



Pangeo

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

Anne Fouilloux, Simula Research Laboratory, Norway
Alexander Knoch, 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

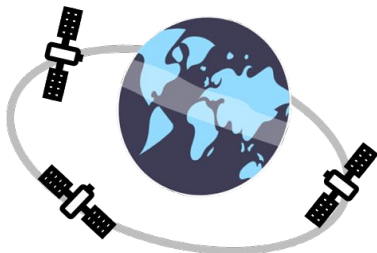
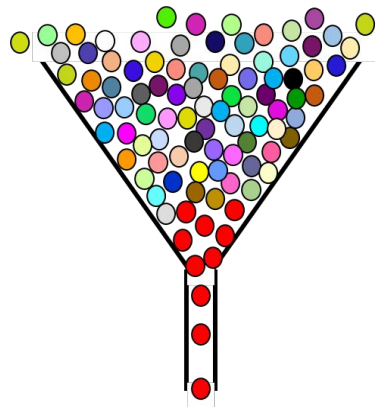
- <https://github.com/pangeo-data/governance/blob/master/governance.md>

Roadmap

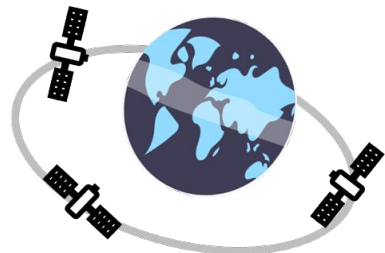
- <https://pangeo.io/roadmap.html>

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

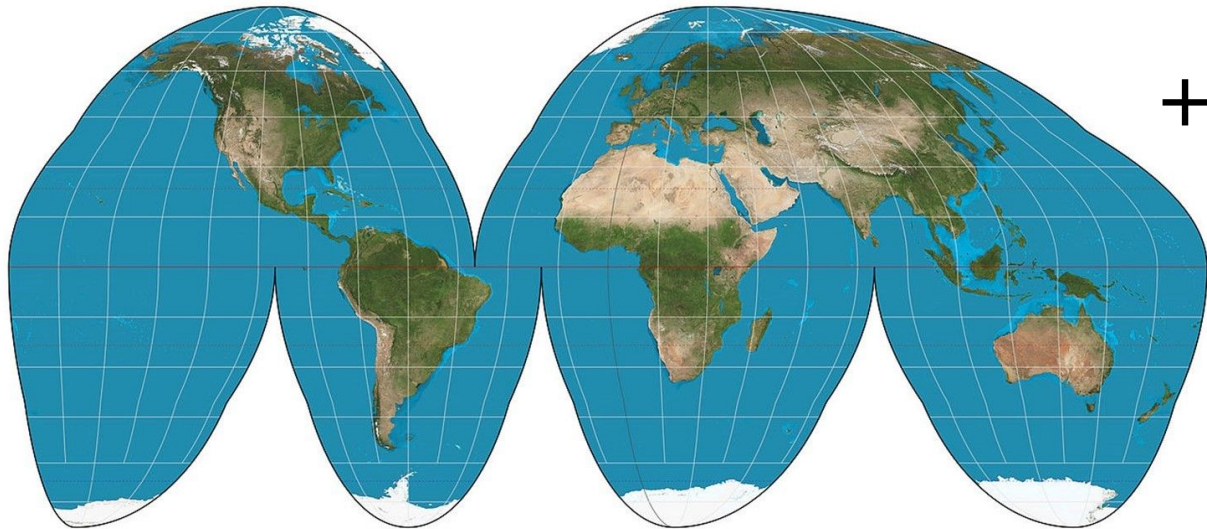
Every Wednesday,
alternating between 12pm
ET and 4pm ET

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

Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

PRINCETON
UNIVERSITY

USGS
science for a changing world

Element 84

Centre for Environmental
Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

ANACONDA.

