Pangeo

An open, inclusive, reproducible and scalable Geoscience data analytic community

Anne Fouilloux, Simula Research Laboratory, Norway
Alexander Kmoch, University of Tartu, Estonia

Outline

- What is Pangeo?
- The Pangeo Project current activities
- The software ecosystem with examples from the community
- How to engage with the Pangeo community

What is Pangeo?

A global community initiative for Big Geoscience Data that promotes open, reproducible, and scalable science

- Open Community
- Open Platform with deployments that can be customized for every needs and for everyone

Code of conduct

- https://github.com/pangeo-data/governance/blob/master/conduct/code_of_conduct.md

Governance

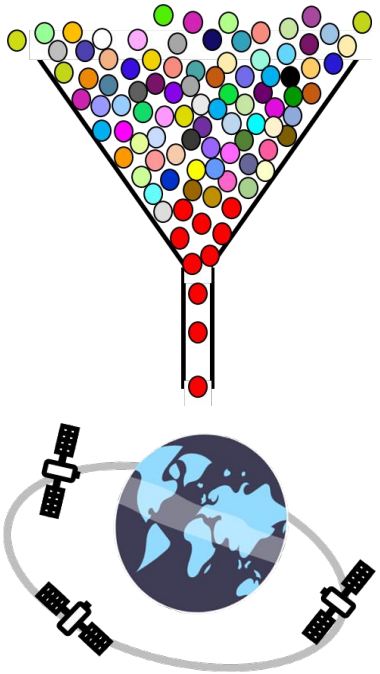
- https://github.com/pangeo-data/governance/blob/master/governance.md

Roadmap

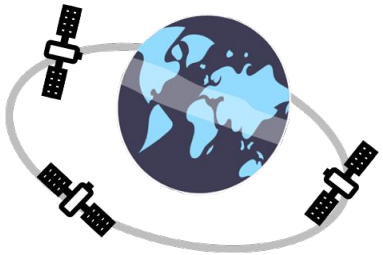
- https://pangeo.io/roadmap.html

Open Community

Pangeo vision: lower barriers to entry to stimulate innovation



Facilitate contributions to increase **diversity**



DEI (Diversity, Equity and Inclusion) is essential

A community of developers, scientists and users

America
Time zone +11-> -2

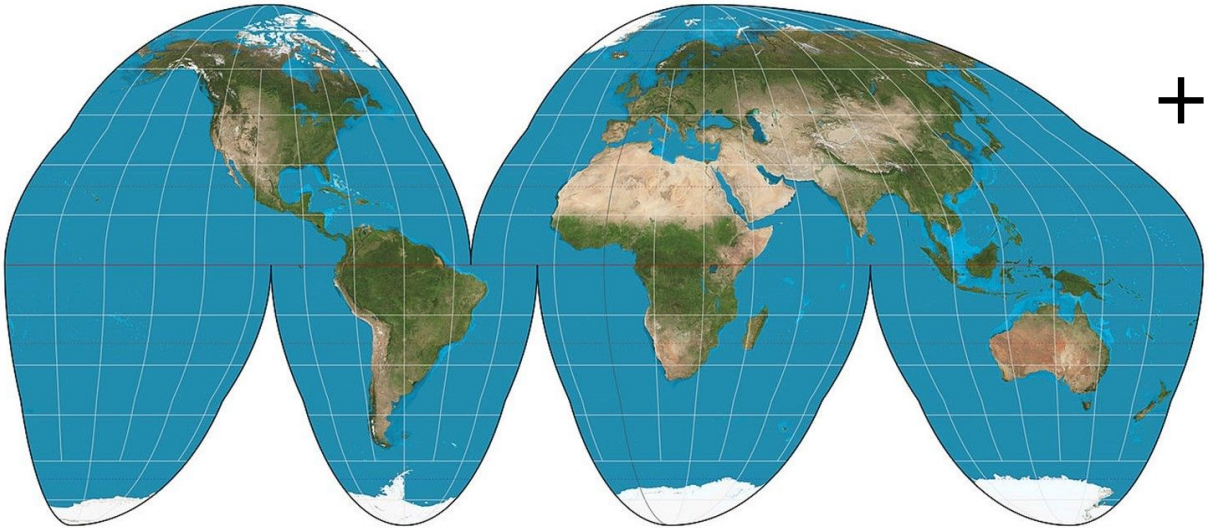
Every Wednesday, alternating between 12pm ET and 4pm ET

