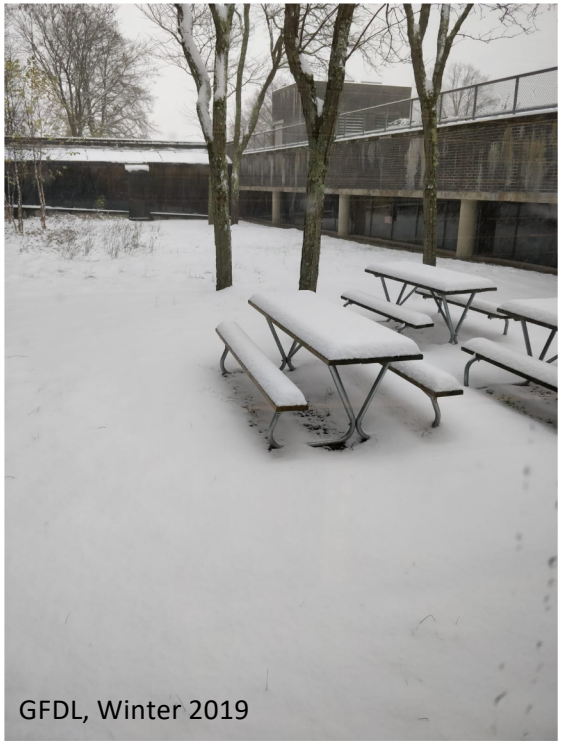
Pangeo and ESGF in the cloud

Aparna Radhakrishnan
On behalf of the

24th Working Group on Coupled Modeling, Dec 9th, 2021

OUTLINE

- Essential ingredient
- What's cooking/baking?
- The finishing touch

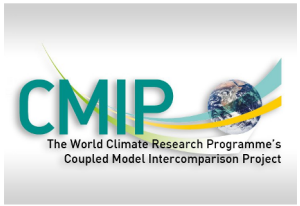


GFDL, Winter 2019

Essential ingredient: Collaboration

Ryan Abernathey, V. Balaji, Julius Busecke, Philip Kershaw, Naomi Naik, Serguei Nikonov, Ana Privette, Kristopher Rand, Ag Stephens, Charles Stern, Hans Vahlenkamp, Mackenzie Blanusa, Anderson Banihirwe, Chris Blanton, Nkeh Perry Boh, Ben Evans, Richard Smith, Rhys Evans, Zac Flamig, Diana Gergel, Thomas Jackson, Rebecca Monge, Natalie O'Leary, Zouberou Sayibou, Martina Stockhause

Pangeo / ESGF Cloud Data Working Group




What's cooking?

Community-driven cloud-based research efforts led by ESGF and Pangeo..

#1 CMIP6 data in the cloud

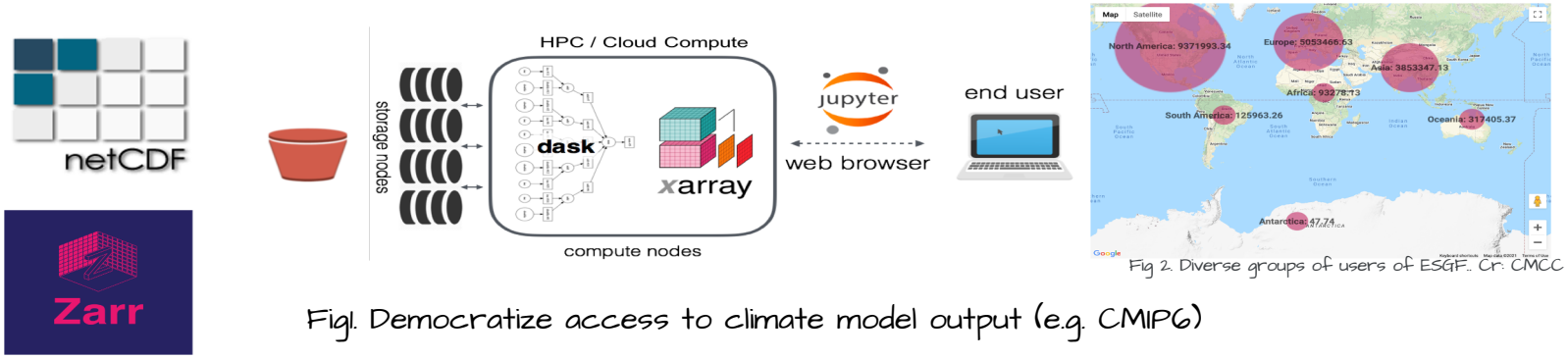


Fig 2. Diverse groups of users of ESGF. Cr: CMCC

Fig1. Democratize access to climate model output (e.g. CMIP6)

