THE STATE OF THE EARTH SYSTEM GRID FEDERATION

LUCA CINQUINI

NASA JET PROPULSION LABORATORY AND CALIFORNIA INSTITUTE OF TECHNOLOGY

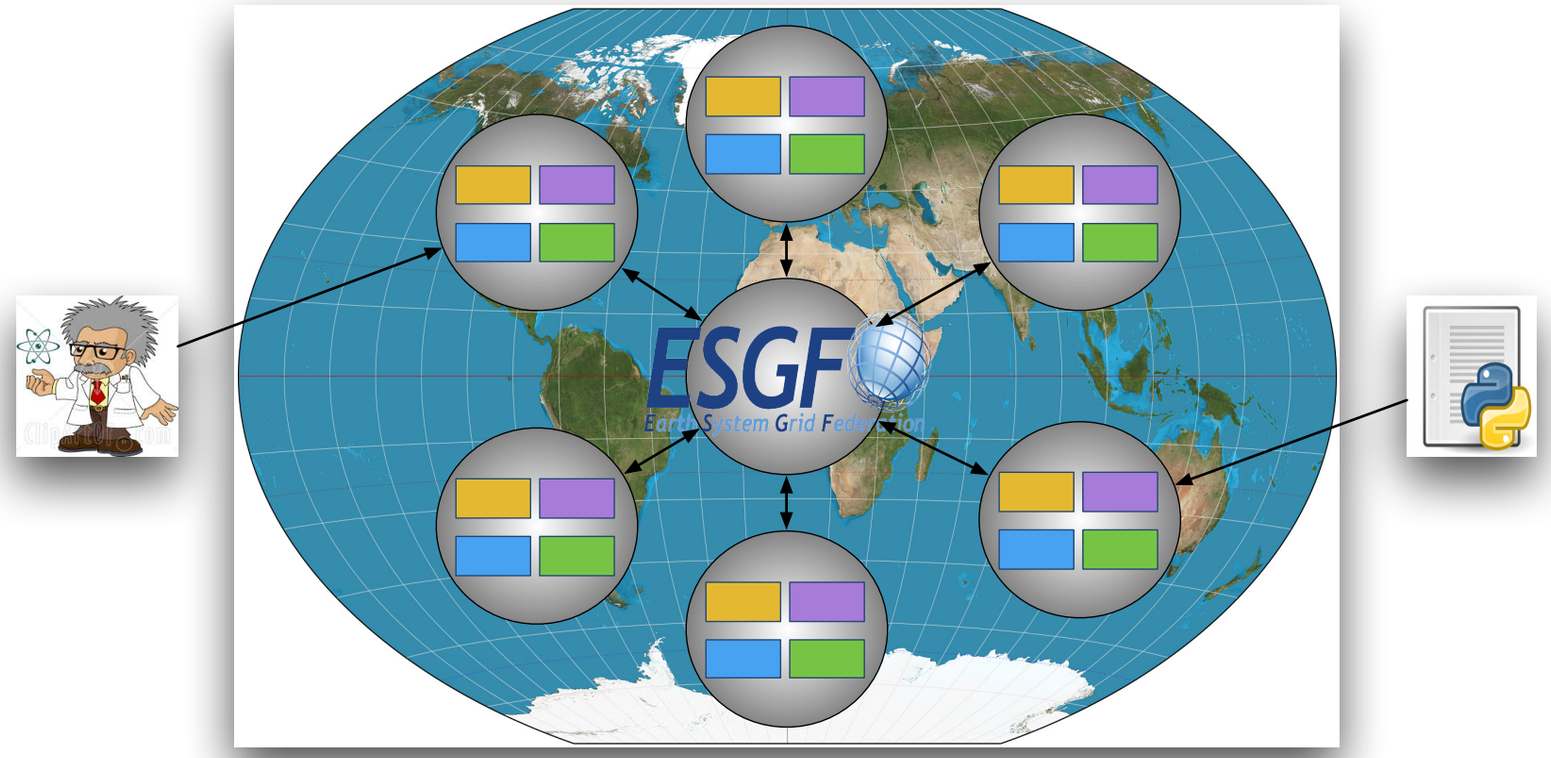
JPL UNLIMITED RELEASE SYSTEM CLEARANCE NUMBER: #18-6942

Introduction

- Recent reports (e.g. 4th National Climate Assessment report) indicate that climate change is accelerating beyond earlier predictions:
  - Enormous economic consequences to the U.S. and world economy
  - Severe weather events (floods, wildfires, tornadoes…)
  - Widespread health effects
  - We are the last generation that can avoid a massive extinction of species
- As the world leading data infrastructure in support off climate change research, ESGF plays an important role in predicting and mitigating its effects on the whole Earth ecosystem