Europe, Africa, West Asia
Time zone -1-> +5

Every Tuesday at 9.30 a.m. CET/CEST here

Australia,East Asia
Time zone -1-> +5

3rd Friday of every month at 1pm Australian Eastern Time



+ Working Group Meetings

- *Machine Learning Working Group*
- *Cloud Operations Working Group*
- *Project Pythia (formerly the Education Working Group)*
- *Pangeo Forge (Cloud Data Platform)*
- *Open Science Meeting - discussions to coordinate open science activities*

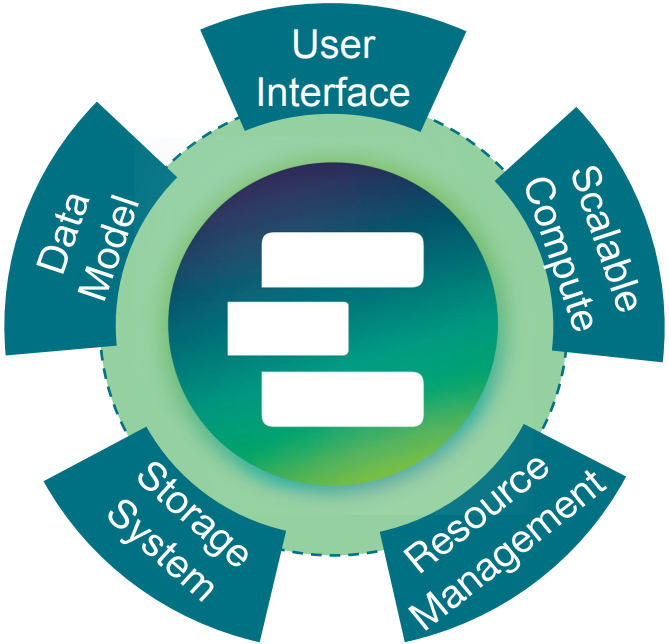
More info & links here - https://pangeo.io/meeting-notes.html

