8th annual ESGF F2F Conference Abstracts

2018 Earth System Grid Federation Face-to-Face Conference
Washington, D.C., USA

Day 1: Tuesday December 4

Steering Committee and ESGF Executive Committee

Infrastructure for the European Network for Earth System modelling: IS-ENES3
Sylvie Joussaume (CNRS-IPSL), sylvie.joussaume@lsce.ipsl.fr
Bryan Lawrence (University of Reading), NCAS, UK
Michael Lautenschlager (DKRZ), lautenschlager@dkrz.de
Francesca Guglielmo (CNRS-IPSL), francesca.guglielmo@lsce.ipsl.fr

The European Network for Earth System modelling (ENES) gathers the European community working on climate modelling. Its infrastructure project, IS-ENES3, will start its 3rd phase in January 2019 for 4 years, engaging 22 partners from 10 countries to support collaboration on software and data in Europe. It will implement its 2017 update of the ENES infrastructure strategy 2012-2022 (Joussaume et al., 2017). IS-ENES3 will support the exploitation of model data from the WCRP coordinated experiments, CMIP and CORDEX. It will maintain and develop the European component of ESGF with the aim of supporting CMIP6. It will support the infrastructure and governance of key metadata and data standards, such as the Climate Forecast conventions for NetCDF and the ES-DOC standards for documentation of models and simulations. It will also develop a new service to ease multi-model data analytics. IS-ENES3 will raise the standard for Earth system model evaluation by gathering more detailed understanding of user requirements, by promoting standards for the science provenance, and by developing a state-of-the-art community European model evaluation framework. IS-ENES3 aims at facilitating the exploitation of model data not only by the Earth system science community but also by the climate change impact community and the climate service community. It will invest in training as well as in the operation and further development of the climate4impact platform and the underlying services to enable customised access to data, documentation, and information about model evaluation to the climate impact community as well as climate service businesses and consultancies. With the support from the ENES data task force, chaired by Michael Lautenschlager, IS-ENES3 will prepare for a long-term sustainable infrastructure in support of climate modelling.

Cyberinfrastructure for Earth Observation and climate services in Canada: review of sustainability projects 2017-2027
Tom Landry (CRIM), tom.landry@crim.ca

Following the completion in 2017 of project PAVICS, the Platform for Analysis and Visualization of Climate Science, several new sustainability projects have been funded in Canada. There is now an opportunity to seek coordination and complementary with ESGF activities and thus to reinforce Canada's contribution in this global effort.

In 2017, a consortium of Canadian institutions was selected as funded recipients for a cyberinfrastructure challenge. The project, codenamed PAVICS NexGen, has already been communicated to ESGF leadership. Through this initiative should arise increased capacities in data computation and management, in support of both climate and EO services. PAVICS NextGen will advance its research platform as well as novel Machine Learning models, tools and techniques.

Two sustainability projects based on PAVICS led by two Canadian universities were funded by CANARIE. The projects will deliver new advanced hydro modelling tools and processes, as well as DeepLearning-based data annotation tools for satellite imagery. In parallel, CRIM continues its implication into Open Geospatial Consortium (OGC) testbeds with intent of implementing, testing and advancing open standards such as Thematic Exploitation Platforms (TEP). Work conducted in OGC Testbed-14 Machine Learning task also envisionning of best practices and interoperable, standardized ML systems for Earth Systems and GeoInt.

CRIM entered final negotiations with Environment and Climate Change Canada (ECCC) and its Canadian Center for Climate Services (CCCS). Through this project and with collaboration of key partners and subcontractors, CRIM will oversee the development and hosting of an interactive, web-based, climate information portal and sectorial modules. Overall, Canadians will benefit from ECCC's contribution through an increased capacity to interact with climate information, data, and tools to inform climate-smart decision-making, thereby increasing resilience to climate change.

Project Requirements

CMIP6's Reliance on ESGF Infrastructure: Present and Future
Karl E. Taylor (PCMDI/LLNL), taylor13@llnl.gov
V. Balaji (Princeton University), balaji@princeton.edu

The success of CMIP6 depends on the capabilities and reliability of ESGF-developed infrastructure. Now that much of the infrastructure is in place and operational, we briefly assess its current state and describe outstanding issues. In a general sense, we review the priorities established at the 2017 ESGF meeting. From the perspective of scientific users of CMIP data, we highlight new capabilities and embellishments proposed or being worked on now which might be most impactful in terms facilitating research.

Obs4MIPs: Progress and Plans
Peter Gleckler (PCMDI/LLNL), Gleckler1@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.

ESGF Compute Services

State of Compute Working Team and Compute Service Certification Process
Charles Doutriaux (LLNL/AIMS), doutriaux1@llnl.gov

In this talk we will review the Earth System Grid Compute Working Team. Individual sub-team achievements will also be highlighted during the talk. The recently introduced CWT certification will be highlighted as well as ESGF latest official stack.

During the 2017 ESGF F2F in San Francisco, concerns were raised about which services will be officially available via the Earth System Grid Federation and how the quality of the services and operational management (or servers) offered could reflect poorly upon ESGF in general. In order to continue fostering creativity and innovation w/o compromising ESGF's reputation the ESGF CWT proposed to establish a set of rules for both the servers and services to obtain an official ESGF certification. This document describe the requirement proposed by the CWT as well as the application and certification processes.

Compute service cluster and deployment
Jason Boutté (LLNL/AIMS), boutte4@llnl.gov

As climate data gets increasingly large and analysis more complex, it's not reasonable to expect the end user to accommodate this. Rather we need to make the move and begin taking the work to the data sources. To accomplish this we've developed the ESGF compute server which is designed to run locally with the ESGF data node. This allows for the computations to be run on a suitable cluster local to the data. Managing the deployment of this software and the resources of the cluster is no simple task. At LLNL we've chosen to containerize with Docker, have Kubernetes manage our cluster resources and Helm handle the deployment of the compute stack. This combination makes for easy deployment and maintenance of the software stack.

The Earth Data Analytics Services (EDAS) Framework
Thomas Maxwell (NASA/NCCS), thomas.maxwell@nasa.gov
Daniel Duffy (NASA/NCCS), daniel.q.duffy@nasa.gov
Laura Carriere (NASA/NCCS), laura.carriere@nasa.gov

Faced with unprecedented growth in earth data volume and demand, NASA has developed the Earth Data Analytic Services (EDAS) framework, a high-performance big data analytics and machine learning framework. This framework enables scientists to execute data processing workflows combining common analysis and forecast operations 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 tools of earth data science, e.g. ESMF, CDAT, NCO, Keras, TensorFlow, etc. EDAS facilitates the construction of high performance parallel workflows by combining canonical analytic operations to enable processing of

huge datasets within limited memory spaces with interactive response times. EDAS services are accessed via a WPS API being developed in collaboration with the ESGF Compute Working Team to support server-side analytics for ESGF. Client packages in Python, Java/Scala, or JavaScript contain everything needed to build and submit EDAS requests. EDAS services include configurable high-performance neural network learning modules designed to operate on the products of EDAS workflows. As a science technology driver, we have explored the capabilities of these services for long-range forecasting of the interannual variation of important regional scale seasonal cycles. Neural networks were trained to forecast All-India Summer Monsoon Rainfall (AISMR) one year in advance using (as input) the top 8-64 principal components of the global surface temperature and 200 hPa geopotential height fields from NASA's MERRA2 and NOAA's Twentieth Century Reanalyses. The promising results from these investigations illustrate the power of easily accessible machine learning services coupled to huge repositories of earth science data. The EDAS architecture brings together the tools, data storage, and high-performance computing required for timely analysis of large-scale data sets, where the data resides, to ultimately produce societal benefits.

Copernicus CP4CDS compute service

Stephan Kindermannn (DKRZ), kindermann@dkrz.de
Ag Stephens (CEDA), ag.stephens@stfc.ac.uk
Sebastien Denvil (IPSL), sebastien.denvil@ipsl.fr

CP4CDS is a project to develop data services in support of the European Copernicus Climate Change Service (C3S). STFC/CEDA (Centre for Environmental Data Analysis, UK), DKRZ and IPSL collaborate to provide a consistent OGC WPS conforming processing interface to a quality assured subset of global (CMIP5) and regional (CORDEX) climate projections (hosted on dedicated ESGF data and index nodes). Special emphasis is put on providing a common packaging framework (for sharing and receiving software environments) as well as a resilient service infrastructure supporting synchronisation across 3 European sites (CEDA, IPSL, DKRZ).
We will present the current developments in the CP4CDS project. This includes:

- A load-balancing solution for compute nodes (also supporting batch processing (cluster) backends)
- An Ansible playbook to install compute nodes.
- A Cookiecutter template to set-up a project specific WPS compute service.
- A Python client to interact with a WPS compute service from the command-line or from IPython notebooks.