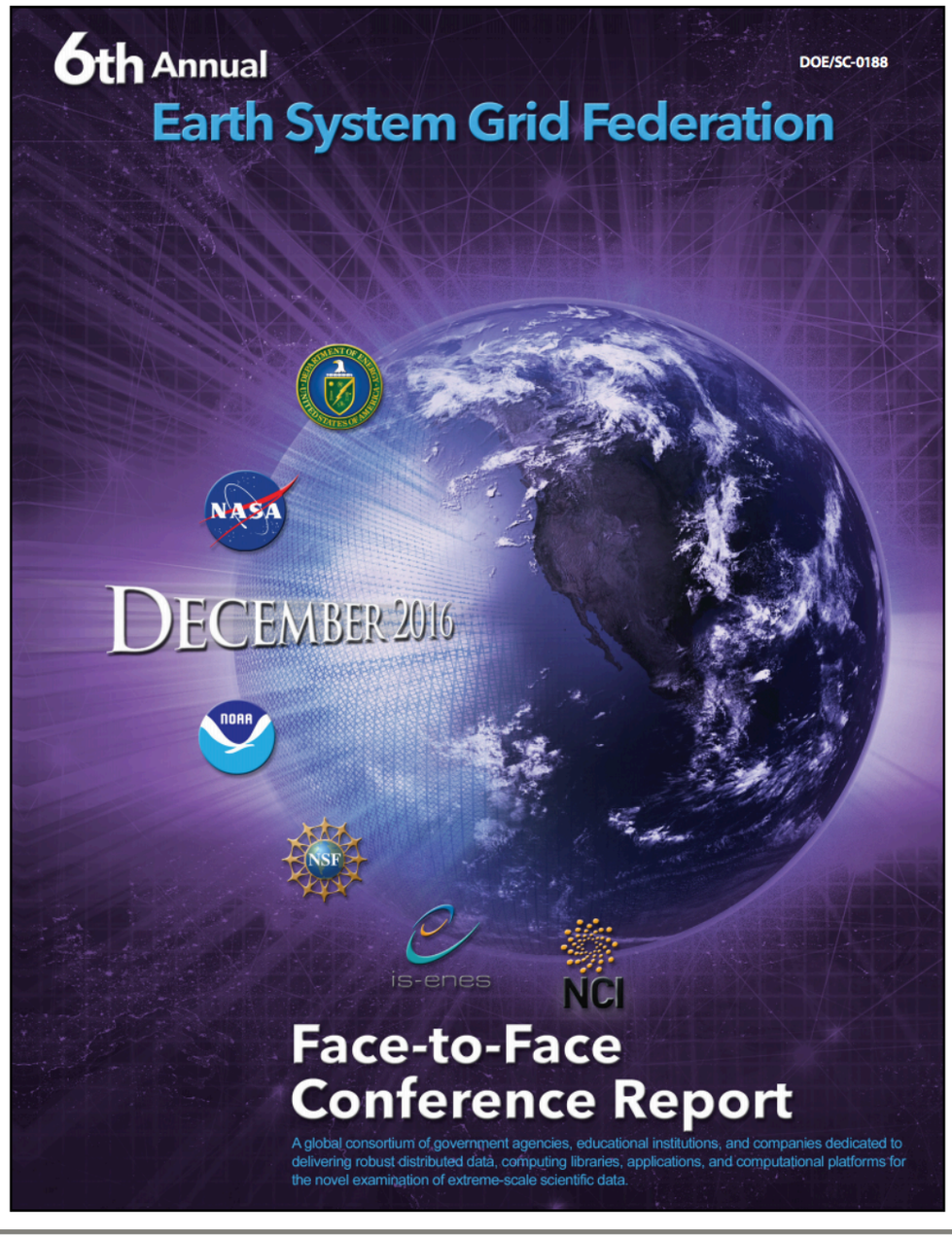
2018 State of ESGF

- ESGF has made constant, solid progress in 2018: improving the reliability of the technical infrastructure, developing new functionality, expanding data holdings and user base
- Most recent stats: 31 nodes, 793,026 datasets, 10,054,190 files, 133 CoG projects, 19,978 users



Review of Action Items from 2017 F2F

- Roadmap established by ESGF-SC and ESGF-XC in 2017 for 2018 (https://esgf.llnl.gov/esgf-media/pdf/2017-ESGF_F2F_Conference_Report.pdf):
  - Short Term Plans for "CMIP6 Preparedness" (0-2 years):
    - Replication
    - Documentation and training for data publishers
    - Software and operations security
    - PID Service
    - Basic data reduction and analysis operations
    - User authentication and authorization
  - Longer Term Plans for ESGF longevity (2-5 years):
    - Server-side computation
    - Installation
    - Cloud computing
    - Programmatic access to data