A Culture of Collaboration

Based on GitHub and papers affiliations

Open Platform

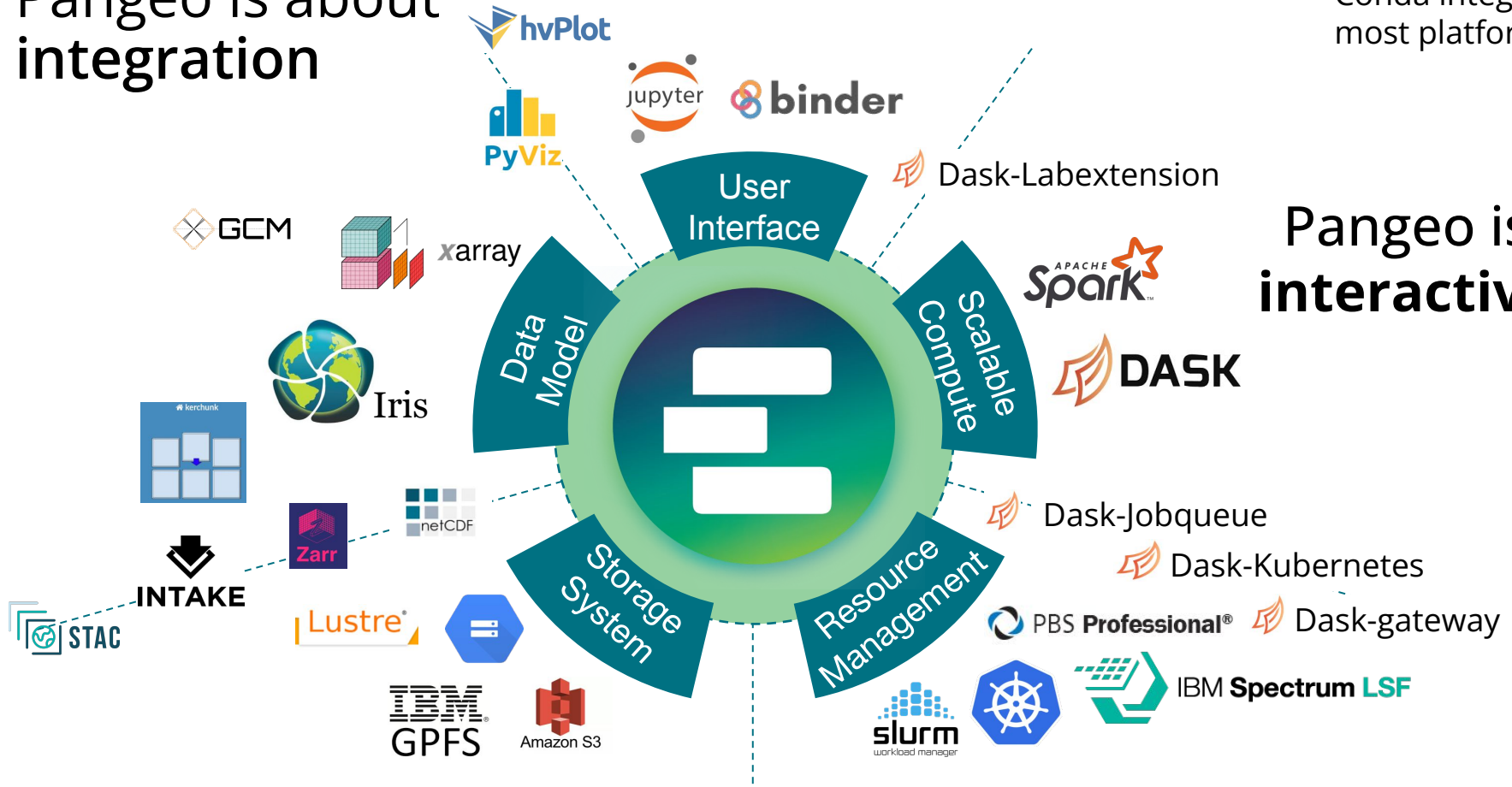
Pangeo **platform** *is scalable*



from laptop to cloud or HPC

Pangeo is about **integration**

Conda integrates most platforms



Pangeo is **interactive**

Pangeo makes it possible to explore geoscience data using HPC or cloud in an interactive manner

Deploy anywhere and for everyone

Pangeo deployments

- Laptop
- Anywhere for anyone!
- Cloud
  - Public Clouds (EOSC)
  - Commercial clouds (AWS, Microsoft, Google)
- EO Platforms
  - Formerly DIAS (CREODIAS ...)
- HPC
  - CNES, IFREMER, PRACE, LUMI, Fugaku, ..

Pangeo eosc Solution

Custom environments

Online content

RoHub

EOSC Community Hub

Data science interface

Cloud infrastructure

Data

Compute

EOSC Authentication for students, researchers, data scientists, etc.

Contribution to Pangeo community from European side (US side: i2c2 Pangeo deployment)

Pangeo eosc

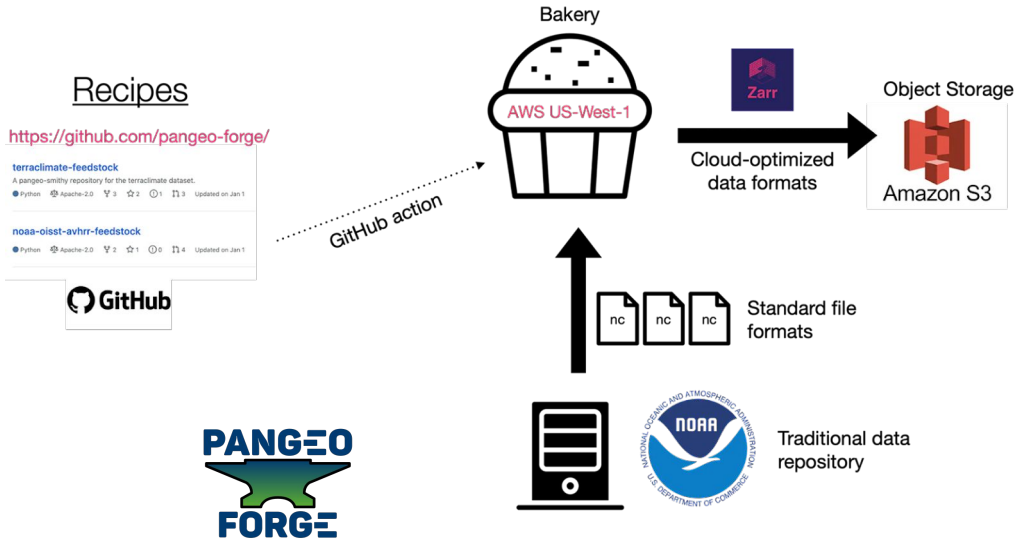
Easy and efficient access to geospatial data

Pangeo-forge for efficient data management within the Pangeo Community

→ Make it easy to extract data from any traditional repository and deposit this data in cloud object storage in an analysis-ready, cloud optimized (ARCO) format;

Two main components:

- Open source Python package for describing and running data pipelines;
- Cloud platform for automatically executing recipes stored in Github repos.

