Partnerships for development of next-generation software for distributed access and analysis of simulated, observed, and reanalysis data from the climate and weather communities.

Joint DOE, NASA, NOAA, IS-ENES, and ANU/NCI Earth System Grid Federation (ESGF) 2017 Face-to-Face Conference Abstracts

Day 1: Tuesday, 5 December 2017
ESGF Steering Committee and Executive Committee

Title and Presenter	Abstract
Department of Energy Office of Biological and Environmental Research Data Management Opening Comments **Justin Hnilo (DOE/BER)** *Justin.Hnilo@science.doe.gov*	The ESGF multi-agency, international software infrastructure has become critical to understanding climate change. Effectively managing the vast volumes of resulting simulation and observation data has become a major challenge for the climate and computational scientists who support climate projections. To manage the massively distributed data volumes, the ESGF connects diverse federated archives from over 21 countries for knowledge discovery. These distributed data archives have aided many Department of Energy (DOE) researchers in producing significant articles and reports, such as those contributing to the Intergovernmental Panel on Climate Change (IPCC) Third and Fifth Assessment Reports (AR3 and AR5, respectively). Today, the ESGF infrastructure houses petabytes of data generated by DOE projects, such as the Energy Exascale Earth System Model (E3SM) and the international Coupled Model Intercomparison Project (CMIP), and securely makes these data available to scientists and nonscientists. In addition, the infrastructure provides data access services for DOE's broad community by conforming to DOE data standards. Data to be ingested, stored, maintained, and served by the infrastructure include DOE observational, experimental, and model-generated information and associated metadata, plus the tools and models directly associated with data generation, value-added products, simple analysis, display, and data serving. Thus, the access the ESGF provides has translated into an impressive volume of new DOE research. Over the next three years, it is estimated that the ESGF distributed archive will grow to tens of petabytes of data storage and bridge the critical gaps between many DOE projects concerning big data issues.
The State of the Earth System Grid Federation **Dean N. Williams (DOE/LLNL/AIMS)** *williams13@llnl.gov* **Luca Cinquini (NASA/JPL)** *luca.cinquini@jpl.nasa.gov*	The ESGF is a multi-institutional, international software infrastructure and development collaboration led by scientists and software engineers worldwide. The ESGF's mission is to facilitate scientific research and discovery on a global scale. The ESGF architecture federates a geographically distributed network of climate modeling and data centers that are independently administered yet united by common protocols and application programming interfaces (APIs). The cornerstone of its interoperability is peer-to-peer messaging, which continuously exchanges information among all nodes through a shared, secure architecture for search and discovery. The ESGF integrates popular open-source application engines with custom components for data publishing, searching, user interface (UI), security, metrics, and messaging to provide petabytes of geophysical data to roughly 25,000 users from over 1,400 sites on six continents. It contains output from the CMIP, used by authors of the Intergovernmental Panel on Climate Change Third, Fifth, and Sixth Assessment Reports, and output from DOE's E3SM project and the European Union's (EU's) Copernicus Programme. Over the next three years, we propose to: 1. sustain and enhance a resilient data infrastructure with friendlier tools for the expanding global scientific community; and 2. prototype new tools that fill important capability gaps in scientific data archiving, access, and analysis.

Title and Presenter	Abstract
	These goals will support a data-sharing ecosystem and, ultimately, provide predictive understanding of couplings and feedbacks among natural-system and anthropogenic processes across a wide range of geophysical spatial scales.

Day 1: Tuesday, 5 December 2017
ESGF Science Drivers: Project Requirements and Feedback

Title and Presenter	Abstract
Coupled Model Intercomparison Project, phase 6 (CMIP6) and the Working Group on Coupled Modeling Infrastructure Panel (WIP) **Karl Taylor (DOE/LLNL/PCMDI)** *taylor13@llnl.gov* **V. Balaji (NOAA/GFDL & Princeton University)** *balaji@princeton.edu*	The WCRP Working Group on Coupled Modelling (WGCM) Infrastructure Panel, referred to as the "WIP", was established to provide clear guidance to ESGF and other projects supporting CMIP6 as to infrastructure needs from the perspective of climate modelling centers and the end-users. The WIP is responsible for oversight of the CMIP6 "data request" and establishing metadata requirements and controlled vocabularies that make it possible to automate management, access, and interaction with the data archive. The WIP also considers the dependencies among various services built to support CMIP6, and guides their development so that they interact smoothly. It also attempts to encourage development of data standards and metadata specifications for closely related projects (e.g., input4MIPs, obs4mips) so that ESGF can provide a more uniform interface to the data produced by them. Following a summary of the current status of CMIP6 and the infrastructure supporting it, we shall identify high priority needs or concerns regarding ESGF's critical contributions to WCRP activities.
An Update on obs4MIPs from an ESGF perspective: progress, plans and challenges **Peter Gleckler (DOE/LLNL/PCMDI)** *gleckler1@llnl.gov* **Duane Waliser (NASA/JPL)** *duane.waliser@jpl.nasa.gov* **Denis Nadeau (DOE/LLNL/AIMS)** *nadeau1@llnl.gov* **Robert Ferraro (NASA/JPL)** *robert.d.ferraro@jpl.nasa.gov* **Karl Taylor (DOE/LLNL/PCMDI)** *taylor13@llnl.gov* **Luca Cinquini (NASA/JPL)** *Luca.Cinquini@jpl.nasa.gov* **Paul Durack (DOE/LLNL/PCMDI)** *durack1@llnl.gov*	During the last year, substantial effort has been devoted to coordinating the use of Climate Model Output Rewriter 3 (CMOR3) in CMIP6 with obs4MIPs. This has included further alignment of the obs4MIPs data specifications with CMIP6. Recently, these metadata specifications have largely been finalized, opening up the potential to include a next generation of obs4MIPs datasets with more enhanced searching capabilities available via the ESGF. Two other recent ESGF-related advancements will be discussed: (1) the inclusion of dataset specific information in the form of a "suitability matrix," and (2) the ability for data providers to include supplemental data and metadata along with their best-estimate contribution to obs4MIPs. After summarizing this progress, this presentation will be describing how obs4MIPs can be further advanced via new ESGF capabilities.
Copernicus and H2020 Programme **Sébastien Denvil (ENES/IPSL)** *sebastien.denvil@ipsl.jussieu.fr* **Michael Lautenschlager (DKRZ)** *lautenschlager@dkrz.de* **Sandro Fiore (CMCC)** *sandro.fiore@cmcc.it* **Francesca Guglielmo (ENES/IPSL)** *francesca.guglielmo@lsce.ipsl.fr* **Martin Juckes (CEDA)**	European Network for Earth System Modelling (ENES) partners are involved in numbers of projects funded by either the Horizon 2020 (H2020) Programme or the Copernicus Programme. Some of those projects will contribute to pieces of the development of the ESGF or will use ESGF results. This talk will introduce both H2020 and the Copernicus Programme and will highlight the major contribution to ESGF activities that are expected by the ENES Data Task Force from currently running projects. Horizon 2020 is the biggest-ever EU research and innovation Programme. Almost €80 billion of funding is available over seven years (2014 to 2020) in addition to the private and national public investment that this money will attract. The goal is to ensure Europe produces world-class science and technology, removes barriers to innovation, and makes it easier for the public and private sectors to work together in delivering solutions to big challenges facing our society.

