2016 ESGF F2F Conference Presentation, Poster and Demonstration Abstracts

2016 Earth System Grid Federation (ESGF) Face-to-Face Conference
(Washington, D.C.)

Day 1: Tuesday, December 6, 2016
ESGF Steering Committee

Title and Presenter	Abstract
The State of the Earth System Grid Federation **Dean N. Williams (DOE/LLNL)** Williams13@llnl.gov	The Earth System Grid Federation (ESGF) has assisted the scientific research community and their projects for over a decade with the dissemination and management of climate simulation and observational data products. Our federated "data cloud" infrastructure houses millions of files and transfers many petabytes of data a year to the community for large-scale knowledge discovery. The large-scale use of the infrastructure has enabled us to amass a great deal of intelligence about the state of our software stack, housed data, and needed capabilities for customer satisfaction. This information allows us to effectively organize, plan, and prioritize the next steps in ESGF software development. To help in our planning phase, we have conducted a survey study around user practices and ESGF node performance and capabilities. From the survey, we also ascertain the extent to which a large sample of projects, regardless of the national or international funding agency, use different types of data quality control, gathering, managing, or sharing methods. The state of ESGF will also feature **federated data usage statistics** generated by the ESGF dashboard and desktop and highlighted ESGF Executive Committee documents, such as the ESGF Policies and Guidelines, the ESGF Strategic Roadmap, the ESGF Software Security Plan, ESGF Implementation Plan, and the ESGF Root Certificate Authorization Policy & Certificate Practices Statement.
Department of Energy Office of Biological and Environmental Research Data Management **Justin Hnilo (DOE/BER)** Justin.Hnilo@science.doe.gov	The Climate and Environmental Sciences Division (CESD) within DOE's Office of Biological and Environmental Research (BER) focuses on advancing a robust predictive understanding of Earth's climate and environmental systems, by exploiting unique modeling, observational, data, and infrastructure assets, developed and managed by BER. CESD's strategic plan includes five goals, each of which contains a modeling, observational, and data management component. Within this plan, there is a special emphasis on leading the nation in developing highly efficient modeling architectures, testbeds, data analytics, and analysis tools, to support the broad climate science community within the context of DOE's mission. The Climate and Environmental Sciences Data Management Activity represents a highly coordinated set of data oriented research activities, with a goal to provide the CESD scientific community with easy and efficient access to all necessary data bases in order to study increasingly complex scientific challenges. Research in support of this activity involves metadata compatibility from disparate research projects; fusion of data derived from laboratory studies, field observatories, model generated output; server-side analysis; development of multi-media analytical tools, including multi-dimensional visualization and efficient storage. Current and future investments will be highlighted.
Infrastructure for the European Network of Earth System Modelling **Sylvie Joussaume (ENES/CNRS-IPSL)** sylvie.joussaume@lsce.ipsl.fr	The European Network for Earth System modeling (ENES) gathers the European community working on climate modeling. Its infrastructure project, IS-ENES, supports the European contribution to ESGF and ES-DOC for the WCRP coordinated experiments for global and regional models, CMIP and CORDEX. The ENES data infrastructure contributes to the development of the ESGF software stack, data quality control, data identification and data citation, data replication and cache maintenance, and dashboard as well as of the metadata tools. It provides support to users of data as well as to providers of data. With the associated European climate modeling groups, it is now preparing for CMIP6 and its large data volume. IS-ENES also aims at facilitating access to model results for the climate impact community by easing the interface to ESGF data through the Climate4Impact portal. This platform provides tools to explore data, compute indices, perform analyses, and provides guidance to users. The ENES community is also engaged to provide access to global projections for the new Copernicus Climate Change Service.
National Aeronautics and Space Administration High-End Computing Program	CMIP datasets had been the most fundamental and critical for climate trend analysis and resiliency applications. As climate models are becoming more comprehensive Earth system models, the requirements for the datasets and data systems are also changing. Under the CMIP6 framework,

Title and Presenter	Abstract
Tsengdar Lee (NASA HQ) tsengdar.j.lee@nasa.gov	there are 21 endorsed "MIPs." How much do we know about the different analyses that will be done for these CMIP6 experiments? How do we enable the research community to interrogate a dataset that is easily PBs in size? We will present the current model development efforts at NASA and our approaches to constrain the models. In addition, we will discuss how NASA is supporting America's National Climate Assessment using the CMIP data.
National Computational Infrastructure **Ben Evans (NCI/ANU)** Ben.Evans@anu.edu.au	How ready is ESGF for the next stage of CMIP activities? With the model data ramping up starting in 2017, the key questions current revolve around production stability, including service pre-service benchmark tests and acceptance, data publishing confirmation, how to ensure data and service quality processes are in place, status report cards, confirmation of software deployment processes and stable and tested release cycle, data management and replication processes, and preparedness for managing user and model group questions. Many nodes will also provide data-intensive environments and fully featured data services that allow users to probe and analyze the data in situ. These services can be extremely valuable, but are they uniform even across the key nodes? Are our loggings and processes in place to ensure that the data is located where it needs to be in response to user demand? I will also touch on some open questions regarding technical architecture and challenges, which we need to consider in addition to the current important and immediate issues for data service.

Day 1: Tuesday, December 6, 2016
ESGF Progress and Interoperability

Title and Presenter	Abstract
CoG User Interface Working Team **Luca Cinquini (NASA/JPL)** Luca.Cinquini@jpl.nasa.gov	Over the past 12 months, the CoG team has worked with the rest of the ESGF collaboration to deploy CoG as the new web front-end of the next generation ESGF software stack. CoG instances are now operational and federated across the ESGF system. Additionally, we have been focusing on implementing several new requirements in support of the upcoming data distribution for the Coupled Model Intercomparison Project, Phase 6 (CMIP6), which will remain our main focus for the next year.
Metadata and Search Working Team **Luca Cinquini (NASA/JPL)** Luca.Cinquini@jpl.nasa.gov	During this past year, the ESGF Search and Publishing services have been expanded to include new functionality needed to support a growing federation of nodes, and the upcoming CMIP6 massive data volumes. New features include publishing to a local (non-shared) index, supporting atomic metadata updates, searching on datasets with date greater or less than a given value, retracting datasets, and improvements in Wget downloads. Additionally, we have been experimenting with using Solr Cloud and a new topology architecture.
Publication Working Team **Sasha Ames (DOE/LLNL)** ames4@llnl.gov	The Publication Working Team has been on track in improving the publisher software and tools to aid in the publication process. The addition of ESGprep marks a major overhaul in the workings of the esg-publisher component, and CMIP6 requirements have necessitated added features for controlled vocabulary and quality checking. We will present details of these efforts and additional features, including the progress in the ingestion service application programmer's interface (API), and ideas for refreshing the current publisher implementation.
Node Manager and Tracking/Feedback Working Team **Sasha Ames (DOE/LLNL)** ames4@llnl.gov	The "old" node manager component that previously ran under Tomcat in the ESGF v1.X is no longer deployed in ESGF v2.X. We have a testable replacement service, based on a two-tier architecture for managing node communication that implements the "registration.xml" for dashboard interoperability and also exports several JSON-based RESTful APIs for gathering node information. The node manager has now been tested at several sites, and we plan to deploy a production version in early 2017. The "Tracking/Feedback" effort focuses on a workflow that comprises several new software service modules whose initial purpose is user notification in the event of the update or retraction of a previously downloaded dataset. We will give a brief overview of the tracking and feedback architecture and software implementation progress. Additionally, we will discuss several additional features, namely, notification for "saved" searches and helpful dataset prediction.
Stats and Dashboard Working Team **Alessandra Nuzzo (ENES/CMCC)** alessandra.nuzzo@cmcc.it **Maria Mirto (ENES/CMCC)**	The Stats and Dashboard Working Team has been on track to improve the metrics modules for the ESGF experiments (mainly, but not only, CMIP5, Obs4MIPS, and CORDEX). Since September 2016, weekly meetings have been mainly focused on requirements validation, testing, and feedback from the testing sites both for the back and the front end and preparation for the coming releases. A short overview about the main results achieved, the status of the activities in the working group, and the plan for the next months will be presented.