These tools are developed within the European Copernicus project and closely relate to the open source Birdhouse OGC WPS ecosystem, and thus are meant as a common ground to support similar projects as well. Currently the processing tools are also used in projects led by CRIM and Ouranos, Canada.
Weblinks:

- https://cp4cds.github.io/
- https://climate.copernicus.eu/
- https://birdhouse.readthedocs.io/

Compute and Machine Learning

High Performance Data Analytics and Machine Learning developments in Ophidia

Sandro Fiore (CMCC Foundation), Sandro.fiore@cmcc.it
Donatello Elia (CMCC Foundation), Donatello.elia@cmcc.it
Giovanni Aloisio (CMCC Foundation and University of Salento), Giovanni.aloisio@unisalento.it

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 I/O server for fast data analysis. It also provides access through standards interfaces like WS-I+ and WPS. Its workflow engine interface allows to implement a variety of analytics and processing chains, parallelism in a transparent way through declarative statements and interleaved mechanisms to cross-link multiple workflows into complex experiments.
A set of HPDA-related recent updates on Ophidia will be presented and demonstrated at the F2F meeting. In particular a set of activities on (i) parallel I/O through specific operators dealing with I/O tasks (e.g. import), (ii) recent ML developments on RNN applied to time series forecasting, and (ii) HPC-based configuration, deployment and test on the Athena HPC cluster running at the CMCC SuperComputing Centre.

OGC Testbed-14: Experiments on Machine Learning and on Earth Observation Exploitation Platforms

Tom Landry (CRIM), tom.landry@crim.ca
David Byrns (CRIM), david.byrns@crim.ca
Jean-François Rajotte (CRIM), Jean-francois.rajotte@crim.ca

Following the completion in 2017 of project PAVICS, the Platform for Analysis and Visualization of Climate Science, several new sustainability projects have been funded in Canada. There is now an opportunity to seek coordination and complementary with ESGF activities and thus to reinforce Canada's contribution in this global effort.

In 2017, a consortium of Canadian institutions was selected as funded recipients for a cyberinfrastructure challenge. The project, codenamed PAVICS NexGen, has already been communicated to ESGF leadership. Through this initiative should arise increased capacities in data computation and management, in support of both climate and EO services. PAVICS NextGen will advance its research platform as well as novel Machine Learning models, tools and techniques.

Two sustainability projects based on PAVICS led by two Canadian universities were funded by CANARIE. The projects will deliver new advanced hydro modelling tools and processes, as well as DeepLearning-based data annotation tools for satellite imagery. In parallel, CRIM continues its implication into Open Geospatial Consortium (OGC) testbeds with intent of implementing, testing and advancing open standards such as Thematic Exploitation Platforms (TEP). Work conducted in OGC Testbed-14 Machine Learning task also envisioning of best practices and interoperable, standardized ML systems for Earth Systems and GeoInt.

CRIM entered final negotiations with Environment and Climate Change Canada (ECCC) and its Canadian Center for Climate Services (CCCS). Through this project and with collaboration of key partners and subcontractors, CRIM will oversee the development and hosting of an interactive, web-based, climate information portal and sectorial modules. Overall, Canadians will benefit from ECCC's contribution through an increased capacity to interact with climate information, data, and tools to inform climate-smart decision-making, thereby increasing resilience to climate change.

Tracking and Predicting Extreme Climate Events using ConvLSTM

Sookyung Kim (LLNL/AIMS), kim79@llnl.gov
Hyojin Kim (LLNL/CASC), kim63@llnl.gov
Mr. Prabhat (LBNL), prabhat@lbl.gov

Tracking and predicting extreme events in spatio-temporal climate data is a major challenge in climate science. Existing approaches to tracking extreme climate events require an appropriate feature selection from physical variables and thresholds by human expertise. To predict extreme climate events, existing methods rely on physics-based climate simulations demanding tremendous computing cost. The recent progress in deep learning provides technical insights by capturing the nonlinear spatio-temporal interactions between a variety of physical variables. We propose two deep-learning-based models to track and predict hurricane trajectories on massive scale climate reanalysis data. First, we address the spatio-temporal tracking as a mapping problem from time-series climate data to time-sequential hurricane heat maps using Convolutional LSTM (ConvLSTM) models. Our result shows that the proposed ConvLSTM-based regression models outperform conventional region-CNN-based detection methods. Second, we present a new trajectory prediction approach as a problem of sequential forecasting from past to future hurricane heat map sequences. Our prediction model using ConvLSTM achieves successful mapping from predicted heat maps to ground truth.

Poster and Demo Session 1: Compute Services and Applications

Web Processing Services Climate Analysis Workflows from a Python Interface

David Huard (Ouranos), huard.david@ouranos.ca
Carsten Ehbrecht (DKRZ), ehbrecht@dkrz.de
Ben Koziol (NOAA), ben.koziol@noaa.gov

Web Processing Services (WPS) is an Open Geospatial Consortium standard defining an interface to pre-defined algorithmic processes available on a network. It specifies how inputs and outputs are communicated between the server and clients. One advantage of WPS is the clear division of labor between the design, implementation, deployment and use of scientific algorithms. Indeed, users can in principle execute complicated processes from their browser by composing an http request, delegating the software installation and hardware maintenance to the server's administrators. In practice, composing WPS requests by hand is unpractical and WPS are generally called through a web frontend or, from the console through clients such as OWSLib. To simplify access to WPS services, the birdhouse project has developed a small package called birdy that provides two simple and intuitive interfaces to WPS processes: a command-line client and a "native" python client. Both use OWSLib and the GetCapabilities, DescribeProcess and Execute operations of the remote WPS server to dynamically create an interface that blends with users' work environment. The native python client dynamically generates a module whose functions look and feel like native Python objects but actually execute remote processes. Scientists and developers can thus blend online WPS functionalities in ordinary scripts, offloading large computation tasks to external dedicated computing resources. Combined with recently added native support for OPeNDAP in PyWPS, we now have an environment where remote resources can be blended seamlessly in local programming environment to facilitate climate analysis workflows, such as computing averaged climate indicators time series on user-defined areas.

Diagnostics Package for Energy Exascale Earth System Model (E3SM_diags)

Chengzhu Zhang (LLNL/AIMS), Zhang40@llnl.gov
Zeshawn Shaheen (LLNL/AIMS), Shaheen2@llnl.gov
Chris Golaz (LLNL/AEED), Golaz1@llnl.gov

A modern, Python-based diagnostics package for evaluating earth system models has been developed by the E3SM project. The goal of this work is to build a comprehensive diagnostics software package as an essential E3SM tool to facilitate the diagnosis of the next generation earth system models. This package is embedded into the E3SM automated process flow to enable seamless transition between model run and diagnostics.
Modeled after NCAR's atmosphere diagnostics package, 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 maps, contour levels) are customizable. Built-in functions to generate derived variables and to select diagnostics regions are supported and can be easily expanded. An updated observational data repository is developed and maintained by this activity.
The architecture of this package follows the Community Diagnostics Package framework, which is also applied by two other DOE funded diagnostics efforts (PCMDI metrics package and ARM diagnostics package), to facilitate effective interactions between different projects.

High-dimensional Cliamte Data Visualization for Advanced Analysis

Jiwoo Lee (LLNL/AIMS), lee1043@llnl.gov
Ken Sperber (LLNL/PCMDI), sperber1@llnl.gov
Peter Gleckler (LLNL/PCMDI), gleckler1@llnl.gov

The Coupled Model Intercomparison Project (CMIP) and its sister projects have generated massive dataset, leading climate researches to big-data intensive territory. One of challenges in for the CMIP is the development of objective metrics and measures of climate model evaluation and interdependency, which efficiently summarize collected PetaByte scale data to give better insight into the data. In the presentation I will introduce two of my recent researches: [1] A newly developed metric for evaluation of simulated climate variability modes, such as Northern Annular Mode (NAM), the North Atlantic Oscillation (NAO), the Pacific North America pattern (PNA), the Southern Annular Mode (SAM), and the Pacific Decadal Oscillation (PDO), and comprehensive data visualization of metrics using a portrait plot, and [2] application of a circular plot to visualize the interdependency of climate models.

CMOR

Chris Mauzey LLNL/AIMS), mauzey1@llnl.gov
Charles Doutriaux LLNL/AIMS), doutriaux1@llnl.gov
Karl E. Taylor PCMDI/LLNL), taylor13@llnl.gov

The Climate Model Output Rewriter (CMOR) is a library used for generating CF-compliant NetCDF files to facilitate model output intercomparison for various Model Intercomparison Projects (MIPs). It is written in C with bindings to Fortran and Python, and is currently supported on Linux and macOS. CMOR also includes PrePARE, a Python-based program used for validating whether NetCDF files conform to the Coupled Model Intercomparison Project Phase 6 (CMIP6) standards for publication into Earth System

Grid Federation (ESGF). Current work on CMOR includes support for unstructured grids, and use of multiple CPU cores for processing many NetCDF files with PrePARE.

