornl.gov)***	The Earth System Grid Federation software infrastructure was created to house climate change model output and to provide tools to access and analyze model output. As Earth system models (ESMs) became more sophisticated, we will need new ways to validate and test models and facilitate collaborations between field scientists, data providers, modelers, and computer scientists. To facilitate ESM validation and testing, CDIAC published part of the AmeriFlux data collection to the ESGF system. Although it does not make sense to publish CDIAC's entire data collection to ESGF (more than 1,500 exist), eleven highly relevant and popular datasets were selected and will be published to ESGF by December 2015. These datasets have different origins and data formats (e.g., gridded, point source, vertical profiles) and thus require considerable effort to publish and create appropriate metadata in order to make them harvestable and visible by the ESGF software stack.

Dali Wang (DOE/ORNL, wangd@ornl.gov)	
Climate Data Management System, version 3 (CDMS3) *Denis Nadeau (DOE/LLNL, nadeau1@llnl.gov) Dean N. Williams (DOE/LLNL, Williams13@llnl.gov) Charles Doutriaux (DOE/LLNL, doutriaux1@llnl.gov) Jeff Painter (DOE/LLNL, painter1@llnl.gov)*	The Climate Data Management System is an object-oriented data management system, specialized for organizing multidimensional, gridded data used in climate analyses for data observation and simulation. The basic unit of computation in CDMS3 is the variable, which consists of a multidimensional array that represents climate information in four dimensions corresponding to: time, pressure level, latitude, and longitude. As models become more precise in their computations, the volume of data generated becomes bigger and difficult to handle due to the limit of computational resources. Models today can produce data at time frequencies of one hourly, three hourly, or six hourly with a spatial footprint close to satellite data. The amount of time for scientists to analyze the data and retrieve useful information is more and more unmanageable. Parallelizing libraries such as CMDS3 would ease the burden of working with such big datasets. Multiple approaches of parallelizing are possible. The most obvious one is embarrassingly parallel or pleasingly parallel programming where each computer node processes one file at a time. A more challenging approach is to send a piece of the data to each node for computation and each node will save results in a file as a slab of data. This is possible with Hierarchical Data Format 5 (HDF5) using the Message Passing Interface (MPI). A final approach would be the use of Open Multi-Processing API (OpenMP) where a master thread is split in multiple threads for different sections of the main code. Each method has advantages and disadvantages. This poster brings to light the benefit for each of these methods and seeks to find an optimal solution to compute climate data analyses in an efficient fashion using one or a mixture of these parallelized methods.
Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT) *Aashish Chaudhary (Kitware, aashish.chaudhary@kitware.com)*	
CDATWeb *Matthew Harris (DOE/LLNL, harris112@llnl.gov) Jonathan Beezley (Kitware, jonathan.beezley@kitware.com)*	What is CDATWEB? A Client/Server Model for UV-CDAT. With the ever-growing size of data, downloading and attempting to visualize on the users hardware is becoming more and more cumbersome. The CDATWEB visualization server enables users to view and analyze simulation output in place rather than locally, which eliminates the need to transfer large datasets over the internet.