Title and Presenter	Abstract
maria.mirto@cmcc.it	
Identity Entitlement Access Management Working Team **Philip Kershaw (ENES/BADC)** philip.kershaw@stfc.ac.uk **Rachana Ananthakrishnan (DOE/ANL)** ranantha@uchicago.edu	Work in the Identity Entitlement Access Management Working Team (IdEA) team this year has focused on integration of OAuth 2.0 for user delegation. Earlier in the year, the Live Access Server (LAS) was linked up with the Centre for Environmental Data Analysis's (CEDA's) OAuth service to implement a delegation flow. The CEDA service has been updated over the course of the year; reviewing the latest Python packages, the decision was made to port the service to use OAuthLib. This allows it to be deployed as a standard Django package and will facilitate future migration to OpenID Connect, which OAuthLib now supports. In addition, the service has been packaged using Ansible to facilitate its integration into ESGF. This will enable easier installation and subsequent roll out by other IdPs in the federation. The next implementation steps involved updating dependent ESGF components so that they can use the new service, including CoG and the ORP in the Data Node. Work has also been done to investigate the steps needed for integration with Globus and the Compute Node. More advanced use cases are also being explored including two-stage delegation for IS-ENES2 Climate4Impacts Portal (KNMI) with the downscaling portal (University of Cantabria).
Compute Working Team **Charles Doutriaux (DOE/LLNL)** doutriaux1@llnl.gov **Daniel Duffy (NASA/GSFC)** daniel.q.duffy@nasa.gov	ESGF's primary goal of the Compute Working Team (CWT) is to facilitate advancements in Earth system science with a primary mission of supporting CMIP activities. In preparation for future climate assessments, the CWT has been working toward a goal of providing server-side analytics capabilities through the development of server-side APIs and client-side (end-user) APIs. This talk will provide an overview of the CWT, current status, and future goals. In addition, we will describe advances made by the CWT on the APIs along with various implementations made over the last year. An overview of projects by NASA, the Ouranos, project and the Euro-Mediterranean Center on Climate Change will be provided. In addition, a demonstration of how the Python client API can be used to access analytics services will be shown.
Errata Service **LEVAVASSEUR Guillaume (ENES/IPSL)** glipsl@ipsl.jussieu.fr **Atef Ben Nasser (ENES/IPSL)** abennasser@ipsl.jussieu.fr **Mark A. Greenslade (ENES/IPSL)** momipsl@ipsl.jussieu.fr **Merret Buurman (ENES/DKRZ)** buurman@dkrz.de **Sébastien Denvil (ENES/IPSL)** sebastien.denvil@ipsl.jussieu.fr **Katharina Berger (ENES/DKRZ)** berger@dkrz.de **Martina Stockhause (ENES/DKRZ)** stockhause@dkrz.de	Recording and tracking the reasons for dataset version changes is important, due to the inherent complexity of the experimental protocols for projects such as CMIP5/6. The currently established system makes it impossible for scientists to know easily whether the data in hands is deprecated and/or replaced and corrected by a newer version. Also it is very difficult to access a description of this issue. The IPSL is finalizing an new ESGF Errata Service in order to: • Provide timely information about known issues. Within the ES-DOC ecosystem, the errata web service front end displays the whole list of known issues. The list can be filtered by several useful parameters, such as the issue severity or status. Three tabs describe each issue, providing (i) the information details, (ii) graphics or pictures to illustrate the issue, and (iii) the list of the affected datasets. • Allow identified and authorized actors to create, update, and close an issue. We developed a piece of software that enables the interaction with the errata service. It can be used to create, update, close, and retrieve issues. The client is basically aimed to be used by publishing teams, so that they can directly describe problems as they are discovered. • Enable users to query about modifications and/or corrections applied to the data in different ways. The errata web service provides an API to query the issue database. The end users can submit one or several files or datasets identifiers to get back all annotations related to each corresponding issue. This search API is also able to retrieve the issues that affect a MIP variable or experiment. To succeed, the Errata Service exploits the Persistent IDentifier (PID) attached to each dataset during the ESGF publication process. The PIDs enable to 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 for the two services. The ESGF implementation of the citation service is coordinated by the ESGF-QCWT. Errata Service development deployment - http://test.errata.es-doc.org/.
Quality Control Working Team: Data Citation Service for CMIP6—Status and Timeline **Martina Stockhause (ENES/DKRZ)** stockhause@dkrz.de	The review of the CMIP6 data citation procedure resulted in the requirement of a citation possibility prior to the long-term archiving of the data in the IPCC DDC (Data Distribution Centre) hosted at DKRZ. The presentation will give an overview over the Data Citation concept, with emphasis on the technical requirements and on the dependencies on other teams' developments. The implementation of the different components will be reviewed to give a status of the service's development and a timeline towards its operability. The ESGF implementation of the citation

Title and Presenter	Abstract
Katharina Berger (ENES/DKRZ) **berger@dkrz.de** ***Guillaume Levavasseur (ENES/IPSL)*** **glipsl@ipsl.jussieu.fr**	service is coordinated by the ESGF-QCWT.
Installation Working Team **Prashanth Dwarakanath (ENES/Liu)** **pchengi@nsc.liu.se**	The installation and maintenance of an ESGF node is unfortunately still a laborious, risky, and over-complicated process. Additionally, the current installation software is not modular enough, and very difficult to evolve in the long-term. On the other hand, easy installation and upgrading of the underlying software stack is critical to ESGF adoption and success. This talk will present a brief analysis of the major shortcomings of the ESGF installation model and outline a few alternatives for transitioning to a more robust and reliable framework.
Docker for ESGF **Luca Cinquini (NASA/JPL)** **Luca.Cinquini@jpl.nasa.gov**	Docker is becoming a mainstream technology for packaging, deploying and operating complex applications on multi-host environments, including in the cloud. This talk will report on an exploratory effort to run an ESGF Node as a set of interacting Docker containers, each running a specific ESGF service. It seems that Docker could be very useful to ESGF in many respects: • It would greatly simplify installing and upgrading an ESGF Node • It would make the installer software much more modular, maintainable, and upgradable • It would allow scaling up of services, as needed, and deployment on the Cloud • It would make adding new services (such as nginx, other Python apps, and other Java web apps) a much simpler process In conclusion—is Docker the future of ESGF?
International Climate Network Working Group **Eli Dart (DOE/ESnet)** **dart@es.net**	The International Climate Network Working Group (ICNWG) is dedicated to improving data transfer performance between the major climate data centers, and from climate data centers to the users of climate model data worldwide. This talk will discuss the efforts of the ICNWG in 2016, and the future in 2017 and beyond.
Data Transfer Working Team **Lukasz Lacinski (DOE/ANL)** **lukasz@uchicago.edu**	The Data Transfer Working Team (DTWT) has worked on both improving data transfer performance and adding new features that simplify transferring datasets to and from the data node. Two new features that can be optionally enabled have been added to the ESGF authorization callout: sharing and write access for users with the publisher role. The DTWT has worked with the CoG User Interface team and Publication team to add Globus URLs as a new method of accessing datasets. The latest ESGF installer provides all of the aforementioned features. DTWT also added a new transport method to Synda that uses Globus Transfer Service.
Security Working Team **George Rumney (NASA/GSFC)** **george.rumney@nasa.gov**	The ESGF Software Security Working Team (SSWT) was established in wake of the multi-site compromise in 2015. Recovery focused on remediation of identified flaws, short-term correction of some software engineering methods, and the creation of a software security plan (http://esgf.llnl.gov/media/pdf/ESGF-Software-Security-Plan-V1.0.pdf). Progress since then has been modest, and significant challenges remain. This talk will highlight the current challenges and near-term goals.
Replication and Versioning Working Team **Stephan Kindermann (ENES/DKRZ)** **kindermann@dkrz.de** **Tobias Weigel (ENES/DKRZ)** **weigel@dkrz.de**	Together with the Data Transfer Working Team and the Networking team (ICNWG), the replication team worked on replication tests between sites as well as on improvements of replication related software components (e.g., Synda). The overall progress in 2016 was mainly hindered by the new set ups of data transmission network–related hardware infrastructure at sites. A short status of the situation is given and plans for 2017 will be discussed. In close collaboration with the publication working team, the versioning procedure for CMIP6 was improved. The PID services team will give an update on the PID-related versioning aspects.
Persistent Identifier Services **Tobias Weigel (ENES/DKRZ)** **weigel@dkrz.de** **Stephan Kindermann (ENES/DKRZ)** **kindermann@dkrz.de** **Katharina Berger (ENES/DKRZ)** **berger@dkrz.de**	The PID Services Team will give a brief update on milestones reached since the last meeting, current status and next activities. Past activities that will be reported on include development of the necessary software components for registering and managing PIDs for CMIP6. The working team has collaborated with the Publisher and Errata teams on this and coordinated the necessary changes with the Climate Model Output Rewriter (CMOR) development team. Upcoming action items include the full operational rollout at multiple sites and the development of dedicated user tools and services or integration of PIDs into existing solutions.
User Working Team	Help pages have completely moved to CoG and have been extended by two new (Authorization for ESGF data access, OPeNDAP). The Wget tutorial has been completely revised.

Title and Presenter	Abstract
Torsten Rathmann (ENES/DKRZ) rathmann@dkrz.de **Matthew Harris (DOE/LLNL)** harris112@llnl.gov	Operative support for users has been continued via mailing list esgf-user@lists.llnl.gov. Compared to the first half of the year 2015, the number of user questions decreased by 47 % (Jan-Jun 2016). More results from support statistics will be shown.