ATM

CSIRO

UNIVERSITY OF MARYLAND BALTIMORE COUNTY
UMBC
1966

Los Alamos
NATIONAL LABORATORY
EST. 1943

simula

NATIONAL OCEANOGRAPHIC AND ATMOSPHERIC ADMINISTRATION
NOAA
U.S. DEPARTMENT OF COMMERCE

British Antarctic Survey

UNIVERSITY OF OXFORD

IMAU
Institute for Marine and Atmospheric research Utrecht

Inria

GEOMAR
Helmholtz Centre for Ocean Research Kiel

UNIVERSITY OF LEEDS

cnes

EOSC
NORDIC

EDOSC-PILLAR
Earth Science VRE

Met Office

IGE

UNIVERSITY OF TORONTO

neic
NORDIC E-INFRASTRUCTURE COLLABORATION

UiO
University of Oslo

The Alan Turing Institute

UNIVERSITY OF SASKATCHEWAN

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CU

climate extremes
ARC centre of excellence

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

unidata

universität innsbruck

Ifremer

LOPS

LEGOS

UNIVERSITY AT ALBANY
State University of New York

W
UNIVERSITY OF WASHINGTON

Net Carbon

INFORMATICS LAB

NVIDIA.

Galaxy PROJECT

HARVARD UNIVERSITY

MIAMI

STOCKHOLMS UNIVERSITET

MIT
Massachusetts Institute of Technology®

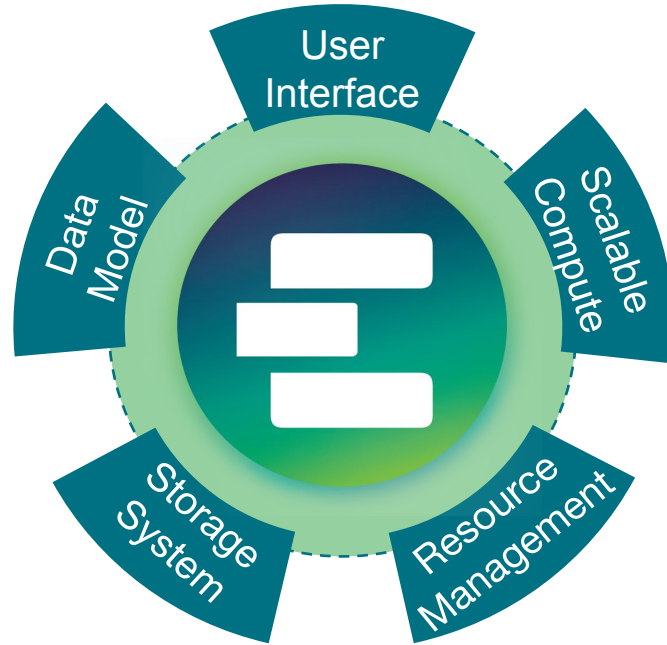
MOSES Langmuir Marine Laboratories

MEMORIAL UNIVERSITY
Microsoft

Based on GitHub and papers affiliations

Open Platform

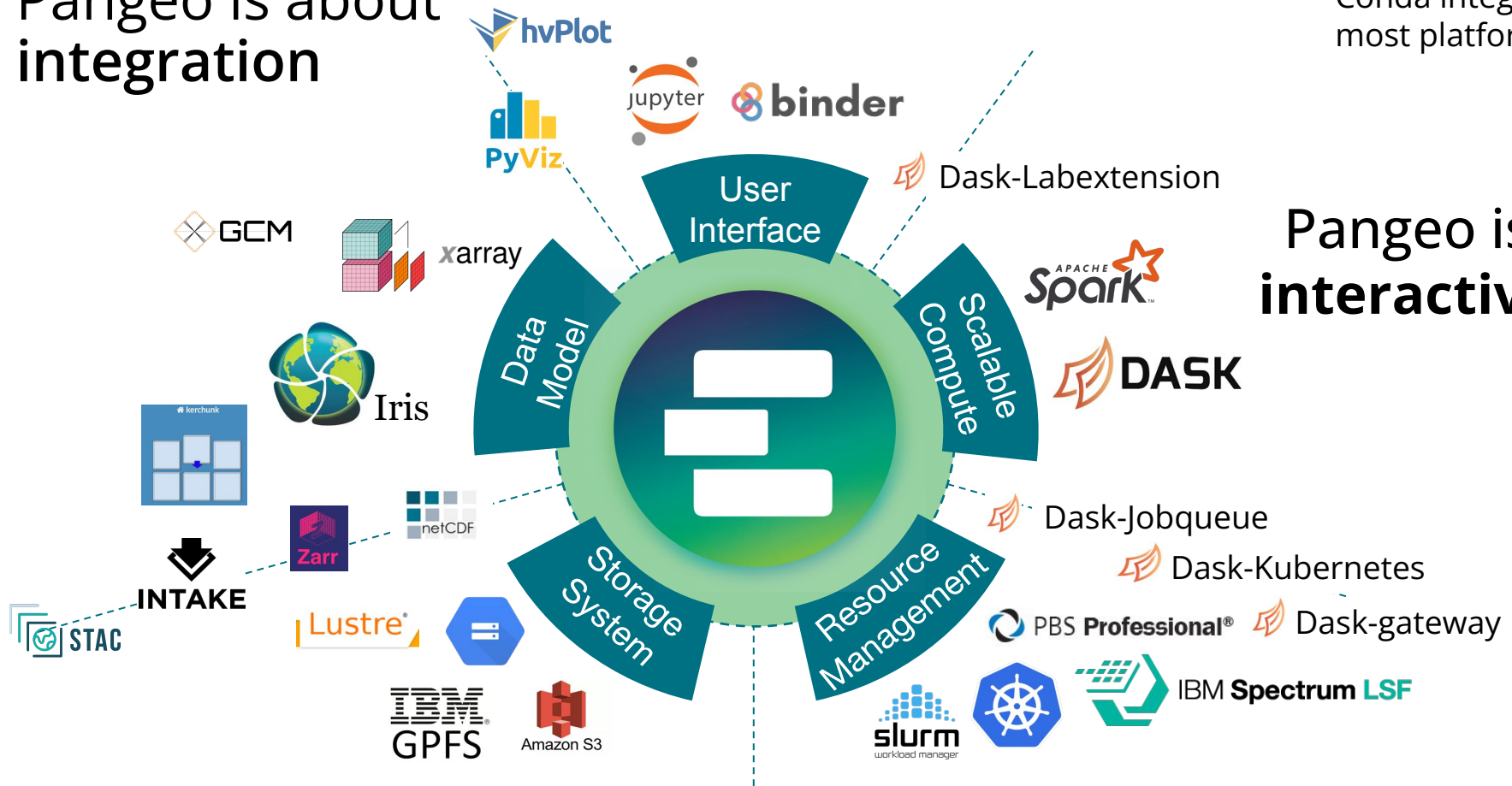
Pangeo **platform** *is scalable*



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



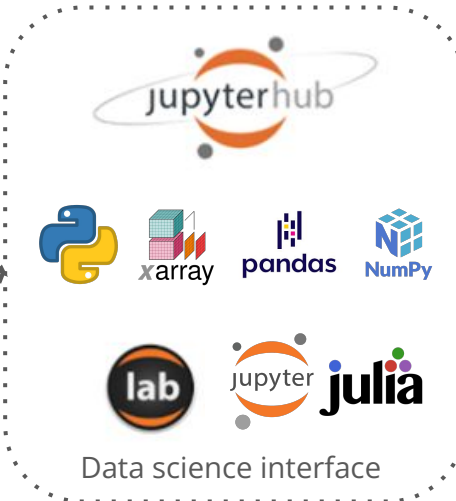
Custom environments



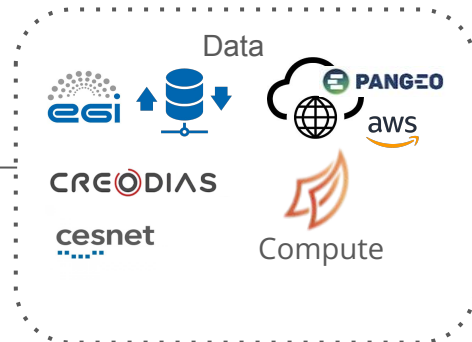