2018 ACCOMPLISHMENTS

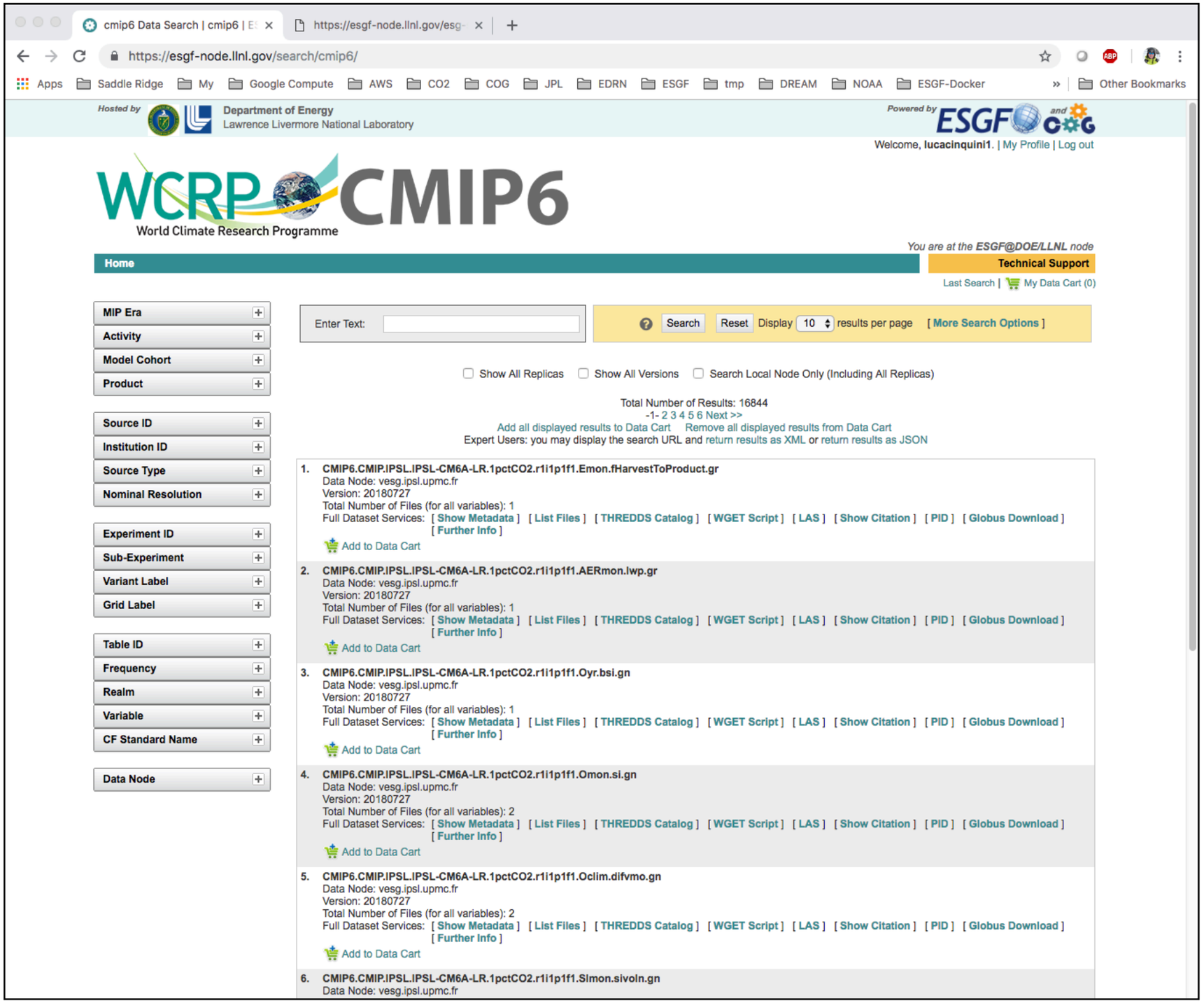
Preparations for CMIP6

- CDNOT coordinated installation and testing of ESGF infrastructure across Nodes (S. Denvil, R. Petrie)
  - 5 “data challenges” held in 2018 to stress-test the system with increasingly larger amounts of CMIP6 test data - very successful
  - Guides for node administrators and data managers
- Replication Working Group is working at managing and improving the replication of core CMIP6 data across “Tier-1” Nodes (S. Kindermann, E. Dart)
  - Replica data are been published at LLNL and DKRZ with GridFTP endpoints
- WIP (WGCM Infrastructure Panel) is overseeing ESGF preparations and providing connections with the CMIP modeling groups (K. Taylor, Balaji)

WCRP CMIP6
World Climate Research Programme

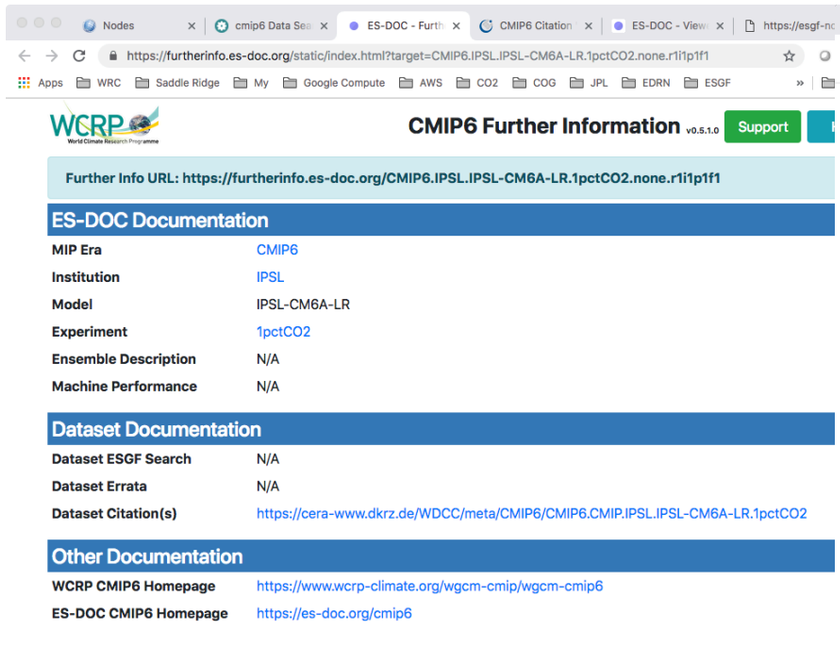
Current CMIP6 data holdings

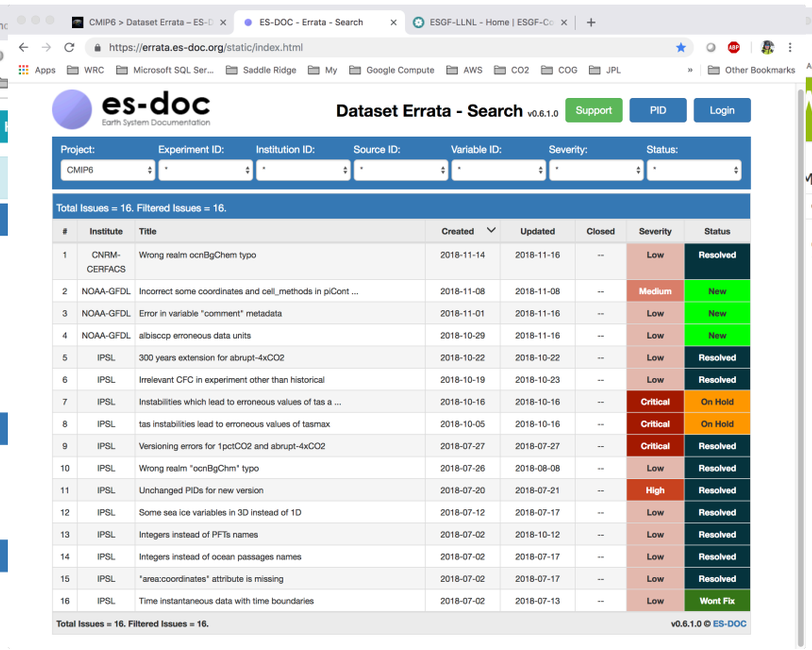
- ESGF opened for CMIP6 data in June 2018
- Currently serving CMIP6 data from 4 Data Nodes: CNRM, GFDL, NCCS, IPSL
- Data replicated at LLNL, DKRZ
- Data holdings:
  - 6 CMIP6 models
  - ~16,844 datasets
  - ~45,930 files