Day 2: Wednesday, December 7, 2016
Advanced Computational Environments and Data Analytics

Title and Presenter	Abstract
Overview of the ESGF Compute Working Team and Target Milestones **Charles Doutriaux (DOE/LLNL)** doutriaux1@llnl.gov **Daniel Duffy (NASA/GSFC)** daniel.q.duffy@nasa.gov **Jason Boutte (DOE/LLNL)** boutte3@llnl.gov **Thomas Maxwell (NASA/GSFC)** thomas.maxwell@nasa.gov **Tom Landry (CRCM)** tom.landry@crim.ca **S. Fiore (ENES/CMCC)** sandro.fiore@cmcc.it **Dean N. Williams (DOE/LLNL)** Williams13@llnl.gov	ESGF's primary goal is to facilitate advancements in earth system science with a primary mission of supporting CMIP activities. In preparation for future climate assessments, CWT has been working toward a goal of providing server-side analytics capabilities through the development of server-side APIs and client-side (end-user) APIs. This talk will provide an overview of the CWT, current status, and future goals. In addition, we will describe advances made by the CWT on the APIs along with various implementations made over the last year. An overview of projects by NASA, the Ouranos, project and the Euro-Mediterranean Center on Climate Change will be provided. In addition, a demonstration of how the Python client API can be used to access analytics services will be shown.
Compute Working Team End-User Application Programmer's Interface ***Jason Boutte (DOE/LLNL)*** ***boutte3@llnl.gov*** ***Charles Doutriaux (DOE/LLNL)*** ***doutriaux1@llnl.gov***	The ESGF Compute Working Team end-user API was created to leverage the power of the Web Processing Service (WPS) interface standard. A WPS server can expose large-scale computational processing to users that are location agnostic, allowing the computations to be performed where the data resides and 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 processes, a well-defined climatology specific API was planned and an object-oriented Python end-user API was created. With the API, users are eased into the use of these WPS processes, allowing them to easily harness the power they provide.
The Climate Data Analytic Services Framework **Thomas Maxwell (NASA/GSFC)** thomas.maxwell@nasa.gov **Daniel Duffy (NASA/GSFC)** daniel.q.duffy@nasa.gov	Faced with unprecedented growth in climate data volume and demand, NASA has developed the Climate Data Analytic Services (CDAS) framework. This framework enables scientists to execute data processing workflows combining common 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 vetted climate data analysis tools (ESMF, CDAT, NCO, etc.). A dynamic caching architecture enables interactive response times. CDAS utilizes Apache Spark for parallelization and a custom array framework for processing huge datasets within limited memory spaces. CDAS services are accessed via a WPS API being developed in collaboration with the ESGF CWT to support server-side analytics for 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. Client packages in Python, Scala, or JavaScript contain everything needed to build and submit CDAS requests. The CDAS architecture brings together the tools, data storage, and high-performance computing required for timely analysis of large-scale datasets, 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 enables decision makers to investigate climate changes around the globe, inspect model trends and variability, and compare multiple reanalysis datasets.
The Ophidia Big Data Analytics Framework **S. Fiore (ENES/CMCC)** sandro.fiore@cmcc.it **C. Doutriaux (DOE/LLNL)** doutriaux1@llnl.gov **J. Boutte (DOE/LLNL)** boutte3@llnl.gov **D. Elia (ENES/CMCC)** donatello.elia@cmcc.it **A. D'Anca (ENES/CMCC)**	The Ophidia project is a research effort on big data analytics facing scientific data analysis challenges in the climate change domain. Ophidia provides declarative, server-side, and parallel data analysis, 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 in 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. Metadata management support is also provided. 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) with a complete set of commands.

Title and Presenter	Abstract
alessandro.danca@cmcc.it C. Palazzo (ENES/CMCC) cosimo.palazzo@cmcc.it D. N. Williams (DOE/LLNL) williams13@llnl.gov G. Aloisio (ENES/CMCC) giovanni.aloisio@unisalento.it	The presentation will give an overview about the new in-memory analytics engine recently released which allows a fast data analysis on large amount of data, outperforming the previous approach based on MySQL servers. Moreover, ongoing activity in the context of the ESGF-CWT Working Team will be presented during the talk.
PAVICS: A Platform to Streamline the Delivery of Climate Services **David Huard (CRCM)** Huard.David@ouranos.ca **Tom Landry (CRCM)** tom.landry@crim.ca	Ouranos is a Montreal-based consortium on regional climatology playing the role of a catalyst for climate adaptation. Beyond creating simulations ensembles with the Canadian Regional Climate Model, we also work on translating climate science into services and products tailored to the needs of decision-makers and scientists from other disciplines. As demand for climate services grows, we felt the need to develop software to speed up and standardize the production of climate scenarios, both for our own needs and those of the climate research community. With funding from the CANARIE research software program, we launched the PAVICS project, one objective of which is to create a web platform to facilitate the distribution of data, streamline standard climate analyses, and serve as a backbone for a variety of tailored web applications and services. Ouranos works closely with CRIM, an IT Applied Research Centre focusing on innovation and collaborative development. In the spirit of the ESGF-Compute Working Team vision, we are working to co-locate the heavy number crunching close to the data stores on the Calcul-Québec (HPC) infrastructure. The system architecture is based on Birdhouse, a collection of independent WPSs manageable as workflows. Birdhouse bundles THREDDS, ncWMS, and OCGIS, as well as identity providers and data sources key to ESGF. PAVICS implements data harvesting, crawling, and updates to SOLR. It also offer search capabilities found in ESGF Search's RESTful API. We are also integrating geospatial management and processing capabilities from GeoServer. This server is used to store region definitions (Bukovski regions, countries and states, watersheds), apply geometrical transformations through WPS (union, buffer, polygonize, rasterize), and request base layers or specific records (WMS, WFS). Beyond a public list of common regions, users will be able to upload custom regions and maintain their own collection for later use. Most of the user interfaces rely on selected WPS services and workflow schema to automatically create the necessary widgets to hold inputs for the climate analyses and show their results. Web-based tools and widgets are developed with modern web frameworks such as React-Redux, OpenLayers 3, Cesium, and Plotly. Services required to build climate scenarios will be created (bias correction, spatial analogs, etc.), as well as tools to build, archive, and run workflows.
Server-Side Computing Services provided by IS-ENES through the Climate4Impact Platform **Christian Pagé (ENES/IPSL)** christian.page@cerfacs.fr **Maarten Plieger (ENES/KNMI)** maarten.plieger@knmi.nl **Wim Som De Cerff (ENES/KNMI)** wim.som.de.cerff@knmi.nl **Manuel Vega (ENES/U. of Cantabria)** manuel.vega@unican.es **Sandro Fiore (ENES/CMCC)** sandro.fiore@cmcc.it	Within the FP7 European projects IS-ENES/IS-ENES2 that work with the European climate model data infrastructure, a web portal tailored for climate change impact communities is being developed, called climate4impact (C4I). It has evolved from a climate web portal to a platform offering standard Open Geospatial Consortium (OGC) services that can be used to build targeted and specific climate data portals. One of the services made available by C4I is server-side computing of climate indices and simple statistics, through the use of the python package icclim developed within the IS-ENES2 and CLIPC European projects. Accessing icclim services is done using OGC WPS processes. This makes possible for users to perform first-step or final analyses and data reduction on the C4I server prior to download and/or visualization. The aims and objectives here aim toward a stronger integration between the developments going on within IS-ENES/C4I/icclim, the Copernicus CLIPC project, and the API that is being developed within the ESGF Computing Working Team. Some possible future integration with EUDAT Services will also be discussed.
CAFE: A framework for collaborative analysis of distributed environmental data **Hao Xu (China/Tsinghua University)** xuhao13@mails.tsinghua.edu.cn ***Sha Li*** **(China/Tsinghua University)**	As the amount of information about our environment expands exponentially on a global scale, researchers are challenged to remain efficient when analyzing data maintained in multiple data centers. In this presentation, we present a new software package named "Collaborative Analysis Framework for Environmental Data" (CAFE). CAFE is dedicated for collaborative analysis of large volumes of distributed environmental data. It is designed to execute analytic functions on the node where the data are stored. Multiple nodes can collaborate with each other to perform complex data analysis. A web-based user interface allows researchers to search for data of interest, submit analytic tasks, check the status of tasks, visualize analysis results, and download those results. Compared with existing web-based environments data analysis systems, CAFE reduced

Title and Presenter	Abstract
lis14@mails.tsinghua.edu.cn **Wenhao Dong** **(China/Tsinghua University)** dongwh12@mails.tsinghua.edu.cn **Wenyu Huang** **(China/Tsinghua University)** huangwenyu@tsinghua.edu.cn **Shiming Xu** **(China/Tsinghua University)** xusm@tsinghua.edu.cn **Yanluan Lin** **(China/Tsinghua University)** yanluan@mail.tsinghua.edu.cn **Bin Wang** **(China/Tsinghua University)** wab@tsinghua.edu.cn **Fanghua Wu** **(China/Tsinghua University)** wufh@cma.gov.cn **Xiaoge Xin** **(China/Tsinghua University)** xinxg@cma.gov.cn **Li Zhang** **(China/Tsinghua University)** zhangli@cma.gov.cn **Zaizhi Wang** **(China/Tsinghua University)** wzz@cma.gov.cn **Tongwen Wu** **(China/Tsinghua University)** twwu@cma.gov.cn **Yuqi Bai** **(China/Tsinghua University)** yuqibai@mail.tsinghua.edu.cn	dramatically the amount of data that had to be transmitted from data centers to researchers. CAFE demonstrates great promise for enabling seamless collaboration among multiple data centers and for facilitating overall research efficiency in scientific data analysis.
Embedded Domain-Specific Language and Runtime System for Progressive Spatiotemporal Data Analysis and Visualization **Cameron Christensen (U. of Utah)** cam@sci.utah.edu **Shusen Liu, Giorgio Scorzelli (DOE/LLNL)** scorzelli2@llnl.gov **Ji-Woo Lee (DOE/LLNL)** lee1043@llnl.gov **Peer-Timo Bremer (DOE/LLNL)** bremer5@llnl.gov **Valerio Pascucci (U. of Utah)** pascucci@sci.utah.edu	As our ability to generate large and complex climate simulation datasets grows, accessing and processing these massive data collections is increasingly becoming the primary bottleneck in scientific analysis. Challenges include retrieving, converting, resampling, and combining remote and often disparately located data ensembles with only limited support from existing tools. In particular, existing solutions predominantly rely on extensive data transfers or large-scale remote computing resources, both of which are inherently offline processes with long delays and substantial repercussions for any mistakes. Such workflows impede scientific discovery by severely limiting the flexible exploration and rapid evaluation of new hypotheses that are crucial to the scientific process. We present an embedded domain-specific language (EDSL) specifically designed for the interactive exploration of large-scale, remote data. Our EDSL 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 computation. This allows, for the first time, interactive remote exploration of massive datasets, such as the 7 km NASA GEOS-5 Nature Run simulation, which previously have only been analyzed offline or at reduced resolution.

