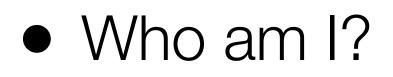
PANGEO A COMMUNITY-DRIVEN EFFORT FOR BIG DATA GEOSCIENCE

WHAT WOULD YOU LIKE TO HAVE AND WHY? Pangeo's vision for scientific computing in the big-data era



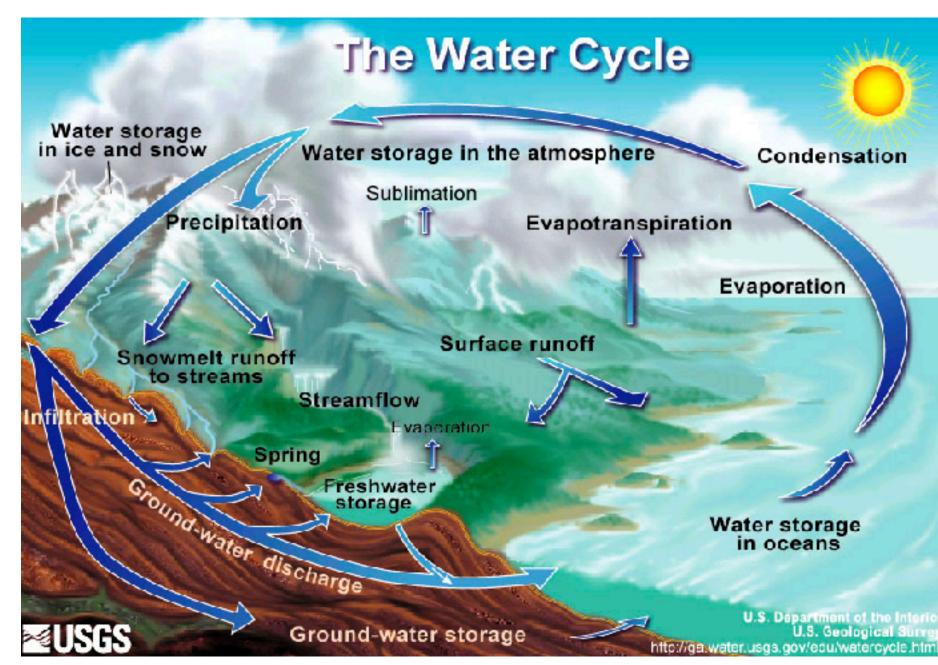
PANGEO'S WEBSITE pangeo-data.org



- ► Joe Hamman, Ph.D., P.E.
- I am a scientist at the National Center for Atmospheric Research
- I study the impacts of climate change on the water cycle.
- I contribute to open-source projects like Pangeo, Xarray, Dask, and Jupyter

PANGEO Ξ

HELLO



Github: @jhamman Twitter: @HammanHydro Web: joehamman.com







Lamont-Doherty Earth Observatory COLUMBIA UNIVERSITY | EARTH INSTITUTE

Ryan Abernathey, Chiara Lepore, Michael Tippet, Naomi Henderson, Richard Seager



Kevin Paul, Joe Hamman, Ryan May, Davide Del Vento



Matthew Rocklin



EARTHCUBE AWARD TEAM EarthCube **Google** Cloud Platform

OTHER CONTRIBUTORS



Met Office



Jacob Tomlinson, Niall Roberts, Alberto Arribas Developing and operating Pangeo environment to support analysis of UK Met office products

Rich Signell









Justin Simcock

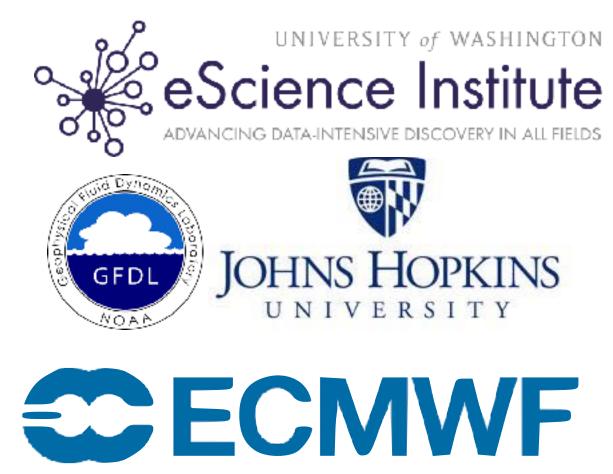
Supporting Pangeo via SWOT mission and funded ACCESS award to UW / NCAR