NetCDF/HDF5 *Ben Evans (NCI/ANU, Ben.Evans@anu.edu.au)*	
PROV *Ben Evans (NCI/ANU, Ben.Evans@anu.edu.au)*	
Climate Forecast (CF) Convention *Karl Taylor (DOE/LLNL, taylor13@llnl.gov)*	
ES-DOC *Mark Greenslade (ENES/IPSL, momipsl@ipsl.jussieu.fr) Sylvia Murphy (NOAA/ERSL, sylvia.murphy@noaa.gov) Allyn Treshansky (NOAA/ERSL, allyn.treshansky@noaa.gov) Cecilia DeLuca (NOAA/ERSL, cecelia.deluca@noaa.gov) Eric Guilyardi (ENES/IPSL, Eric.Guilyardi@locean-ipsl.upmc.fr) Sébastien Denvil (ENES/IPSL, sebastien.denvil@ipsl.jussieu.fr) Bryan Lawrence (ENES/STFC, bryan.lawrence@ncas.ac.uk)*	During the course of 2015 the Earth System Documentation (ES-DOC) project began its preparations for CMIP6 (Coupled Model Inter-comparison Project 6) by further extending the ES-DOC tooling eco-system in support of Earth System Model (ESM) documentation creation, search, viewing & comparison. The ES-DOC online questionnaire, the ES-DOC desktop notebook, and the ES-DOC python toolkit will serve as multiple complementary pathways to generating CMIP6 documentation. It is envisaged that institutes will leverage these tools at different points of the CMIP6 life-cycle. Institutes will be particularly interested to know that the documentation burden will be either streamlined or completely automated. As all the tools are tightly integrated with the ES-DOC web-service, institutes can be confident that the latency between documentation creation & publishing will be reduced to a minimum. Published documents will be viewable with the online ES-DOC viewer (accessible via citable URL's). Model inter-comparison scenarios will be supported using the ES-DOC online comparator tool. The comparator is being extended to: • Support comparison of both Model descriptions & Simulation runs; • Greatly streamline the effort involved in compiling official tables. The entire ES-DOC eco-system is open source and built upon open standards such as the Metafor Common Information Model (version 1 and 2).
Agreement on Data Management and Publication Workflow *Guillaume Levavasseur*	The ESGF publication workflow depends strongly on the data management of each datanode. Consequently, the high flexibility of the publication command-line allows partner institutes to build their own publication workflows according to their local data policies. Unfortunately, without common use of the publication tools the ESGF archive became difficult to use and manage, especially for projects containing

(ENES/IPSL, glipsl@ipsl.jussieu.fr) Ag Stephens (ENES/BADC, ag.stephens@stfc.ac.uk)	thousands of datasets (e.g., CMIP5). To ensure a high data quality for CMIP6, the IS-ENES Data Task Force is investigating around the ESGF publication workflow, taking into account as many use cases of existing data management from ESGF partners as possible. We defined and agreed on several points: • The role and tasks at each datanode have to be clearly defined and declared. Who provides, manages and/or publishes the data? • A modular design could be useful to manage the metadata redundancy between the Postgres database, the THREDDS catalogs and the Solr index. • A review of publication tools is required in order to avoid incorrect use of the publisher and to follow CMIP6 versioning requirements. • We need a publication test suite. We aim to promote best practice in publishing our CMIP6 data, using enforcement that will be implemented in the publisher code, improving the end-user experience through new ESGF services (e.g., PID, errata, etc.).
Data Citation Service ***Martina Stockhause (ENES/DKRZ, stockhause@dkrz.de) Katharina Berger (ENES/DKRZ, berger@dkrz.de)***	The review of the CMIP6 data citation procedure resulted in the requirement of a citation possibility prior to the long-term archival of the data at the IPCC DDC (Data Distribution Centre) hosted at DKRZ. A concept for a new citation module was developed and described in the WIP paper “CMIP6 Data Citation and Long-Term Archival” [1]. It consists of a repository, a GUI for data ingest and an API for data access [2]. This new component has to be integrated in the overall CMIP6 infrastructure. Several connections exist to the long-term archival, the ESGF development (esp. CoG portal, data versioning and data replication), and the other components providing additional information on the data (e.g., CIM documents) quality information and other annotations. The poster gives a short summary of the citation concept for CMIP6 and the relations between the concept for data citation and data long-term archival to other CMIP6 infrastructure components. The focus will lie on the implementation of the citation concept and the technical integration of the citation module into the ESGF infrastructure, which is a part of the ESGF-QCWT efforts [3]. References: [1] M. Stockhause, F. Toussaint, M. Lautenschlager (2015): CMIP6 Data Citation and LTA. Submitted as WIP white paper. http://www.earthsystemcog.org/projects/wip/resources [2] Data Citation Service: http://cmip6cite.wdc-climate.de [3] ESGF Quality Control Working Team (ESGF-QCWT). https://acme-climate.atlassian.net/wiki/display/ESGF/ESGF-QCWT+Charge
PCMDI’s Metrics Package ***Paul Durack (DOE/LLNL, durack1@llnl.gov) Peter Gleckler (DOE/LLNL, gleckler1@llnl.gov)***	
UV-CDAT Metrics (UVCMetrics) ***Jeff Painter (DOE/LLNL, painter1@llnl.gov) Brian Smith (DOE/ORNL, smithbe@ornl.gov)***	UVCMetrics is a new framework for climate scientists to analyze, verify, and compare output from multiple model runs (or observation sets). UVCMetrics implements the functionality of most of the NCAR NCL-based land and atmosphere diagnostics in a more flexible, extensible Python-based framework which utilizes CDAT and VCS to plot (or summarize in tabular form) typical diagnostic plots. UVCMetrics has full command line support and individual diagnostics can be run from within UVCDAT. Data produced by the package is in standard format (e.g. netCDF, XML, and PNG files) and can be further analyzed or manipulated by scientists in UVCDAT or other tools. The framework supports recreation of the NCAR plots in their entirety or by an individual plot set. Additional variables, regions, seasons, or variable options can be added easily to expand the diagnostics available. Entirely new plot sets can be added as well, and the framework also supports “loose coupling” where existing scripts written in NCO, R, etc. can be integrated. Current efforts are focusing on multiple levels of parallelization - both computation of individual diagnostics and running multiple separate diagnostic computations in parallel.