Day 2: Wednesday, December 7, 2016
Coordinated Efforts with Community Software Projects

Title and Presenter	Abstract
CMIP6 Standards Enabling Management, Search and Interpretation of Model Output	As specifications for CMIP6 model output and metadata become finalized, the requirements for ESGF have become clearer. A brief review of the CMIP6 requirements will emphasize areas perceived to be possibly problematic. A variety of ESGF software services will rely on certain

Title and Presenter	Abstract
Karl Taylor (DOE/LLNL/PCMDI) taylor13@llnl.gov **Paul J. Durack (DOE/LLNL/PCMDI)** pauldurack@llnl.gov **Denis Nadeau (DOE/LLNL)** nadeau1@llnl.gov **Sasha Ames (DOE/LLNL)** ames4@llnl.gov	global attributes that identify and describe essential aspects of model simulations. These global attributes must be drawn from controlled vocabularies (CVs), which have been defined, for example, for model and experiment names, grid descriptions, and frequencies. An overview will be provided describing how the global attributes will be used to construct file names and directory structures, as well as their use in defining the CMIP6 Data Reference Syntax, which enables faceted searches and links to model and experiment documentation. Despite the growth of CMIP (now with nearly 250 experiments planned), the strictly enforced data requirements, the expanded capabilities of the CMIP-supporting infrastructure, and the increased emphasis on transparency (e.g., via web-based services for sharing code and exposing issues) together promise to serve an expanding community of scientists and stakeholders with an interest in climate and climate change.

CMIP6 ESGF Tier 1 & Tier 2 Nodes
 	
Sébastien Denvil (ENES/IPSL, sebastien.denvil@ipsl.jussieu.fr***)*** **Michael Lautenschlager (ENES/DKRZ)** lautenschlager@dkrz.de
	The ESGF Executive Committee tasked Michael Lautenschlager and Sébastien Denvil to collect and discuss Tier 1 and Tier 2 data node requirements for ESGF. After discussions and iterations within the ENES Data Task Force, we came up with an initial plan. We collected feedbacks from group like the CDNOT (CMIP Data Node Operation Team), and we will present during this talk the main outcomes of this process.
	Tier 1 data node requirement for the ESGF infrastructure will cover the level of service (90–95% uptime), installation of the full software stack, contribution to development and maintenance, support for Tier 2 data nodes, and support for data providers. The level of service needs to be at the core of those requirements. For example:
	• NAGIOS-like monitoring for certificates (host and Globus certificates)
	• Use NAGIOS to monitor ESGF nodes and to guard against:
	  ○ Expired certificates
	  ○ http/https endpoints unavailable
	  ○ GridFTP endpoints unavailable
	  ○ Etc.
	• Tier 1 node shall take the responsibility to monitor data node publishing to them
	• Tier 1 node shall monitor themselves
	For Tier 1, requirements for data projects (i.e., CMIP6) will cover the following major items:
	• Spinning disks and compute resources contribution to the Data Project for data replication and analyses purposes.
	• Tier 1 will have to work on optimization of nominal bandwidth of 10 GBit/s that will result in 30–50% for real bandwidth for replication. This, together with the specification of the core dataset, defines the CMIP6 replication strategy.
	• Tier 2 will have to warranty a bandwidth of 12 GBit/s for data provision. CMIP5 experience shows that each data node provides 10 times the data it hosts over a period of 4 years, and the average available network bandwidth should cover this.
	• Single Tier 1: about 20 PB for long-term archiving of reference data from the CMIP6 data (volume not clear yet)
	• Tier 1: tapes to fill the storage gap in case of insufficient disc space for initial data publication and data replication
	There is an ongoing proposition to enable ESGF to exclude a data node that does not satisfy all the CMIP6 requirements or a data node that will degrade the federation usability. The implementation of this is currently under discussion, but we can anticipate that when the governance will be set it will be the responsibility of Tier 1 nodes to enforce and to operate the decision.

CMIP6 "Impact" on Scientific Community **Sergey Nikonov (NOAA/GFDL)** serguei.nikonov@noaa.gov **V. Balaji (NOAA/GFDL)** balaji@princeton.edu **Aparna Radhakrishnan (NOAA/GFDL)** aparna.radhakrishnan@noaa.gov **Hans Vahlenkamp (NOAA/GFDL)** hans.vahlenkamp@noaa.gov	The results of resource estimations of the forthcoming CMIP6 are shown. The analysis is done based on an XML database designed and populated with MIPs requests by Martin Juckes (CEDA). The main goal is to show impact of CMIP6 on both sides of climate community—data producers and data analyzers. The results characterize the volume of output and corresponding efforts demanded for publishing planned experiments. The total amount of generated data from all participating modeling centers was estimated and compared with volume of CMIP5. There was also an attempt to assess scientific human resources being spent for quality control of published data and for analyzing and utilizing the CMIP6 outcome.
Control Vocabulary Software Designed for CMIP6	CMIP6 contains more activities and many more experimentations than its predecessor CMIP5. In order to compare this increase of model outputs, a standard has been created to ensure

Title and Presenter	Abstract
Denis Nadeau (DOE/LLNL) nadeau1@llnl.gov **Karl Taylor (DOE/LLNL)** taylor13@llnl.gov **Sasha Ames (DOE/LLNL)** ames4@llnl.gov	homogeneity of information. This standard creates the ability to understand and exchange data between different earth science groups. The Climate Forecast (CF-1) compliance already insures interoperability between different visualization and analysis software, but an extension of CF-1 is necessary to accommodate CMIP6 outputs. Variable names, global attributes, and variables attributes need to be set to facilitate comparison between similar geophysical variables coming from different provenances. The CMIP6_CV Python program insures the control of the different attributes needed before publication of CMIP6 and distribution of resulting NetCDF model outputs. CMIP6_CV ensures that all required attributes are present, and if one is missing, the program will not allow publication of the file. Other missing attributes or wrong attributes can sometime be created or replaced automatically by the CMIP6_CV Python program and warn users about the changes that have been made. CMIP6_CV establishes common ground between model outputs, which facilitates analyses for scientists studying climate changes. CMIP6_CV is very flexible and can be used by similar projects that necessitate a control vocabulary.
Developing a Vocabulary Management System for Data Reference Syntax using Linked Data Technologies in the Climate Information Platform for Copernicus Project **Ruth Petrie (ENES/CEDA)** ruth.petrie@stfc.ac.uk **Phil Kershaw (ENES/CEDA)** philip.kershaw@stfc.ac.uk **Ag Stephens (ENES/CEDA)** ag.stephens@stfc.ac.uk **Antony Wilson (ENES/CEDA)** antony.wilson@stfcac.uk	CEDA host data centers managing a large and varied archive of climate and Earth observation data. CEDA is the lead partner in the Climate Information Platform for Copernicus (CLIPC) project. One aim of the CLIPC project is to be a single point of access for a variety of climate data records. Within CLIPC many highly heterogeneous datasets were published through ESGF, such as satellite and in-situ observational data and climate impact indicators. Within each of these communities, different descriptive metadata is required to construct a useful Data Reference Syntax (DRS) when compared with the traditional model-based data published through ESGF. The European Space Agency (ESA) Climate Change Initiative (CCI) project is the most mature of these, having a dedicated ESA-CCI Open Data Portal. A CCI DRS was developed to provide a single authoritative source for cataloguing and searching the CCI data, and this has been successfully deployed for the ESA-CCI Open Data Portal and the CLIPC portal. The use of the Simple Knowledge Organization System (SKOS) and Web Ontology Language (OWL) to represent the DRS are a natural fit, providing controlled vocabularies as well as a way to represent relationships between similar terms used in different communities. The CLIPC portal supports data discovery based on the OGC CSW specification and ESGF's powerful faceted search. These services provide complementary content at different levels of granularity, and therefore a common data model was needed. Key terms are defined in vocabularies serialized in SKOS and OWL and are accessible from a central vocabulary server, which can be queried from applications consuming metadata content. Exploiting the vocabulary service, it has been possible to develop an innovative solution tagging ISO19115 records for the CSW with the equivalent vocabulary terms used for the ESGF faceted search system. SKOS provides a tool to manage CVs with semantic relationships and arbitrary tagging of datasets. In this way, it has been possible to create enhanced metadata records and a search interface, combining CSW and ESGF search results driven by a faceted search interface managed and populated from the vocabulary server.
DKRZ ESGF-Related Infrastructure and CMIP6 Services **Stephan Kindermann (ENES/DKRZ)** kindermann@dkrz.de **Michael Lautenschlager (ENES/DKRZ)** lautenschlager@dkrz.de **Stephanie Legutke (ENES/DKRZ)** legutke@dkrz.de **Katharina Berger (ENES/DKRZ)** berger@dkrz.de **Martina Stockhause (ENES/DKRZ)** stockhause@dkrz.de	The DKRZ will coordinate the German ESGF-related activities, as well as the national CMIP6 contribution. Besides hosting ESGF nodes and providing support for CMIP6 data ingest, data publication, long-term archiving, and data citation, the DKRZ in engaging in a set of new activities to support the national and international climate community, including: • Integration of CMOR with CDO to support climate modelers in generating CMIP6-compliant data. • Establishment of a national CMIP data pool acting as a replica cache of often-needed CMIP5- and CMIP6-related data, which can be exploited for efficient data analysis and evaluation. • Development of a generic data quality assurance tool supporting CMIP6 data quality checking (going beyond "pure CMIP6 convention compliance checking," e.g. including CF compliance checking and outlier detection). • Establishment of a persistent identification infrastructure integrated with ESGF and supporting CMIP6. • Extension of the data citation service on long-term archived data, with a citation possibility for the evolving CMIP6 data. • Integration of the CMIP data pool and ancillary metadata into the IPCC DDC AR6 reference data archive and improvements in the integration of the IPCC DDC in IPCC's assessment process. • Development of a Web Processing Service framework to support future data processing service provisioning, including supporting conda sw packaging and Docker.