Leveraging CAFÉ to enable a collaborative computing paradigm for climate model intercomparision
Yuqi Bai (Department of Earth System Science, Tsinghua University) yuqibai@tsinghua.edu.cn
Bin Wang (State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences wab@tsinghua.edu.cn
Tongwen Wu (Beijing Climate Center, China Meteorological Administration) twwu@cma.gov.cn

As the amount of environmental data expands exponentially worldwide, researchers are challenged to efficiently analyze data maintained in multiple data centers. Because distributed data access, server-side analysis, multinode collaboration, and extensible analytic functions are still research gaps in this field, we introduces a collaborative analysis framework for gridded environmental data, i.e. CAFE. Multiple CAFE nodes can collaborate to perform complex data analysis. Analytic functions are performed near where data are stored. A web-based user interface allows researchers to search for data of interest, submit analytic tasks, check the status of tasks, visualize the analysis results, and download the resulting data files.

A four-node prototype system was established to demonstrate the feasibility of CAFÉ. All CMIP5 Sea Ice climate model data are used as a use case to show the advantage of CAFÉ: climate model intercomparision work could be fulfilled without downloading the model output data first and then archiving them locally.

CAFE facilitates overall research efficiency by dramatically lowering the amount of data that must be transmitted from data centers to researchers for analysis. The results of this study may lead to the further development of collaborative computing paradigm for environmental data analysis.

Day 2: Wednesday December 5

CMIP6 Services in ESGF

Status and Outlook for the CMIP Data Request
Martin Juckes (CEDA), Martin.juckes@stfc.ac.uk
Karl Taylor (LLNL/PCMDI), taylor13@llnl.gov
Matthew Mizielinski (UK Met Office), matthew.mizielinski@metoffice.gov.uk

The Data Request (DReq) of the Coupled Model Intercomparison Project Phase 6 (CMIP6) defines all the quantities from the CMIP6 simulations which are to be archived. The fundamentals of the request are inherited from the success of past CMIP activities, with strong foundations in the Climate and Forecast (CF) Metadata Conventions. CMIP6 brings a new scale and complexity through the diversification into more than 20 endorsed Model Intercomparison Projects (MIPs). The request includes both quantities which are specialised and only of interest to a single endorsed Model Intercomparison Project (MIP) and generic quantities.

The data request can be described in terms of 4 sections: parameter specifications, experiment details, scientific objectives and, finally, the cross linkage. The parameter specifications provide precise definitions of parameters and specify attribute values and directives to define the file metadata required in the CMIP archive. The remaining sections provide details of the subsets of the data which are required for different purposes and different experiments. A significant degree of selectivity was required in order to avoid unnecessary expansion of the volume of data which would need to be generated.

The request is provided through a number of channels in order to match a range of requirements. A versioned XML document provides a robust long term reference, while a web view, a command line tool and a python API provide greater flexibility.

The presentation will review the background, process, architecture, content and lessons learned from the CMIP6 data request and report on initial ideas for CMIP7, based on a workshop held earlier in 2018 on rationalisation of the process. It will also look at the level of support which might be needed.

The ES-DOC documentation workflow for CMIP6

David Hassell (National Centre for Atmospheric Science, UK), david.hassell@ncas.ac.uk

Earth System Documentation (ES-DOC) is a multi-national, multi-institutional collaboration that supplies services in support of earth system documentation creation, analysis and dissemination. Given the wide variety of users and the need for traceability, the CMIP6 project and its results will be fully documented and made accessible via ES-DOC. Viewed from a documentation perspective, the CMIP6 process could be described as follows:

- An ensemble of related simulations is run for a particular experiment. Each simulation is an integration of a configured
- model that has a specified conformance to the numerical requirements of an experiment.
- The simulations are run a machine, with an observed performance, and produce output datasets that are archived on ESGF.
- Throughout the process, particular people and organizations are involved, and published work needs to be cited.
- Individual output datasets are linked to all relevant documentation relevant to the dataset's creation.

The Common Information Model (CIM) is a metadata standard used by the climate research community and others to describe the artifacts and processes they work with. ES-DOC use this framework to document the CMIP6 experiments, ensembles, simulations, models, conformance, responsible parties and citations. Much of the documentation is, wholly
or partially, automatically created from existing resources without the need for human intervention. In particular, the simulation documents are entirely created during ESGF publication from the attributes found in published netCDF datasets. This ensures that all CMIP6 simulations will be documented, compared with fewer than 50% of CMIP5 simulations. The link from an individual dataset to the full documentation of the modelling process which created it is provided by an ES-DOC service that ensures the "further_info_url" dataset attribute resolves to an on-line landing page containing all of the relevant information.

Data Citation Service – status and first experiences

Martina Stockhause DKRZ stockhause@dkrz.de
Michael Lautenschlager DKRZ lautenschlager@dkrz.de

In the introduction, a short summary of the data citation concept is given using the operational input4MIPs data references as example. Next, the individual service components (GUIs and APIs) for ESGF partners, data creators, and data users (data citers) will be presented along with best practices and recommendations.
First experiences with the citation service are analyzed and priorities for future work derived, e.g. general consolidation of individual services, improvement of user support and its integration into the general ESGF support infrastructure, as well as service improvements in terms of usability and implementation of Scholix results.

ESGF Errata Service – Production release and forms

Bennasser Atef (IPSL/CNRS), abennasser@ipsl.jussieu.fr
Levavasseur Guillaume (IPSL/UPMC), glipsl@ipsl.fr
Greenslade Mark (IPSL), momipsl@ipsl.fr

In June 2018, the IPSL moved the ESGF Errata Service to production phase, the services are already available at https://errata.es-doc.org/.
As a part of the ES-DOC ecosystem, the Errata Service exploits the Persistent IDentifier (PID) attached to each dataset and file during the ESGF publication process. Consequently, IPSL is closely working with DKRZ on the required connections and APIs between the errata system and handle services.

The Errata Service offers a user-friendly front-end and a dedicated API to provide timely information about known issues. ESGF users can query about modifications and/or corrections applied to the data in different ways:

- Through the centralized and filtered list of ESGF known issues
- Through the PID lookup interface to get the version history of a (set of) file/dataset(s).

CMIP6, CMIP5 and CORDEX projects are currently supported by the Errata Service, with more to come in the next months. It allows identified and authorized actors of the corresponding modeling groups to create, update and close issues using either a lightweight CLI and/or a new and easy form.
The documentation has been fully revised to guide any user through the errata procedure: https://es-doc.github.io/esdoc-errata-client/

Synda

Bennasser Atef (IPSL/CNRS), abennasser@ipsl.jussieu.fr
Levavasseur Guillaume (IPSL/UPMC), glipsl@ipsl.fr
Denvil Sébastien (IPSL), sdipsl@ipsl.fr

SYNDA is a command-line alternative to the ESGF web front end. SYNDA is entering a new phase of development and should be well out of the bottleneck it has been.
This new phase will include obligatory maintenance tasks varying from bug fixes and release streamlining to including much expected features by the ESGF community. The order of processing will of course be down to priority with CMIP6 replication in mind.
Current main features include:

• Simple data installation using an apt-get–like command
• Support for every ESGF project (e.g., CMIP6, CMIP5, CORDEX...)

• Parallel downloads, incremental process (download only what is new)
• Transfer priority, download management and scheduling, and history stored in a database
• GridFTP is already available but it's on the list of priority tasks to optimize and make better use of the protocol advantages.
• Installation using a docker container and/or Red Hat Package Manager
SYNDA can download files from the ESGF archive in an easy way based on a list of facets (e.g., variables, experiments, and ensemble members). The program evolves together with the ESGF archive back-end functionalities.
This talk will walk through SYNDA's main features from the perspectives of replication and replica publication. Also, ESGF currently only supports an "offline, on-demand" replication procedure, by which dedicated replication sites pull replica sets from ESGF sites, reorganize them to fit into their internal ESGF data organization structure, and publish them as "replicas" into the ESGF data federation.
Future developments include (Non exclusive/exhaustive list):
• CMIP6 syntax related issues
• Different synda package distribution (conda, docker...)
• Synda automatic fallbacks in case of protocol failures, etc.