Yuvi Panda, Chris Holdgraf Spending lots of time helping us make things work on the cloud



Deploying Pangeo on AWS to support analysis of coastal ocean modeling

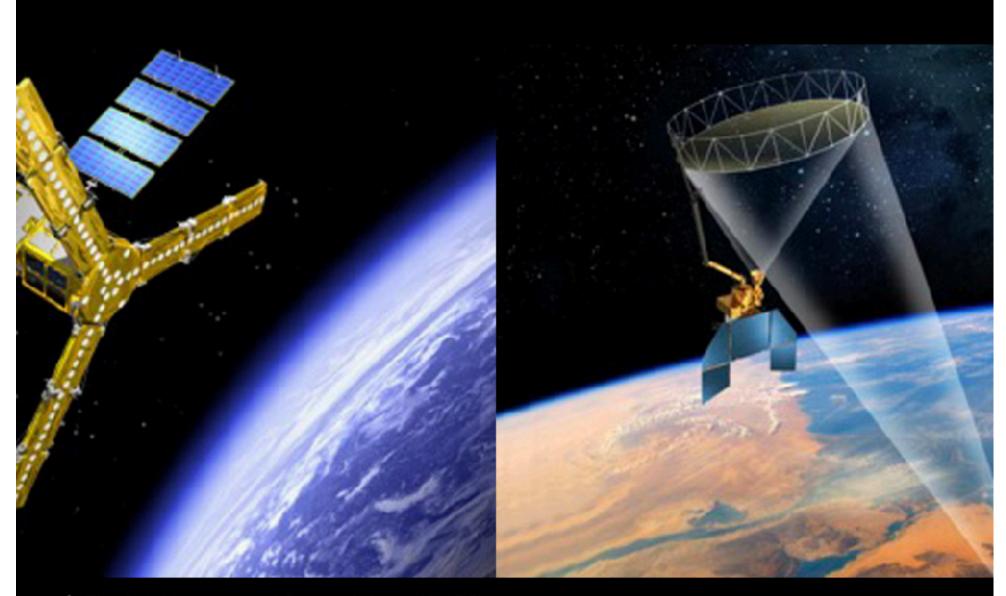
Operating Pangeo in the cloud to support Climate Impact Lab research and analysis



BIG DATA IN THE GEOSCIENCES We use our observations to test our models... and our models to test our

We use our observations to test or observations

Observations



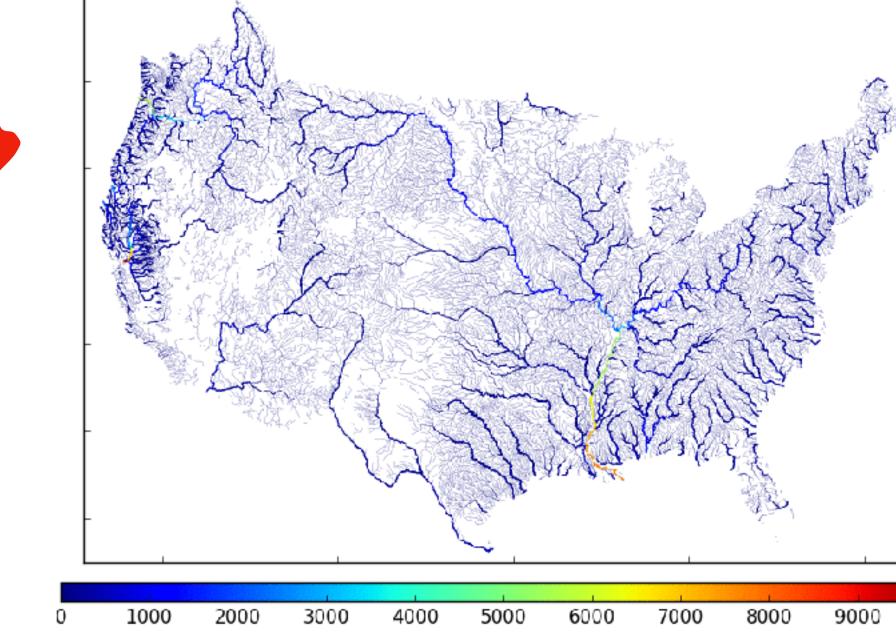
Left: The Soil Moisture and Ocean Salinity (SMOS) www.smos-mode.eu

Right: The Soil Moisture Active/Passive (SMAP) mission www.jpl.nasa.gov



Simulations

Monthly Streamflow [cms] 2000-1







tools for the big-data geosciences can be developed, distributed, and sustained.