Title and Presenter	Abstract
	The talk will summarize the current status of these activities as well as next steps and plans.
The IPCC DDC in the Context of CMIP6 **Martina Stockhause (ENES/DKRZ)** **stockhause@dkrz.de** **Michael Lautenschlager (ENES/DKRZ)** **lautenschlager@dkrz.de** **Stephan Kindermann (ENES/DKRZ)** **kindermann@dkrz.de**	The CMIP6 data underlying the IPCC AR6 of WG1 will be transferred in the long-term archive of the IPCC DDC (Data Distribution Centre) at DKRZ to build the Reference Data Archive of the global climate model output. Apart from the data, different pieces of data-related information are to be integrated in the archive in order to enrich the data documentation for the interdisciplinary long-term use. The second task of the DDC within CMIP6 is the support of the IPCC authors by opening DKRZ's CMIP Data Pool for IPCC authors. The CMIP6 data subset in the CMIP Data Pool will be the source for the Reference Data Archive for AR6. The presentation will give an overview over the different connections between IPCC DDC and CMIP6. A detailed description of the transfer of data and metadata from ESGF and repositories of ancillary metadata, such as errata, citation, model descriptions etc., will be given, with special emphasis on the requirements for ancillary metadata providers.
Persistent Identifiers in CMIP6 **Merret Buurman (ENES/DKRZ)** **buurman@dkrz.de** **Tobias Weigel (ENES/DKRZ)** **weigel@dkrz.de** **Stephan Kindermann (ENES/DKRZ)** **kindermann@dkrz.de** **Katharina Berger (ENES/DKRZ)** **berger@dkrz.de** **Michael Lautenschlager (ENES/DKRZ)** **lautenschlager@dkrz.de**	All CMIP6 files and datasets in ESGF will receive a persistent identifier (PID). A PID is a string that can be resolved to a landing page that shows some minimal metadata, for example, information about data storage locations, checksums, etc. In CMIP6, persistent identifiers (PIDs) will additionally be used to record relationships between data objects (e.g., which dataset version consists of which file sets), or which dataset version is replaced by which new version. Also, information on replication sites and dataset errata[1] will be stored. These metadata will be available even after un-publication of the data from the ESGF data nodes, so researchers are able to find some metadata on a data object they have been using even if the data were outdated. In particular, they can find out that it is outdated and which new version replaces it. Technically, the system behind the CMIP6 PIDs is the Handle System[2], which also underlie the Digital Object Identifier (DOI) system but does not aim for citability. The handles are registered at the ESGF Handle Service during the ESGF publication process, using the esgfpid library[3] called by the ESGF publisher. To cushion temporary publication peaks, a message queuing system ensures that the publication process is not delayed and no PID registration request is lost. Users can view and access the PIDs from the CoG front end, where PIDs are displayed for every dataset and file. Furthermore, the PID strings are contained in the files' NetCDF headers. Permanently bound to the file, they will help researchers find information about data they have used/found/received for years to come and potentially beyond the scope of ESGF. PIDs will also provide a sustainable foundation for data management tools and "intelligent" client-side tools (e.g., exploiting the versioning and replication information). The talk will outline the current status of the technical PID infrastructure as well as its integration with the ESGF publisher and first test results. Furthermore, it will be outlined how PIDs will support CMIP6 versioning and CMIP6 errata annotation. [1] See abstract on QCWT Errata Service by Levavasseur et al. [2] http://www.handle.net/. [3] https://github.com/IS-ENES-Data/esgf-pid.
ES-DOC and ESGF Errata Services **Atef Ben Nasser (ENES/IPSL)** **abennasser@ipsl.jussieu.fr** **Mark Greenslade (ENES/IPSL)** **momipsl@ipsl.jussieu.fr**	The Earth System DOCumentation has started in the documentation of the CMIP6 project, putting into good use the experiences gathered from CMIP5. Using further formalization and a clear set of use cases, the process has been streamlined and rendered less of a burden as a large chunk of it has been automated, a beta period has been scheduled and each and every step thoroughly documented. The ES-DOC is ready for community review as of November 2016, the beta testing phase will take place during the period October 2016 through February 2017. The full community release is scheduled for March 2017. In the context of overseeing the quality of data, the ESGF Errata service has been encapsulated in the ES-DOC structure and built on top of the Handler service that will be deployed in the next release cycle. Consuming PIDs from Handler Service, the ESGF Errata service is guided by a specifically built algorithm that extracts metadata regarding the issues that may or may not affect the quality of datasets/files and cause newer version to be published. This new structure has been designed keeping in mind usability by end users specialized in the publishing process or other scientists requiring feedback on reliability of needed data. The expected outcome from both ES-DOC and the Errata service project is to increase the quality of data. Providing this critical information for end-users requires the application of a well-defined process and is ensured by exploring incoming features of the ESGF ecosystem.

Title and Presenter	Abstract
National Computational Infrastructure's Research Data Services: Providing High-Quality Data to Enable Climate and Weather Science **Claire Trenham (NCI/ANU)** claire.trenham@anu.edu.au **Kelsey Druken (NCI/ANU)** Kelsey.druken@anu.edu.au **Adam Steer (NCI/ANU)** adam.steer@anu.edu.au **Jon Smillie (NCI/ANU)** jon.smillie@anu.edu.au **Jingbo Wang (NCI/ANU)** jingbo.wang@anu.edu.au **Ben Evans (NCI/ANU)** Ben.Evans@anu.edu.au	The National Computational Infrastructure (NCI) hosts over 10 PB of broad-based, nationally significant research data collections, including climate and weather, water and ocean, satellite Earth observations, reanalysis, elevation and bathymetry, geodetic, other geosciences, astronomy, social sciences, and bioinformatics. We have a particular focus on Earth systems data (including CMIP) as part of the National Environmental Research Data Interoperability Platform (NERDIP). The data is, where possible, stored in a standard format (NetCDF) in clear directory structures. Quality is assured by compliance to metadata standards and tested against common tools and protocols, and data management plans are created to provide full data collection metadata and provenance information. As well as being available on NCI's high performance file system, open datasets are published through our Research Data Services, including THREDDS. These data services are accessible from anywhere in the world to allow broader access than just within NCI's high -performance environment, and data is therefore available to a wider community of scientists and visualization specialists via remote file access, download, or OGC-compliant services. NCI is a node of the ESGF for the publication and replication of CMIP and other international climate data, which enables bulk data transfer for greater access to the CMIP data by the Australian climate research community, and distribution of their modeled data. A powerful use of these data facilities is to provide high-performance access to data enabling advanced virtual laboratories, particularly the Climate and Weather Science Laboratory, as well as a number of other community virtual laboratories and portals. NCI also makes the data available via our interactive Virtual Desktop Infrastructure (VDI) and Raijin supercomputer. The VDI enables climate science, including multi-model intercomparison, and detection and attribution work, by providing access to the data collocated with programming, analysis, and visualization tools (including Python, UV-CDAT, VisTrails). VDI also provides remote batch submission capability to the high-performance computing (HPC) infrastructure. We provide a metadata catalog of our data holdings, through which anyone can search our data collections and datasets and find information on how to access the data. In particular, the catalogue information shows location on NCI's file systems and via web data services as available. NCI supports the widespread use of these datasets, without the need for scientists to move data to their local workstations, by enabling the data to be accessed remotely from anywhere, via standard web protocols, or for Australian researchers, directly at NCI in a value-added virtual environment equipped with the typical datasets and tools a climate or weather scientist is likely to need.
Automating Data Synchronization, Checking, Ingestion and Publication for CMIP6 **Ag Stephens (ENES/CEDA)** ag.stephens@stfc.ac.uk **Alan Iwi (ENES/CEDA)** alan.iwi@stfc.ac.uk	CEDA is responsible for providing the ESGF UK data node. This involves publication of Met Office Hadley Centre datasets provided for CMIP6. Following lessons learnt from CMIP5, we have developed an automated system for remotely synchronizing the contents of the Met Office MASS tape archive to CEDA. This builds upon a RabbitMQ message service that prompts for actions such as "publish-to-ESGF" and "withdraw-from-ESGF." The ingestion pipeline at CEDA is complex because it requires access to multiple services running across a range of platforms. In order to automate the pipeline, we have developed a simple client-server architecture in which a collection of distributed workers query a centralized database for instructions to manage their own processes and workloads. This approach allows each independent worker to run his/her own controller under a different user ID with access to specific resources relevant to its stage in the processing chain (e.g. "sync," "validate," "ingest," "publish"). Individual (client) workers have no knowledge of other workers because all states and decisions about which controller should be run on each dataset are managed through the database (server). The data model uses the ESGF Dataset as its unit of currency and the system records each "do" (and, when problems occur, "undo") event that takes place across all platforms. A Django web application provides queryable views of important components such as files, ESGF datasets, and events, as well as a "Global Pause" feature that can be activated to quickly halt all clients for an important fix or change. This modular architecture allows pipelines to be added or modified without redesign of the underlying framework, making the tool ideal for a range of automated processing chains in big data management.
Input4MIPs: Boundary Condition and Forcing Datasets for CMIP6 **Paul J. Durack (DOE/LLNL/PCMDI)** pauldurack@llnl.gov **Karl E. Taylor (DOE/LLNL/PCMDI)** taylor13@llnl.gov	Input4MIPs (input datasets for Model Intercomparison Projects) is an activity to make available via ESGF the boundary condition and forcing datasets needed for the 6th Coupled Model Intercomparison Project (CMIP6). Various datasets are needed for the pre-industrial control (piControl), AMIP, and historical simulations, and additional datasets are needed for many of the 17 CMIP6-endorsed MIPs experiments. Earlier versions of many of these datasets were used in CMIP5. Unlike model data generated from CMIP6 experiments and standardized using CMOR, the formats