ESGF Dashboard: data usage statistics on the Earth System Grid Federation
Alessandra Nuzzo (CMCC Foundation), alessandra.nuzzo@cmcc.it
Maria Mirto (CMCC Foundation), maria.mirto@cmcc.it
Paola Nassisi (CMCC Foundation), paola.nassisi@cmcc.it

The monitoring of the Earth System Grid Federation gained high relevance in the last months, especially as far as the publication process of the CMIP6 project data is concerned.
During the last year, in particular, the esgf-dashboard efforts were focused on the CDNOT (CMIP6 Data Node Operation Team) activities with the aim to efficiently gather data usage metrics over the data nodes joining the testbed and to contribute to the delivery of an operative and stable environment for the publication and management of the CMIP6 project data.

Therefore, the dashboard user interface, deployed at the CMCC (Euro-Mediterranean Center on Climate Change) Supercomputing Center, also gathers and provides "Cross-project" and "Project-specific" statistics sections for CMIP6 data, with a rich set of charts and reports, allowing users and system managers to visualize the status of the infrastructure through a set of smart and attractive web gadgets. Moreover, a view of the total amount of data published and available through the ESGF infrastructure offers users the possibility to monitor the status of the data archive of the entire federation.

New widgets were released last year on the user interface to improve the users' experience and give them the possibility to get data usage statistics over a wider range of data perspectives (information by project, host, experiment, model, etc.).

CMIP6 Experience

NCI production experience with CMIP6
Kate Snow (National Computational Infrastructure), kate.snow@anu.edu.au
Ben Evans (National Computational Infrastructure), ben.evans@anu.edu.au
Chris Allen (National Computational Infrastructure), chris.allen@anu.edu.au

We will present on two components from our experience at NCI:

1. A brief summary of our CMIP5/CMIP6 status:
- Publication of Australian CMIP6 and re-publication of CMIP5 Data.
- Priority downloaded variables for CMIP6 Australian climate community.

2. ESGF Technical infrastructure in production:
- CMIP5 & 6 data transfer speeds and replication experience.
- Re-publication workflow for CMIP5 and CMIP6.
- Maintaining our ESGF production node: NCI's deployment strategy and enhancements.
- Update the user and operational experience for CMIP6 so far.

Finally, an alternative way for users to get federated statistics has been made available through the deployment of the esgf-stats-api service on the collector node, providing aggregated statistics over the different data nodes of the federation.

During the talk, last year activities will be illustrated. The future work planning that will be presented will also take into account the forthcoming start of the IS-ENES3 project.

input4MIPs: ESGF experiences of a 2-year-old project

Paul J. Durack (LLNL/PCMDI), pauldurack@llnl.gov
Karl E. Taylor (LLNL/PCMDI), taylor13@llnl.gov
Sasha K. Ames (LLNL/AIMS), ames4@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 CMIP6. The project was initiated in 2016 and has subsequently becoming the primary source of input data for CMIP6 modeling centers. Various datasets are needed for the pre-industrial control (piControl), AMIP, historical, 1 percent compounding CO2 and abrupt four times CO2 simulations, and additional datasets are needed for many of the 23 CMIP6-endorsed model intercomparison projects (MIPs) experiments that comprise the CMIP6 project. Unlike model data generated from CMIP6 experiments and standardized using CMOR or similar tools, the formats of these contributed datasets are various 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 as to how the publishing step was augmented to deal with these highly variable data formats. These will be useful to consider as ESGF further evolves to address the requirements of obs4MIPs and other new international projects.

ESGF Service Interoperability

Containerising ESGF for CP4CDS

Matt Pryor (Centre for Environmental Data Analysis, STFC), matt.pryor@stfc.ac.uk
Philip Kershaw (Centre for Environmental Data Analysis, STFC), philip.kershaw@stfc.ac.uk

The CP4CDS data services will provide global climate projections, initially from CMIP5, to the Copernicus Climate Change Service (C3S) Climate Data Store (CDS). In turn the CDS provides a single, freely

available, public interface to a range of climate-related observations and simulations in support of users across all sectors. The CP4CDS data services will be provided by CEDA and other project partners using ESGF components. The components must be deployed in a highly available and resilient manner - to achieve this, they will be deployed at multiple sites as no single site can guarantee the required availability. Rather than behaving as a federation, the sites will be load-balanced under a single domain, appearing to the user as a single deployment. Containers, with an orchestration system such as Kubernetes, have many attractive features for building a highly available system, e.g. the ability to scale-out in response to demand and increased resiliency through health-checks and self-healing. Containers improve portability and simplify installation, as each container bundles the correct versions of any dependencies. They also encourage modularity, with each container performing a specific job and working together to provide an integrated system. Working with the ESGF Container Working Team, the ESGF components have been containerised. This has been a challenge, as it is a significant departure from the traditional installer – e.g. it has been difficult to extract the configuration for each individual app, as a traditional installation bundles many apps together with shared config files. However, we now have a proof-of-concept ESGF node that can run using either Docker Compose or Kubernetes. In order to meet our availability goals for CP4CDS, we will also make use of Public Cloud. Currently, Amazon's Route 53 is used for DNS load-balancing. We are also investigating the possibility of using Kubernetes on Google Cloud Platform to provide the required availability for the index node (unfortunately, storage is still too expensive for us to host a data node in the public cloud).

The application of ESGF to develop an Open Data Portal for ESA Climate Change Initiative

Philip Kershaw (CEDA, RAL Space, STFC Rutherford Appleton Laboratory), Philip.kershaw@stfc.ac.uk
Victoria Bennett (CEDA, RAL Space, STFC Rutherford Appleton Laboratory), Victoria.Bennett@stfc.ac.uk
Antony Wilson (Scientific Computing Department, STFC Rutherford Appleton Laboratory), Antony.Wilson@stfc.ac.uk

The ESA (European Space Agency) Climate Change Initiative (CCI) is a major programme to develop Essential Climate Variables (ECV) derived from satellite data. This was first initiated as response to the UNFCCC (United Nations Framework Convention on Climate Change) call to create a database of ECVs and also to more fully exploit long-term data archives of earth observation datasets held by ESA and its member states. As part of this programme, ESA commissioned the development of the Open Data Portal (http://cci.esa.int/data/), a single point of access for the data providing it in a consistent and harmonised form in order to support its broad dissemination. To facilitate this goal, ESA in their requirements explicitly specified the use of ESGF, along with other international collaborations and the application of various standards and protocols relevant to the climate science user community.
CEDA won the contract to develop the Portal as part of a consortium led by commercial supplier Telespazio VEGA UK. With responsibility for the data archive and metadata catalogue, CEDA was tasked with a challenging integration and development activity, bringing together varied datasets from across the initial 13 ECVs funded for the programme into a single system. We highlight specific innovations related to ESGF: the use of Linked Data technologies in the development of controlled vocabularies and services to support a DRS (Data Reference Syntax) for CCI, the integration of ESGF Search with OGC (Open Geospatial Consortium) CSW (Catalogue Services for the Web) search services and the extensive use of aggregation capabilities to make long time series data for OPeNDAP and OGC WMS (Web Map Service) using THREDDS Data Server.

Climate Data in Geospatial Information Systems: ESGF and OGC coordination opportunities

George Percivall (Open Geospatial Consortium), gpercivall@opengeospatial.org

The ESGF enterprise for management, dissemination, and analysis of model output and observational data can expand to additional application communities through interaction with the Open Geospatial Consortium (OGC). OGC's mission is to advance the development and use of international standards and supporting services for geospatial interoperability. To accomplish this mission, OGC serves as the global forum for the collaboration of geospatial data / solution providers and users. Coordination of OGC and ESGF is discussed here in three horizons: Now, Next and After-Next. Now, ESGF activities are using the OGC Web Processing Standard (WPS). WPS and other OGC Standards provide machine-to-machine interoperability for geographic information systems [1]. Multiple implementations of WPS to ESGF have been made already with contributions to refinement of the WPS standard. OGC's Testbed 13 built on ESGF/WPS interoperability to make climate prediction readily available for agriculture crop prediction [2]. Next, ESGF needs can be factored into the next version of WPS currently under development. OGC Web Service standards are currently undergoing an evolution based on resource-oriented architectures and the use of OpenAPI tools. ESGF participation in the revision process would advance the capabilities of WPS for model outputs interoperability. Additional relevant OGC activities can be considered by ESGF, e.g., OGC netCDF, OGC HDF5, security, grid/cloud, and the OGC Earth science WGs. Based on recent DoE award, OGC working with CRIM will develop extensions of an ESGF Hybrid Climate Data Research Platform. After-Next, as part of the Technology Trends activity, OGC is developing a technology roadmap that examines opportunities for the coordination of predictive modeling, e.g., as accessed by ESGF, with other modeling, simulation and prediction technologies. An anticipated outcome of the Roadmap is the environmental simulation enhanced by the best of scientific Considering these activities, increased coordination and possible partnership between OGC & ESGF would be a good basis and motivation for both communities to identify common ground and find ways to work together. [1] https://ieeexplore.ieee.org/document/7570342 [2] http://docs.opengeospatial.org/per/17-022.html

ESGF in Analysis Applications

C4I: Leveraging and Easing End Users' CMIP6 Access by Interfacing Infrastructures

Christian Pagé (CERFACS), christian.page@cerfacs.fr
Wim Som de Cerff (KNMI), Wim.Som.de.Cerff@knmi.nl
Maarten Plieger (KNMI), maarten.plieger@knmi.nl
Alessandro Spinuso (KNMI), alessandro.spinuso@knmi.nl

Proper climate data access is getting to another dimension with CMIP6, as data volumes are increasing very fast compared to CMIP5, resulting in difficulties to process and analyze needed data for research and applications. This is especially true for end users. The whole climate data archive is expected to reach a volume of an estimated 30 Pb in 2018 and up to potentially 2000 Pb in 2022. On-demand data processing solutions as close as possible to the data storage are emerging and are absolutely needed, thanks to newly developed standards, provenance and infrastructures.

