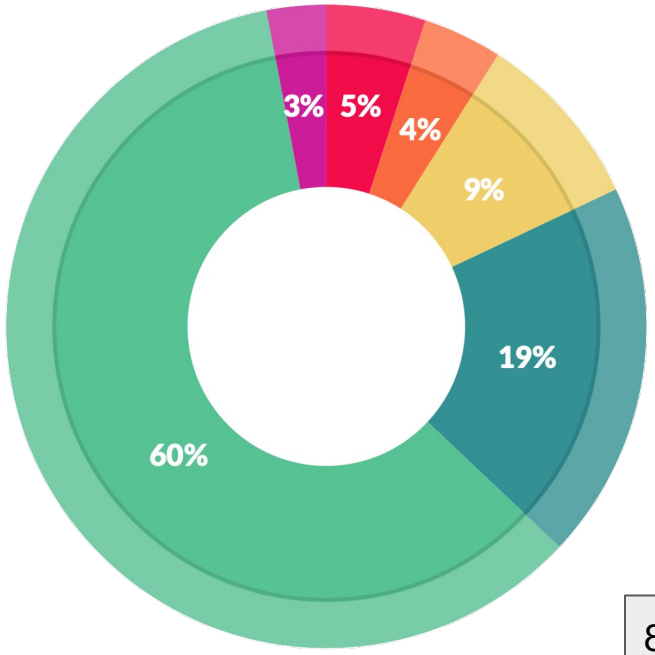
Introduction: Community-developed data exploration for earth system and beyond

Aparna Radhakrishnan

August 2024

The need for data exploration



Ref: How do data scientists spend their time?
Crowdflower Data Science Report (2016)

What data scientists spend the most time doing

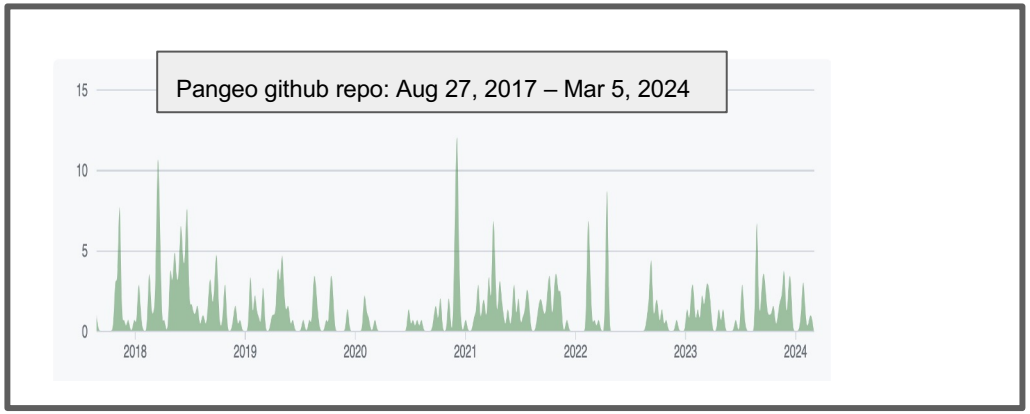
- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

80% of their time on preparing and managing data for analysis.

Acknowledging community collaborations

PANGEO

A community platform for Big Data geoscience

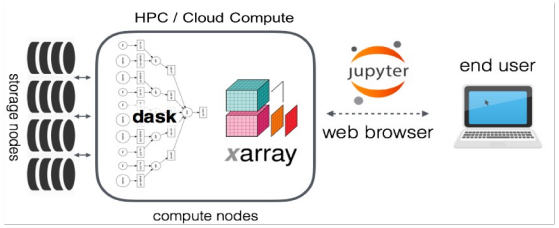


CMIP6 Hackathon (2019)

CMIP

select-OBS

How can search for data?
How can I load the data in my multi-model analysis?



[2020-2021]