Online content



EOSC Community Hub



Cloud infrastructure



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



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



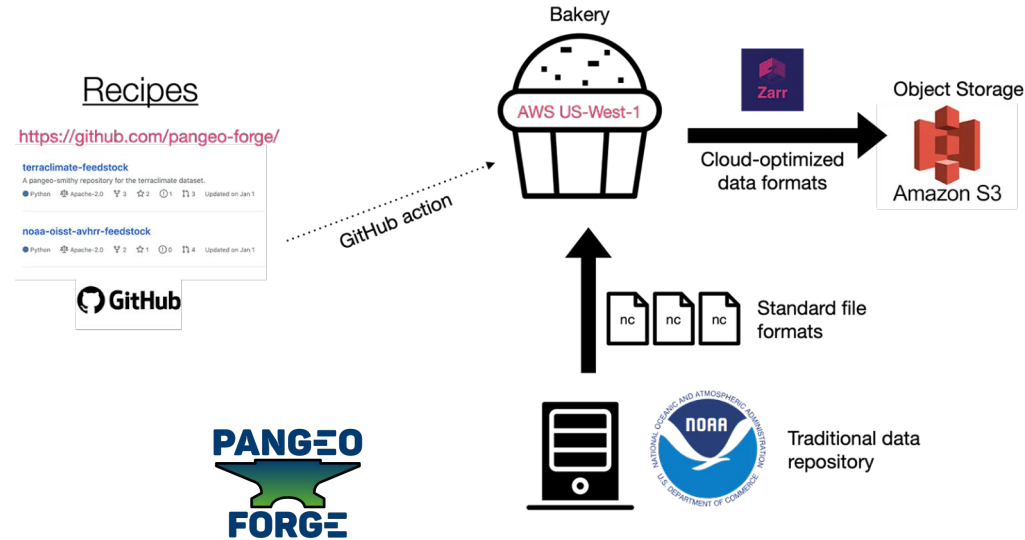
Easy and efficient
access to geospatial
data

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:

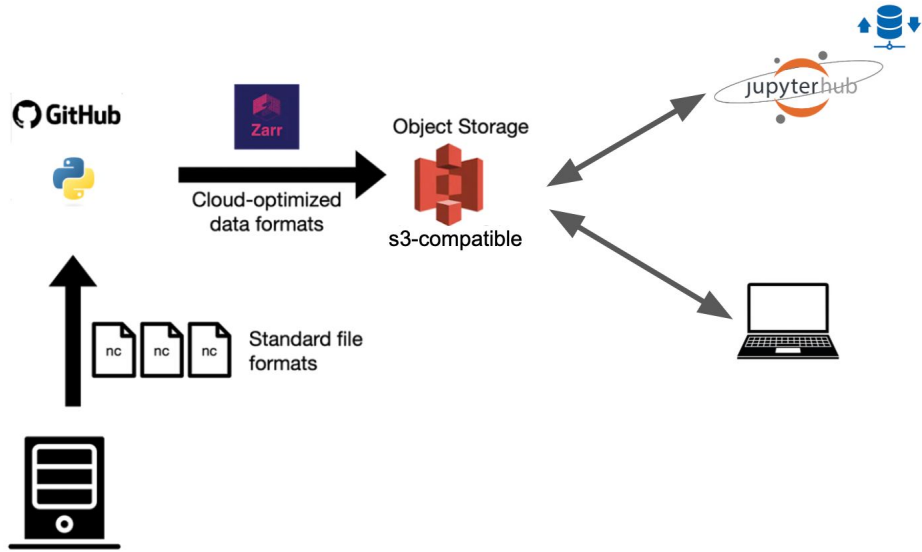
- Open source Python package for describing and running data pipelines;
- Cloud platform for automatically executing recipes stored in Github repos.



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.



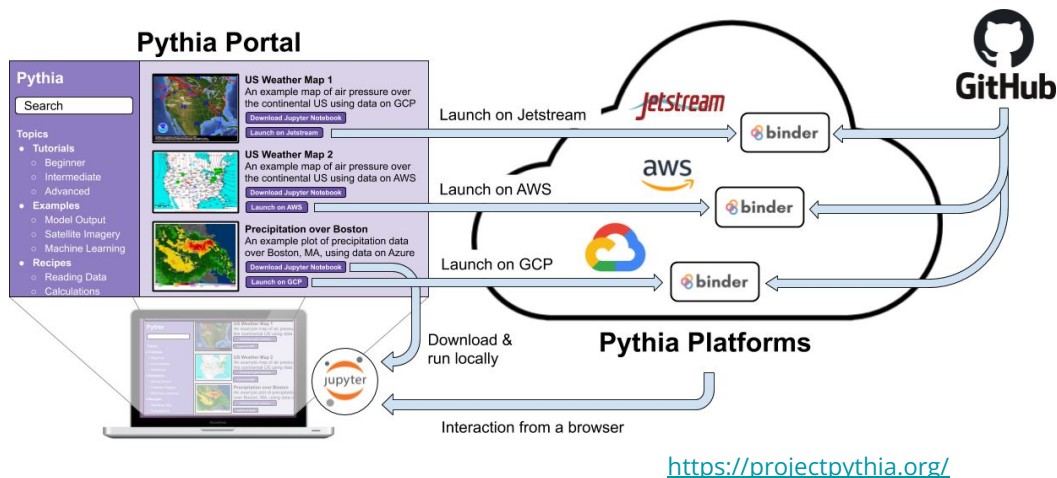
Team-up with other initiatives to
onboard new members and
increase diversity