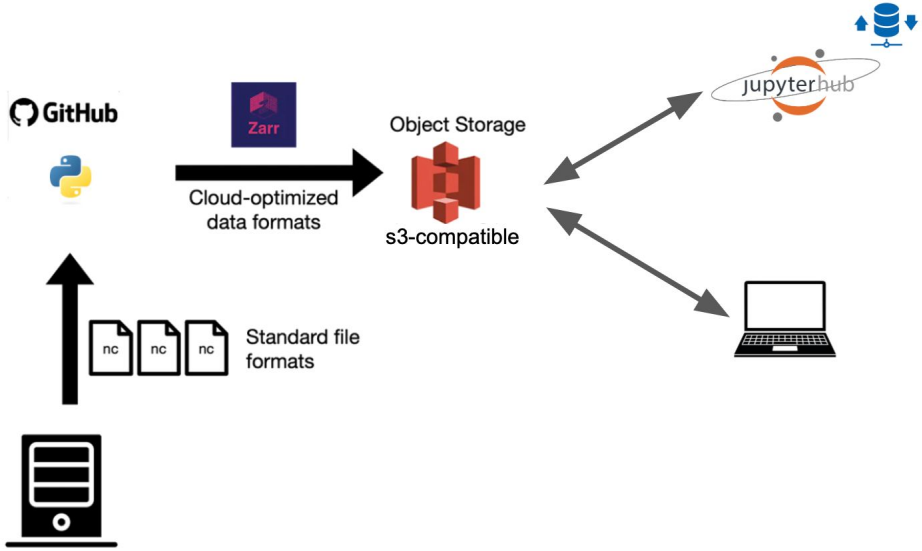


https://pangeo-forge.org/

Bring your own data

Flexible to customise for own needs:

- **Small to large** amount of data;
- Still possible to create Analysis Ready Cloud Optimised data yourself;
- **Private** or **public** s3-compatible buckets;
- Read/write data from/to **local storage** too.



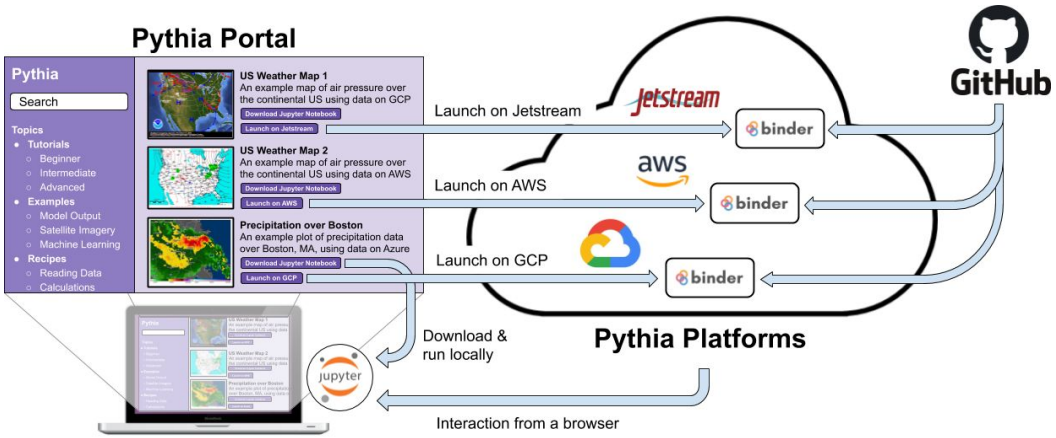
STAC and Pangeo

- Pangeo-forge supports the creation of analysis-ready cloud optimized (ARCO) data in cloud object storage from "classical" data repositories;
- STAC is used to create catalog and goes beyond the Pangeo ecosystem;
- Work is ongoing to figure out the best way to expose Pangeo-Forge-generated data assets via STAC catalogs.

Team-up with other initiatives to onboard new members and increase diversity

Team-up with Project Pythia

A Community Learning Resource for Geoscientists



https://projectpythia.org/

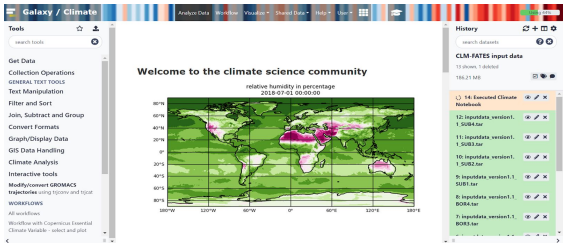
This work supported through the National Science Foundation Award #2026899.