• Vision:

- Open and collaborative development
- Tools for scaling computations from small to very large datasets
- Frameworks for moving scientific analysis to the data
- Welcoming and inclusive development culture



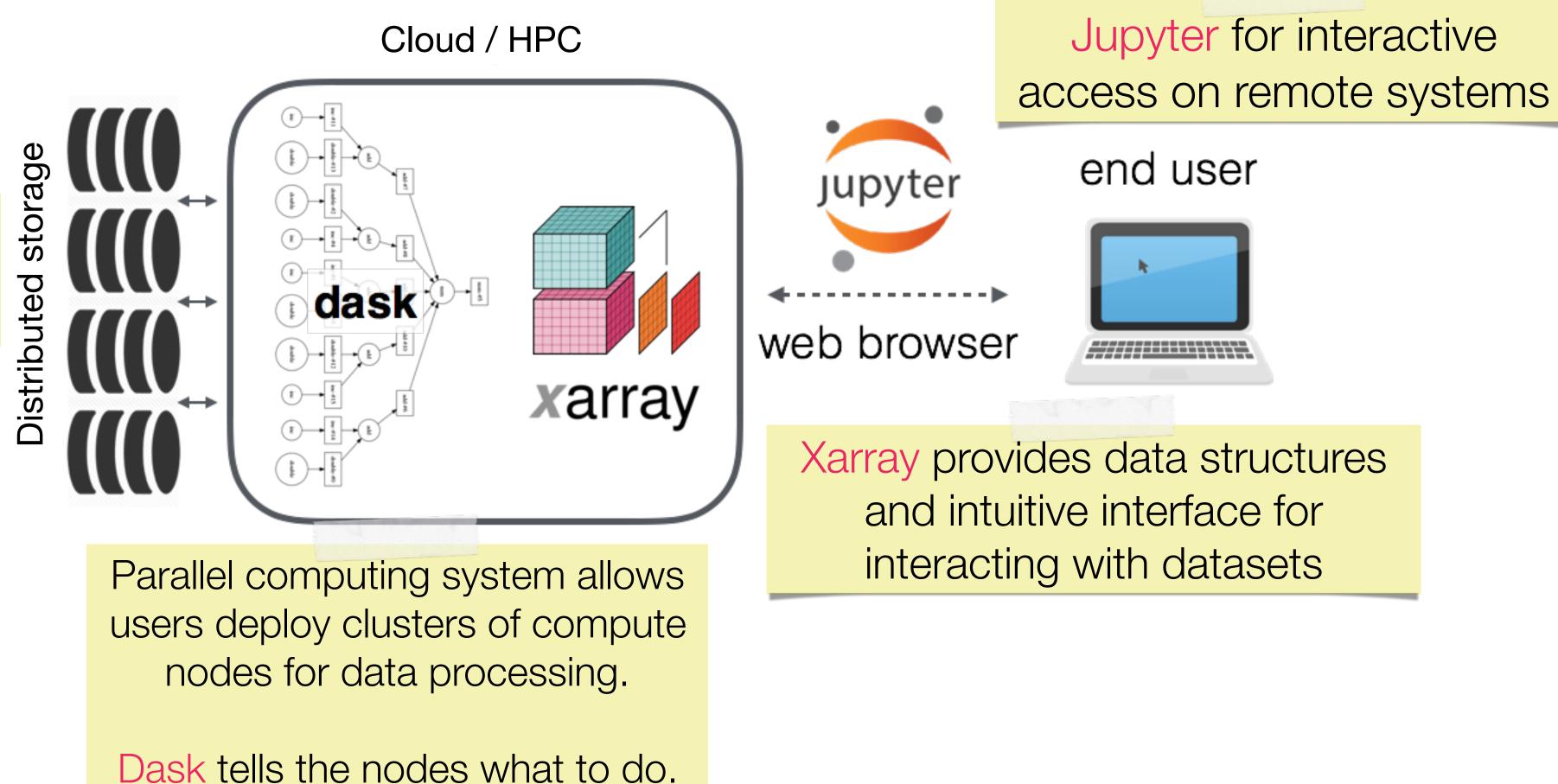
WHAT IS PANGE0?

Pangeo is a community working to develop software and infrastructure to enable big-data geoscience.