New ESGF Services for CMIP6

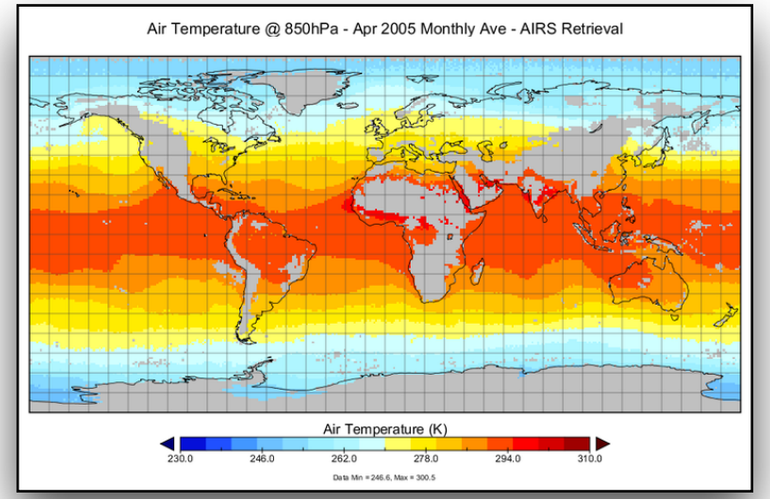
- New ESGF services provide enhanced functionality in support of CMIP6:
  - PID ("Persistent Identifiers") service: assigns PIDs to datasets and files at time of publication for long-term identification
  - ES-DOC: landing pages for datasets, models, experiments, CMIP6
  - Errata Service: central catalog for datasets that had to be retracted for various reasons
  - DOI Data Citation pages at WDC: provide information on how to cite the data, license, content, and related datasets (forcing).



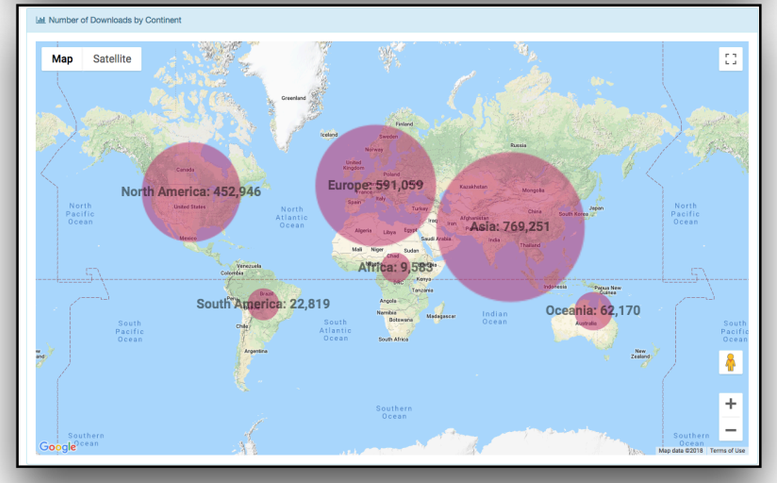


Other ESGF Development

- Obs4MIPs serving more observational data, better (R. Ferraro, P. Gleckler, D. Waliser and P. Durack)
  - New specification for dir structure, filenames, search facets that is aligned with CMIP6 (ODSv2.1)
  - "Dataset Indicators" matrix captures the "maturity level" for model evaluation
- Installation Working Team: classic shell-based installer and new Python-based installer (P. Dwarakanath, S. Ames, W. Hill)
- Idea Working Team: transitioning the current ESGF Security infrastructure (based on OpenID 2.0) to more current industry standards: OAuth2 and OpenID-Connect (P. Kershaw)
- ESGF publisher and ESG prep - several upgrades to support CMIP6 and improve performance (S. Ames, G. Levavasseur)
- Dashboard team - integrating the information provider into the installer and supporting the central metrics aggregator site, also developing custom view for CMIP6 (S. Fiore, A. Nuzzo, M. Mirto)



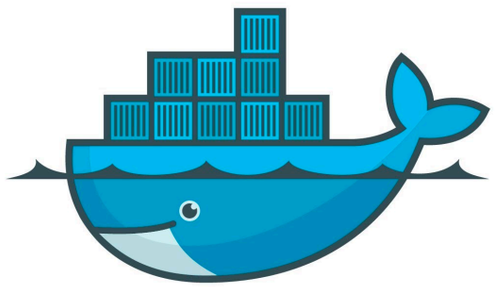
Technical Requirements		Dataset Suitability and Maturity			Comparison Complexity
Meets obs4MIPs data technical requirements	Includes obs4MIPs technical note information	Closeness or robustness of measurement to observed reference quantity	Maturity with respect to climate model evaluation	Provision for robust uncertainty information	Complexity of Model Observation Comparison
Data suitably processed with CMOR and/or consistent with obs4MIPs standards	Complete technical note information provided	Firmly established and/or validated methodology	Multiple peer-reviewed examples of application to CMIP climate model evaluation	Uncertainty information provided per retrieval/grid point	Comparison can be made directly with CMIP model output variable
Largely complete with minor metadata inconsistencies	Technical note information incomplete and/or could be improved	Indirect means of calculation or observations only providing partial constraint (e.g. ocean surface latent heat flux)	One peer-reviewed example of application to CMIP climate and/or examples of other sorts of model evaluation.	General uncertainty information given relative to the methodology and dataset as a whole - backed by actual field/in-situ validation exercises	Comparison requires some simple post processing of CMIP output variable(s) (e.g. vertical integral or ratio of two variables)
Non-compliant. Should be removed from database!	Technical note not provided	Largely model-derived quantity (e.g. LAI, root zone soil moisture, NPP)	As of DATE-TBS, no significant application to climate model evlauation	No uncertainty information provided	Comparison requires complex processing of CMIP output (e.g. "simulator", budget calculation)