Aspiration goal: Be the goto resource for learning the *Scientific Python Ecosystem*

- ★ Geoscience focused
- ★ From beginner to the power user
- ★ Tutorials, videos, examples, on-line courses, and sample data
- ★ Community owned!

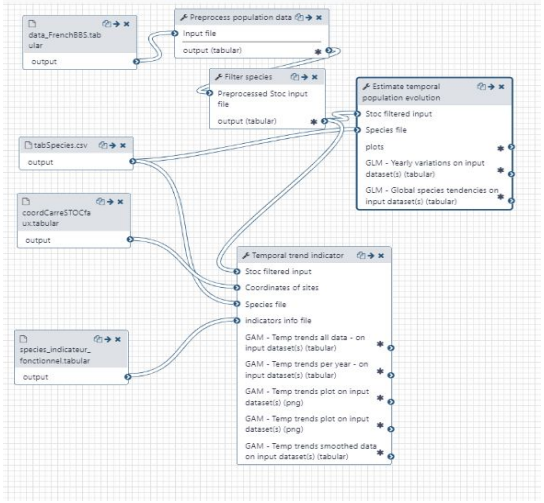
Source: Clyne et al (2022). ARM/ASR Open Science Virtual Workshop May 10, 2022

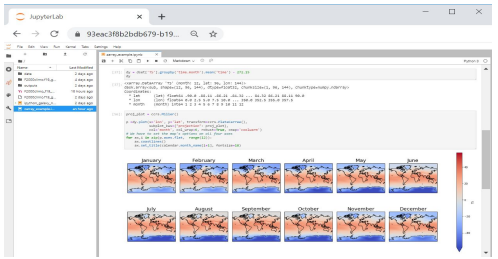
Team-up with Galaxy Europe



Galaxy is an open-source community and platform for FAIR data analysis. It offers:

- **Graphical User Interface (GUI) for users with no programming skills**
- Workflow editor to create and run fully reproducible data analysis
- Compute & Storage to everyone (free registration)
- Self Paced Learning material with the Galaxy Training Network
- Training Infrastructure as a Service (TiaaS) free and ready to use with private queues where only training's jobs run





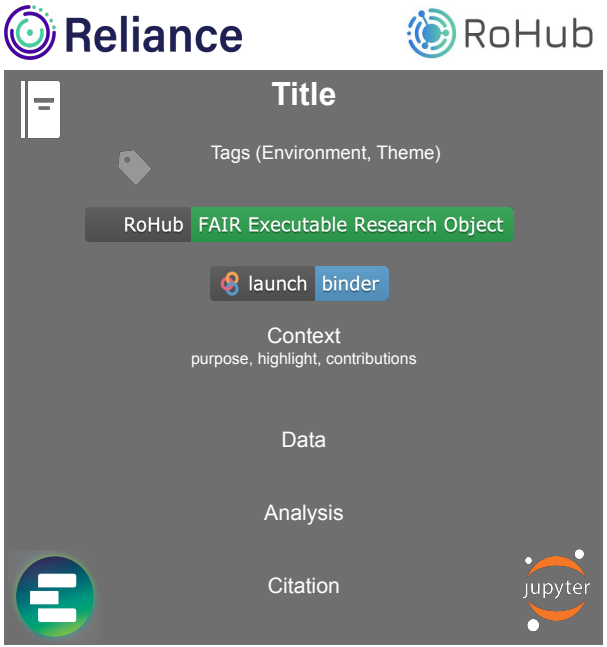
Galaxy PROJECT

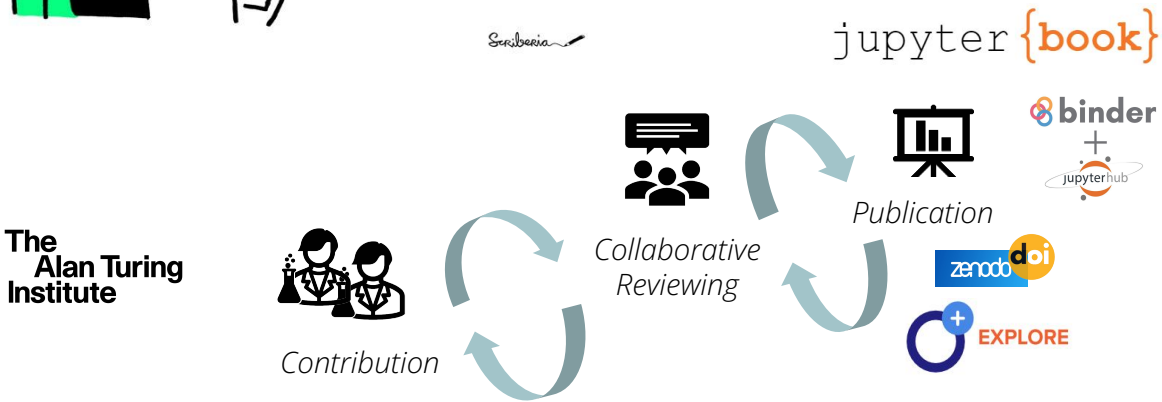
TIaaS

Team-up with the Environmental Data Science Book

https://edsbook.org/welcome.html

*Living, open and community-driven online resource to **showcase and support the publication** of data, research and open-source tools.*