• **Mission:** To cultivate an ecosystem in which the next generation of open-source analysis

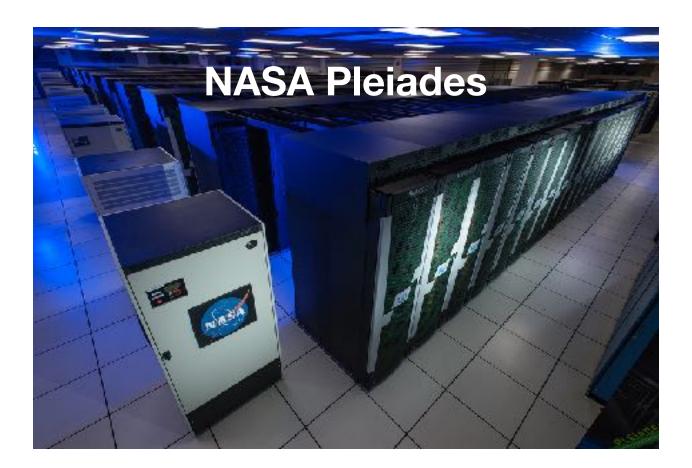
PANGEO ARCHITECTURE

"Analysis Ready Data" stored on globally-available distributed storage.





PANGEO DEPLOYMENTS



NCAR Cheyenne



(SCALE USING JOB QUEUE SYSTEM)



<u>HTTP://PANGEO-DATA.ORG/DEPLOYMENTS.HTML</u>

PANGEO.PYDATA.ORG



Over 500 unique users since March!

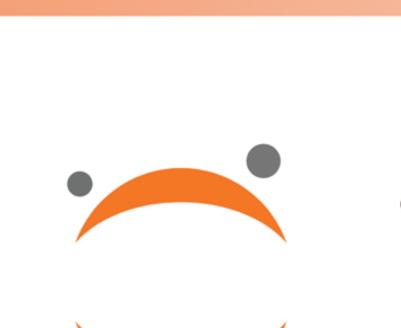
Google Cloud Platform



(SCALE USING KUBERNETES)

<u>PANGEO.PYDATA.ORG</u>

1 0 +



PANGEO

Ċ

pangeo.pydata.org

Pangeo is a community effort for big data geoscience. This JupyterHub is a multi-user server for interactive data analysis running on Google Cloud Platform.

Acknowledgments: Pangeo is supported, in part, by the National Science Foundation (NSF) and the Earth Cube Program. Google provided compute credits on Google Compute Engine.

For assistance, please open a Github Issue.

SIGN IN WITH GITHUB



Google Cloud Platform



• What is pangeo.pydata.org?

- Multi-user JupyterHub running on Google Cloud Platform
- Zero-to-jupyterhub deployment using Kubernetes
- Dask scales using "Dask-Kubernetes"

• Why the cloud?

- Highly scalable (storage, compute, user access)
- Easy to customize
- Cost effective

WHAT WOULD YOU LIKE TO HAVE AND WHY?

- 1. Scalable data-proximate computing
- 2. Cloud optimized data formats
- 3. Machine readable catalogs
- 4. Transparent data portals
- 6. On demand derived data products



5. Helpful scientific IT administrators with cloud-native experience

CLOUD OPTIMIZED DATA FORMATS FOR MULTI-DIMENSIONAL DATA

- ► WHAT DO WE WANT?
 - **Self-describing**: data and metadata packaged together
 - **On-demand**: data can be read/used in its current form from anywhere
 - Analysis-ready: no pre-processing required
- ► WHY THE CLOUD?
 - Too big to move: assume data is to be used but not copied
 - **Easy to share:** reduces duplicate storage
 - **Scalable:** storage and throughput during computation



COG CLOUD OPTIMIZED GEOTIFF

GeoTiff files stored in cloud object store support http byte range requests.



We don't know what the Cloud Optimized NetCDF will be.

MACHINE READABLE DATA CATALOGS

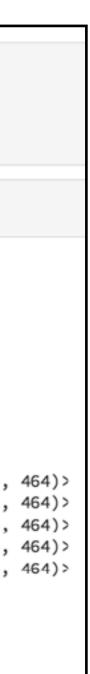
- Data discovery and access is too hard.
- We need better ways to index/ search existing metadata catalogs
- It should be much easier to use catalogs to get actual data
- As a scientist, 3 lines of code is about the right amount of effort to starting working with a dataset.