ESGF NEW DIRECTIONS

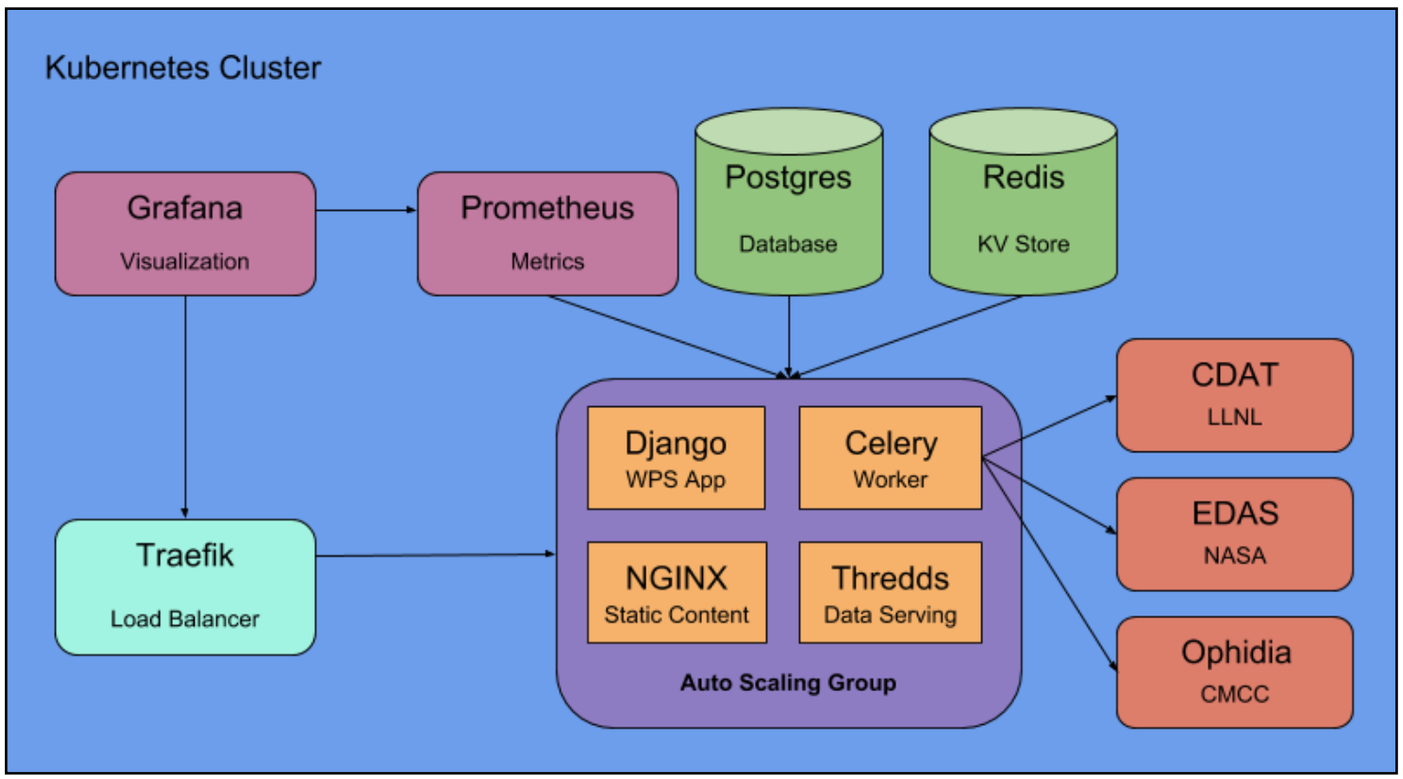
Containerization

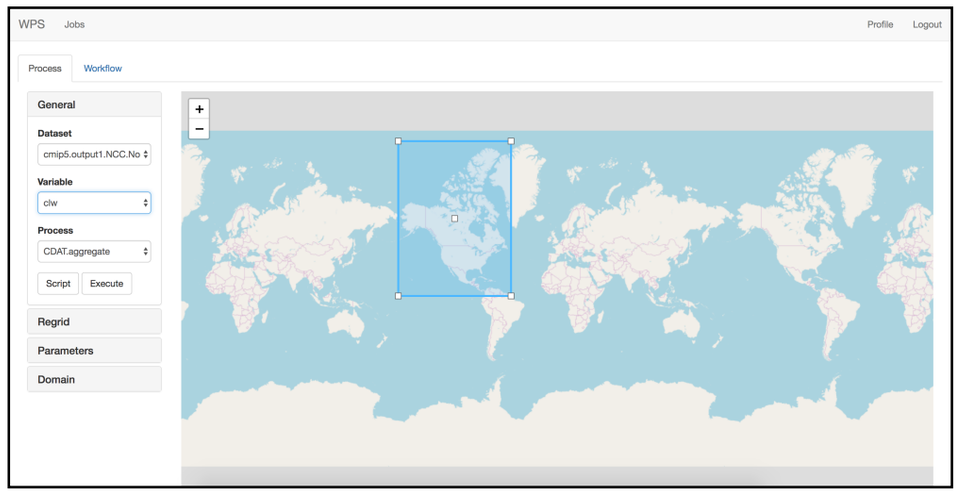
- ESGF/Docker: alternative architecture for ESGF Node where all services are packaged, deployed and managed as Docker containers
- Advantages of container based architecture (“micro-services”):
  - Easier to deploy and test
  - More flexible
  - More scalable
  - Easier to evolve
- ESGF/Docker first release in September 2018
  - Stable but not feature complete (no Globus)
  - Based on Docker, Kubernetes and Helm