Initiative led by Alejandro Coca-Castro

GeoPython community

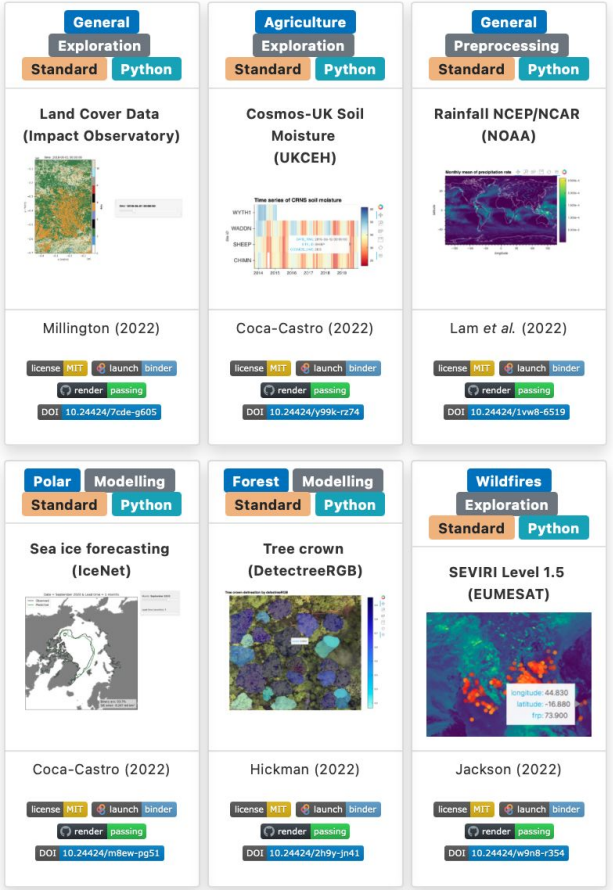
- Geopandas
- GDAL
- Pysal

Digital open books, building on Jupyter notebooks etc.

- “Geocomputation with Python” (or R) by Michael Dorman, Anita Graser, Jakub Nowosad, Robin Lovelace: https://py.geocompx.org/
- “Geographic Data Science with Python”, by Sergio J. Rey, Dani Arribas-Bel and Levi J. Wolf: https://geographicdata.science