Title and Presenter	Abstract
martin.juckes@stfc.ac.uk **Stephan Kindermann (DKRZ)** kindermann@dkrz.de **Michael Kolax (SMHI)** *Michael.Kolax@smhi.se* **Wim Som de Cerff (KNMI)** *wim.som.de.cerff@knmi.nl*	Within H2020, three types of activities will be supported to make world-class research infrastructures accessible to all researchers in Europe and to fully exploit these resources' potential for scientific advancement and innovation. • The first activities are targeted to the development of new world-class research infrastructures. Support will be provided for the implementation and operation of the research infrastructures listed on the European Strategy Forum on Research Infrastructures (ESFRI) roadmap. Support will cover the preparatory phase of new ESFRI projects and the implementation and operation phases of prioritized ESFRI projects. Further world-class facilities will also be part of this action. • The second set of activities aims at optimizing the use of national facilities by integrating them into networks and opening their doors to all European researchers. This is a continuity of the so-called Integrating Activities under FP7. • The third action will support further deployment and development of ICT-based e-infrastructures which are essential to enable access to distant resources, remote collaboration, and massive data processing in all scientific fields. Copernicus has been specifically designed to meet user requirements. Through satellite-based, in situ observations and simulations, the services deliver data at a global level which can also be used for local and regional needs. This is essential to help us better understand our planet as well as sustainably manage the environment we live in. Copernicus is served by a set of dedicated satellites (the Sentinel families) and contributing missions (existing commercial and public satellites). The Sentinel satellites are specifically designed to meet the needs of the Copernicus services and their users. Since the launch of Sentinel-1A in 2014, the EU plans to place a constellation of almost 20 more satellites in orbit before 2030. The main users of Copernicus services are policymakers and public authorities who need the information to develop environmental legislation and policies or to make critical decisions in the event of an emergency, such as a natural disaster or a humanitarian crisis. Based on the Copernicus services and on the data collected through the Sentinels and the contributing missions, many value-added services can be tailored to specific public or commercial needs, resulting in new business opportunities. These value-adding activities are streamlined through **six thematic streams of Copernicus services**. One of them is the Copernicus Climate Change Service (C3S).
Collaborative REAnalysis Technical Environment Intercomparison Project (CREATE-IP) **Jerry Potter (NASA/GSFC)** *gerald.potter@nasa.gov* **Laura Carriere (NASA/GSFC)** *laura.carriere@nasa.gov* **Judy Hertz (NASA/GSFC)** *judy.hertz@nasa.gov*	In light of the recent extreme weather events, it has become increasingly evident that quick and easy access to multiple up-to-date modern reanalysis products is useful to a variety of researchers. The Collaborative REAnalysis Technical Environment (CREATE) service offers reanalysis data repackaged in a form that is easily accessible through the ESGF. The data adhere to the standard CMIP metadata requirements, and the data sets are extended every two months through automation. In addition to monthly and six-hour data for selected variables, CREATE provides five of the major reanalyses regridded to a standard grid, along with the ensemble average and standard deviation. These data were processed using the Earth Data Analytics Services (EDAS), a high-performance big data analytics framework developed at the NASA Center for Climate Simulation (NCCS) for the ESGF, and the data are provided at six-hour and monthly intervals with a common set of vertical levels. CREATE data are also made available through CREATE-V, a web tool that leverages the common data format to provide visualization and comparison features and utilizes EDAS to display plots of monthly anomalies and yearly cycles at a user-selected location. Using multiple reanalyses, NASA will use the CREATE service and products to explore examples of the recent droughts, floods, and hurricanes and study longer term climate trends.
Energy Exascale Earth System Model (E3SM) Workflow **Dean N. Williams (DOE/LLNL/AIMS/E3SM)** *williams13@llnl.gov* **Valentine Anantharaj (DOE/ORNL/E3SM)** *anantharajvg@ornl.gov*	The advanced model development, testing, and execution infrastructure has been designed to strongly accelerate the model development and testing cycle for the new DOE E3SM model by automating labor-intensive tasks, providing intelligent support for complex tasks, and reducing duplication of effort through collaborative Workflow Group support. The Workflow Group has two important assignments: (1) advance model development by developing, testing, and executing an end-to-end infrastructure that automates labor-intensive tasks; and (2) provide intelligent support for complex tasks in model development through scientific model component (i.e., atmosphere, land, ocean, and sea ice) collaboration.

Title and Presenter	Abstract
Dave Bader (DOE/LLNL/E3SM) *bader2@llnl.gov* **Renata McCoy (DOE/LLNL/AIMS/E3SM)** *mccory20@llnl.gov*	To achieve our primary objectives, the team was split into several epic subtasks: (1) E3SM Workbench and Process Flow; (2) Data Management; (3) Analysis and Visualization; (4) Diagnostics; (5) Provenance Capture; and (6) Hardware Infrastructure. These open-source projects have grown in scope as requirements have shifted or completely changed over the course of the project. The tools and experience resulting from their development provides the foundation on which the end-to-end model testbed infrastructure will be based. As the global view of the E3SM project expands across the component model space, the usefulness and urgency of the workflow software becomes more apparent. The end goal of every quarter for the Workflow Group is to advance a step closer to reducing the level of effort to successfully run the E3SM model, archive output, generate diagnostics, and share the results of both the model output and diagnostics results with E3SM colleagues.

Day 1: Tuesday, 5 December 2017
Poster and Live Demonstration Session

Title and Presenter	Abstract
The Earth Data Analytics Services (EDAS) Framework **Thomas Maxwell (NASA/GSFC)** *Thomas.maxwell@nasa.gov* **Dan Duffy (NASA/GSFC)** *Daniel.q.duffy@nasa.gov* *Poster & Demo*	Faced with unprecedented growth in Earth data volume and demand, the National Aeronautics and Space Administration (NASA) has developed the EDAS framework, a high-performance big data analytics framework built on Apache Spark. This framework enables scientists to execute data processing workflows combining common analysis operations close to the massive data stores at NASA. The data are accessed in standard formats (e.g., network common data form [netCDF], hierarchical data format [HDF]) in a Portable Operating System Interface (POSIX) file system and processed using vetted Earth data analysis tools (e.g., Earth System Modeling Framework [ESMF], Community Data Analysis Tools [CDAT], netCDF operators [NCO]). EDAS utilizes a dynamic caching architecture, a custom distributed array framework, and a streaming parallel in-memory workflow for efficiently processing huge datasets within limited memory spaces with interactive response times. EDAS services are accessed via a web processing service (WPS) API being developed in collaboration with the ESGF Compute Working Team (CWT) to support server-side analytics for the ESGF. The API can be accessed using direct web service calls, a Python script, a Unix-like shell client, or a JavaScript-based web application. New analytic operations can be developed in Python, Java, or Scala (with support for other languages planned). Client packages in Python, Java/Scala, or JavaScript contain everything needed to build and submit EDAS requests. The EDAS architecture brings together the tools, data storage, and high-performance computing capabilities required for timely analysis of large-scale data sets, where the data reside, to ultimately produce societal benefits. It is currently deployed at NASA in support of the CREATE project, which centralizes numerous global reanalysis datasets onto a single advanced data analytics platform. This service enables decision makers to compare multiple reanalysis datasets and investigate trends, variability, and anomalies in Earth system dynamics around the globe.
PAVICS: A platform for the Analysis and Visualization of Climate Science – Toward Inter-operable Multidisciplinary Workflows **D. Huard (Ouranos)** *huard.david@ouranos.ca* **T. Landry (CRIM)** *tom.landry@crim.ca* **D. Byrns (CRIM)** *David.Byrns@crim.ca* **B. Gauvin-St-Denis (Ouranos)** *GauvinSt-Denis.Blaise@ouranos.ca* *Poster*	Climate services comprise the necessary data and expertise to describe current and future climate conditions and their potential impact on human and environmental systems. Climate services are by their nature interdisciplinary, and an important bottleneck in delivering relevant and timely climate services lies at the interface between disciplines; differences in jargon, conventions, data formats, and programming languages act as barriers to effective collaborations. Here we describe how a scientific model—one in which researchers publish not only papers and data but also “expertise” in the form of online interoperable services—has the potential to drastically reduce the friction across disciplines. This scientific model could be exploited by scientific gateways such as the Power Analytics and Visualization for Climate Science (PAVICS) to widen the scope and relevance of climate services. The PAVICS platform is built from modular components that target the various requirements of climate data analysis. The data components host and catalog netCDF data as well as geographical information layers. The analysis and processing components are made available as atomic operations through WPS, which can be combined into workflows and executed on a distributed network of heterogeneous computing resources. The visualization components range from Open Geospatial Consortium (OGC) standards (e.g., WMS, web coverage service [WCS], web feature service) to a complete front-end for searching the data, launching workflows, and interacting with