Compute Node

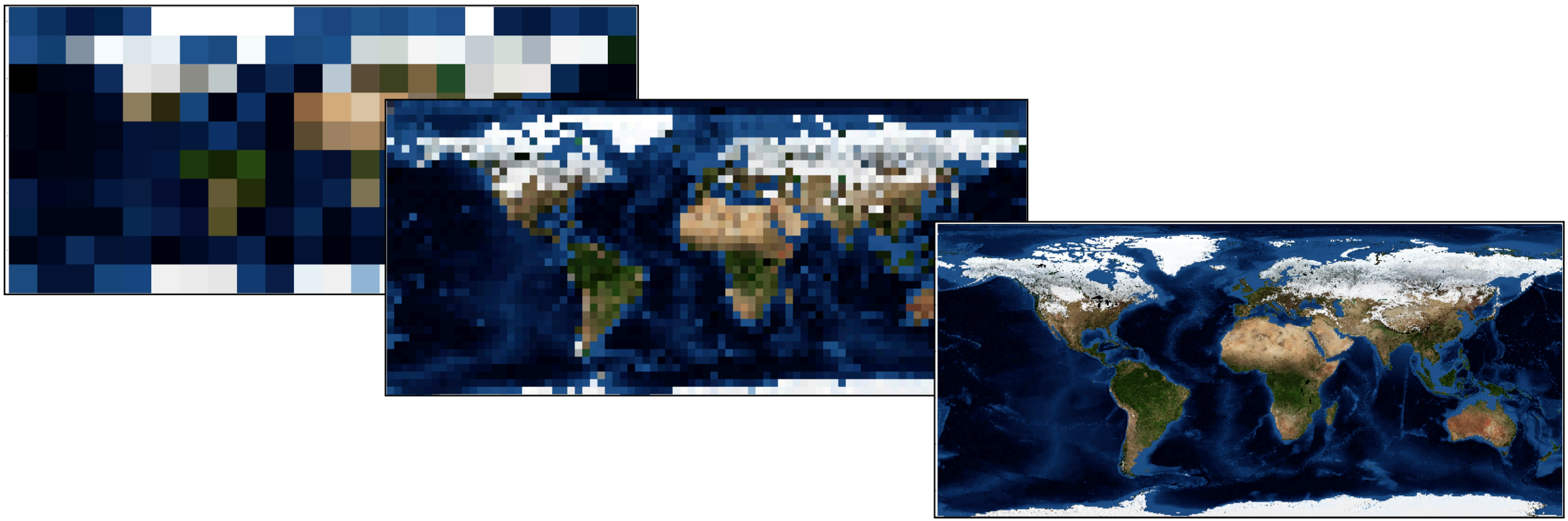
- The ESGF Compute Working Team has made great progress in developing scalable computing capabilities for ESGF (J. Boutte, C. Doutriaux, T. Maxwell, T. Landry)
- Architecture of compute node was designed from the ground up as a system of interacting Docker containers => highly scalable - both horizontally and vertically
- ESGF Compute API implemented by 3 back-ends (sub-set, average, min/max, etc.)
- Status: converted to Kubernetes+Helm, ready to be deployed w/ ESG/Docker stack





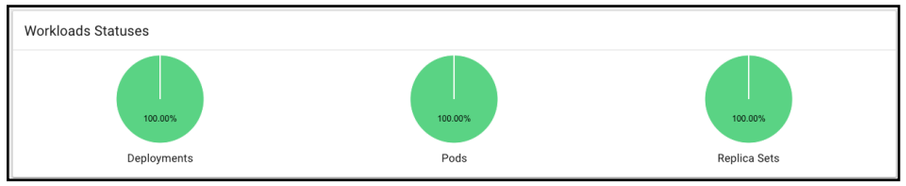
Visus

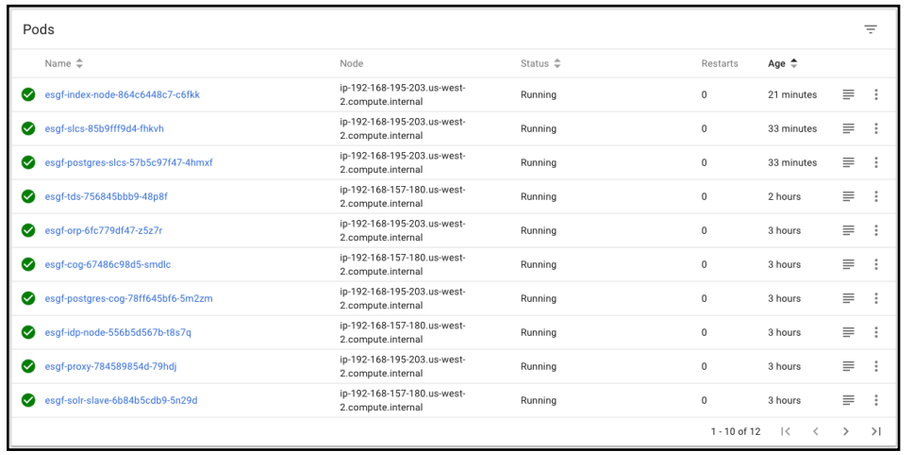
- Visualization engine developed by University of Utah (S. Petruzza)
  - Streaming climate datasets at multiple resolutions
  - Data converted on the fly from NetCDF to IDX
  - Finalizing the integration with ESGF/Docker