ESMValTool ***Stephan Kindermann (ENES/DKRZ, kindermann@dkrz.de)***	The Earth System Model Evaluation Tool (ESMValTool) effort provides a community diagnostic and performance metrics tool for routine evaluation of Earth System Models (especially in CMIP6). The priority of the effort so far has been to target specific scientific themes focusing on selected Essential Climate Variables and a range of known systematic biases common to ESMs. To support CMIP6 it is necessary to deploy ESMValTool based processing services "near to" ESGF nodes, providing fast access to large amouts of model data (local as well as replicated). An approach for the ESGF integration has been developed and is currently in a testing phase. This poster will provide an overview of the ESMValTool as well as the ESGF integration solution, which was implemented. The implementation exploits local ESGF caches as well as a Synda tool based replication from remote sites. Also first experiments were done do integrate the ESMValTool in a WPS (Web Processing Service) framework.
ESGF-QCWT: CMIP6 Errata as a New ESGF Service ***Guillaume Levavasseur (ENES/IPSL, glipsl@ipsl.jussieu.fr)***	Because of the experimental protocol inherent complexity of project like CMIP5 or CMIP6 it becomes important to record and to track reasons for datasets version changes. During CMIP5 it was impossible for scientists making use of data sets hosted by ESGF to know easily whether they were using a dataset having a known problem and whether this known problem were corrected by a newer version. Also very difficult was to have access to a description of this issue. To move towards a better errata system is motivated by key requirements:

Sébastien Denvil (ENES/IPSL, sebastien.denvil@ipsl.jussieu.fr)	• Provide timely information about newly discovered issues. Because errors cannot entirely be eliminated, we should implement a centralized public interface to data providers, so that they can directly describe problems when they are discovered. • Provide known issues information prior to download. The user has to be informed of known issues before downloading through the ESGF search interface. • Enable users to interrogate a database to determine whether modifications and/or corrections have been applied to data they have downloaded. This service could rely on unique file identifiers so end users can discover whether files of interest to them 1) have been affected by known issues, 2) have been withdrawn, and/or 3) have been modified or corrected. • Develop as part of the errata system a capability to notify end users of updates to files of interest to them. The Quality Control Working Team aims to define and to establish a stable and coordinated procedure to collect and give access to errata information related to data sets hosted by ESGF.