Examples from the community

Gallery of FAIR and Open Executable Jupyter Notebooks



https://edsbook.org/notebooks/gallery.html

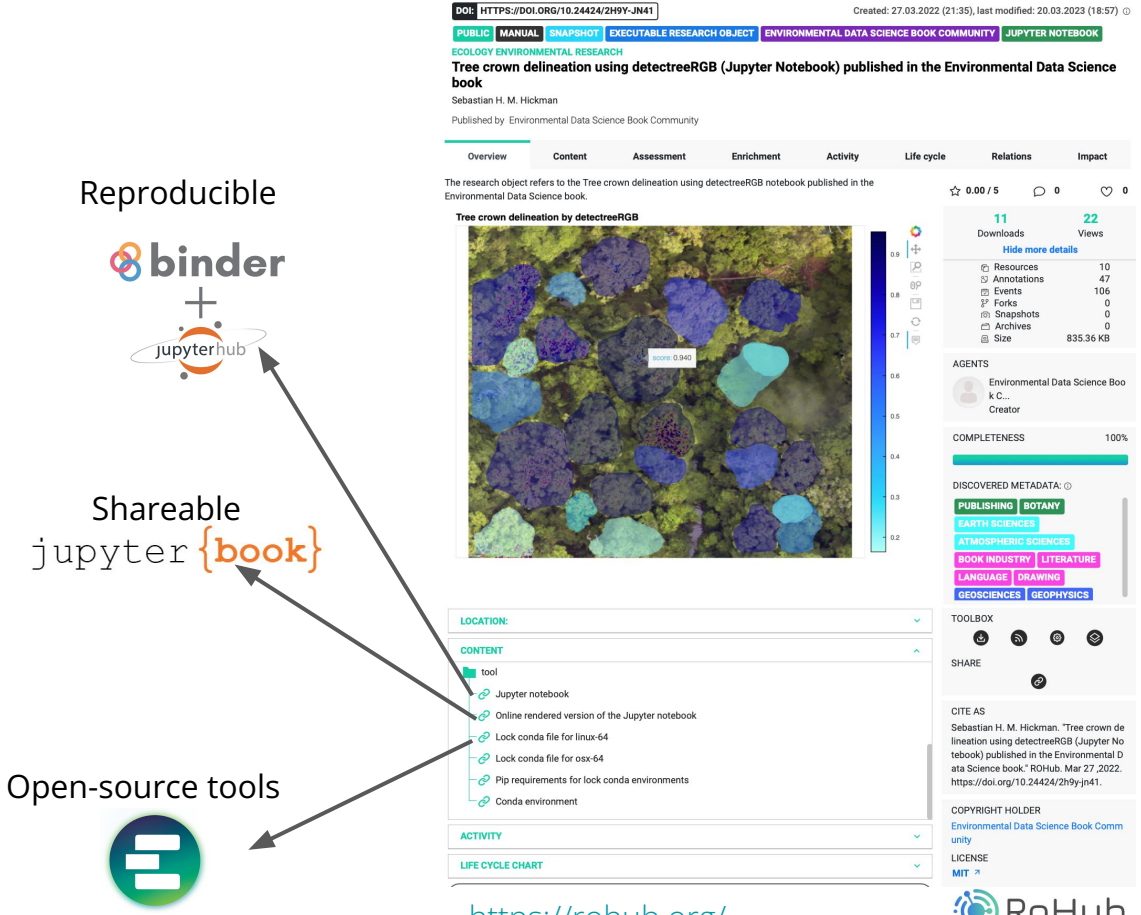
Reproducible

binder + jupyterhub

Shareable

jupyter {book}

Open-source tools

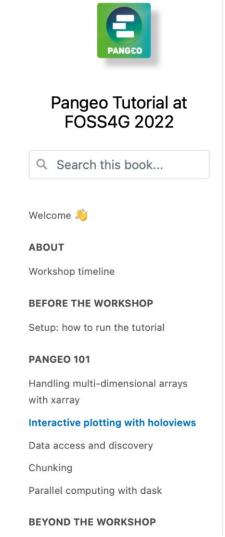


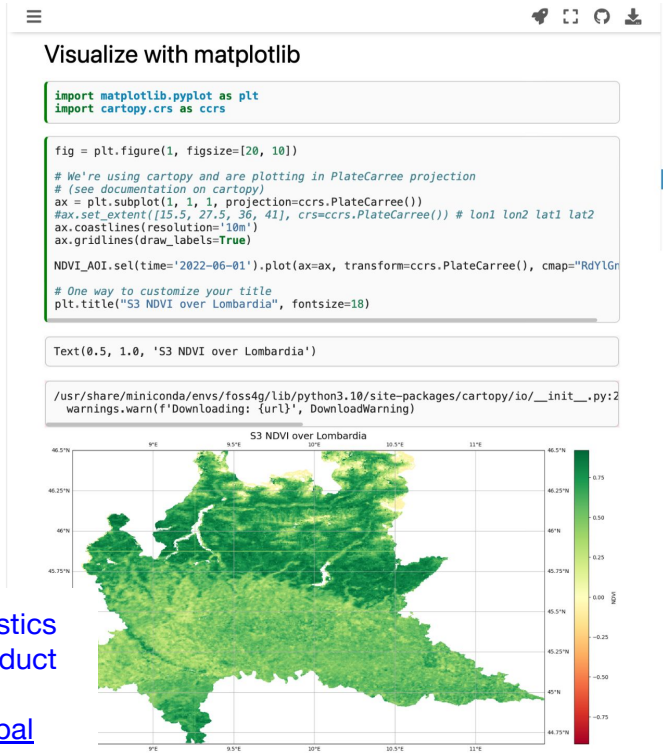
https://rohub.org/

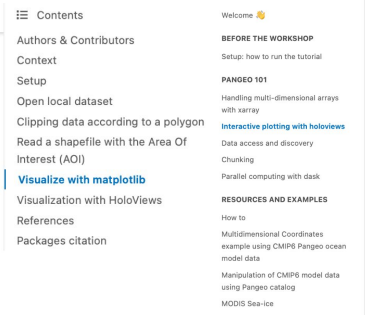
Xarray to access data

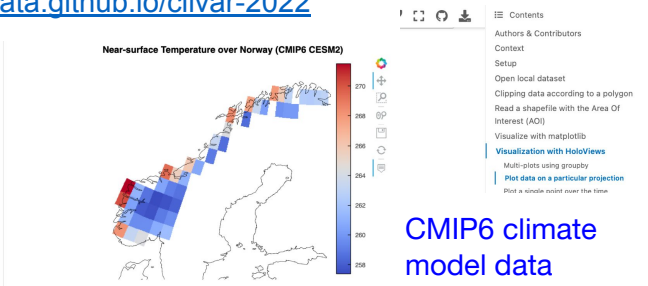
https://docs.xarray.dev/en/stable/

https://pangeo-data.github.io/clivar-2022

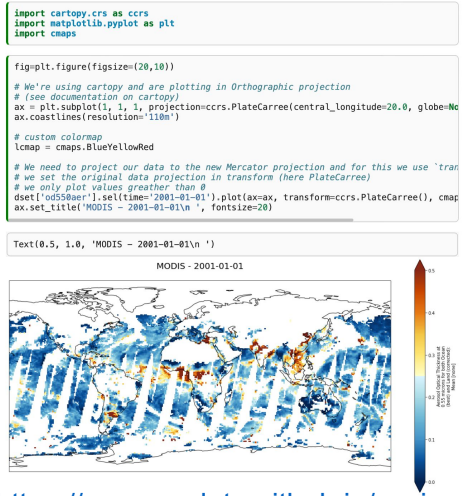








CMIP6 climate model data



Long Term Statistics (1999-2019) product provided by the Copernicus Global Land Service

Xarray data model is used by many higher-level libraries to manipulate gridded data

https://pangeo-data.github.io/foss4g-2022

https://pangeo-data.github.io/escience-2022

Pangeo Showcase Webinar Series

To learn from each other. It it part of “America” community calls. 15 minute talks are recorded, given a DOI and made available on the Pangeo YouTube Channel;

SPRING 2023 SHOWCASE

Date	Speaker	Title
2023-02-01 4PM EST	Tom Nicholas, Columbia University	Xarray-Datatree: Hierarchical Data Structures for Multi-Model Science DOI 10.5281/zenodo.7679730
2023-02-08 12PM EST	Alex Kerney, Gulf of Maine Research Institute	Mental Health for Geoscientists DOI 10.5281/zenodo.7679821
2023-02-15 4PM EST	Tim Crone, Columbia University	Lessons learned teaching Pangeo in the classroom DOI 10.5281/zenodo.7680128
2023-02-22 12PM EST	Ramon Ramirez-Linan, Navteca	D’explorer Explore cloud datasets from your notebooks (Replacement to S3 Explorer from IBM) DOI 10.5281/zenodo.7680210