Team-up with Project Pythia



A Community Learning Resource for Geoscientists

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



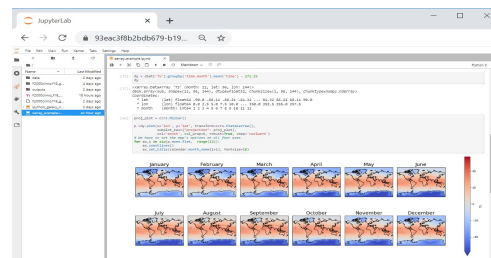
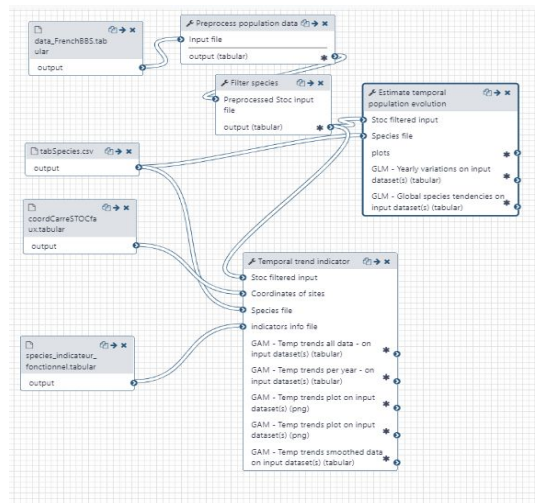
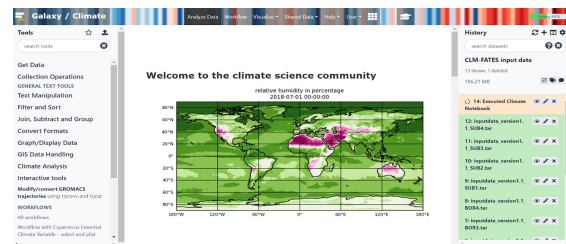
Aspiration goal: Be the goto resource for learning the *Scientific Python Ecosystem*

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

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



Team-up with the Environmental Data Science Book

<https://edsbook.org/welcome.html>

Reproducible, scalable, & shareable
ENVIRONMENTAL DATA SCIENCE



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



Title

Tags (Environment, Theme)

RoHub FAIR Executable Research Object

launch binder

Context
purpose, highlight, contributions

Data

Analysis

Citation

jupyter {book}

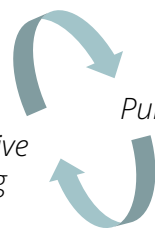
The Alan Turing Institute



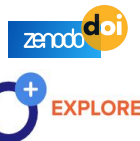
Contribution



Collaborative
Reviewing



Publication



Initiative led by Alejandro Coca-Castro

CC-BY image by Turing Way and Scriberia, CC-BY 4.0, EDS Book community, @EnvDSBook

GeoPython community

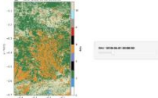
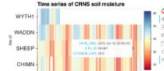
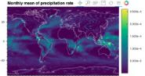
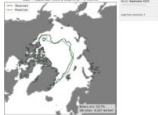
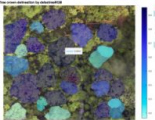
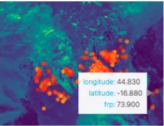
- Geopandas
- GDAL
- Pysal