- Efforts to host CMIP6 data in the cloud (E.g. Zarr/Pangeo in AWS,GCP, NetCDF/ESGF under ASDI,)
- Increased scope for collaboration and unified APIs
- Make data more accessible
- Reduce need for Dark repos
- Recognize the stars behind the cloud: Original work from modeling centres, additional efforts to make data more usable

#1 CMIP6 data in the cloud

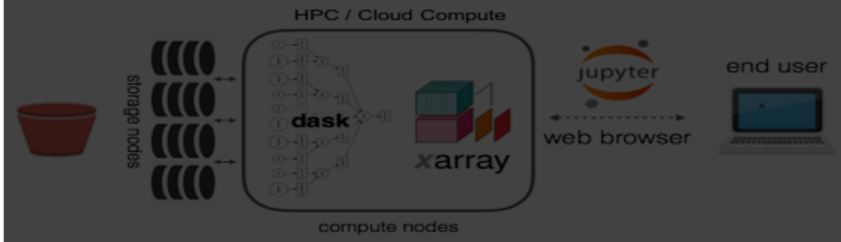


Fig 2. Diverse groups of users of ESGF. Cr: CMCC

Fig1. Democratize access to climate model output (e.g. CMIP6)

- Efforts to host CMIP6 data in the cloud (E.g. Zarr/Pangeo in AWS,GCP, NetCDF/ESGF under ASDI,)
- Reduce need for Dark repos
- Increased scope for collaboration and unified APIs
- Make data more accessible
- Recognize the stars behind the cloud: Original work from modeling centres, additional efforts to make data more usable

#2 Data exploration

Intake-esm API

E.g. s3://esgf-world/CMIP6/AerChemMIP/NOAA-GFDL/GFDL-ESM4/hist-piNTCF/r1i1p1f1/Amon/tas/gr1/v20180701/tas_Amon_GFDL-ESM4_hist-piNTCF_r1i1p1f1_gr1_185001-194912.nc

Intake-esm


```
exp_filter = ['historical']
table_id_filter = 'Amon'
variable_id_filter = "tas"
cat = col.search(experiment_id=exp_filter,
                 table_id=table_id_filter,
                 variable_id=variable_id_filter)
```

catalog with 55 dataset(s) from 1872 asset(s):

Many thanks: CF/CMOR, Directory Reference Syntax (DRS) established by the ESGF community makes cataloguing possible.

Taylor et al,2017, CMIP6-CV

STAC (Coming soon)

Spatiotemporal Asset Catalogs

- Big, open community, shared API, shared interests
- Flexible dev with STAC extensions, heterogeneous data model
- Supports Indexing and searching

Free-text search

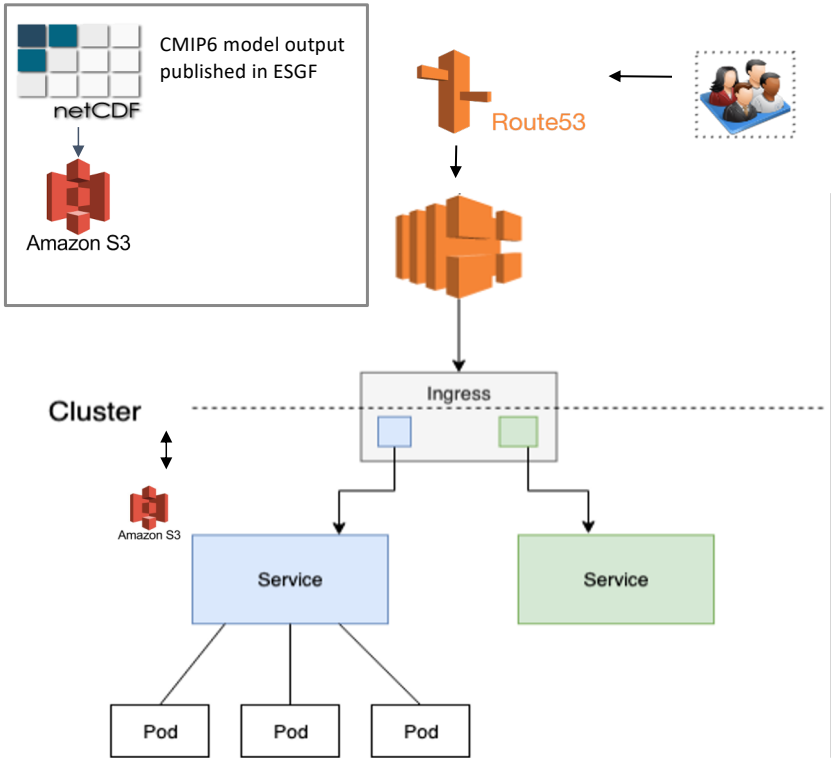
```
# All these queries end up with the same search result:
result = Client.search(q="AerChemMIP")
result = Client.search("AerChemMIP")
result = Client.search("aerchemmip")
result = Client.search("AerChem*") # the star, *, is a wildcard symbol.

# It is also possible to add other arguments to the free text search:
result = Client.search("aerchem*", datetime="2000-11-01T00:00Z/..")
```