Title and Presenter	Abstract
	maps of the results. Each component can easily be deployed and executed as an independent service through the use of Docker. Permissions on data and processes are managed via a RESTful interface and are enforced systematically with a token-based service. PAVICS includes various components from Birdhouse, a collection of WPS developed by the German Climate Computing Centre (DKRZ) and Institut Pierre Simon Laplace (IPSL). Further connectivity is made with the ESGF nodes, and local results are made searchable using the same API terminology.
PAVICS: A Platform for the Analysis and Visualization of Climate Science **D. Huard (Ouranos)** *huard.david@ouranos.ca* **D. Byrns (CRIM)** *David.Byrns@crim.ca* **T. Landry (CRIM)** *tom.landry@crim.ca* *Demo*	The PAVICS project aims to provide climate scientists and climate service providers with tools that simplify the creation of value-added products from raw climate data sets. It includes a Thematic Real-time Environmental Distributed Data Services (THREDDS) data server, a search engine that connects with the ESGF database, a GeoServer instance to supply geographical layers, a diverse set of analytical services and a workflow engine to chain them together, along with a graphical workspace interface that can overlay geographic information system (GIS) layers and netCDF gridded data sets. The project relies heavily on projects orbiting the ESGF landscape, namely Birdhouse, OpenClimate GIS (OCGIS), and Indice Calculation CLIMate (ICCLIM). This demonstration shows how users create a project, define a data ensemble, subset the data over multiple geographical regions, compute climate indices, and create graphics displaying the results without any knowledge of netCDF. The demonstration will also cover how administrators manage services and permissions, as well as upload additional netCDF data and GIS layers via OGC standards. We will show JavaScript-based UI components pertaining to experience management, search, thematic layer management, web mapping, scientific workflows, and user workspaces. We will also discuss how the services that are planned by the ESGF CWT could be integrated in the platform.
OGC Testbed-13 Earth Observation Clouds **T. Landry (CRIM)** *tom.landry@crim.ca* **D. Byrns (CRIM)** *David.Byrns@crim.ca* *Poster*	On a yearly basis, Earth Observation (EO) satellites already generate petabytes of raw data. Resources required to process and store that data are quickly increasing due to higher resolutions, larger number of bands, and growing satellites and constellation count. The cloud computing landscape is well suited to cover most requirements of EO data and its applications. OGC's thirteenth testbed initiative (TB-13) aims to clarify cloud API interoperability and application portability as key elements in cloud computing research. The goal of participants of the EO Clouds (EOC) thread of TB-13 is to develop an integrated solution compatible with the European Space Agency's (ESA's) Thematic Exploitation Platforms (TEP) and the Canadian Forestry Service's operation, which is part of Natural Resources Canada (NRCan). Centre de Recherche Informatique de Montréal (CRIM) is mandated by OGC to deliver a cloud-enabled application that extracts forest features or biophysical parameters from Radarsat-2 Synthetic Aperture Radar (SAR) to estimate forest biomass in Canada, in accordance with NRCan's specifications. CRIM implemented the SAR decompositions using ESA's Sentinel Application Platform (SNAP) graph processing tool and packaged the application using Docker, in an OpenStack environment. In that implementation, WCS and WMS capabilities were provided by GeoServer 2.11, while WPS requests were served by PyWPS 4. Participants in TB-13 are required to participate in the elaboration of an Engineering Report (ER) to be presented to designated OGC Working Groups. In order to address sponsors' requirements for the EOC thread, CRIM provided engineering content and prototypes on cloud computing, remote sensing, authentication and security, asynchronous processing, workflow execution, and job management. For TB-13, U.S. Department of Energy (DOE) provided OGC with specifications related to ESGF. Helped by ESGF Compute Working Team (CWT), CRIM explored cloud implementations for climate processes and initiated integration in both PAVICS and Birdhouse. An implementation goal was the development of an integrated solution compatible for both NRCan and ESGF. Both TB-13 deliverables will be made available via managed services at CRIM to other testbed participants and their authorized affiliates for a period of one calendar year. Ideas on the upcoming testbed at OGC proposes collaborative testbed experiments in federated environments. Initial findings indicates that ESGF would be an appropriate case study. OGC demonstration event for TB-13 is planned for the second week of December, in Reston, Virginia.
Using the ESGF CWT-API in the Context of the EUDAT-EGI e-Infrastructure and the ENES climate4impact Platform **Christian Pagé (CERFACS)**	Supporting data analytics in climate research with respect to data access is a challenge due to increasing data volumes. Several international and European initiatives have emerged and provide standalone solutions that offer potential for interoperability. In Europe, the IS-ENES (https://is.enes.org) consortium has developed a platform to ease access to climate data for the climate impact community (C4I: https://climate4impact.eu). It exposes data from ESGF data nodes as well as any OPeNDAP server. It provides UIs, wizards, and services for search and discovery,

Title and Presenter	Abstract
christian.page@cerfacs.fr **Xavier Pivan (CERFACS)** *xavier.pivan@cerfacs.fr* **Asela Rajapakse (MPI-M)** *asela.rajapakse@mpimet.mpg.de* **Wim Som de Cerff (KNMI)** *wim.som.de.cerff@knmi.nl* **Maarten Plieger (KNMI)** *maarten.plieger@knmi.nl* **Ernst de Vreede (KNMI)** *ernst.de.vreede@knmi.nl* **Alessandro Spinuso (KNMI)** *alessandro.spinuso@knmi.nl* **Lars Barring (SMHI)** *Lars.Barring@smhi.se* **Antonio Cofino (University of Cantabria)** *antonio.cofino@unican.es* **Alessandro d'Anca (CMCC)** *alessandro.danca@cmcc.it* **Sandro Fiore (CMCC)** *sandro.fiore@cmcc.it* *Poster & Demo*	visualization, processing, and downloading. Also in Europe, an emerging e-infrastructure is being designed and built for several scientific domains, led by EUDAT (https://eudat.eu) and EGI (https://egi.eu), which will form the basis of the future European Open Science Cloud (EOSC) to support scientific researchers. This e-infrastructure provides services within the EUDAT Collaborative Data Infrastructure (CDI). The ENES climate community is participating in the EUDAT CDI. Within the EUDAT project, work has been done to integrate these existing e-infrastructures. The goal is to develop interoperable interfaces. 1. A first-level prototype has been completed that deploys the Generic Execution Framework (GEF) Docker backend onto the EGI FedCloud to perform computations and feeds the results into the EUDAT CDI. 2. The second-level prototype involves integrating the GEF backend and the ESGF CWT-API. The GEF backend pulls data from the ESGF infrastructure through the CWT-API so that data reduction is achieved through on-demand calculations. Furthermore, complex calculations are then executed on the EGI FedCloud, and the results are fed back into EUDAT B2 services. This raises the authentication and authorization integration between the ESGF and EUDAT/EGI. A first solution would be to use the token-based approach of C4I. 3. The third-level prototype is the same as the second one, except that the GEF is executed by the C4I platform on demand by the user, and the final results are fed back into the C4I user space for visualization, storage, or download. A variant of this third-level prototype that could be implemented is getting the data directly from one or several ESGF data nodes, using the C4I Search WPS to locate the data files. A demo of the third-level prototype will be presented as well.
Managing Growth and Complexity - Technologies to Meet the Challenges of Operating Data, Services, and Infrastructure at Scale **Phil Kershaw (ENES/CEDA)** *philip.kershaw@stfc.ac.uk* **Jonathan Churchill (ENES/CEDA)** *jonathan.churchill@stfc.ac.uk* **Alan Iwi (ENES/CEDA)** *alan.iwi@stfc.ac.uk* **Bryan Lawrence (University of Reading)** *bryan.lawrence@ncas.ac.uk* **Neil Massey (ENES/CEDA)** *neil.massey@stfc.ac.uk* **Sam Pepler (ENES/CEDA)** *sam.pepler@stfc.ac.uk* **Matt Pritchard (ENES/CEDA)** *matt.pritchard@stfc.ac.uk* **Matt Pryor (ENES/CEDA)** *Matt.Pryor@stfc.ac.uk* **Ag Stephens (ENES/CEDA)** *ag.stephens@stfc.ac.uk* *Poster*	The Centre for Environmental Data Analysis (CEDA) hosts data and services for a wide range of communities and with international collaboration efforts such as the ESGF. Since 2012, the underlying computing infrastructure has been provided by JASMIN, a shared computing platform for the environmental sciences community consisting of a high-performance, high-volume storage system to host key datasets co-located with computing resources for processing and analysis. Operational experiences to date and lessons learnt are informing decisions about JASMIN's and ESGF's future technical direction. With the success of the system, both the quantity of data stored and the number of supported users have grown. As part of its technical evolution, a program of work is underway to address the challenges associated with this growth. Here we highlight two specific technologies being used as part of that program of work: object storage and containers. These will bring fundamental changes to how we operate JASMIN and their consequent impact on our existing service infrastructure like the ESGF. A number of factors are driving the adoption of object storage, but there are issues with adopting it in the JASMIN environment. We will briefly discuss these factors and issues before introducing our plans to migrate to object storage for JASMIN. These plans include the development of domain-specific software customized to exploit HDF5/netCDF4 data held in object stores, providing the user community efficient access consistent with existing familiar interfaces. We will also describe the application of the container technologies Docker and Kubernetes to underpin the provision and operation of new services including climate services for the EU Copernicus program—which re-uses the ESGF application stack—and the development of a new Cluster-as-a-Service concept for JASMIN's cloud: the dynamic provision of clusters to host new compute and analysis applications such as Jupyter Notebooks, Dask, and PySpark.
Ophidia: An Interoperable 'Big Data' Framework for Climate Change Analytics Experiments **Sandro Fiore (CMCC)**	The Ophidia project provides a complete environment for scientific data analysis on multidimensional datasets. It exploits data distribution and supports array-based primitives for mathematical and statistical operations, analytics jobs management and scheduling, and a native in-memory input/output (I/O) server for fast data analysis. It also provides access through standard interfaces like SOAP, GSI/VOMS, and OGC-WPS.