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

<p>General Exploration</p> <p>Standard Python</p> <p>Land Cover Data (Impact Observatory)</p>  <p>Millington (2022)</p> <p>license MIT launch binder</p> <p>render passing</p> <p>DOI 10.24424/7cde-g605</p>	<p>Agriculture Exploration</p> <p>Standard Python</p> <p>Cosmos-UK Soil Moisture (UKCEH)</p>  <p>Coca-Castro (2022)</p> <p>license MIT launch binder</p> <p>render passing</p> <p>DOI 10.24424/y99k-r274</p>	<p>General Preprocessing</p> <p>Standard Python</p> <p>Rainfall NCEP/NCAR (NOAA)</p>  <p>Lam et al. (2022)</p> <p>license MIT launch binder</p> <p>render passing</p> <p>DOI 10.24424/1vw8-6519</p>
<p>Polar Modelling</p> <p>Standard Python</p> <p>Sea ice forecasting (IceNet)</p>  <p>Coca-Castro (2022)</p> <p>license MIT launch binder</p> <p>render passing</p> <p>DOI 10.24424/m8ew-pg51</p>	<p>Forest Modelling</p> <p>Standard Python</p> <p>Tree crown (DetectreeRGB)</p>  <p>Hickman (2022)</p> <p>license MIT launch binder</p> <p>render passing</p> <p>DOI 10.24424/2h9y-jn41</p>	<p>Wildfires Exploration</p> <p>Standard Python</p> <p>SEVIRI Level 1.5 (EUMESAT)</p>  <p>Jackson (2022)</p> <p>license MIT launch binder</p> <p>render passing</p> <p>DOI 10.24424/w9n8-r354</p>

Reproducible



Shareable
jupyter {book}

Open-source tools



DOI: <https://doi.org/10.24424/2h9y-jn41> Created: 27.03.2022 (21:35), last modified: 20.03.2023 (18:57) ⓘ

PUBLIC **MANUAL** **SNAPSHOT** **EXECUTABLE RESEARCH OBJECT** **ENVIRONMENTAL DATA SCIENCE BOOK COMMUNITY** **JUPYTER NOTEBOOK**

ECOLOGICAL ENVIRONMENTAL RESEARCH

Tree crown delineation using detectreeRGB (Jupyter Notebook) published in the Environmental Data Science book

Sebastian H. M. Hickman
Published by Environmental Data Science Book Community

Overview Content Assessment Enrichment Activity Life cycle Relations Impact

The research object refers to the Tree crown delineation using detectreeRGB notebook published in the Environmental Data Science book.

11 Downloads 22 Views

Hide more details

- Resources 10
- Annotations 47
- Events 106
- Forks 0
- Snapshots 0
- Archives 0
- Size 835.36 KB

AGENTS

Environmental Data Science Book Community
k C...
Creator

COMPLETENESS 100%

DISCOVERED METADATA: ⓘ

- PUBLISHING
- BOTANY
- EARTH SCIENCES
- ATMOSPHERIC SCIENCES
- BOOK INDUSTRY
- LITERATURE
- LANGUAGE
- DRAWING
- GEOSCIENCES
- GEOPHYSICS

LOCATION:

CONTENT

- tool
- Jupyter notebook
- Online rendered version of the Jupyter notebook
- Lock conda file for linux-64
- Lock conda file for osx-64
- Pip requirements for lock conda environments
- Conda environment

ACTIVITY

LIFE CYCLE CHART

TOOLBOX

SHARE

CITE AS

Sebastian H. M. Hickman. "Tree crown delineation using detectreeRGB (Jupyter Notebook) published in the Environmental Data Science book." RoHub. Mar 27, 2022. <https://doi.org/10.24424/2h9y-jn41>.

COPYRIGHT HOLDER

Environmental Data Science Book Community

LICENSE

MIT ^{2.0}

Xarray to access data

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

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



Pangeo Tutorial at FOSS4G 2022

Search this book...