Faceted search

```
result = Client.search(
    filter={
        "institution_id": ["CNRM-CERF"]
    }
)
```

Smith et al, AGU 2021

#3 ESGF in the cloud

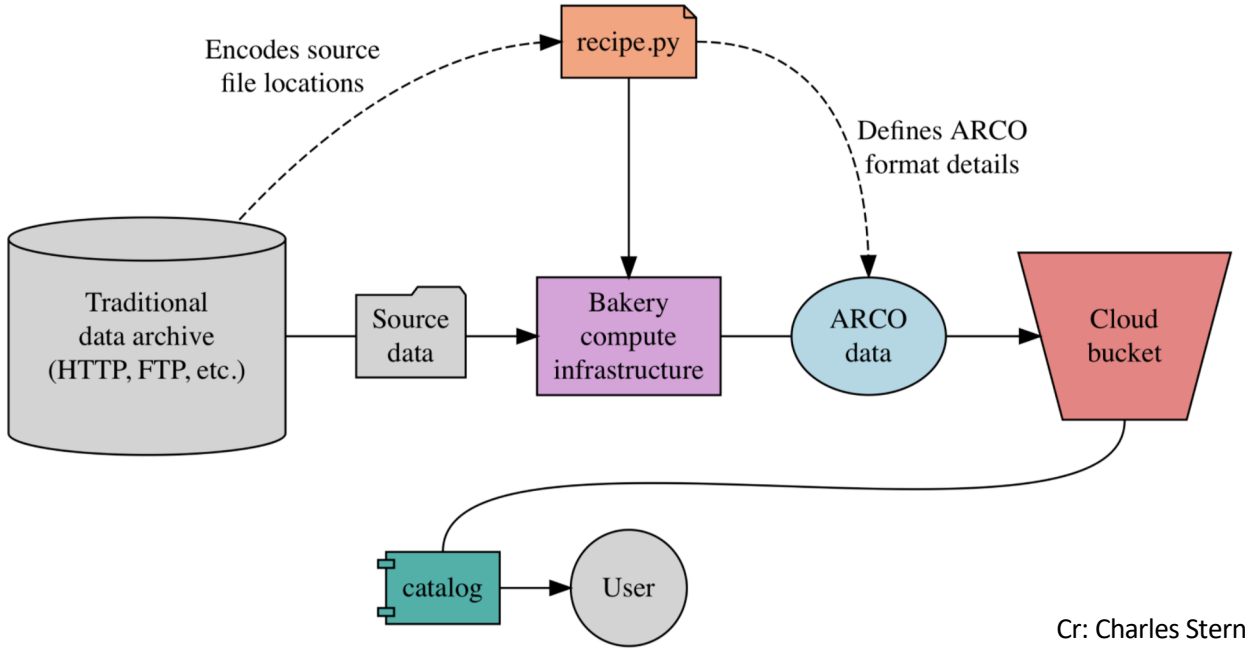


User requests resolved by AWS Route53. A load balancer ingress controller handles requests to the EKS services and PODs. The cluster in this figure may be an EKS cluster with an autoscaling group. The cluster related services are all within a secure virtual private cloud. S3 (NetCDF CMIP6 bucket) storage is mounted as a filesystem via goofys on to the EKS cluster.