Title and Presenter	Abstract
sandro.fiore@cmcc.it **Charles Doutriaux (DOE/LLNL/AIMS)** *doutriaux1@llnl.gov* **Cosimo Palazzo (CMCC)** *cosimo.palazzo@cmcc.it* **Alessandro d'Anca (CMCC)** *alessandro.danca@cmcc.it* **Zeshawn Shaheen (DOE/LLNL/AIMS)** *shaheen2@llnl.gov* **Donatello Elia (CMCC)** *donatello.elia@cmcc.it* **Jason Boutte (DOE/LLNL/AIMS)** *boutte3@llnl.gov* **Valentine Anantharaj (ORNL/E3SM)** *anantharajvg@ornl.gov* **Dean N. Williams (DOE/LLNL/AIMS)** *williams13@llnl.gov* **Giovanni Aloisio (CMCC)** *giovanni.aloisio@unisalento.it* *Poster & Demo*	In the climate change domain, the Ophidia framework has been applied to support the implementation of real use cases on multi-model analysis, climate indicators, and processing chains for operational environments in different European projects. A recent effort concerns a new interface implementing the ESGF WPS Extension Specification. In this regard, a complete Python-based interface has been developed to support Ophidia workflows submission by means of Python clients and applications. Authentication and authorization are guaranteed through a token-based approach. The remote submissions exploit the Ophidia workflow engine interface which exposes several constructs to implement different features (e.g., loops, automated workflows, arguments, interleaved mechanisms, parallelism). The Ophidia stack along with the ESGF WPS compliant interface has been installed in OphidiaLab, a new multi-user environment for scientific data analysis deployed at the CMCC Supercomputing Center. The demonstration will focus on the new OphidiaLab environment, the WPS-enabled interface, the workflow capabilities provided by Ophidia, and some use cases from European projects like EUBra-BIGSEA and INDIGO-DataCloud.
Federated Data Usage Statistics in the Earth System Grid Federation **Alessandra Nuzzo (ENES/CMCC)** *alessandra.nuzzo@cmcc.it* **Maria Mirto (CMCC)** *maria.mirto@cmcc.it* **Paola Nassisi (CMCC)** *paola.nassisi@cmcc.it* **Katharina Berger (DKRZ)** *berger@dkrz.de* **Torsten Rathmann (DKRZ)** *rathmann@dkrz.de* **Luca Cinquini (NASA/JPL)** *Luca.Cinquini@jpl.nasa.gov* **Sébastien Denvil (ENES/IPSL)** *sebastien.denvil@ipsl.jussieu.fr* **Sandro Fiore (CMCC)** *sandro.fiore@cmcc.it* **Dean N. Williams (DOE/LLNL/AIMS)** *williams13@llnl.gov* **Giovanni Aloisio (CMCC)** *giovanni.aloisio@unisalento.it* *Poster & Demo*	The federated monitoring system plays an important role in the context of the ESGF. This task is accomplished by the ESGF-dashboard component, which is composed by a backend and a front-end module: the former dedicated to managing data usage statistics at single site and federation level and the latter providing a flexible and usable web interface. The main goal of the ESGF-dashboard is to provide a distributed and scalable monitoring framework responsible for capturing usage metrics at the single site level and at the global ESGF level. The backend component of the ESGF-dashboard, included into the software stack of the ESGF data node, has a main role of collecting and storing a high volume of heterogeneous metrics, covering measures such as downloads and clients' statistics, aggregated cross and project-specific download statistics. With respect to the previous version of the dashboard front-end, its final implementation moves away from the previous desktop metaphor to approach a brand new one closer to the dashboard concept with a stronger usability. The new UI, already deployed in production, provides a rich set of charts and reports through a web interface, allowing users and system managers to visualize the status of the infrastructure through a set of smart and attractive web gadgets. The key challenges of such concept are to communicate the most important information in a straightforward way and allow users to view specific details at the same time. The collection of federated statistics is accomplished through a RESTful API that retrieves and aggregates metrics from all data nodes across the federation.
WPS-Based Processing Services for the Copernicus Climate Change Service (C3S) **Stephan Kindermann (DKRZ)**	The C3S will integrate global and regional climate projections into the Climate Data Store (CDS) [1]. The CDS will also provide consistent access to in situ and satellite-based climate observations, reanalysis data and multi-model seasonal forecasts. On the data access side, the ESGF and its data services (search, authentication, download, and subset) will provide the interface layer between C3S and the model data archives at DKRZ, the Science & Technology Facilities Council (STFC),

Title and Presenter	Abstract
kindermann@dkrz.de **Carsten Ehbrecht (DKRZ)** *ehbrecht@dkrz.de* **Ag Stephens (CEDA)** ag.stephens@stfc.ac.uk **Björn Brötz (DKRZ)** *Bjoern.Broetz@dlr.de* **Wim Som de Cerff (KNMI)** *wim.som.de.cerff@knmi.nl* **Maarten Plieger (KNMI)** *maarten.plieger@knmi.nl* **Sébastien Denvil (ENES/IPSL)** *sebastien.denvil@ipsl.jussieu.fr* *Poster & Demo*	and IPSL. To provide data near processing services, a new service component will be developed and deployed near to the data archives. These services are supporting OGC WPS standardized interfaces and thus are supported by a wide range of different client tools and applications. We will provide an overview of the status of the Copernicus processing approach, including software packaging and deployment as well WPS development and deployment, which is based on the Birdhouse [2] open-source initiative. With respect to the processing codes to be made available via WPS, we are concentrating on climate data evaluation packages developed as part of the Copernicus C3S-Magic project. Key cornerstones of the approach presented are: • A generic software packaging and deployment solution based on Conda and Docker • A generic WPS component system supporting the flexible generation and deployment of WPS standardized web services (Birdhouse based) • Support for parallel processing clusters based on different batch systems (e.g., SLURM, GridEngine) A demo of a first test deployment will be presented. [1] https://www.ecmwf.int/en/about/what-we-do/environmental-services/copernicus-climate-change-service [2] http://birdhouse.readthedocs.io/en/latest/
Diagnostics Package for the E3SM Model **Chengzhu Zhang (DOE/LLNL/AIMS)** *zhang40@llnl.gov* **Zeshawn Shaheen (DOE/LLNL/AIMS)** *shaheen2@llnl.gov* **Chris Golaz (DOE/E3SM)** *golaz1@llnl.gov* **Jerry Potter (NASA/GSFC)** *gerald.potter@nasa.gov* *Poster & Demo*	A new E3SM diagnostics package has been developed by the E3SM Workflow team to build a comprehensive diagnostics software that facilitates the diagnosis of the next-generation Earth system models. This package is embedded into the E3SM Automated Workflow for seamless transition between model run and diagnostics. This software is designed in a flexible, modular, and object-oriented fashion, enabling users to manipulate different processes in a diagnostics workflow. Numerous configuration options for metrics computation (i.e., regridding options) and visualization (i.e., graphical backend, color map, contour levels) are customizable. Built-in functions to generate derived variables and select diagnostics regions are supported and can be easily expanded. The architecture of this package follows the Community Diagnostics Package (CDP) framework, which is also applied by two other DOE-funded diagnostics efforts (Program for Climate Model Diagnosis and Intercomparison [PCMDI] metrics package and Atmospheric Radiation Measurement [ARM] diagnostics package), to facilitate effective interactions between different projects.
ESGF Errata Service **Guillaume Levavasseur (ENES/IPSL)** *glipsl@ipsl.jussieu.fr* **Atef Ben-nasser (ENES/IPSL)** *abennasser@ipsl.fr* **Mark A. Greenslade (ENES/IPSL)** *momipsl@ipsl.jussieu.fr* *Demo*	Due to the inherent complexity of the experimental protocols of projects such as CMIP5 and CMIP6, it becomes important to record and track reasons for dataset version changes. The IPSL is finalizing a new ESGF Errata Service, currently under test phase at http://test.errata.es-doc.org/, in order to: • provide timely information about known issues within the Earth System Documentation (ES-DOC) ecosystem; • allow identified and authorized actors to create, update, and close an issue using lightweight client (https://es-doc.github.io/esdoc-errata-client/); and • enable users to query about modifications and/or corrections applied to the data in different ways through a dedicated API (https://es-doc.github.io/esdoc-errata-client/api.html). The Errata Service exploits the Persistent IDentifier (PID) attached to each dataset during the ESGF publication process. The PIDs request the Handle Service to get the version history of a (set of) file/dataset(s). Consequently, IPSL is closely working with DKRZ on the required connections and APIs between both services. A first demonstration of the service has been very well received from the ESGF developer community. A release candidate of the service is currently delivering to potential users with a goal of deploying into production before the end of the year. This release will include two improvements: • pyessv Controlled Vocabulary Manager, and