Enabling in-situ analytics in the Community Earth System Model via a Functional Partitioning Framework ***Valentine Anantharaj (DOE/ORNL, anantharajvg@ornl.gov)***	Efficient resource utilization is critical for improved end-to-end computing and workflow of scientific applications. Heterogeneous node architectures, such as the GPU-enabled Titan supercomputer at the Oak Ridge Leadership Computing Facility (OLCF), present us with further challenges. In many HPC applications on Titan, the accelerators are the primary compute engines while the CPUs orchestrate the offloading of work onto the accelerators, and moving the output back to the main memory. On the other hand, applications that do not exploit GPUs, the CPU usage is dominant while the GPUs idle. We utilized Heterogeneous Functional Partitioning (HFP) runtime framework that can optimize usage of resources on a compute node to expedite an application's end-to-end workflow. This approach is different from existing techniques for in-situ analyses in that it provides a framework for on-the-fly analysis on-node by dynamically exploiting under-utilized resources therein. We have implemented in the Community Earth System Model (CESM) a new concurrent diagnostic processing capability enabled by the HFP framework. Various single variate statistics, such as means and distributions, are computed in-situ by launching HFP tasks on the GPU via the node local HFP daemon. Since our current configuration of CESM does not use GPU resources heavily, we can move these tasks to GPU using the HFP framework. Each rank running the atmospheric model in CESM pushes the variables of of interest via HFP function calls to the HFP daemon. This node local daemon is responsible for receiving the data from main program and launching the designated analytics tasks on the GPU. We have implemented these analytics tasks in C and use OpenACC directives to enable GPU acceleration. This methodology is also advantageous while executing GPU-enabled configurations of CESM when the CPUs will be idle during portions of the runtime. In our implementation results, we demonstrate that it is more efficient to use HFP framework to offload the tasks to GPUs instead of doing it in the main application. We observe increased resource utilization and overall productivity in this approach by using HFP framework for end-to-end workflow.
The OPTIRAD Project: cloud-hosting the IPython Notebook to provide a collaborative data analysis environment for the earth sciences community ***Phil Kershaw (ENES/CEDA, philip.kershaw@stfc.ac.uk) Bryan Lawrence (ENES/STFC, bryan.lawrence@ncas.ac.uk), et al.***	We report on experiences deploying the IPython Notebook on the JASMIN science cloud for the OPTIRAD project and its evolution towards a generic collaborative tool for data analysis in the earth sciences community. The system has been developed in the context of OPTIRAD (OPTImisation environment for joint retrieval of multi-sensor RADiances), a project funded by the European Space Agency focused on data assimilation of earth observation products for land surface applications. This domain presents a number of challenges which have provided drivers for the solution developed: the use of computationally expensive processing algorithms, access to large volume earth observation datasets and the need for shared working within its user community. The IPython Notebook has been gaining traction in recent years as a teaching tool for scientific computing and data analysis. It provides an interactive Python shell hosted in an intuitive, user-friendly web-based interface together with functionality to save and share sessions. The IPython development community is very active and recent work has led to the creation of a new package JupyterHub. The name Jupyter reflects the fact that the notebook can now support other languages in addition to Python such as the statistical package R. JupyterHub builds on the baseline functionality of the notebook but incorporates the ability to support multiple user sessions fronted with the required authentication and access control. These developments are significant since they provide the key capabilities needed to enable notebooks to be hosted via a cloud service. Use of cloud technology makes a powerful combination, enabling the Notebook to take advantage of cloud computing's key attributes of scalability, elasticity and resource pooling. In this way it can address the needs of long-tail science users of Big Data: an intuitive interactive interface with which to access powerful compute and storage resources. We describe how the Notebook has been used in combination with the package ipyparallel to provide a Python-based API to parallel compute capability. Use of Docker containers and the Swarm cluster management system is facilitating scaling of resources to meet demand. We look at how this and other developments are informing the future evolution of the system.
A NASA Climate Model Data Services (CDS) End-to-End System to Support Reanalysis Intercomparison ***Jerry Potter (NASA/GSFC, jpotter@ucdavis.edu)***	Scientists engaged in reanalyses—essentially re-forecasts of past weather using the latest forecast models—are interested in reproducing the success of the Coupled Model Intercomparison Project (CMIP5). They are studying reanalysis differences and uncertainties to improve reanalysis techniques. Reanalysis data also allows interdisciplinary scientists to compare their datasets (e.g., biodiversity, water planning, wind power) with 30 or more years of gridded climate data. These research efforts require large sets of monthly and hourly data, formatted identically to facilitate comparisons. NASA's Climate Model Data Services (CDS) is collaborating with the world's five major reanalysis projects to collect this data and present it through Distribution, Visualization, Analytics, and Knowledge services, resulting in the Collaborative REAnalysis Technical Environment-Intercomparison Project (CREATE-IP).