Welcome

ABOUT

Workshop timeline

BEFORE THE WORKSHOP

Setup: how to run the tutorial

PANGEO 101

Handling multi-dimensional arrays with xarray

Interactive plotting with holoviews

Data access and discovery

Chunking

Parallel computing with dask

BEYOND THE WORKSHOP

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

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

Visualize with matplotlib

```
import matplotlib.pyplot as plt
import cartopy.crs as ccrs

fig = plt.figure(1, figsize=[20, 10])

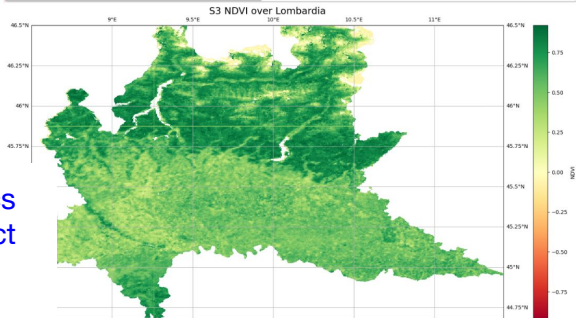
# We're using cartopy and are plotting in PlateCarree projection
# (see documentation on cartopy)
ax = plt.subplot(1, 1, projection=ccrs.PlateCarree())
#ax.set_extent([15.5, 27.5, 36, 41], crs=ccrs.PlateCarree()) # lon1 lon2 lat1 lat2
ax.coastlines(resolution='10m')
ax.gridlines(draw_labels=True)

NDVI_AOI.sel(time='2022-06-01').plot(ax=ax, transform=ccrs.PlateCarree(), cmap='RdYlGn')

# One way to customize your title
plt.title("S3 NDVI over Lombardia", fontsize=18)

Text(0.5, 1.0, 'S3 NDVI over Lombardia')
```

A map of Lombardia, Italy, showing NDVI data for June 1, 2022. The map uses a PlateCarree projection and a RdYlGn color map. The title is "S3 NDVI over Lombardia".



Contents

- Authors & Contributors
- Context
- Setup
- Open local dataset
- Clipping data according to a polygon
- Read a shapefile with the Area Of Interest (AOI)
- Visualize with matplotlib
- Visualization with HoloViews
- References
- Packages citation

```
import cartopy.crs as ccrs
import matplotlib.pyplot as plt
import cmaps

fig=plt.figure(figsize=(20,10))

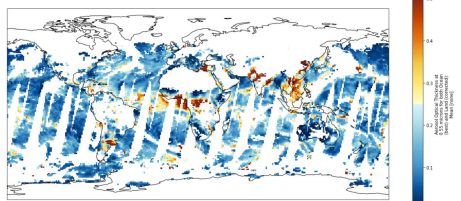
# We're using cartopy and are plotting in Orthographic projection
# (see documentation on cartopy)
ax = plt.subplot(1, 1, projection=ccrs.PlateCarree(central_longitude=20.0, globe=False))
ax.coastlines(resolution='10m')

# custom colormap
lcnmap = cmaps.BlueYellowRed

# We need to project our data to the new Mercator projection and for this we use 'transform'
# we set the original data projection in transform (here PlateCarree)
# we only plot values greater than 0
fig['ndvi'] = ndvi.sel(time='2001-01-01').plot(ax=ax, transform=ccrs.PlateCarree(), cmap=
ax.set_title('MODIS - 2001-01-01', fontsize=20)

Text(0.5, 1.0, 'MODIS - 2001-01-01')
```

A map of the world showing MODIS data for January 1, 2001. The map uses a PlateCarree projection and a BlueYellowRed color map. The title is "MODIS - 2001-01-01".



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



Pangeo Tutorial at CLIVAR CMIP6 Bootcamp 2022

Search this book...

Welcome

BEFORE THE WORKSHOP

Setup: how to run the tutorial

PANGEO 101

Handling multi-dimensional arrays with xarray

Clipping data according to a polygon

Data access and discovery

Chunking

Parallel computing with dask

RESOURCES AND EXAMPLES

How to

Multidimensional Coordinates

example using CMIP6 Pangeson ocean model data

Manipulation of CMIP6 model data using Pangeo catalog

MODIS Sea-ice

A map of Norway showing near-surface temperature data from the CMIP6 CESM2 model. The map uses a PlateCarree projection and a color map ranging from 270 to 298. The title is "Near-surface Temperature over Norway (CMIP6 CESM2)".

Plot a single point over the time dimension

```
tas_AOI.isel(member_id=0).sel(lat=60, lon=10.75, method='nearest').hvplot(yLim=(260, 290))

WARNING:param_CurvePlot02669: Converting cftime.datetime from a non-standard calendar

lat = 59.84 [degrees_north], lon = 11.25 [degre...
```

A line plot showing Surface Air Temperature (K) over time. The x-axis is labeled "time" and ranges from 1900 to 2000. The y-axis is labeled "Surface Air Temperature (K)" and ranges from 270 to 290. The plot shows a highly variable time series.

CMIP6 climate model data

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

Pangeo Showcase Webinar Series

To learn from each other. It is 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!