Title and Presenter	Abstract
	• issue registration support for any ESGF project.
DREAM Data Services for Biological Data and Beyond **Sasha Ames (DOE/LLNL/AIMS)** *ames4@llnl.gov* **Luca Cinquini (NASA/JPL)** *Luca.Cinquini@jpl.nasa.gov* **Dean N. Williams (DOE/LLNL/AIMS)** *williams13@llnl.gov* *Poster & Demo*	In this poster and demo, we introduce an alternate data service for DREAM. The THREDDS data server (TDS) has been very effective for serving the netCDF data published in the ESGF. However, we need a service more specific for alternate data (e.g., ASCII-based) in other domains, such as the FASTA format used in bioinformatics to represent genomic and protein sequences. This service will allow a variety of content types to interoperate properly with a user’s web browser. We will also show how non-netCDF data are published. Future work for the service will include random access for FASTA.
Community Data Analysis Tools **Charles Doutriaux (DOE/LLNL/AIMS)** *doutriaux1@llnl.gov* **Denis Nadeau (DOE/LLNL/AIMS)** *nadeau1@llnl.gov* **Dan Lipsa (Kitware)** *dan.lipsa@Kitware.com* **Dean N. Williams (DOE/LLNL/AIMS)** *williams13@llnl.gov* **Aashish Chaudhary (Kitware)** *aashish.chaudhary@kitware.com* *Poster & Demo*	The CDAT is an open-source, Python-based suite of tools designed to provide many of the basic capabilities needed for validating, comparing, and diagnosing scientific data, with an emphasis on climate model behavior. It can be controlled either interactively or via a script file, or control can alternate between these modes during a session. Its strengths are that it allows users to: (1) build end-to-end complex data analysis and visualization workflows that use predefined components for data transformations; (2) collect data from disparate data sources; and (3) ingest user-defined local and remote processing steps. CDAT’s success can also be measured by its expanding use. It is now integrated with the international ESGF peer-to-peer enterprise system as a front-end access mechanism to acquire data for analysis and visualization and as a prototype backend tool to reduce data sets and return visualization products. It is also expanding into other DOE-, National Oceanic and Atmospheric Administration (NOAA)– and NASA-funded projects as the cornerstone of interagency proposed projects. DOE’s E3SM project aims to use CDAT to deliver new capabilities that will further facilitate interactive and visual exploration and diagnostics of simulation and observational output. This project shares a collaborative vision for large-scale visualization and analysis of climate data and is working to organize and expand CDAT’s capabilities. The design of CDAT incorporates the following requirements: • Interactive and batch operations. • Workflow analysis and provenance management. • Parallel visualization and analysis tools (exploiting parallel I/O). • Local and remote visualization and data access. • Comparative visualization and statistical analyses. • Robust tools for regridding, projection, data subsetting, and aggregation. • Support for unstructured grids and non-gridded observational data, including geospatial formats often used for observational data sets. The CDAT offers capabilities for climate scientists to manage big data analytics, sensitivity analyses, heterogeneous data sources, and multiple disciplinary domains, incorporating existing software components in combinations that were previously difficult or even impossible. The CDAT framework addresses challenges in analysis and visualization and incorporates new opportunities, including parallelism for better efficiency, higher speed, and more accurate scientific inferences. Today, the open-source CDAT provides hundreds of users access to increasing 1D, 2D, and 3D analysis and visualization products on many different operating system platforms (i.e., Linux/Unix, Windows, and Mac OSX).
Visual Community Data Analysis Tools (vCDAT) **Matthew Harris (DOE/LLNL/AIMS)** *harris112@llnl.gov* **Dan Lipsa (Kitware)** *dan.lipsa@kitware.com* **James Crean (DOE/LLNL/AIMS)**	Parallel computing, workflows and provenance, exploratory analysis, big data processing for analysis, interactive analysis and visualization, and web informatics are some of the key features of the overall CDAT framework. To support these features, CDAT utilizes core technologies from open-source toolkits such as VTK, R, NumPy, SciPy, and a host of others. In its current format, the Visual CDAT (vCDAT) sits on the web server and provides a Python-based API, which provides the ability to read data from local or remote sources, run analysis algorithms on local or remote computing resources in serial or parallel mode, and visualize algorithm output in a thick client (e.g., desktop graphical UI [GUI]) or a smart client (e.g., web browser). CDAT can use this computing server-side horsepower of a cluster or a supercomputer. The ability to connect to other instances of CDAT compute nodes is under development. On the client side, the deprecated

Title and Presenter	Abstract
crean2@llnl.gov **Matthew Ma (Kitware)** *matthew.ma@Kitware.com* **Charles Doutriaux (DOE/LLNL/AIMS)** *doutriaux1@llnl.gov* **Dean N. Williams (DOE/LLNL/AIMS)** *williams13@llnl.gov* **Aashish Chaudhary (Kitware)** *aashish.chaudhary@kitware.com* *Poster & Demo*	desktop GUI used the CDAT Python API, whereas communication between its smart client replacement and the Python framework uses the latest in web technologies, such as web-sockets and a RESTful API. Our web-based analysis and visualization system, vCDAT, uses the traditional client–server architecture concept within the web-based model. It is similar to the thick-client concept in that the vCDAT smart clients are Internet-connected devices that allow a user's local applications to interact with server-based applications through the use of web services. This allows for more analysis and visualization interaction and software customization but without the hassle of software downloads and installation.
Integrating ES-DOC with the ESG Publisher **Alan Iwi (ENES/CEDA)** *alan.iwi@stfc.ac.uk* **David Hassell (NCAS/UoR)** *david.hassell@ncas.ac.uk* **Mark A. Greenslade (ENES/IPSL)** *momipsl@ipsl.jussieu.fr* **Ag Stephens (ENES/CEDA)** *ag.stephens@stfc.ac.uk* *Poster & Demo*	The ES-DOC ecosystem has the capacity to capture and deliver essential information about climate modeling activities. Within CMIP6, scientists are describing their models and experiments in detail using a rich semantic model (Common Information Model [CIM] 2). Additionally, ES-DOC requires information about ensemble runs and each individual simulation. The extensive global metadata in CMIP6 netCDF data files will provide enough information to allow the ensemble and simulation records to be generated by scanning the file system directly. A command-line tool and Python library, cdf2cim, has been developed to manage the file-scanning, serialization to JSON, and upload to the ES-DOC server. The ESG Publisher captures information from data files to generate aggregations and metadata summaries suitable for publishing to various sources, including THREDDS and the ESGF Search system. Since all CMIP6 data (in the ESGF) will pass through the Publisher, it was considered appropriate to interface with cdf2cim in order to generate CIM 2 for ES-DOC. Staff at IPSL, the National Centre for Atmospheric Science (NCAS)/University of Reading (UoR), and STFC CEDA collaborated on building an extension to the Publisher that automates the generation of CIM 2 metadata and sends it to the server. We will describe how data node managers will work with these tools for CMIP6. This includes the use of GitHub tokens to authenticate with the ES-DOC server. This solution further integrates the publication of data and metadata from detailed climate simulations. Beyond CMIP6, this approach would be applicable to other similar projects. ESG Publisher: https://esgf.github.io/esg-publisher/ CIM 2: https://es-doc.org/cim CDF2CIM: https://es-doc.org/utility-library-cdf2cim
Compute Working Team End-User Application Programming Interface **Jason Boutte (DOE/LLNL/AIMS)** *boutte3@llnl.gov* **Charles Doutriaux (DOE/LLNL/AIMS)** *doutriaux1@llnl.gov* *Poster & Demo*	The ESGF CWT end-user API was created to leverage the power of the WPS interface standard. A WPS server can expose large-scale computational processes and data reduction that are location agnostic, allowing the computations and reductions to be performed where the data reside, thus saving bandwidth and time. In order to execute a WPS process, a user would normally be confronted with lengthy and intricate URLs. To simplify the task of using a WPS process, a well-defined climatology-specific API was planned and an object-oriented Python end-user API and server implementation were created. With the API, users are eased into adoption of these WPS processes.