In Europe several initiatives are taking place to support scientific on-demand data analytics at the European scale. They offer the huge potential of interoperability, as for example the DARE e-science platform (http://project-dare.eu), designed for efficient and traceable development of complex experiments and domain-specific services on the Cloud. Also, 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). The platform is based on existing standards (ISO and OGC), such as

WPS (Web Processing Service). DARE will integrate services from the EUDAT CDI, enabling generic access and cross-domain interoperability, as well as providing compliance and integration with the future EOSC platform, and interfacing with the ESGF Compute Working Team (CWT) Computing Nodes.

C4I is an alternative interface to COG with respect to the access to CMIP6 data, as described in the official CMIP6 Users' Guide. C4I provides computing services, abstracting several APIs and platforms that will provide processing services, such as the ESGF CWT Computing Nodes, the DARE Platform, EGI, EUDAT, etc. It also helps the user cope well with large datasets and fragmented data files.

In this presentation an overview of C4I focussed on processing services will be presented. The integration of several processing backends will be discussed, as well as the DARE Platform architecture along with its provenance system and integration with the ESGF CWT.

This work is funded by the H2020-DARE project, and has been previously funded by FP7-IS-ENES2, FP7-IS-ENES, and FP7-CLIPC.

The ENES Climate Analytics Service (ECAS)

Sofiane Bendoukha (DKRZ), bendoukha@dkrz.de
Alessandro d'Anca (CMCC), alessandro.danca@cmcc.it
Sandro Fiore (CMCC), sandro.fiore@cmcc.it
Tobias Weigel (DKRZ), weigel@dkrz.de

In the Horizon 2020 project EOSC-hub [1], the major components of the European Open Science Cloud (EOSC) are getting integrated, based on prior work of large e-infrastructures (EUDAT, EGI and Indigo DataCloud) and domain-specific research infrastructures such as ENES.

Within EOSC-hub, DKRZ and CMCC develop a service called ECAS, the ENES Climate Analytics Service [2], based on offering the computing capabilities of the Ophidia framework to a wider range of users also beyond ENES by building a user-friendly data ingest and sharing workflow with other EOSC components. The fully integrated ECAS environment will provide a JupyterHub-based workbench, where users can access CMIP5 and CMIP6 data via ESGF, design and re-use Ophidia workflows, and then share their results via EUDAT B2DROP and B2SHARE. A main goal of ECAS is to empower users to exploit server-side computing capabilities, triggering a cultural change that will both open up a wider range of datasets and shared workflows to users and reduce the load on data replication services.

Currently, early versions of the integrated service are available at DKRZ [3] and CMCC [4], with the final service installations expected end of next year. Training for early adopters is part of the ECAS concept, and a first training on ECAS and Ophidia has been given in September 2018 as part of the ESiWACE workshop on workflows in Brussels [5]. New example workflows supported by ECAS and used in training include calculation of various climate indices, such as tropical nights, frost days and daily temperature ranges [6].

[1] https://www.eosc-hub.eu
[2] https://portal.enes.org/data/data-metadata-service/processing/ecas/enes-climate-analytics-service-ecas
[3] https://ecaslab.dkrz.de

[4] https://ophidialab.cmcc.it
[5] https://www.esiwace.eu/events/esiwace-workshop-on-workflows-1/esiwace-workshop-on-workflows
[6] https://github.com/ECAS-Lab/ecas-notebooks

Pangeo: a flexible climate analytics infrastructure for climate data on ESGF

Ryan Abernathy (Columbia University), rpa@ldeo.columbia.edu

Pangeo (https://pangeo.io/) is a climate data analytics ecosystem built around widely-used python infrastructure for scientific computation and analysis. The infrastructure includes a Jupyter notebook front end for interactive and collaborative analytics development; scientific computing middleware based on xarrays/scipy/numpy, and a DASK backend for scalable analytics that can be deployed on the cloud or on dedicated computing clusters. In this talk we present examples of Pangeo running in conjunction with the cloud-based invocation of an ESGF data node (see talks by Cinquini and Nikonov). We demonstrate task parallel analysis of CMIP6 data hosted on an ESGF data node hosted on the Google Cloud Platform.

Fully integrated solution for interactive data conversion, streaming, analysis and visualization of scientific data using containers and web-based technologies

Steve Petruzza (SCI Institute – University of Utah), spetruzza@sci.utah.edu
Cameron Christensen (SCI Institute – University of Utah), cam@sci.utah.edu
Valerio Pascucci (SCI Institute – University of Utah), pascucci@sci.utah.edu

The increasing resolution of data from large simulations and acquisitions is a major challenge for access, sharing, visualization and analysis. To improve data access and enable multiresolution streaming we utilize the IDX format.
While the IDX format enables excellent data streaming performance, it only stores information about the grid and the layout of the data on disk. But, formats such as NetCDF include a rich set of metadata used for visualization and analysis purposes. Therefore we created a new metadata format called XIDX (i.e., eXtensible IDX), which enables independence of metadata from any underlying formats such as binary, xml, text, HDF5, or NetCDF. It is XML-based and supports all metadata information currently contained in the NetCDF and HDF5 formats. Users who want to convert their datasets to a more efficient streaming format such as IDX can now preserve and utilize all metadata information contained in the original datasets.
A new Docker-based deployment is now available, that includes the ViSUS server, on-demand data conversion, and a convenient web-based viewer for easy access and visualization of massive datasets. Using the on-demand component of this container, requested datasets are transparently converted to IDX and immediately accessible through the integrated server for streaming analysis and visualization. All metadata are mapped to the new XIDX format and available via the built-in server. The included web viewer enables interactive visualization of both 2D and 3D time series datasets, streaming data at different resolutions. The data can also be explicitly requested and downloaded by users for their own purposes. Furthermore, specific web viewer setups can be easily shared using auto-generated hyperlinks. This container with streaming data server, conversion service and web viewer has also been successfully tested within the new ESGF containerized federation.

Finally, we released the OpenViSUS framework to facilitate adoption and integration of our entire data streaming ecosystem, providing benefits to multiple scientific communities. In this same direction, we enabled deployment of the system as a Python module that can easily be installed (e.g., via pip) and used from standalone applications or through new web-based technologies such as Jupyter notebook and lab.

vCDAT a Web Based Data Visualization Suite for a Broader Community
Carlos Downie (LLNL/AIMS), Downie4@llnl.gov
Charles Doutriaux (LLNL/AIMS), Doutriaux1@llnl.gov
Sterling Baldwin (LLNL/AIMS), Baldwin32@llnl.gov

vCDAT is the graphical frontend of the Community Data Analysis Tools (CDAT). One of the challenges and reasons for building vCDAT 1.0 is to provide a set of visualization tools that are more user friendly and remove the hassles associated with installation of dependencies, while continuing to incorporate user feedback and build upon the existing functionalities of the Ultra-scale Visualization Climate Data Analysis Tools (UVCDAT). To address the challenge, vCDAT 1.0 was built as an in-browser end-user experience that leverages CDAT's core services for the server-side backend and manages the connection between client and server by using the vcsjs library. For future efforts, we envision a deeper python/client embedded experience that takes advantage of JupyterLab technology and fully connects vCDAT with functionalities offered by the Compute Working Team (CWT). By utilizing the latest web-based technologies, the open-source vCDAT project can provide the powerful analysis, diagnosis, and visualization capabilities provided by CDAT to a broader community via a easily accessible, browser-based platform.