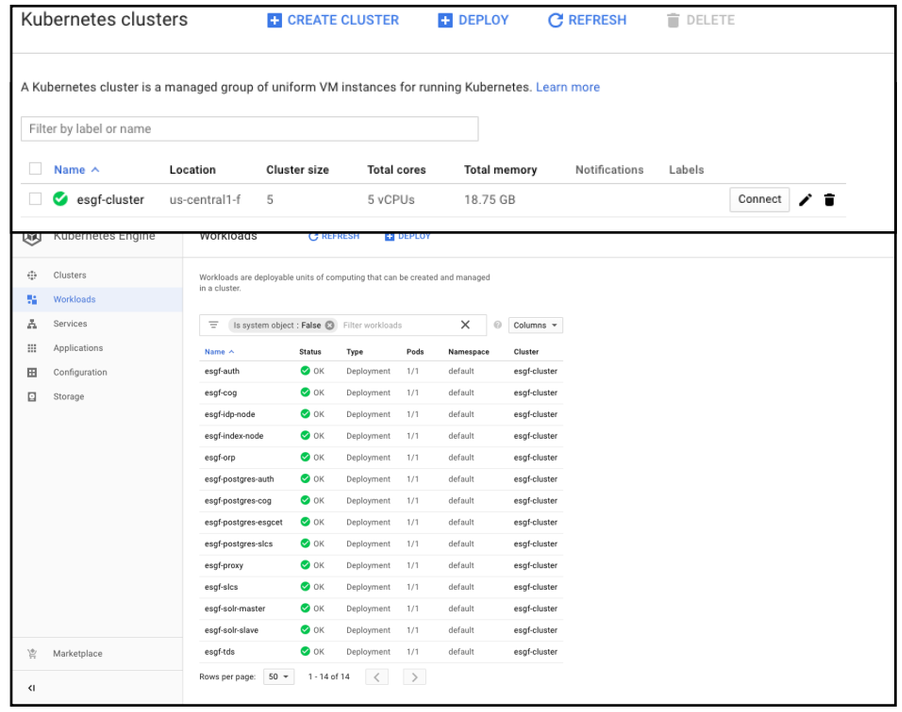
Moving to the Cloud

- ESGF is experimenting with moving its services -all or in part- to the Cloud
- Cloud advantages: practically unlimited scalability, high availability, managed resources
- Cloud challenges: cost model, new architecture designs
  - How to deploy on a cluster of nodes, how to persist data, how to plan for failure
- Several efforts undergoing:
  - ESGF/Docker with Kubernetes is immediately suitable for Cloud deployment
  - GFDL is running a prototype node on Google GKE, published some CMIP6 data, enabling access to Pangeo via openDAP
  - GSFC/JPL planning to deploy a single ESGF/NASA node on AWS GovCloud
  - New Index Node architecture based on Solr Cloud, stable deployment on AWS for several months





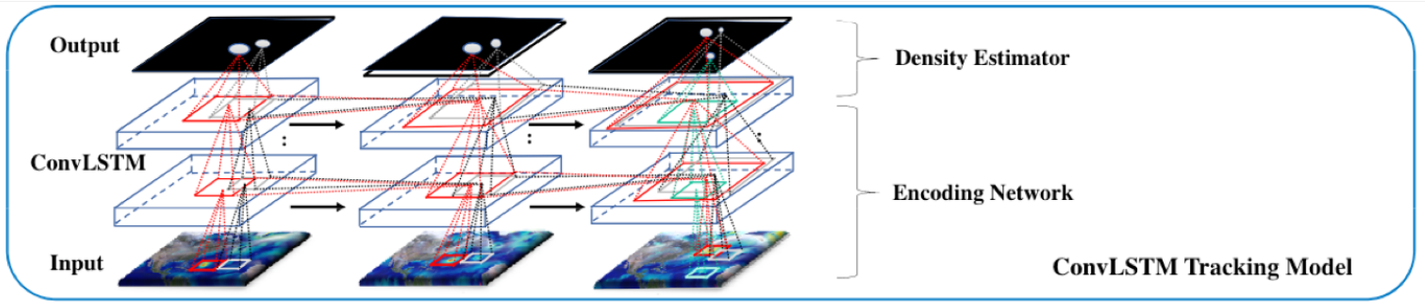
AWS EKS

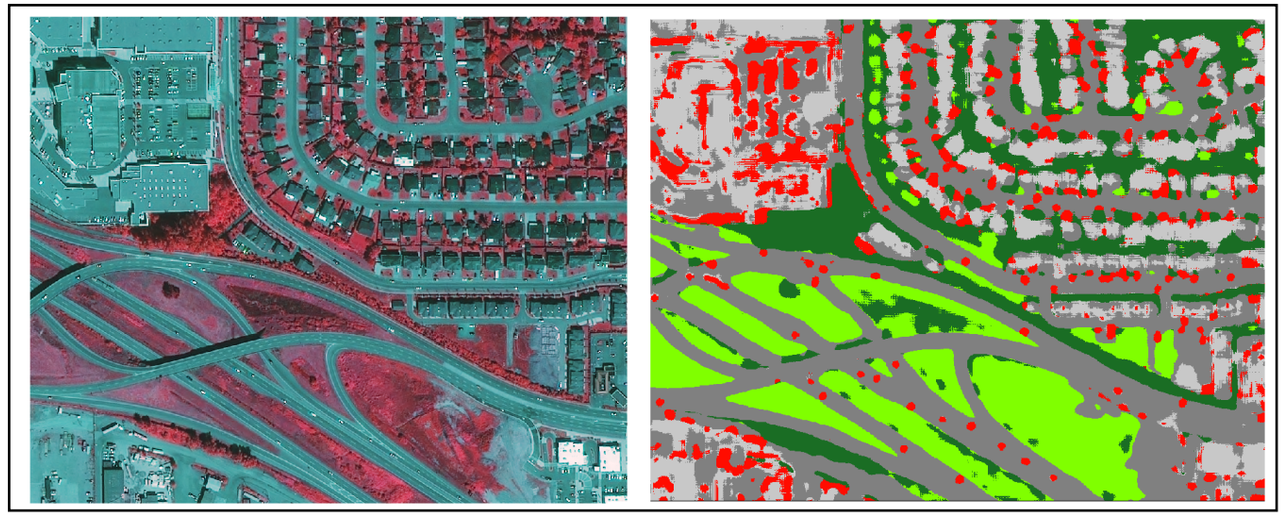


GCP GKE

Machine Learning

- How to mine the vast amounts of data held by ESGF to make reasonable predictions on future global climate and weather events?
- LLNL: “Deep Hurricane Tracker” model analyzes patterns in climate simulation data to predict hurricane tracks (S. Kim)
- CCMC: High Performance Data Analytics and Machine Learning using Ophidia - an infrastructure for executing declarative, parallel, server side analytics workflows (S. Fiore)
- CRIM: working with OGC to advance ML&DL capabilities for high resolution satellite images (T. Landry)