Title and Presenter	Abstract
A Compliance-Checking Framework for CMIP7 **Ag Stephens (ENES/CEDA)** *ag.stephens@stfc.ac.uk* **Antony Wilson (STFC)** *antony.wilson@stfc.ac.uk* **Guillaume Levavasseur (ENES/IPSL)** *glipsl@ipsl.jussieu.fr* *Demo*	The activity known as "compliance-checking" is distinct from "quality control/assurance" in that it is not concerned with the scientific credibility of the results but aims to check that data files adhere to a set of rules associated with a given project. There are many tools for compliance-checking that perform a very useful function for specific projects. However, it is commonplace for the written specification for a project to diverge from the software implementation. A prototype compliance-checking "framework" is described that draws on the positives of its predecessors and is suggested as a model suitable for future MIPs. Built on the existing IOOS compliance-checker, which employs a plugin architecture per project, the framework attempts to provide a clear separation of concerns between the code implementing the "checks" and the project-specific configuration. In the main library, each check is encoded in a Python class, including documentation (to describe what the check does), messages (to report successes/failures), and modifier parameters (so that each check has some defined flexibility). On the configuration side, a separate library (compliance-check-maker) is concerned with generating the specific checks to be employed by a project. Each project writes a set of YAML files to describe a group of checks (e.g., to check global attributes or file names). The code then generates a Python plugin for the IOOS checker as well as a compliance specification document for the project. The latter is an essential feature of the framework, generating both the checks and specification from a single information source. The modular approach employed by the framework makes it highly adaptable to different use cases, and the community is encouraged to add to the collection of supported checks. As the tool develops, it should take hours, rather than days or months, to configure a set of checks for a new project. IOOS compliance-checker: https://github.com/ioos/compliance-checker
Google Earth Engine and Project Jupyter **Tyler Erickson (Google)** *tylere@google.com* *Demo*	The volume of Earth science data generated from models and by sensors (particularly those on satellites) continues to increase. For many analyses, managing this large volume of data is a barrier to progress, as it is difficult to explore and analyze large volumes of data using the traditional paradigm of first downloading datasets to a local computer. Furthermore, methods are needed that communicate Earth science algorithms that operate on large datasets in an easily understandable and replicable way. This demo will highlight two technologies: • Google Earth Engine – a cloud-based geospatial analysis platform that provides access to petabytes of Earth science data and hundreds of geospatial operators via a JavaScript or Python API. • Project Jupyter – an open-source project that supports interactive data science and scientific computing, including the Jupyter Notebook, a web-based environment that supports documents that combine code and computational results with text narrative, mathematics, images, and other media. The technologies will be demonstrated by calculating climate indices from downscaled climate projections based on CCl/CLIVAR/JCOMM Expert Team on Climate Change Detection and Indices (ETCCDI).

Day 2: Wednesday, 6 December 2017
ESGF Focus Areas

Title and Presenter	Abstract
International Climate Network Working Group, Replication / Versioning and Data Transfer Working Team Plenary **Eli Dart (DOE/ESnet)** *dart@es.net*	Efficient CMIP6 data analysis depends on the transfer and replication of high-volume data sets to data centers around the world. These data centers manage replica pools to support their user communities by, for example, redistributing the data or by providing data near processing facilities. The data transfer and replication are integrated into a complex workflow involving file systems, local networks, wide area networks, as well as dedicated data transfer nodes (DTNs), which are integrated into a data pipeline managed by dedicated data replication software installed at sites.

Title and Presenter	Abstract
Lukasz Lacinski (DOE/ANL) *lukasz@uchicago.edu* **Stephan Kindermann (ENES/DKRZ)** *kindermann@dkrz.de*	This session will provide an overview of the current status of the overall CMIP6 data replication pipeline and its different (technical and organizational) aspects. The session especially concentrates on: • Data transfer and replication: o Status and progress so far o Current problems • CMIP6 replication strategy: o Status of current discussion o Planning of the international data replication to well-established “CMIP6 data hubs” such that, for instance, transatlantic connections can be exploited efficiently • Short-term action planning to support CMIP6 initially o Improve single-stream bandwidths to CMIP6 data servers from DTNs o Configuration issues at sites o Data publication to support download via DTNs • Long-term action planning to support CMIP6+ in the future o Exploit Globus Transfer in replication pipeline o Expand DTN deployments to match data scale
Compute and Data Analytics Working Team Plenary **Charles Doutriaux (DOE/LLNL/AIMS)** *doutriaux1@llnl.gov* **Daniel Duffy (NASA/GSFC)** *daniel.q.duffy@nasa.gov*	**Charles Doutriaux and Daniel Duffy—Presentation on server-side computing progress** **Abstract** The ESGF’s main goal is to facilitate advancements in Earth system science with a primary mission of supporting CMIP activities. In preparation for emerging data analysis needs, such as future climate assessments, the CWT has been working to provide data-proximal analytics capabilities through the development of server-side APIs and client-side (end-user) APIs. This talk will provide a brief overview of ongoing development projects focused on the implementation of ESGF server-side analytics and discuss future goals of the working team. **Cameron Christensen, Giorgio Scorzelli, Peer-Timo Bremer, Shusen Liu, Ji-Woo Lee, Brian Summa, Valerio Pascucci - Interactive Analysis and Visualization of Arbitrarily Large, Disparately Located Climate Data Ensembles Using a Progressive Runtime Server, On-Demand Data Conversion, and an Embedded Domain Specific Language Suitable for Incremental Computation** **Abstract** Massive datasets are becoming more common due to increasingly detailed simulations and higher resolution acquisition devices. Yet accessing and processing these huge data collections for scientific analysis is still a significant challenge. Solutions that rely on extensive data transfers are increasingly untenable and often impossible due to lack of sufficient storage at the client site as well as insufficient bandwidth to conduct such large transfers, that in some cases could entail petabytes of data. Large-scale remote computing resources can be useful, but utilizing such systems typically entails some form of offline batch processing with long delays, data replications, and substantial cost for any mistakes. Both types of workflows can severely limit the flexible exploration and rapid evaluation of new hypotheses that are crucial to the scientific process and thereby impede scientific discovery. To facilitate interactive analysis and visualization of these data ensembles, we introduce a dynamic runtime system suitable for progressive computation and interactive visualization of arbitrarily large, disparately located spatiotemporal datasets. This system is based on the streaming IDX data format, which utilizes an hierarchical z-order to facilitate fast loading of coarse resolution data as well as better spatial locality for more efficient sub-region reads. We provide an on-demand IDX data server to enable access to existing datasets, designed to provide streaming hierarchical versions of equivalent NetCDF climate data volumes in a user-directed manner such that specific timestep fields are converted just-in-time. This permits the bulk of the data to remain on the server and facilitates interactive analysis and visualization by immediately sending results for specific data requests. Initial conversions are cached for future use, amortizing the cost across successive requests.