Poster Session 2: CMIP6 and Cloud

ESGF PID services status and future perspectives
Tobias Weigel (DKRZ), weigel@dkrz.de
Stephan Kindermann (DKRZ), kindermann@dkrz.de
Katharina Berger (DKRZ), berger@dkrz.de

The PID services for CMIP6 consist of the following components:

- An agreed workflow, described in the WIP papers, where each CMIP6 file must have a specific tracking_id format (hdl:21.14100/<UUID>). CMOR has been extended and configured for CMIP6 to check this.
- A Python library (esgf-pid) and configuration for the ESGF publisher that submits PID requests to the message queue.
- The RabbitMQ message queue system, consisting of 3 entry nodes (DKRZ, PCMDI, IPSL) and 1 exit node (DKRZ).
- The Handle server for managing PIDs with a component that interprets RabbitMQ messages (consumer).

All components are stable and in operation. The most significant irregularity in 2018 was undetected downtime of 2 of the 3 exit nodes. The servers at IPSL and PCMDI were fixed and monitoring was put in place at DKRZ to detect such issues early on.
Throughout the CMIP6 data challenges, the performance at the exit node was always sufficient to process all incoming requests. No messages were queued for longer than the smallest monitoring period

(5 minutes). Monitoring was extended throughout the year, and the execution rate (Handles/second) data available until now shows a maximum at 150 H/s, with no messages piling up. Nonetheless, in case Handle server performance turns out to be a limiting factor in the future, several contingency strategies exist, including configuration of the PID server to write to disk asynchronously, database optimization and introducing multiple exit nodes.
A future development perspective, based on the stable operational experience, is to support replication with PIDs. This would rely on creating dynamically additional PIDs for higher hierarchy levels of the DRS syntax, offering additional granularities that include those most common for replication. A mechanism for replication agents to create individual collections can complement this. Both would rely on the scalable message queue already in place. Replication can then be supported by packaging the relevant data by referencing it via PIDs, and letting replication scripts make use of additional PID metadata. The details of the collection levels and the extended PID metadata schema will have to be defined. Implementation and deployment may be of average complexity as the replication sites are few and well known.

Cooperation with Google for deploying ESGF Node on Google Cloud Platform

Serguei Nikonov (Princeton University/NOAA GFDL), serguei.nikonov@noaa.gov
Hans Vahlenkamp (UCAR/NOAA GFDL)
Aparna Radhakrishnan (Engility Corporation/NOAA GFDL)
Karan Bhatia (Google Cloud)
V. Balaji (Princeton University/NOAA GFDL)
Luca Cinquini (NASA/JPL), luca.cinquini@jpl.nasa.gov

Cooperation of GFDL and Google purposes the dual objects - containerization of ESGF and demonstrating reliable exploiting an ESGF Node running on Google Cloud Platform. Nowadays, it's a common place that deployment of large heterogeneous complex system like ESGF via Docker Containers technology is main stream in modern IT. GFDL made use of an opportunity of cooperation with Google to investigate this technology in practice, namely how migration onto Google Kubernetes cluster will simplify the non-trivial ESGF installation and maintenance process. As mitigating difficulties in deployment is not enough reason to adopt this approach, the second declared target was to get solid evidence of full functionality of ESGF Node, its acceptable performance and reliability in main actions - publishing data (CMIP6) and deploying data access services. Achieving these goals will motivate the scientific community to search for ways of economically affordable cooperation with clouds providers to migrate to new technologies.

An automated heterogeneous pipeline for managing CMIP6 ingestion, replication and publication at CEDA

Ruth Petrie (STFC CEDA), ruth.petrie@stfc.ac.uk
Ag Stephens (STFC CEDA), ag.stephens@stfc.ac.uk
Alan Iwi (STFC CEDA), alan.iwi@stfc.ac.uk

CEDA manages the UK nodes for the dissemination of climate simulations through the Earth System Grid Federation (ESGF). This involves publication of Met Office Hadley Centre (MOHC) datasets to be provided for CMIP6. We have developed an automated pipeline for remotely synchronizing the contents of the Met Office MASS tape archive to CEDA followed by subsequent publication to ESGF. The ingestion-to-publication pipeline is complex because it requires access to multiple services running

across a range of platforms. Deployed within a client-server architecture, a collection of distributed workers query the centralized database for instructions in order to manage their own processes and workloads. Each independent worker runs its 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 control over the rest of the pipeline and can operate in sandboxed isolation. Since it may be necessary to reverse an operation the workers are equipped with both “DO” and “UNDO” actions and all events are recorded in a relational database. A Django application provides queryable web 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. The entire design is modular, so new processing stages and chains can be added as the system grows to cater for emerging projects and requirements.

How can the Coordinated Regional Climate Downscaling Experiment (CORDEX) community facilitate dissemination and analysis of the next generation climate simulations in the Cloud?

Huikyo Lee (Jet Propulsion Laboratory, California Institute of Technology), huikyo.lee@jpl.nasa.gov
Jeongmin Han (APEC Climate Center), goal@apcc21.org
Jiwoo Lee (LLNL/AIMS), lee1043@llnl.gov

We describe a successful use case of processing Earth Science Data from NASA on Amazon Web Service (AWS) and potential benefit of utilizing AWS as a platform to support the Coordinated Regional Downscaling Experiment (CORDEX). CORDEX is an international modeling effort that parallels the Coupled Model Intercomparison Project (CMIP) but with a focus on regional-scale climate change. To complement CMIP, which is based on global climate model simulations at relatively coarse resolutions, CORDEX aims to improve our understanding of climate variability and changes at regional scales by providing higher resolution regional climate model (RCM) simulations for 14 domains around the world compared to CMIP global climate models. The Earth System Grid Federation (ESGF) already hosts a massive amount of RCM output for CORDEX. As more RCMs with high resolution participate in CORDEX by downscaling CMIP6 output, the ESGF infrastructure would face a challenge of providing a common framework where users in different CORDEX domains (e.g. North/South America, Europe, East Asia, and so on) can easily access the RCM simulations conducted for their domains of interest. Especially, it is a high priority to facilitate evaluation of the next-generation CORDEX RCM simulations within each domain. Analysis of multiple RCMs with relatively high resolution also requires the appropriate architectural framework capable of manipulating large datasets. We believe that the geographically separated AWS regions can provide optimal infrastructure for hosting the high-resolution CORDEX RCM simulations, because most of the demand for RCM simulations for each CORDEX domain is from countries within the domain. Our training at a recent CORDEX training workshop demonstrated the value of maintaining a large amount of datasets from the NASA Earth Exchange (NEX) in AWS S3. The Regional Climate Model Evaluation System (RCMES) is a software package that offers a variety of tools to evaluate the CORDEX RCMs. The diagnostic tools included in RCMES have been also proven to be useful for processing and analyzing NEX datasets on AWS. RCMES will provide a complete start-to-finish workflow to access errors in RCM simulations over the 14 CORDEX domains using observational data from the RCMES database and other data sources available on AWS.

Data long-term archival in the IPCC DDC and DDC Support
Martina Stockhause (DKRZ), stockhause@dkrz.de
Michael Lautenschlager (DKRZ), lautenschlager@dkrz.de
DDC Support Group (IPCC WGI/WGII TSUs and DDCs at CEDA and DKRZ),
https://cedadev.github.io/ipcc_ddc/

Driven by experiences within IPCC AR5 the role of the DDC (Data Distribution Centre) was reviewed resulting in a closer co-operation between Working Groups and DDC, which led to the formation of the informal DDC Support group. The main tasks of the DDC Support group will be presented. The long-term archival (LTA) procedure established for CMIP5 was successfully applied in several other federated Earth System Modeling projects, e.g. CORDEX or HAPPI-MIP. This procedure needs to be adapted to additional or renewed infrastructure components for CMIP6, e.g. ES-DOC including Errata. The responsibility for connecting these pieces of information to the data has moved to the service providers. The LTA will access the links published in the ESGF index to enrich metadata. Required curation beyond CMIP6, include paper references and detected errors, the second requiring a stable and reliable Errata API for machine access.

Day 3: Thursday December 6

Topics from ESGF Working Teams

Next Generation ESGF Search Services
Luca Cinquini (NASA/JPL), luca.cinquini@jpl.nasa.gov

The ESGF search services, which retrieve results from all nodes in the federation in a timely manner, is arguably one of the most useful features of the entire ESGF software stack. Nonetheless, the ESGF search is affected by a few shortcomings, such as a lot of manual configuration, the possibility of replica failure, and untested scalability to increasingly larger number of records and index nodes.
This talk will present a new proposed architecture for the ESGF search services, based on newer technologies such as Solr Cloud, Docker containers, and Kubernetes, which promises to be able to support a larger federation, and much larger metadata holdings, for years to come. Additionally, we will describe software tools that enable records migration and synchronization across nodes in the federation. We will conclude with a roadmap for deploying this new architecture within a few months.