Title and Presenter	Abstract
Sasha Ames (DOE/LLNL) ames4@llnl.gov	of these contributed datasets vary and at often times test the limits of the ESGF infrastructure. This presentation will highlight some of the use cases encountered during collation and publishing of the Input4MIPs data, and will provide some insights into how the publishing step was augmented to deal with these highly variable data formats. These will be useful to consider as the ESGF system further evolves to address the requirements of Obs4MIPs and other large international projects.
An Update on ESGF Needs for Obs4MIPs **Peter Gleckler (DOE/LLNL/PCMDI)** gleckler1@llnl.gov	Obs4MIPs has advanced considerably since its inception nearly five years ago. It is now formally recognized as a WCRP project, with oversight provided by the WCRP Data Advisory Council Task Team (WDAC TT). There are currently 7 ESGF nodes serving obs4MIPs data, contributed by 15 institutions, with a current inventory of over 80 large-scale gridded observational products. Scientists have recently proposed adding 100 new datasets to Obs4MIPs. The WDAC TT has identified several ways the ESGF could be enhanced to greatly facilitate the planned expansion of Obs4MIPs, which would insure it can include a broader observational community. This presentation will describe two enhancements the Obs4MIPs WDAC TT would like to convey to the ESGF community.
Recent Climate4Impact Developments: Provenance in Processing and Connection to the CLIPC Portal **Maarten Plieger (ENES/KNMI)** maarten.plieger@knmi.nl **Christian Pagé (ENES/IPSL)** christian.page@cerfacs.fr **Sandro Fiore (ENES/CMCC)** sandro.fiore@cmcc.it	The aim of Climate4Impact is to enhance the use of research data and to support other climate portals. It has been developed within the European projects IS-ENES, IS-ENES2, and CLIPC. Climate4Impact is connected to ESGF using certificate-based authentication, ESGF search, OpenID, OPeNDAP and THREDDS catalogs. Climate4Impact offers web interfaces for searching, visualizing, analyzing, processing, and downloading datasets. Climate4Impact exposes open standards like WMS, WCS, and WPS using open-source tools. Processing services include climate indicator calculations, country-based statistics and polygon extraction by GeoJSON. Provenance integration is achieved using the W3C PROV standard for fully traceable provenance. The PROV document is stored in NetCDF files and can be visualized. The provenance module traces data usage statistics in a database, which is interesting for data providers. Climate4Impact has a personal basket where users can upload their own data and do research with the provided tools. The basket supports formats like NetCDF, GeoJSON and CSV. The basket has an access token mechanism to make data sharing and command line access to web services easier. This enables client side scripting of the Climate4Impact portal possible and makes it possible to connect third-party portals, like the EU FP7 CLIPC portal. The CLIPC portal uses web services from Climate4Impact. It has an appealing front end built in openlayers3 and is targeted to boundary workers. The web services, provenance integration, and connection with the CLIPC portal are detailed in this presentation.
Federated Data Usage Statistics in the Earth System Grid Federation A. Nuzzo (ENES/CMCC) alessandra.nuzzo@cmcc.it M. Mirto (ENES/CMCC) maria.mirto@cmcc.it P. Nassisi (ENES/CMCC) paola.nassisi@cmcc.it K. Berger (ENES/DKRZ) berger@dkrz.de T. Rathmann (ENES/DKRZ) rathmann@dkrz.de **L. Cinquini (NASA/JPL)** Luca.Cinquini@jpl.nasa.gov **S. Devil (ENES/IPSL)** sebastien.denvil@ipsl.jussieu.fr S. Fiore (ENES/CMCC) sandro.fiore@cmcc.it **D. N. Williams (DOE/LLNL)** williams13@llnl.gov **G. Aloisio (ENES/CMCC)** giovanni.aloisio@unisalento.it	Monitoring ESGF is a challenging topic. From an infrastructural standpoint, two components (the Dashboard and the Desktop) provide the proper environment for capturing (i) usage metrics, as well as (ii) system status information at local (node) and global (institution and/or federation) level. All the metrics collected by the ESGF monitoring infrastructure are stored in a system catalog that has been extended to support a large set of information about the data usage statistics. More specifically, the Dashboard provides coarse- and fine-grained data usage statistics. Regarding the coarse-grained statistics, information like the data downloaded (GB/TB), number of downloads, number of distinct files and users, downloads by user and Identity Provider, client statistics (country/continent distribution) are provided. Moreover, the fine-grained statistics are related to (i) cross projects, such as the number of downloads per project and host and by time and (ii) specific projects such as CMIP5 and Obs4MIPs download data. In this case, the number of downloads, number of successful downloads, downloaded data, and timeframe are provided with the possibility to know the top ten datasets, experiments, variables, grouped by (i) experiment/model, and (ii) experiment/model by time. The fine-grained statistics are available for single and federated data nodes. The federation is allowed by a specific protocol, which has been implemented for gathering the metrics from each data node (classified as leaf node, which is responsible for each site, and collector node, which gathers the data from registered leaf nodes). To this end, several data marts have been created to allow fast access to this information. Project-specific views provide a deep insight about the statistics related to specific projects.
Large-Scale Data Analytics	Defining and implementing experiments with hundreds of data analytics operators can be a real

Title and Presenter	Abstract
Workflow Support for Climate Change Experiments **S. Fiore (ENES/CMCC)** sandro.fiore@cmcc.it ***C. Doutriaux (DOE/LLNL)*** doutriaux1@llnl.gov **C. Palazzo (ENES/CMCC)** cosimo.palazzo@cmcc.it **A. D'Anca (ENES/CMCC)** cosimo.palazzo@cmcc.it **Z. Shaheen (DOE/LLNL)** shaheen2@llnl.gov **D. Elia (ENES/CMCC)** cosimo.palazzo@cmcc.it ***J. Boutte (DOE/LLNL)*** boutte3@llnl.gov **V. Anantharaj (DOE/ORNL)** anantharajvg@ornl.gov **D. N. Williams (DOE/LLNL)** williams13@llnl.gov **G. Aloisio (ENES/CMCC)** giovanni.aloisio@unisalento.it	challenge in many practical scientific use cases, such as multi-model analysis, climate indicators, processing chains for operational environments, etc. This is usually done via scripts (e.g. bash) on the client side and requires climate scientists to implement and replicate workflow-like control logic aspects (which may be error-prone too) in their scripts, along with the expected application-level part. High-level solutions leveraging workflow-enabled big data analytics frameworks for e-science could help scientists in defining and implementing the workflows related to their experiments by exploiting a more declarative, efficient, and powerful approach. This talk will present key needs and challenges regarding *big data analytics workflow management for e-science* and will then provide some insights about the implementation of real use cases implemented in some European projects (e.g. BIGSEA, CLIPC, INDIGO). All the proposed use cases have been implemented exploiting the Ophidia big data analytics framework. The software stack includes an internal workflow management system, which coordinates, orchestrates, and optimizes the execution of multiple scientific data analytics and visualization tasks. Real-time workflow monitoring execution is also supported through a graphical user interface. The provided data analytics workflow engine supports conditional sections, parallel loops, and massive statements for high-throughput experiments. Specific emphasis will be devoted to a large-scale climate model intercomparison data analysis experiment (e.g. precipitation trend analysis) performed in the context of the H2020 INDIGO-DataCloud project. The use case exploits the INDIGO capabilities in terms of software framework deployed on cloud, UV-CDAT for data visualization, and Ophidia to run multi-model data analysis.