Title and Presenter	Abstract
	Our system includes an embedded domain-specific language (EDSL) that allows users to express a wide range of data analysis operations in a simple and abstract manner. The underlying runtime system transparently resolves issues such as remote data access and resampling while at the same time maintaining interactivity through progressive and interruptible processing. Computations involving large amounts of data can be performed remotely in an incremental fashion that dramatically reduces data movement, while the client receives updates progressively, thereby remaining robust to fluctuating network latency or limited bandwidth. This system is integrated with the ESGF software stack using a docker-based deployment, and facilitates interactive, incremental analysis and visualization of massive remote datasets up to petabytes in size.
Identity Entitlement Access Working Team Plenary **Phil Kershaw (ENES/CEDA)** *philip.kershaw@stfc.ac.uk* **Lukasz Lacinski (DOE/ANL)** *lukasz@uchicago.edu*	Over the past year, the Identity Entitlement Access (IdEA) working team has focused on the packaging of OAuth 2.0 support into the ESGF release in order to replace the legacy OpenID 2.0 system and bring new capabilities to applications and services in the operational federation. This work has centered on two main aspects: (1) the incorporation of the standalone OAuth 2.0 Authorization and Resource services implementation from CEDA, and (2) work to embed OAuth 2.0 with the various dependent ESGF services—the compute, index, and data nodes. We will describe the new core identity services including the Short-Lived Credential Service (SLCS), which with OAuth 2.0 provides a means to get delegated certificates. We will also describe the provision of the various integration hooks to other ESGF services: the refactoring needed to integrate OAuth with the ORP—the access control filter system overlaying the THREDDS Data Server—and the development of a generic Python OAuth Client package by Argonne National Laboratory (ANL). In addition to integration with the ESGF installer, work with the Jet Propulsion Laboratory (JPL) has been undertaken to run the OAuth, SLCS, and dependent services as Docker containers. We will review the roadmap for making this and other functionality operational and outline our plans for the future evolution of the IdEA architecture.
Status Update and Future Planning for the ESGF UI, Search, and Dashboard Working Teams Plenary **Luca Cinquini (NASA/JPL)** *Luca.Cinquini@jpl.nasa.gov* **Guillaume Levavasseur (ENES/IPSL)** *glipsl@ipsl.jussieu.fr* **Alessandra Nuzzo (ENES/CMCC)** *alessandra.nuzzo@cmcc.it*	This presentation will provide a progress report and future roadmap for the recently unified working group that includes the CoG UI, the search backend services, and the Dashboard and metrics functionality. CoG development has been focused on integrating new features in support of critical community projects such as the upcoming CMIP6 and the ongoing obs4MIPs. If funding is provided, we plan to completely re-factor the CoG software to enhance its modularity, functionality, security, and extensibility. Also, because the front-end is more and more requested by non-scientific users from different backgrounds, future efforts must lead to a friendlier interface with intuitive layouts and helpful tutorials. The backend search services have been mostly stable, with some development effort again focused on supporting CMIP6 features, as well as addressing newly discovered security vulnerabilities. Future work must address necessary Solr upgrades and perhaps moving to Solr Cloud. The dashboard UI application has been completely re-written since the ESGF shut-down. The dashboard is deployed as an information provider as part of each data node, and since the last release it includes a RESTful API. Federation-level metrics are provided by aggregator applications that will be deployed at selected Tier 1 sites.
Installation and Software Security Working Team Plenary **William Hill (DOE/LLNL/AIMS)** *hill119@llnl.gov* **Sasha Ames (DOE/LLNL/AIMS)** *ames4@llnl.gov* **Prashanth Dwarakanath (ENES/Liu)** *pchengi@nsc.liu.se* **Luca Cinquini (NASA/JPL)** *Luca.Cinquini@jpl.nasa.gov* **George Rumney (NASA/GSFC)**	**William Hill (DOE/LLNL/AIMS) Sasha Ames (DOE/LLNL/AIMS) and Prashanth Dwarakanath (ENES/Liu) – Software Installation** **Abstract:** This presentation will provide an update on the current state of the ESGF installation process and the future direction of the installer. The installer version 2.5.13 is currently written as a collection of Bash scripts. Steady progress has been made in porting these Bash scripts over to Python. Refactoring the script to Python will benefit both the ESGF developers and users. Python enhances the codebase’s readability, maintainability, and testability, thus speeding up the development cycle for future releases. The code will be written to be more modular in structure. Additionally, the refactor opens the possibility for creating a web interface for the installer.

Title and Presenter	Abstract
george.rumney@nasa.gov	**Luca Cinquini (NASA/JPL) – Software Container (i.e., Docker)** **Abstract:** This presentation will report on the current state of the effort to design and implement a next-generation ESGF architecture based on Docker containers. Such a model presents great advantages with respect to the current "monolithic" architecture supported by the shell-based installer, such as easier to install and upgrade, scalable onto multiple hosts, and deployable both on internal clusters and commercial Cloud. This work has been so far supported by the DREAM project, and is now joining forces with the new European Copernicus project. **George Rumney (NASA/GSFC), Daniel Duffy (NASA/GSFC), Luca Cinquini (NASA/JPL), and Dean N. Williams (DOE/LLNL/AIMS) – Software Security Working Team Overview** **Abstract:** The Executive Committee of the ESGF chartered a Software Security Working Team (SSWT) to oversee the security of the ESGF software stack and to provide guidance for a continuous improvement path consistent with federal controls. The SSWT maintains the security review procedure for all ESGF software releases and is responsible for ensuring best practices are maintained across the federation. For more information, the software security plan can be found at http://esgf.llnl.gov/media/pdf/ESGF-Software-Security-Plan-V1.0.pdf. While progress has been made, significant challenges remain within such a complex software stack. This short talk will highlight those challenges, the need for more involvement across the community, and near-term goals.

Day 3: Thursday, 7 December 2017
Coordinated Efforts with Community Software Projects

Title and Presenter	Abstract
Publication, Quality Control, Metadata, and Provenance Capture Working Team Plenary **Sasha Ames (DOE/LLNL/AIMS)** *ames4@llnl.gov* **Heinz-Dieter Hollweg (ENES/DKRZ)** *hollweg@dkrz.de* **Bibi Raju** (PNNL) mailto:bibi.raju@pnnl.gov	**Sasha Ames (DOE/LLNL/AIMS) – Publication Progress** **Abstract:** The ESGF publication team supports many different projects, although the most recent focus has been on CMIP6 readiness. This talk will give an overview of the current state of the process and our future directions. CMIP6 is an example of a project where the Publisher is enabled for attribute-controlled vocabulary checking. Another example project, input4MIPs, make use of several unconventional features. These projects introduced the PID assignment functionality aspect of the process. Python 3 conversion will become an action item for the next calendar year. **Heinz-Dieter Hollweg (ENES/DKRZ) – Quality Control Progress** **Abstract:** The current state of the QA-DKRZ tool is presented. A Best Practices procedure for installation/update is given as well as configuration, operating the tool for projects like CMIP6 and eventually summarizing the results. **Bibi Raju (PNNL) - Provenance Data harvest and Scientific Results Reproducibility** **Abstract:** Data provenance provides a way for scientists to observe how experimental data originates, conveys process history, and explains influential factors such as experimental rationale and associated environmental factors from system metrics measured at runtime. PNNL developed a provenance harvester that is capable of extracting already existing file based information produced by applications. File based information is extracted and transformed into an intermediate data format inspired in part by W3C CSV on the Web recommendations, called the Harvester Provenance Application Interface (HAPI) syntax. This syntax provides a general means to pre-stage provenance into messages that are both human readable and capable of being

Title and Presenter	Abstract
	written to a provenance store, Provenance Environment (ProvEn). The harvested provenance data can later be retrieved from ProvEn store and can be used for various purposes. This extracted information greatly helps to reproduce a simulation either by the same user or a different user in the same host environment. The harvested provenance information can be also used to compare different application runs. HAPI is being applied to harvest provenance from climate ensemble runs for Energy Exascale Earth System Model (E3SM) project funded under the U.S. Department of Energy's Office of Biological and Environmental Research (BER) Earth System Modeling (ESM) program. E3SM informally provides provenance in a native form through configuration files, directory structures, and log files that contain success/failure indicators, code traces, and performance measurements. HAPI is a generic format and can be applied to harvest provenance from relational database tables as well as other scientific applications that log provenance related information.
Machine Learning Plenary **Sookyung Kim (DOE/LLNL/AIMS)** *kim79@llnl.gov* **Sebastien Denvil (ENES/IPSL)** *sebastien.denvil@ipsl.fr* **Philip Kershaw (ENES/CEDA)** *philip.kershaw@stfc.ac.uk* **Tom Landry (CRIM)** *tom.landry@crim.ca*	**Sookyung Kim (DOE/LLNL/AIMS) – Community Machine Learning** **Abstract:** This presentation will report on the progress of current effort to leverage deep learning techniques to detect, localize, and track extreme climate events using the ESGF framework. Specifically, we present recent results of the system we developed to detect and locate extreme climate events by Convolutional Neural Networks (CNNs). Our system can capture the pattern of extreme climate events from pre-existing coarse reanalysis data corresponding to only 16,000 grid points—without an expensive downscaling process and with fewer than 5 hours to training using 5-layered CNNs. As the use case of our framework, we tested tropical cyclone detection with labeled reanalysis data and achieved 99.98% of detection accuracy with localization accuracy within 4.5 degrees of longitude/latitude. In addition, we will introduce the prototype of the deep learning system to track the extreme climate events by considering spatiotemporal evolution of an event using long short-term memory (LSTM), which can track the event in time-series reanalysis data. **Sébastien Denvil (ENES/IPSL), Sandro Fiore (ENES/CMCC), Philip Kershaw (ENES/CEDA) – Presentation on Copernicus and H2020 Programme Machine Learning Efforts** **Abstract:** The past ten years has been the witness of ancient Machine Learning (ML) and Deep Learning (DL) algorithms awaken. This trend was accompanied by large disruption in so called "*big data*" technology (cloud, GPU, Docker and alike). There are many applications of ML in the field of Earth observations. Those algorithms are very well placed to fill gaps in observations. In the field of modelization, there are a few examples of ML usage for model parameters tuning but the full potential of ML with respect to climate modeling has not yet been fully realized. This talk will discuss how those trends in ML and DL and how the can be leveraged in the climate community and the role ESGF could play. **Tom Landry (CRIM) – Imagery, text and geospatial Machine Learning applications in Montreal's booming ML landscape** **Abstract:** Montreal's technological landscape is currently transformed by an Artificial Intelligence revolution. Several major world-class tech corporations recently opened laboratories and offices in the city. New research chairs, super-clusters and specialized institutes are living proof of increasing provincial and federal public investments in the domain. Prospects for startups are also getting better; talent pool is large and venture capital is receptive. Centre de recherche informatique de Montreal (CRIM) is an applied research center positioned in the middle of academia and industry. Three research teams at CRIM - Vision and Imaging (VISI), Emerging Technologies and Data Science (TESD) and Speech and Text (PATX) - has