PANGEO

```
In [4]: # Load with intake catalog service
        import intake
        cat = intake.Catalog('https://raw.githubusercontent.com/pangeo-data/pangeo/master/gce/catalog.yaml')
        ds = cat.gmet_v1.read_chunked()
In [5]: # Print dataset
        ds
Out[5]: <xarray.Dataset>
                       (ensemble: 100, lat: 224, lon: 464, time: 12054)
        Dimensions:
        Coordinates:
          * ensemble
                       (ensemble) int64 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 ...
          * lat
                       (lat) float54 25.06 25.19 25.31 25.44 25.56 25.69 25.81 25.94 ...
                       (lon) float54 -124.9 -124.8 -124.7 -124.6 -124.4 -124.3 ...
          * lon
          * time
                       (time) datetime64[ns] 1980-01-01 1980-01-02 1980-01-03 ...
        Data variables:
            elevation (lat, lon) float64 dask.array<shape=(224, 464), chunksize=(224, 464)>
                       (lat, lon) int32 dask.array<shape=(224, 464), chunksize=(224, 464)>
            mask
                       (ensemble, time, lat, lon) flcat64 dask.array<shape=(100, 12054, 224, 464), chunksize=(1, 366, 224, 464)>
            рср
                       (ensemble, time, lat, lon) float64 dask.array<shape=(100, 12054, 224, 464), chunksize=(1, 366, 224, 464)>
            t_max
                       (ensemble, time, lat, lon) float64 dask.array<shape=(100, 12054, 224, 464), chunksize=(1, 366, 224, 464)>
            t_mean
                       (ensemble, time, lat, lon) float64 dask.array<shape=(100, 12054, 224, 464), chunksize=(1, 366, 224, 464)>
            t_min
                       (ensemble, time, lat, lon) flcat64 dask.array<shape=(100, 12054, 224, 464), chunksize=(1, 366, 224, 464)>
            t_range
        Attributes:
            history:
                                       Version 1.0 of ensemble dataset, created Decem...
            institution:
                                       National Center for Atmospheric Research (NCAR...
            nco_openmp_thread_number:
            references:
                                       Newman et al. 2015: Gridded Ensemble Precipita...
                                       Generated using version 1.0 of CONUS ensemble ...
            source:
                                       CONUS daily 12-km gridded ensemble precipitati...
            title:
```

Example Python snippet demonstrating the use of Intake (<u>https://</u> intake.readthedocs.io) catalog package with Xarray, Jupyter, and GoogleCloud



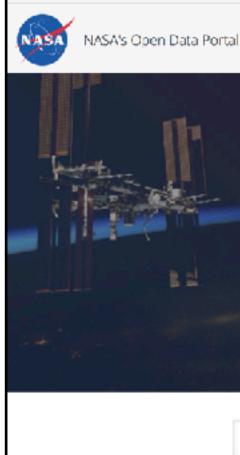




TRANSPARENT DATA PORTALS

► WHAT DO WE WANT?

- Intuitive: organization needs to easy to understand
- **Simple**: prioritize direct/easy access to datasets instead of fancy interfaces
- ► WHY?
 - Manual searches and check boxes don't scale: portals should supply machine readable data catalogs
 - Easy to automate: Batch queries and retrievals



About

This registry exists to help people discover and share datasets that are available via AWS resources. Learn more about sharing data on AWS.

Search datasets (currently 64 matching datasets)

Add to this registry



t Q 🐱 f 🛩 🔠 🖸 Data Catalog About Developer Resources Suggest a Dataset

Welcome to NASA's Data Portal

This site is a continually growing catalog of publicly available NASA datasets, APIs, visualizations, and more!

Q Search

https://open.nasa.gov/open-data/

Registry of Open Data on AWS

See all usage examples for datasets listed in this registry.

Search datasets

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the Registry of Open Data on AWS GitHub repository.

Landsat 8

earth observation

satellite imagery gis natural resource sustainability disaster response

An ongoing collection of satellite imagery of all land on Earth produced by the Landsat 8 satellite.

Details →

Usage examples

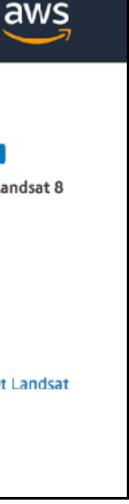
- Apps for exploring and analyzing Landsat imagery on the fly by Esri
- Sentinel Playground for Landsat by Sinergise
- Development Seed Geolambda by Matthew Hanson
- Using Vector tiles and AWS Lambda, we can build a really simple API to get Landsat. and Sentinel images by Remote Pixel
- FME Landsat-8 on AWS Reader by Safe Software