2023-03-01 4PM EST	Tasha Snow, Colorado School of Mines	CryoCloud: Accelerating discovery for NASA Cryosphere communities with open cloud infrastructure DOI 10.5281/zenodo.7857296
2023-03-08 12PM EST	Isa Elegbede, Lagos State University	Data inclusivity and user needs for the global south DOI 10.5281/zenodo.7857353
2023-03-22 12PM EDT	James A. Bednar, Anaconda Inc.	SOSA: The Scalable Open-Source Analysis Stack DOI 10.5281/zenodo.7857369
2023-04-05 12PM EST	Alejandro Coca-Castro, The Alan Turing Institute	Environmental Data Science Book: a computational notebook community showcasing open and reproducible environmental science DOI 10.5281/zenodo.7857378

Summary

http://pangeo.io

- **No vendor lock-in;**
- Easy to start and **deployment on laptops, cloud and HPC;**
- “Reference” deployments on different cloud infrastructures;
- **Team up with other initiatives.** It can help to increase DEI (Diversity, Equity Inclusion):
  - Educational material and deliver trainings (Pythia, Galaxy);
  - Training infrastructure as a Service (Pythia, Galaxy, EOSC);
  - Use Pangeo from GUI (no programming skills required) on Galaxy Europe.
- **Contribute to easy creation of data in analysis-ready, cloud optimized (ARCO) format (pangeo-forge);**
- Promote the work done by the Pangeo Community and other Geosciences initiatives (Pangeo Show & Tell/Showcase/Pangeo discourse);
- Pangeo heavily used in industry;
- **Spin-off** (often from Pangeo community members) and many startups & companies using Pangeo software stack and contributing to Pangeo ecosystem.

Thank you for your attention!