Title and Presenter	Abstract
	been delivering and transferring Machine Learning (ML) expertise and applications for several years. We will present of few of our ML projects that are susceptible to be of use for ESGF. VISI demonstrated ML in target detection, classification, super-resolution and filtering from several sensing types: either Synthetic Aperture Radar (SAR), optical satellites, LIDAR, aerial images or video sequences. TESD produced a highly scalable grid-density clustering algorithm for Spark MLLib and unsupervised ML on climate products with SciSpark. The team also works closely with City of Montreal in Smart City scenarios for both descriptive and predictive analytics. PATX's Natural Language Understanding (NLU) expertise recently allowed them to propose future work for climate and EO data. This future work includes metadata annotation for Active Learning, Query Understanding Interface (QUI) and workflow recommendations.
Diagnostics Plenary **Zeshawn Shaheen (DOE/LLNL/AIMS)** *shaheen2@llnl.gov* **Tom Landry (CRIM)** *tom.landry@crim.ca*	**Zeshawn Shaheen (DOE/LLNL/AIMS) – Community Diagnostics Package** **Abstract:** Scientific code is often created for a single, narrowly focused goal. Such code is inflexible and over time may cause progress on a project to reach an impasse. The Analytics and Informatics Management Systems (AIMS) team at Lawrence Livermore National Laboratory (LLNL) is developing the CDP, a framework for creating new diagnostics packages in a generalized manner. Designed in an object-oriented method, CDP allows for a modular implementation of the components required for running diagnostics. The design of CDP consists of modules to handle the user-defined parameters, metrics, provenance, file I/O, output of results, and algorithms for calculating the diagnostics.
	Tom Landry (CRIM) – Presentation on Canada Diagnostics **Abstract:** Our computation framework is largely reliant on Birdhouse and its extensive WPS logging and monitoring capabilities. Most services and processes are called by a client web platform offering its users tools and workspaces. Data access and computation are also currently conducted separately on a Spark cluster at CRIM. Different jobs schedulers were used—for instance, SLURM on high-performance computers and RabbitMQ on hybrid clouds. In order to interoperate between systems, the ESGF CWT API is being evaluated. New implementations are deployed on a staging testbed infrastructure at CRIM, composed of a dozen Open Stack virtual machines. In that testbed, CRIM conducted several Technical Interaction Experiments (TIEs) in the OGC TB-13 EOC thread. Production systems are deployed on Ouranos infrastructure, on a bare metal server located at Calcul Quebec premises. Both CRIM and Ouranos infrastructure are tied to the CANARIE network, the national backbone of Canada’s ultra-high-speed National Research and Education Network (NREN). All services and major constitutive elements of PAVICS are placed in CANARIE’s science gateway registry, where a mandatory REST API documents the component, compiles usage statistics, collects reliability metrics, and notifies administrators of downtime. We improved Birdhouse workflow capabilities and added several data validation tests. Specific workflows are executed regularly to test system integrity of both CRIM and Ouranos data and processing resources. Any change in the outcome of the integration workflows triggers a warning in the team’s #pavics Slack channel, better exposing the system’s state to a concentration of developers. To help manage and monitor its numerous Docker instances, PAVICS uses a lightweight Docker host management tool called Portainer. Queues are monitored and controlled with the Flower library.
	S. Denvil (IPSL), M. Lautenschlager (DKRZ), S. Fiore (CMCC), F. Guglielmo (IPSL), M. Juckes (CEDA), S. Kindermann (DKRZ), M. Kolax (SMHI), C. Pagé (CERFACS,), W. Som de Cerff (KNMI) – Presentation on Copernicus and H2020 Programme Diagnostics Needs and Overview **Abstract:** A diverse set of software, tools, frameworks, and technologies are available around the globe for computing diagnostics relevant for climate modeling simulation results analysis: Birdhouse, Climate4Impact, NCO, ESMValTools, climate data operators, Climaf, vCDAT, WPS, MAGICS, and so on. Diagnostics themselves can be tailored towards climate modelers as well as to scientific researchers interested in climate data products but who are not climate modelers

Title and Presenter	Abstract
	themselves. It can also be tailored to scientific researchers (or non-researchers) from other domains who need those products to assess the impacts of climate change on ecosystems, on economic activities, or for other applications. The fact that such a large diversity exists makes it hard to think that only one diagnostic toolbox “to rule them all” can emerge. The focus of our talk will be to summarize the various needs we can anticipate, to describe diagnostic toolboxes we have at our disposal, and to identify potential gaps. Copernicus, H2020, and national initiatives, past or present, have largely contributed ideas and software packages on this topic. We will give an overview of related ENES activities, together with the implications and potential benefits for EGSF.
CMIP6 Data Node Operations Team (CDNOT) Plenary **Sébastien Denvil (ENES/IPSL)** *sebastien.denvil@ipsl.jussieu.fr*	A CMIP6 Data Node Operations Team (CDNOT) has been appointed by the WGCM Infrastructure Panel (WIP) and will include representatives from groups responsible for CMIP6 ESGF data nodes. The CDNOT is charged with implementing a federation of data nodes responsive to the requirements of the evolving CMIP6 process as articulated by the WIP. The CDNOT scope covers functions and resource issues (hardware, networks, people) related to: installing, configuring and operating all the nodes and node services in the CMIP6 data federation and the policies and processes involved in both managing data on a data node (including acquisition, quality assurance, citation, versioning, and publishing) and providing access services. The talk will summarize ongoing and foreseen actions and will describe the groups objectives.
Node Manager and Tracking / Feedback Notification Plenary **Sasha Ames (DOE/LLNL/AIMS)** *ames4@llnl.gov* **Tobias Weigel (ENES/DKRZ)** *weigel@dkrz.de*	**Sasha Ames (DOE/LLNL/AIMS) – Node Manager** **Abstract:** After operating for several years without a node manager software component, this module has been redeployed with software stack in v2.5.x. This gives the ESGF a registry of components running at several sites and feeds information into the dashboard UI module. Moreover, there are several additional APIs running from the node manager, namely node status pages and one to distribute PID server (RabbitMQ) credentials. We aim to tighten security and to determine additional APIs for the node manager to contain. We will also give an update regarding a subscription service for user notification. **Tobias Weigel (ENES/DKRZ) – PID Services and Tracking/Feedback** **Abstract:** The goal of the ESGF PID services is to record PIDs (Handles) for all files and datasets in CMIP6 and, more recently, for obs4MIPs and CORDEX (Coordinated Regional Climate Downscaling Experiment). PID services for the ESGF consist of multiple components: 1. A message queue federation, based on RabbitMQ installations at DKRZ, IPSL, and PCMDI, which provides failover and load-balancing capacities in view of massive CMIP6 data object numbers. 2. A Python library (esgf-pid) used by the ESG Publisher to create and dispatch PID operation messages. 3. A Java servlet (queue consumer) which runs locally at DKRZ, connected to the PID servers (Handle server deployments) that execute the actual PID operations. This talk will explain these components and their current status and embedding in the overall ESGF workflow, which also touches on concerns of CMOR and the CoG UI. We will also explain a special part of the PID services: the custom collection-building facility geared towards end-users. The talk will also showcase existing and future developments toward providing data tracking through PIDs and providing end-users with tools that help them understand the state of data at hand (e.g., whether new versions are available). Relevance to other efforts such as RDA, EUDAT, and EOSC will be highlighted.
User Support and Documentation Plenary **Matthew Harris (DOE/LLNL/AIMS)**	The ESGF Support Working Team is a collection of people from around the globe who aim to give ESGF users the best experience possible. This team includes representatives from Tier 1 and Tier 2 data and modeling centers, respectively. We have learned a lot over the last few years as the ESGF has had transitions and changes in both the wiki and website. We will cover our experiences and

Title and Presenter	Abstract
harris112@llnl.gov	the direction our group should go to maintain the quality of user experience. The ESGF Documentation Working Team is responsible for managing the documentation generated by other working teams. The team manages esgf.llnl.gov, which offers additional documents, such as for sponsors and committees, acknowledgments, governance, publications, tutorials, supported projects, wikis, and much more. We will cover the usability of our documentation and the direction moving forward using Sphinx and Read the Docs.