ESGF Installer 3.0
William Hill (LLNL/AIMS), Hill119@llnl.gov
Nathan Carlson (LLNL/AIMS Carlson60@llnl.gov
Lina Muryanto (LLNL/AIMS Muryanto1@llnl.gov

Previous versions of the ESGF installer has been marred by inconsistent performance and hard to maintain code. The IWT at LLNL has taken on the arduous task of completely refactoring the installation scripts from the ground up. The code has been rearchitected and implemented in Python. The refactor has provided several benefits including more maintainable code, an automated testing suite, improved performance, and more robust documentation. We will discuss the challenges we faced, the new structure of the code, and the new processes we put in place to improve code and documentation quality.

ESGF Test Suite and Jenkins at IPSL

Sébastien Gardoll (IPSL/CNRS), sgardoll@ipsl.fr
Matt Pryor (CEDA), Matt.Pryor@stfc.ac.uk

ESGF-test-suite is a stable command line testing suite based on Nose, a Python2/3 testing framework and is available at https://github.com/ESGF/esgf-test-suite. As leverage of regression and integration testing, ESGF-test-suite aims to aggregate all the integration tests of the ESGF stack into one location and automate their running. Currently it is able to verify that every high level functionality of an ESGF node is up and running (stats, orp, cog logging, thredds, http and gridftp downloading, etc.). It can manage authentication with myproxy and IDP, and is able to perform browser-based tests (selenium). Thanks to the Nose engine, tests are selectable so as to run sets of tests. Moreover, presets are provided and organized like a mind map. Nose functionalities (and so test reporting) are pretty extensible via plugin system. ESGF-test-suite dependencies are packaged into a singularity image (containerization system) that makes it easy to deploy. As examples, ESGF-test-suite validates the ESGF-docker images in their automated procedure of release and it is used to monitor the ESGF nodes at IPSL.

ESGF-Jenkins is an instance of Jenkins, a continuous integration platform. This service, which is hosted and maintained at the IPSL, is offered to the ESGF community as leverage for continuous integration and best practices. It consists of a cluster of three nodes: a master and its workers. The master holds a web frontend (https://esgf-build.ipsl.upmc.fr/jenkins/) that displays a nice control panel and facilitates the creation of jobs. They consist of various types of task as running periodically a program or reacting to specific events from Internet like a push into a Github repository. Jenkins naturally comes with a good security level that includes a credential manager. It is pretty extensible and interconnectable via its plugin system and its RESTful API that can be remotely and securely run from tailored applications. ESGF-Jenkins comes with a set of plugins like Github, Docker and Slack connectors and some other tools (Maven, Gradle, etc.). Every node (including the master node) of ESGF-Jenkins are able to run up to two jobs at the same time and every worker can run up to one set of docker and singularity images, at the same time. This infrastructure is easily scalable and support failover.
As an example, ESGF-Jenkins runs a job that builds and tests the images of ESGF-Docker. This job, versioned in the ESGF-Docker repository as Jenkinsfile, is a good start to understand what can be done with Jenkins in terms of interconnection, docker, testing, and error handling. It implements one of the following behaviors:
Every pull request, triggered in the ESGF-Docker repository, that targets the devel or master branch, is built and tested by the Jenkins job (the github repository is pulled locally on ESGF-Jenkins). A successful build gives a set of ESGF-Docker images. The tests consist of running ESGF-test-suite against living ESGF-Docker containers. If all the tests passed successfully, Jenkins sends green light to Github, allowing the user to merge his/her pull request into the targeted branch.

Working Teams Update: Publication and ESGF Services

Sasha Ames (LLNL/AIMS), sasha@llnl.gov

We present an overview of ESGF activities concerning the ESGF publisher software. While this software package did not undergo any major revisions, there were a series of feature updates, some of which were crucial for the CMIP6 publication process. Specifically, through iterations of testing during the "Data Challenges" conducted by the CDNOT, we refined the procedure for which the integrated PrePARE module properly sources its input based on the requirements of the model data and present state of the "data specs" and controlled vocabulary. Moreover, experience with the publisher suggests several modifications that are needed and areas for investigation. For instance input4MIPs presents a project in

which extended metadata could be better integrated into the publisher for indexing. A key issue was identified through the E3SM publishing process, where scalability pushes our systems to their limits.

The second part of the presentation will cover several complementary ESGF web services. The ESGF Node Manager is transitioning to a registry service that is simplified without an active (daemon) process. Development of user notification service has progressed with the completion of a prototype capable of generating email notification based on user preferences out of a database based on real changes to the Solr index. We have implemented a simulator of publication related events (new updated, and retracted datasets) that run at regular intervals in order to create a reasonable testing environment for notifications, and so far are pleased with the results. Future work involves the implementation of web UI components to enable subscriptions and scalability improvements to the back-end search algorithm.

Review of development activities for ESGF IdEA – Identity, Entitlement and Access Management

Philip Kershaw (CEDA, RAL Space, STFC Rutherford Appleton Laboratory) Philip.kershaw@stfc.ac.uk

We report on the work of the IdEA Working Team over the past twelve months. Steady progress has been made integrating OAuth into ESGF with releases including first the SLCS (Short-Lived Credential Service) and OAuth server components and more recently, the development of the new 'AuthClient' to replace the legacy OpenID Relying Party used to secure access to the Data Node. The SLCS provides a replacement for MyProxy and in conjunction with OAuth allows clients to obtained delegated user certificates. Work to integrate the AuthClient with CoG will complete the capability needed and allow the whole federation to migrate to OAuth from OpenID 2.0. As part of this work, OAuth and the SLCS will be used to improve the current wget script system, embedding a user certificate in the script and so dispensing with the need for users to explicitly authenticate when they download files. In the mid to long term, further work is needed to review and improve the whole system for user-scripted data download. We will be considering alternatives to wget and in addition, we will need to find solutions to collectively manage download of data with open and restricted access policies. Looking more broadly, we will lay out proposals for the future architecture of the IdEA system to allow token-based download for scripted access and also, the effective management of client trust with a network of OAuth providers. This will also consider, cross-domain efforts in the research community to establish a standard blueprint for identity federations as proposed by the AARC project (https://aarc-project.eu/wp-content/uploads/2017/04/AARC-BPA-2017.pdf).

Pyessv: A Simple Controlled Vocabulary Service

Mark Greenslade (IPSL), momipsl@ipsl.fr
Guillaume Levavasseur (IPSL), glipsl@ipsl.fr
Philip Kershaw (UKRI STFC, Rutherford Appleton Laboratory), philip.kershaw@stfc.ac.uk

The ES-DOC tooling eco-system is diverse and extensively leverages controlled vocabularies. Such vocabularies are used to validate documents: generate URL's; render user interface(s); initialize GitHub repositories; read/write spreadsheets, check data compliance, organize data on file systems, drive automated jobs … etc.
At the heart of the ES-DOC tooling chain's approach to controlled vocabularies is a lightweight but robust python library called pyessv – python earth sciences standard vocabularies. The pyessv library defines a vocabulary as 'a scoped collection of terms governed by an authority'.
With the pyessv library it is relatively trivial to create and save vocabulary collections. Associated with the library is an archive, a repository of vocabularies stored in a simple normalized JSON format. The

archive is populated with vocabularies derived from an array of sources: WCRP CMIP6 json files; ESG-F publisher ini files, ES-DOC CMIP6 specializations … etc.
The vocabularies hosted within the archive are also available over web-service endpoint(s) – this streamlines the experience of using pyessv. The pyessv library and web-service are used extensively by the ESG-F Dataset Errata system.
This presentation is designed to open up a discussion by showcasing the pyessv library and it's current feature set. Such a discussion will inevitably touch upon potential usage with the ESG-F context, plus a roadmap of future features such as SKOS/OWL support.

Community Data Management Systems for CMIP6

Denis Nadeau (LLNL/AIMS), nadeau1@llnl.gov
Dean N. Williams (LLNL/AIMS), williams13@llnl.gov
Charles Doutriaux (LLNL/AIMS), doutriaux1@llnl.gov

The "Community Data Management System (CDMS)" was designed in the mid to late 90's. Its original intent was to automatically locate and extract metadata (i.e., variables, dimensions, grids, etc.) from collections of simulation runs and analysis files. CDMS has been redesigned to run in python 3 and allows regridding using the latest Earth System Modeling Framework (ESMF) Application Programming Interface (API). CDMS is now using DASK which provides enhanced parallelism and analytic in a cluster or by taking advantage of a multi-core machine. CDMS can read Coupled Model Intercomparison Project Phase 6 (CMIP6) directly from Earth System Grid Federation (ESGF) using SSL certificates. CDMS documentation can be found on "read the docs" web site which includes its latest API. CDMS aims to incorporate 21st century technologies and integrate additional geoscience domains. In addition to conforming to the latest community standards and protocols, the new CDAT ingest package.