See 13 usage examples →

https://registry.opendata.aws



Sign In



PANGEO-DATA.ORG

PANGEO Blog Site - Page -



PANGEO

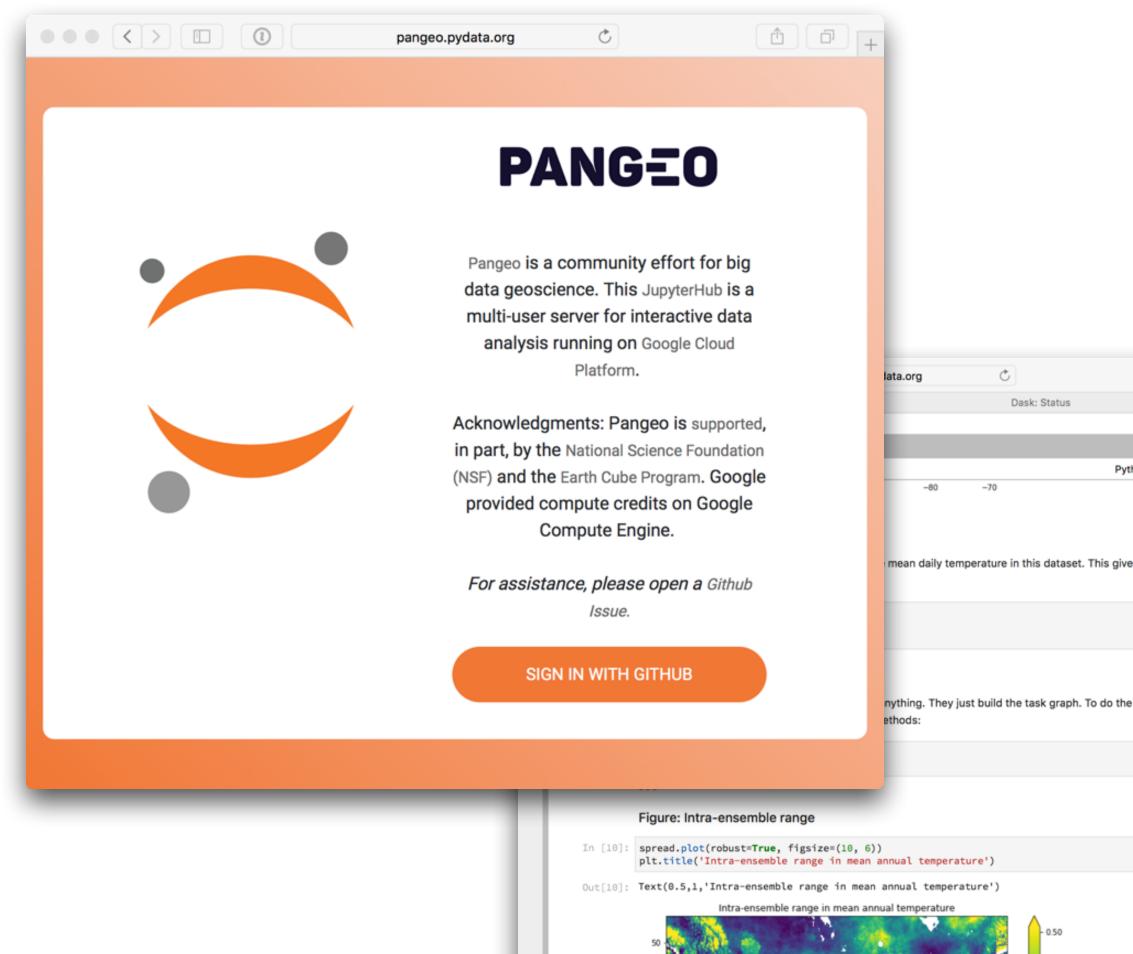
A community platform for Big Data geoscience

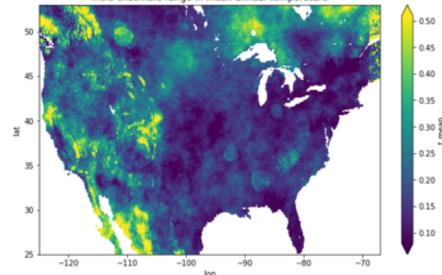
OUR GOALS

- 1. Foster collaboration around the open source scientific python ecosystem for ocean / atmosphere / land / climate science.
- 2. Support the development with domain-specific geoscience packages.
- 3. Improve scalability of these tools to to handle petabyte-scale datasets on HPC and cloud platforms.

PANGEO

PANGEO.PYDATA.ORG







t ø	
+	
hon [default]	
es us a sense	