Informal Pangeo/ESGF Cloud Data working group

Columbia University

Centre for Environmental Data Analysis

SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

SAIC UCAR

PRIVATE SECTOR

NCAR

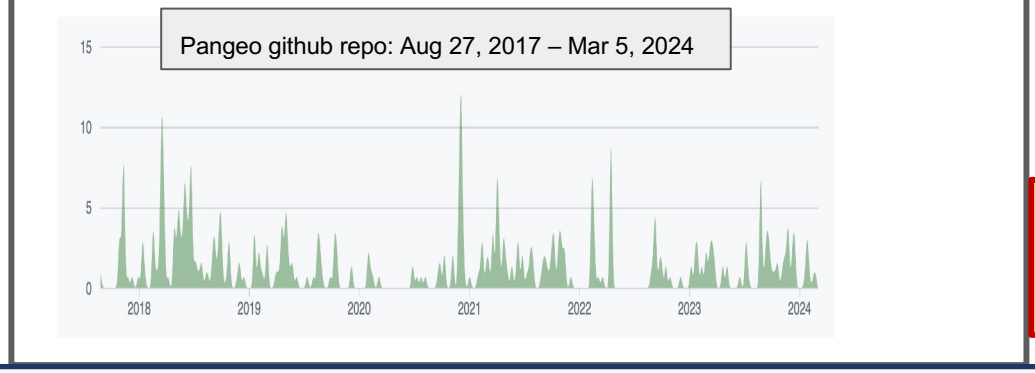
DKRZ
DEUTSCHES KLIMARECHENZENTRUM

NOAA Model Diagnostics Task Force

Acknowledging community collaborations

PANGEO

A community platform for Big Data geoscience



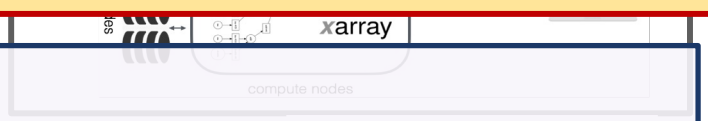
CMIP6 Hackathon (2019)

CMIP

select-OBS

How can search for data?
How can I load the data in my multi-model analysis?

Intake-esm was introduced



[2020-2021]

Informal Pangeo/ESGF Cloud Data working group

Centre for Environmental Data Analysis

SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

SAIC UCAR PRIVATE SECTOR

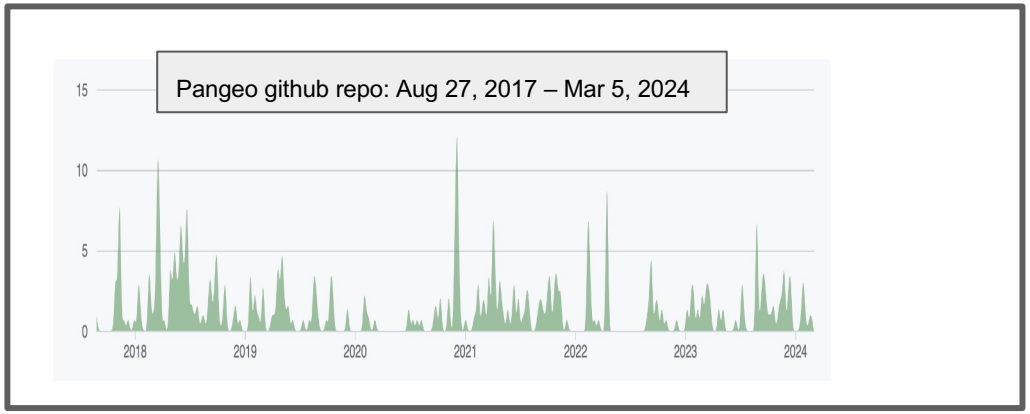
NCAR DKRZ

NOAA Model Diagnostics Task Force

Acknowledging community collaborations

PANGEO

A community platform for Big Data geoscience

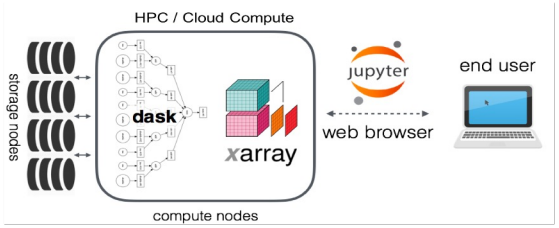


CMIP6 Hackathon (2019)

CMIP

select-OBS

How can search for data?
How can I load the data in my multi-model analysis?



[2020-2021]

Informal Pangeo/ESGF Cloud Data working group

Columbia University

Centre for Environmental Data Analysis

SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

SAIC UCAR

PRIVATE SECTOR

NCAR

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

NOAA Model Diagnostics Task Force

Data catalogs? intake-esm?

Catalog Specification

- what we expect to find inside and how to open the “datasets”/objects?
- Provides metadata about the catalog
- Identifies how multiple files can be aggregated into a single “dataset”
- Extensible metadata
- Single JSON File

Catalog

- Tells us more about the data collection
  - Path to the files (objects), and associated metadata.
- User-defined granularity
- CSV File

Intake-esm: Opens possibilities to QUERY and ANALYZE

- Provides a pythonic way to “query” for information about data collections.
- Loads the results in an xarray dataset object