Common Input Arguments with CDP

Zeshawn Shaheen (LLNL/AIMS), shaheen2@llnl.gov
Charles Doutriaux (LLNL/AIMS), doutriaux1@llnl.gov
Peter Gleckler (LLNL/PCMDI), gleckler1@llnl.gov

Many modeling centers are interested in using multiple internally or externally developed analysis packages. They and other potential users can experience substantial benefit if independently developed packages can be operated in a somewhat similar manner. The common input options provided by the Common Input Arguments (CIA), a part of the Community Diagnostics Package (CDP), offers a starting point for creators of analysis packages to integrate this feature into their software. These options can be easily be modified or appended as they are being used. This feature, combined with CDP, serves to incrementally building a framework that can facilitate inter-operability of different analysis packages.

Rethinking of ESGF Architecture

Philip Kershaw (CEDA, RAL Space, STFC Rutherford Appleton Laboratory) Philip.kershaw@stfc.ac.uk
Ben Evans (National Computational Infrastructure), ben.evans@anu.edu.au

It is now nearly ten years since the original ESGF system was designed. At that time, the development was driven by the need to create a globally distributed archive for downloading CMIP data. Since that time, ESGF has grown and diversified in the number of projects and disciplines it supports; the

operational requirements are clearer for the ESGF to support an international federated archive of this size; many of the ESGF nodes now have other functions beyond CMIP; and there has also been changing landscape of data repository, science and user needs. In light of these changes, there is a need for a fundamental rethink of ESGF's architecture, the suitability of now legacy software that make up the components in the system, developments in informatics for the Earth sciences and wider changes in the world of IT.

Poster Session 3: ESGF Working teams and next-generation ideas

ESGF Installer validation with Jenkins

Lina Muryanto (LLNL/AIMS), Muryanto1@llnl.gov
Sasha Ames (LLNL/AIMS), Ames4@llnl.gov

IWT at LLNL has implemented an automated validation of ESGF installer 2.x and upcoming ESGF installer 3.0 using Jenkins automation server. The Jenkins Pipeline starts with a VM snapshot which has the minimum setup needed to clone from github. It clones the esgf-installer git repo to the VM, runs the esgf_bootstrap, prepares the auto install config file, and launches esg-autoinstall, takes care of post install steps, and run esgf-test-suite and publisher tests. One big benefit of the validation of ESGF installer is that the process can be repeated consistently with just a click of a button. In addition to that, the pipeline can be easily duplicated to validate esgf-installer that is on a git branch that has not been merged to master. This ensures the stability of the installer code on the master branch. One of the challenges in implementing this automation is to ensure that every step, stdout and stderr are logged for post install review.

Applying object oriented design principles to installation processes

Nathan Carlson (LLNL/AIMS), carlson60@llnl.gov
William Hill (LLNL/AIMS), hill119@llnl.gov
Sasha Ames (LLNL/AIMS), sasha@llnl.gov

The ESGF Node installer has undergone a translation from the Bash programing language to the Python programming language. This translation came with several design changes, including the logical distribution of functionality into Python modules, more resilient and flexible processing of all forms of input, better exception handling of installation processes and several other improvements. It did not, however, change the overall monolithic design of the installer. This work explores and implements common patterns and utilities, not only in the ESGF Node installation, but in installation and system configuration in general. Using the Python programming language, it follows the object-oriented design principles of modularity, inheritance, abstraction and polymorphism to provide a powerful set of tools and features. Features include on-demand status checking for installation, initialization and run status of components, as well as the ability to install, initialize, and control specified sets of components. The addition of components is designed to be as simple and flexible as possible for developers as the ESGF Node itself evolves. The overall effect is a better experience for users and developers alike.

Towards a blockchain enabled ESG-F

Mark Greenslade (Institut Pierre Simon Laplace, Paris, France), momipsl@ipsl.fr

The CMIP6 dataset archive will be vast, distributed and will require secure, simple, transparent access. Re-architecting the current infrastructure & concomitant applications is a non-trivial technological task. Surveying the current technological landscape, one is struck by the extent to which distributed ledger technologies, commonly referred to as blockchain, provides a foundational basis towards the re-architecture of such an infrastructure.

Blockchain's are secure by design and exemplify an economically strong distributed computing system with high fault tolerance. They are suitable for the logging of events, storing records, managing identities, processing transactions, documenting provenance, voting (i.e. governance) and orchestrating complex workflows. These capabilities can be chained together to deliver complex sets of use cases such as:

- The publication lifecycle of scientific datasets;
- The automated tracking of dataset downloads;
- The registration of both institutional and corporate users;
- The registration of canonical entities such as vocabularies;

This talk will present a high-level overview of the technology and proceed to illustrate how certain aspects of the current ESG-F stack may evolve. Furthermore, the presenter will situate this evolution in the context of Web 3.0.

Enhancing the ESGF data node to load balance a distributed cluster of THREDDS instances

Ezequiel Cimadevilla Álvarez ezequiel.cimadevilla@unican.es
Pablo Celaya Crespo pablo.celaya@alumnos.unican.es
Antonio S. Cofiño antonio.cofino@unican.es
Santander Meteorology Group, Dep. Of Applied Mathematics and Computational Sciences, University of Cantabria, Spain

The THREDDS data server (TDS) is designed to work as a standalone web application serving data from a single server instance. This design makes TDS performance to scale badly and management difficulties arise when a huge catalog tree has to be maintained or when the TDS has to deal with overloads, causing a degraded or faulty service. Currently, the ESGF-node includes a gateway to the services running in the ESGF-node. One of these services is the TDS web application which runs in the ESGF-node sharing existing resources on the host. The ESGF-node design only considers one TDS instance running besides to the rest of ESGF-node services. Also, this TDS instance deploys the complete catalog hierarchy automatically generated by the esg-publisher, which can become difficult to maintain and to scale if lots of datasets and collections are generated. In this contribution, we show a way of deploying a load balanced and automatic provisioned cluster of TDS instances. The definition of the desired infrastructure is declared in a YAML file for Ansible (Infrastructure as Code, IaC), which uses roles and playbooks, that will automatically deploy the cluster of TDS instances and catalogs. This definition of the deployment infrastructure follows the TDS Deployment Model, which is composed by Collections, Replicas and Instances deployed in Hosts conforming Clusters. A Collection is a hierarchy of THREDDS file catalogs that can be deployed to a regular TDS instance on its own. TDS instances are Apache Tomcat server instances, accessed from the outside through a gateway (i.e. reverse proxy), running the TDS web application in a load balanced way. We refer to every publication of the collection in the TDS instances as a replica. Following this TDS Deployment Model, a sysadmin can define both instances and hosts where each replica will be deployed conforming a cluster. The TDS Deployment Model and its

implementation have been tested with the current deployment of a ESGF-node in order to extend the data node to act as full functional gateway of a distributed cluster of TDS instances. We conclude that it is feasible to add automatically deployed load balancing support, catalog partitioning and integration with the current publication workflow to the architecture of the ESGF data node.

Nonactive ESGF Catalog Migration

Yingshuo Shen (NASA NCCS), yingshuo.shen@nasa.gov
Luca Cinquini (NASA JPL), Luca.Cinquini@jpl.nasa.gov
Laura Carriere (NASA NCCS), Laura.Carriere@nasa.gov
Dan Duffy (NASA NCCS), daniel.q.duffy@nasa.gov

ESGF Publication involves adding metadata into the database, generating THREDDS catalogs from the database, and publishing the catalogs to the index node for searching. Catalogs of an ongoing project are often updated thus they are active catalogs. At NCCS, we have two active projects CMIP6 and CREATE-IP. Catalogs of completed projects such as CMIP5 and NEX are nonactive but they can't be removed from our ESGF data node unless they are unpublished. Nonactive catalogs are still written into the master catalog each time when there is a publication event. We propose to migrate these nonactive catalogs to a separate THREDDS server. Doing so, we can avoid unnecessarily writing these catalogs into the master catalog during each publication. Further, we can take this opportunity to fix all the broken time aggregation links that resulted from the use of the same time range (e.g. 1-365 for daily data) in each yearly file rather than the CMIP compliant use of a time range based on the model simulation start time. This problem might be unique to GISS but we can have it fixed without reprocessing or republishing all the CMIP5 data while we are fully busy with CMIP6. In addition, we can add more services such as WMS (to support GIS users) and netCDF Subset Services. LAS link from the index node can also be deleted since we have the service disabled due to security concerns. We propose to use ESGF Web Service (harvest/unharvest) to do this migration. We have successfully migrated 112 catalogs out of total 4680 CMIP5 catalogs. All the dataset IDs and versions will remain unchanged. It is totally seamless to the ESGF users.