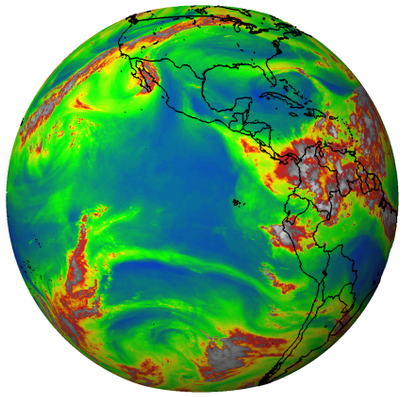


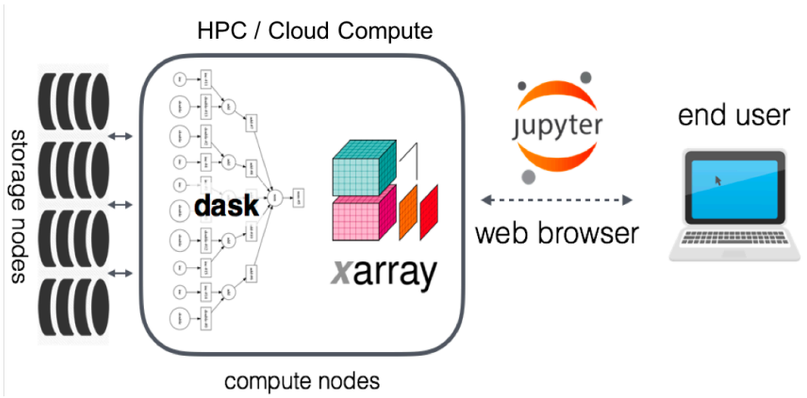


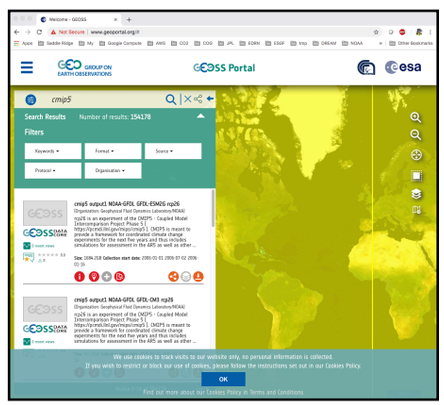
KEY CHALLENGES

Key Challenges for 2019 and Beyond

- Scalability
  - A container based architecture is highly scalable, but the underlying applications must be scalable
  - Publishing services, data catalogs (TDS), and display of search results
- Easy Data access
  - Improve the wget scripts, provide alternative clients
  - Better support for subsetting at the source, over space and time aggregations
- Server side distributed computing
  - Deploy the Compute Node operationally
  - Enable workflows that span multiple steps, at multiple sites
- Interoperability
  - NASA DAACs, ESA, Copernicus
  - Pangeo
  - GEOSS: Global Earth Observations System of Systems







Dear Colleagues:

My apologies for not attending this year's 2018 ESGF Conference. However, I know you are in great hands with the SC and XC committees. I am proud to be a part of this international team and you should be proud of another great year of contributions to the ESGF and climate change community efforts! Deadlines are being met and climate scientists are using ESGF to address one of the most pressing challenges of the day and the future. Without you NONE of this would be possible.

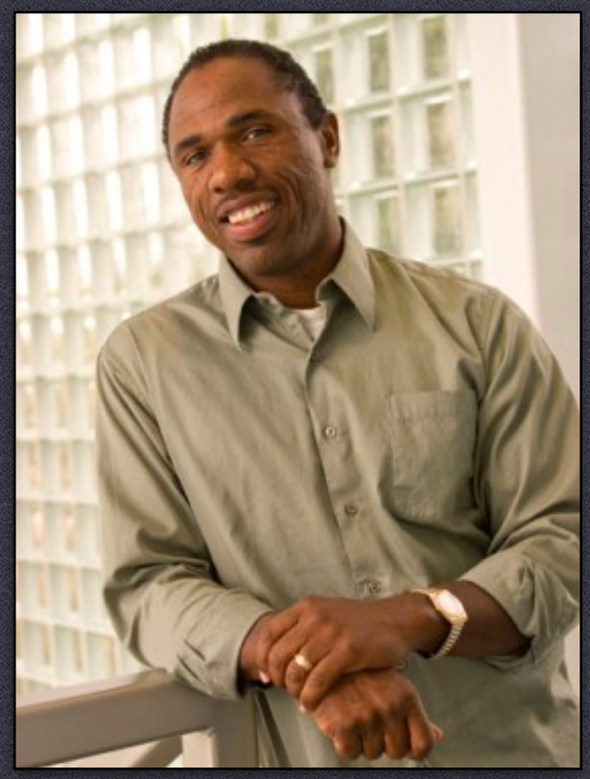
Again, I thank each and everyone of you for your contributions to ESGF and making it what it is today, "The leading climate simulation distributed data warehouse".

I hope to see everyone soon as I recover from my illness.

Best regards,

Dean N. Williams

ESGF Chair



Earth System Grid Federation