Day 3: Wednesday, December 7, 2016
Coordinated Efforts with Community Software Projects

Title and Presenter	Abstract
THREDDS Data Server: OPeNDAP and Other Tales from the Server-Side **Sean Arms (NSF/Unidata)** sarms@ucar.edu	This talk is geared towards informing ESGF of the status of data services in the TDS. OPeNDAP will be discussed in terms of both DAP2 and DAP4. Other services provided by the TDS, such as ncWMS and the NetCDF Subset Service, will be discussed, highlighting how they may benefit ESGF users.
A Hybrid Provenance Capture Approach to Scientific Workflow Reproducibility and Performance Optimization **Todd Elsethagen (DOE/PNNL)** todd.Elsethagen@pnnl.gov **Eric Stephan (DOE/PNNL)** Eric.Stephan@pnnl.gov **Bibi Raju (DOE/PNNL)** bibi.raju@pnnl.gov	As HPC infrastructures continue to grow in capability and complexity, so do the applications that they serve. HPC and distributed-area computing (DAC) (e.g. grid and cloud) users are looking increasingly toward workflow solutions to orchestrate their complex application coupling and pre- and post-processing needs. To gain insight and a more quantitative understanding of workflow performance, the Provenance Environment (ProvEn) architecture includes not only the capture of traditional provenance information, but also the capture and integration of system environment metrics helping to give context and explanation for a workflow's execution. This presentation describes how ESGF will use ProvEn to support reproducibility, data lineage, and performance optimization.
QA/QC at the DKRZ **Heinz-Dieter Hollweg *(ENES/DKRZ)*** hollweg@dkrz.de	Between the states of climate datasets being compliant or rejectable there exists a grey range of being almost compliant to project rules. Thus, it is important to provide annotations describing compliance deviations in a brief but helpful way. Just as important is to enable modelers to perform pre-checks before submitting a large amount of data, whether by a locally installed quality assurance tool or via WPS. The QA-DKRZ tool is presented with some technical aspects: installation and running as WPS, respectively, selection of sub-sets out of rather large data volumes, an interface to run external components or tools within QA-DKRZ (e.g., the CMOR checker), and the annotation model used to generate concise results. Eventually, some experiences gained during CMIP5 and CORDEX are discussed about the "interaction" between submitters of data and QA results.
Web Processing Services and ESGF: The Birdhouse System **Stephan Kindermann**	Provisioning web processing services near large ESGF sites supports efficient future data analysis activities. To support the exposure of data analysis and data evaluation code in the form of OGC WPS, a modular set of easily installable and deployable components are being developed and bundled in the "birdhouse" framework (http://birdhouse.readthedocs.io/en/latest/). Individual

Title and Presenter	Abstract
(ENES/DKRZ) **kindermann@dkrz.de** **Carsten Ehbrecht** **(ENES/DKRZ)** **ehbrecht@dkrz.de** **Nils Hempelmann (ENES/IPSL)** **nils.hempelmann@lsce.ipsl.fr**	processing services as well as generic infrastructural services (the "birds") are supported by a generic "birdhouse," providing a generic installation and deployment solution (supporting e.g. Docker-based hosting solutions). The current status of the system will be described and an overview is given of existing services and services in active development, especially supporting the climate impact community (a short demo will be given). Special emphasis will be put to the unique aspects and open issues that supporting efficient computing at sites providing large ESGF replica caches entails.
Synda (synchro-data) **Sébastien Denvil (ENES/IPSL)** **sebastien.denvil@ipsl.jussieu.fr** **Raciazek Jérôme** **(ENES/IPSL)** ***jripsl@ipsl.jussieu.fr*** **LEVAVASSEUR Guillaume** **(ENES/IPSL)** **glipsl@ipsl.jussieu.fr**	Synda is a command line tool to search and download files from the ESGF archive. Since its inception in 2011, Synda has essentially been used for the replication use case when a large bulk of data needs to remain in sync between a local archive and the ESGF system. New features have been added to support a broader set of use cases including the first of them: easily grabbing a small or a large number of files or datasets from ESGF. 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 back-end functionalities. This talk will walk through Synda main features and supported use case. We will also expose how we plan to support an automatic replication workflow for CMIP6. Current main features are listed below: • Simple data installation using an apt-get like command • Support every ESGF project (CMIP5, CORDEX, SPECS, etc.) • Parallel downloads, incremental process (download only what is new) • Transfer priority, download management and scheduling, history stored in a database • GridFTP enabled, fallback position to HTTP when needed hooks available for automatic publication upon datasets download completion.
Globus Update **Rick Wagner (University of Chicago and DOE/ANL)** **rick@globus.org**	ESGF uses Globus for managed data transfer and sharing, both for replication between nodes and for users to transfer the data. Globus is software-as-a-service for research data management and provides high speed and secure data transfer, data sharing directly from existing storage systems, and data publication. Developed and operated by the University of Chicago, Globus has become a preferred service for moving and sharing data between and among a wide variety of storage systems at research laboratories, campus computing resources, and national facilities across the United States. This presentation will cover Globus integration with ESGF components and services, relevant recent updates to Globus, and potential methods for leveraging these new features by ESGF.
BASEJumper: Publishing HPSS Datasets via ESGF **Sam Fries (DOE/LLNL)** **fries2@llnl.gov** **Sasha Ames (DOE/LLNL)** **ames4@llnl.gov** **Alex Sim (DOE/LBNL)** **asim@lbl.gov**	The capacity of hard disk space has not kept pace with the volume of output created by climate models. To store model output, High Performance Storage Systems (HPSSs) are required. These tape archives are notoriously slow, and getting permission to access them can be tricky and time consuming. In order to facilitate climate modelers and consumers of model output, the Analytics and Informatics Managements Systems team at Lawrence Livermore National Laboratory (in collaboration with Lawrence Berkeley National Laboratory) has created a system that allows archived data to be requested and retrieved via ESGF. This system (BASEJumper, named after the Berkeley Archive Storage Encapsulation library) uses existing ESGF services, provides all of the normal metadata required for a dataset, and uses ESGF's access control mechanisms to safeguard the data. It uses a two-stage design to allow it to move around firewalls and many layers of security to prevent denial of service attacks on HPSS resources.

Day 3: Wednesday, December 7, 2016
Live Demonstration Session

Title and Presenter	Abstract
ESGF Ingestion Service Overview **Lukasz Lacinski (DOE/ANL)** **lukasz@uchicago.edu**	This talk will present an overview of the Ingestion service, which provides a remote interface to the ESGF publication command line tools. The interface is provided as a RESTful API, integrated with the ESGF authentication and authorization. Additionally, the API can manage dataset transfers through Globus Transfer service and reorganize dataset files on the data node before publication. The API supports three different publication workflows independence on a location of dataset files being published: the local data node, a remote file systems accessible through a Globus endpoint, and a remote storage with dataset accessible through HTTP.
Compute Working Team Server-side Demonstration	Over the last year, the ESGF-CWT made significant progress. An API was established to communicate with ESGF-CWT's WPS servers. Additionally, an end-user Python-based API was

Title and Presenter	Abstract
C. Doutriaux (DOE/LLNL) doutriaux1@llnl.gov **Jason Boutte (DOE/LLNL)** boutte3@llnl.gov	developed. In this talk, we will demonstrate how the user API can be used to call various ESGF-CWT WPS servers, all implementing a similar workflow in various fashions. Specifically a multi-model ensemble average will be computed on the server(s)-side and employed by the end user.
Live Demo of Visualization and Processing Services in the Climate4Impact Portal **Maarten Plieger (ENES/KNMI)** maarten.plieger@knmi.nl	The aim of Climate4Impact is to enhance the use of climate research data and to support other climate impact portals. This live demonstration shows how Climate4Impact enables researchers to use climate data in their research. Researchers are spending considerable amounts of their time on data gathering, conversion, integration, and interpretation. Parts of this process have already been done before and do not need to be repeated or re-invented again. Climate4Impact facilitates this and lifts the burdens from researchers, thus increasing the time available for real research. To facilitate this, Climate4Impact offers several web processing services and wizards, including an averager, subsetter, regridder, reformatter, combine tool, and tool for polygon subsetting. The following topics will be demonstrated: 1. Login: Using OpenID to access Climate4Impact and the ESGF data nodes. 2. Discovery: Faceted search using the ESGF search API. 3. Visualization: Visualize data using OGC Web Map Services. 4. Convert and subset: Transform data into other formats and geographical projections, using OCG Web Processing Services. 5. Import the obtained data in a GIS system, like QGIS. 6. Other processing services, like averaging, polygon subsetting, and combining. Climate4Impact is developed through the IS-ENES, IS-ENES2 and CLIPC projects, which receive funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration.