- Background: Initial NetCDF data transfer to the cloud used the GFDL Unified Data Archive*
- Deployed the containerized future software architecture of ESGF in Amazon (Cr: ASDI, CEDA)
- ESGF cloud node prototype is up and running
- Data publication/replication to the cloud via ESGF data publisher.
- Published data discoverable via synda (cr: IPSL),THREDDS and direct S3 access; eventually ESGF-search API
- Future: ESGF node federation

*GFDL Unified Data Archive is GFDL's centralized data repository with IPCC chapter fields, GFDL user-requested CMIP model output and other projects such as OBS4MIP

#4 Pangeo-forge

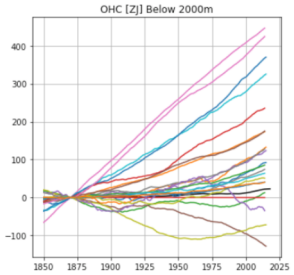


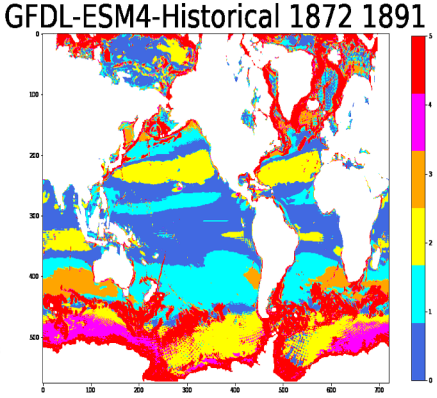
Cr: Charles Stern

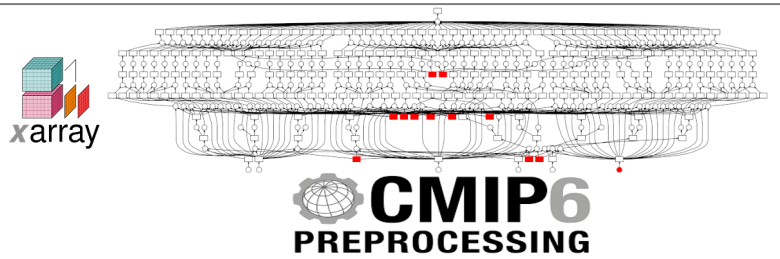
Provenance: pangeo-forge-recipes provides logic for transforming all of these source files into a single consolidated zarr store.
Future: Automated workflow pipeline in the cloud to produce Zarr data from cloud-hosted NetCDF using Pangeo-forge recipes.

#5 Bring efficient analysis to data

Despite the CF/CMOR/DRS standards, there are several non-uniform dimension names, etc in CMIP6 model output making model intercomparison time-consuming.....







- Preprocessing for dataset homogenization
- Based on the xarray datamodel
- Integrates with the existing pangeo stack where possible (xgcm, xesmf, etc)
- Lightweight, dask friendly
- Works on or off the cloud

Cr: Julius Busecke

Researchers including young scientists have enjoyed working in the cloud with amazing results in short time.

The Finishing Touch, TODO

- Automated workflow pipelines for data version consistency checks, cataloguing and usage reporting
- Process for tracking and servicing community requests for CMIP6 data in the cloud
- Documentation on different cloud-data efforts
    - Data discoverability
    - Data citations
    - Errata information
- Awareness of cloud optimized solutions and pathways for researchers- How?
- Sustainability of cloud efforts, beyond CMIP6 - How? Process?
    - Requested 3PB of additional S3 allocation (used approx 40% of the present allocation).
    - Balance availability of NetCDf (cloud-optimized) and Zarr CMIP6 model output in the cloud

REFERENCES

- intake-esm https://intake-esm.readthedocs.io/en/latest/
- CMIP6 Controlled Vocabulary: https://github.com/WCRP-CMIP/CMIP6_CVs
- intake-esm https://intake-esm.readthedocs.io/en/latest/
- Example notebooks: https://github.com/pangeo-data/pangeo-example-notebooks
- https://github.com/aradhakrishnanGFDL/gfdl-aws-analysis/blob/master/examples/intake-esm-s3-nc-simple-access.ipynb
- Pangeo documentation: https://pangeo-data.github.io/pangeo-cmip6-cloud/
- Pangeo-forge: https://github.com/cisaacstern/pangeo-forge-slides, documentation: https://pangeo-forge.readthedocs.io/
- CMIP6 registry in AWS
- CMIP6 preprocessing

THANK YOU

Contact: aparna.radhakrishnan@princeton.edu