Data catalogs? intake-esm?


```
{
  "esmcat_version": "0.1.0",
  "id": "sample",
  "description": "This is a very basic sample ESM catalog.",
  "catalog_file": "sample_catalog.csv",
  "attributes": [
   {
     colum_name": "experiment_id",
     "vocabulary": "https://raw.githubusercontent.com/WCRP-CMIP/CMIP6_CVs/master/CMIP6_CV.json"
   } ,
   {
     "column_name": "variable_id",
     "vocabulary": ""
   },
   {
     "column_name": "path",
     "vocabulary": ""
   }
  ],
"assets": {
   "column_name": "path",
   "format": "netcdf"
 }
}
```


experiment_id,variable_id, path

cmdev-test,ts,tsfilename.1900.nc
cmdev-test,ts,tsfilename.1904.nc
cmdev-test,ts,tsfilename.1905.nc
cmdev-test,thetao,thetaofilename.1900.nc
cmdev-test,thetao,thetaoilename.1904.nc
cmdev-test,thetao,thetaofilename.1905.nc

```
col = intake.open_esm_datastore(path_to_catalog_specification
```

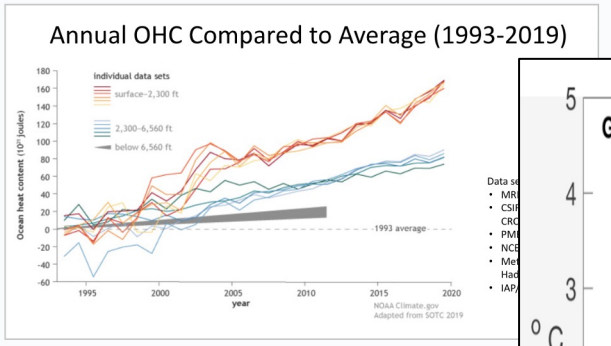
catalog with 2 dataset(s) from 6 asset(s):

```
exp_filter = [omdev-test]
variable_id_filter = "ts"
cat = col.search(experiment_id=exp_filter,
                 variable_id=variable_id_filter)
```

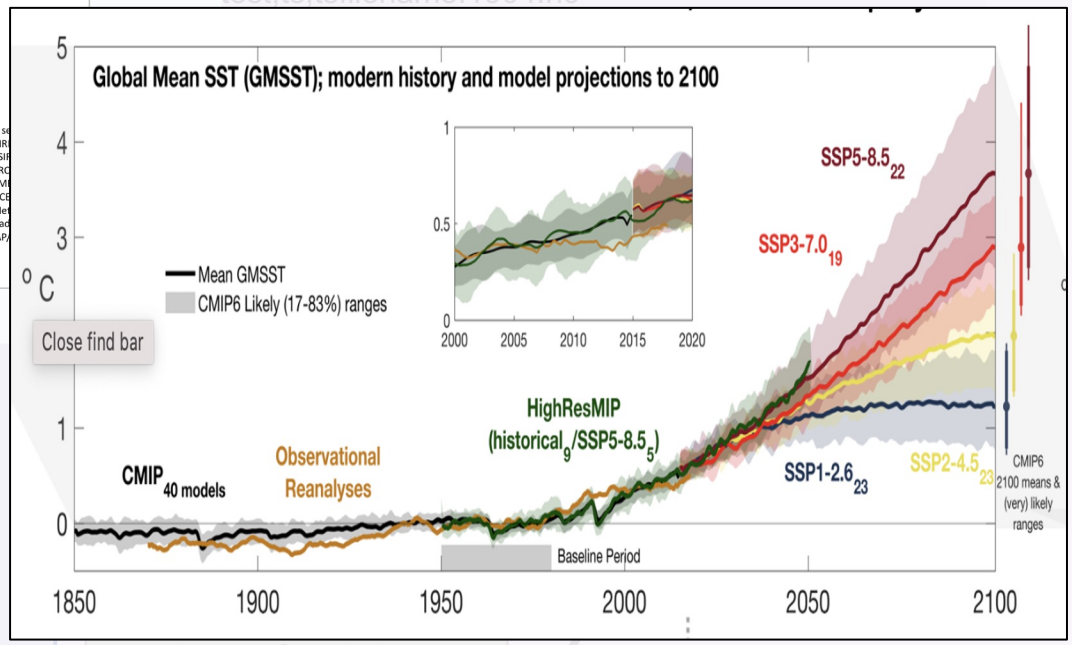
catalog with 1 dataset(s) from 3 asset(s):

Few lines of code later.. →

Data catalogs? intake-esm?



Cr. CIMES internship, Mackenzie Blanusa



Cr. IPCC and xMIP from Julius Busecke

A notebook example that “uses” the catalog

Example notebook [1] (Github reference)

Example notebook [2]

Example catalog specification (json), catalog (csv)

Please contribute notebook examples that use GFDL generated catalogs and intake-esm [Issue page]
(Homework) Binder link Once the link loads, go to notebooks and run the demo-search-explore cell by cell.
GitHub reference)

HOW TO BUILD A CATALOG? Please follow the GFDL Catalog builder video tutorial and docs.

More examples from the community:
MDTF example notebook (GitHub reference)
AWS S3 ASDI analysis collection
DKRZ documentation and references
Pangeo gallery
Student notebook