Day 3: Thursday, December 8, 2016
Poster Session

Title and Presenter	Abstract
ADAGUC Open-Source Visualization in Climate4Impact Using OGC Standards **Maarten Plieger (ENES/KNMI)** maarten.plieger@knmi.nl **Ernst de Vreede (ENES/KNMI)** ernst.de.vreede@knmi.nl	ADAGUC is an open-source geographical information system to visualize NetCDF, HDF5, and GeoJSON over the web. The software consists of a server-side C++ application and a client-side JavaScript application. The software provides several features to access and visualize data over the web; it uses the OGC WMS and WCS standards for data dissemination. Web clients like Google Maps, OpenLayers, and Leaflet are supported and can directly use the exposed web services. ADAGUC is used in projects like Climate4Impact to visualize datasets stored in ESGF. ADAGUC can visualize remotely published NetCDF files by adding the OPeNDAP resource as parameter to the web service request. This enables direct visualization of any OPeNDAP enabled resource over the web. Checking the variable standard_name and units does graphic styling of data. OGC Web Coverage Services (WCSs) are available and can be used for data reprojection, subsetting, and conversion to other formats. Access to OPeNDAP services is done efficiently; multiple requests are aggregated into one and only the domain of interest is requested. This allows for easy, quick, and interactive visualization of OPeNDAP-enabled datasets. ADAGUC has a number of data converters and data postprocessors to support various data conventions. Supported file formats are "true color NetCDF" for satellite imagery, structured grids, curvilinear grids, satellite swaths, point observations, point time series, and polygons stored in GeoJSON. Datasets consisting of several NetCDF files can be aggregated into a single dataset and are offered over WMS, WCS, and OPeNDAP. ADAGUC can be used as a component for WPS to subset data and convert GeoJSON to grids. Results and lessons learned will be presented at the conference.
Community Data Management System **Denis Nadeau (DOE/LLNL)** nadeau1@llnl.gov	The Analytics and Informatics Managements Systems (AIMS) team will completely redesign and transform the "Climate Data Management System Version 2" (CDMS2) into the "Community Data Management System" (CDMS). Designed in the mid to late 90s, CDMS's original intent was to automatically locate and extract metadata (variables, dimensions, grids, etc.) from collections of simulation runs and analysis files. Since then it has grown to include multiple regridders, time

Title and Presenter	Abstract
Charles Doutriaux (DOE/LLNL) doutriaux1@llnl.gov **Dean N. Williams (DOE/LLNL)** williams13@llnl.gov	components, masked arrays, and more. However, with the rapid changes in technology, it is time for an upgrade to broaden its scope and design to include newer "community" data file formats (such as HDF5 and IDX), the latest "community" Numerical Python packages (such as NumPy 3.0 and Numba), and the latest "community" of regridders that combine the manipulation of simulation, observation, reanalysis, and point datasets. CDMS aims to incorporate 21st century technologies and integrate additional geoscience domains. In addition to conforming to the latest community standards and protocols, it will include the new CDAT ingest package.
Community Diagnostics Package **Zeshawn Shaheen (DOE/LLNL)** shaheen2@llnl.gov **Charles Doutriaux (DOE/LLNL)** doutriaux1@llnl.gov **Samuel Fries (DOE/LLNL)** fries2@llnl.gov	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 AIMS group and LLNL are developing the Community Diagnostics Package (CDP), a framework for creating new climate diagnostic 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.
Earth System Model Development and Analysis using FRE-Curator and Live Access Servers: On-demand Analysis of Climate Model Output with Data Provenance **Aparna Radhakrishnan (NOAA/GFDL)** aparna.radhakrishnan@noaa.gov **V. Balaji (NOAA/GFDL)** balaji@princeton.edu **Roland Schweitzer (NOAA/GFDL)** roland.schweitzer@noaa.gov **Serguei Nikonov (NOAA/GFDL)** serguei.nikonov@noaa.gov **Kevin O'Brien, (NOAA/PMEL)** kevin.m.o'brien@noaa.gov **Hans Vahlenkamp (NOAA/PMEL)** hans.vahlenkamp@noaa.gov	There are distinct phases in the development cycle of an Earth system model. During the model development phase, scientists make changes to code and parameters and require rapid access to results for evaluation. During the production phase, scientists may make an ensemble of runs with different settings and produce large quantities of output that must be further analyzed and quality controlled for scientific papers and submission to international projects such as CMIP. During this phase, provenance is a key concern: being able to track back from outputs to inputs. We will discuss one of the paths taken at GFDL in delivering tools across this life cycle, offering on-demand analysis of data by integrating the use of GFDL's in-house FRE-Curator, Unidata's THREDDS, and NOAA PMEL's Live Access Server (LAS). Experience over this life cycle suggests that a major difficulty in developing analysis capabilities is only partially the scientific content. It is often devoted to answering the questions "Where is the data?" and "How do I get to it?" "FRE-Curator" is a database-centric paradigm used at NOAA GFDL to ingest information about the model runs into an RDBMS (Curator database). The components of FRE-Curator are integrated into Flexible Runtime Environment workflow and can be invoked during climate model simulation. The front end to FRE-Curator, known as the Model Development Database Interface (MDBI) provides an in-house web-based access to GFDL experiments: metadata, analysis output, and more. In order to provide on-demand visualization, MDBI uses LAS, which is a highly configurable web server designed to provide flexible access to geo-referenced scientific data that makes use of OPeNDAP. Model output saved in GFDL's tape archive, the size of the database and experiments, continuous model development initiatives with more dynamic configurations add complexity and challenges in providing an on-demand visualization experience to our GFDL users.
Toward a High-Performance Data Analysis Platform for Impact Analysis **Wim Som De Cerff (ENES/KNMI)** wim.som.de.cerff@knmi.nl **Sandro Fiore (ENES/CMCC)** sandro.fiore@cmcc.it **Maarten Plieger (ENES/KNMI)** maarten.plieger@knmi.nl	The aim of Climate4Impact is to enhance the use of climate research data and to enhance the interaction with climate effect/impact communities. The portal is based on impact use cases from different European countries and is evaluated by a user panel of use case owners. In the data analytics landscape, Ophidia provides a big data framework for e-science, focusing on the analysis of large-scale n-dimensional datasets. Ophidia provides data operators to manipulate data in the form of data cubes, and array-based primitives to perform data analysis on large scientific data arrays (e.g. statistical analysis, predicate evaluation, subsetting, reduction, compression). KNMI and CMCC are working together toward a high-performance data analysis platform for impact analysis by integrating and properly extending/adapting Climate4Impact and Ophidia. A key point to be addressed is the interoperability with ESGF, in terms of security and access interface, which means working closely with the ESGF-CWT and the ESGF-IdEA working teams. In order to support users for their analytics and scientific operations on large amount of data, the Climate4Impact portal will interact with Ophidia for the back-end processing capabilities. The Ophidia WPS interface and Climate4Impact services will allow easy front-end controlling, visualizing, and tracking of remote Ophidia task submissions. Moreover, the Ophidia workflow engine in addition to the visualization capabilities of Adaguc will lead to near-real-time output of production and visualization of complex experiments.
Climate Model Output Rewriter	Version 3.2 of the Climate Model Output Rewriter (CMOR) has been released to handle a state-of-

Title and Presenter	Abstract
Version 3.2 for CMIP6 **Denis Nadeau (DOE/LLNL)** **nadeau1@llnl.gov** **Karl Taylor (DOE/LLNL)** **taylor13@llnl.gov** ***Charles Doutriaux (DOE/LLNL)*** **doutriaux1@llnl.gov** **Dean N. Williams (DOE/LLNL)** **williams13@llnl.gov**	the-art MIPs. The Working Group Coupled Model Infrastructure Panel (WIP) has created an exhaustive Data Request database for CMIP6 which is used by CMOR to re-write model output. The files created by CMOR also follow the CF-1 metadata conventions to promote the processing and sharing of CMIP6 data. The latest version of CMOR 3.2 incorporates a "Control Vocabulary" API to line up with continuously growing CMIP6 requirements from the WIP. This API has also been incorporated into the ESGF publisher in order to validate every published files for the CMIP6 project. Delineating new input table structure confines CMOR to very strict CMIP6 requirements, which empowers each model to maintain value delivery. CMOR is a very robust program and can work with different type of grids, different projections, LIDAR tracks, or ship transects. Its high flexibility allows customization of global attributes to accommodate growth in capability needed by different MIPs, such as Obs4MIPs, Input4MIPs, or CREATES.
QoS-based Dynamic and Elastic Scenarios in the Cloud for Data Analytics in the BIGSEA Project **Authors: Donatello Elia (ENES/CMCC)** ***donatello.elia@cmcc.it*** **ENES/CMCC, Sandro Fiore (ENES/CMCC)** ***sandro.fiore@cmcc.it*** **ENES/CMCC, Giovanni Aloisio (ENES/CMCC)** ***giovanni.aloisio@unisalento.it***	EUrope-BRAzil Collaboration on BIG Data Scientific REsearch through Cloud-Centric Applications (EUBra-BIGSEA) is a project funded under the third coordinated call for Europe and Brazil. It targets the development of cloud services for big data applications to ensure quality of service (QoS), security, and data privacy. The integrated and fast big data ecosystem is the central component, as it addresses the data management aspects of the EUBra-BIGSEA platform. Its key elements are the integration of different classes of big data technologies, such as Ophidia framework or Spark; the dynamicity and elasticity of the environment based on QoS policies; and a secured-by-design architecture. The ecosystem joins these aspects in a cloud environment to tackle massive data processing scenarios like the ones proposed in the BIGSEA project. In particular, this poster shows how Ophidia has been integrated into a smart city context, to deal with weather forecasting data in cloud QoS-based elastic and dynamic scenarios. The generality of the approach makes its adoption straightforward in the ESGF-based context, with special regard to the computing/analysis part, where different user needs and workloads could benefit from these new developments.