PANGEO

A COMMUNITY-DRIVEN EFFORT FOR
BIG DATA GEOSCIENCE
Who am I?

- Joe Hamman, Ph.D., P.E.
- I am a scientist at the National Center for Atmospheric Research (RAL & CGD)
- I study the impacts of climate change on the water cycle.
- I am a core developer of Xarray
- I am a founding member of the Pangeo project

Github: @jhamman
Twitter: @HammanHydro
Web: joehamman.com
Scientific Python for Data Science

Credit: Stephan Hoyer, Jake Vanderplas (SciPy 2015)
Xarray Dataset: Multidimensional Variables with Coordinates and Metadata

Data variables 
*used for computation*

Coordinates 
*describe data*

Indexes 
*align data*

“netCDF meets pandas.DataFrame”

Attributes 
*metadata ignored by operations*

Credit: Stephan Hoyer
XARRAY

http://xarray.pydata.org

• label-based indexing and arithmetic

• interoperability with the core scientific Python packages (e.g., pandas, NumPy, Matplotlib)

• out-of-core computation on datasets that don’t fit into memory (thanks dask!)

• wide range of input/output (I/O) options: netCDF, HDF, geoTIFF, zarr

• advanced multi-dimensional data manipulation tools such as group-by and resampling
XARRAY UPDATES

  - More flexible grids/indexing
  - More flexible arrays/computing
  - More flexible storage backends

- **NumFOCUS Sponsorship**
  - [https://numfocus.org/project/xarray](https://numfocus.org/project/xarray)

- **New Contributors**
  - New core devs: Spencer Clark and Deepak Cherian
Assuming that ice forms part of the solid matrix, the effective saturation of soils, directions, and the terms on the RHS of equation (15), the terms forms in the larger pore spaces. In this work, we follow the approach adopted by Liquid water flow in partially frozen soils is driven by strong capillary pressure gradients that develop as ice that there is no volume expansion during freezing [2015a], in which same constitutive functions can be used to relate two terms of the Darcy flux are the capillary and gravity fluxes, define rainfall, infiltration-excess runoff and saturation-excess runoff, respectively. Within the soil profile, the water content. For example, the vertical fluxes of liquid water can be parameterized as a Darcy flux, with infiltration into equations (12) and (13) where equation (14) the solid precipitation flux occurs only at the top of the snowpack (snowfall over bare ground and at times when the canopy is completely covered with snow). Note that in

\[ q^\text{soil} = \begin{cases} q_{\text{rain}} - q_{\text{ix}} - q_{\text{sx}} & z=0 \\ -K^\text{soil} \frac{\partial \psi}{\partial z} + K^\text{soil} & z > 0 \end{cases} \]

To accommodate both unsaturated and saturated flow through soils, the fluxes on the RHS of equation (15) the generalized Clapeyron equation is combined with the water retention curve to separate the total water content.
FRAGMENTATION PROBLEMS

1. Software
   - Few tangible incentives to share source code (funding agencies, journals)
   - Lack of extensible development patterns; often it is easier to “home grow” your own solution, rather than using someone else’s.
     - Result is that most geoscientific research is effectively unrepeatable and prone to failure.

2. Data sprawl
   - Inefficiencies of many copies of the same datasets (“dark replicas”)
     - Lessons learned from the CMIP archives (CMIP3 was duplicated > 30x)

3. Local vs. High-performance vs. Cloud Computing
   - Traditional scientific computing workflows are difficult to port from a laptop, to HPC, to the cloud
PANGEO PROJECT GOALS

• Foster collaboration around the open source scientific Python ecosystem for ocean / atmosphere / land / climate science.

• Support the development with domain-specific geoscience packages.

• Improve scalability of these tools to handle petabyte-scale datasets on HPC and cloud platforms.
PANGEO COLLABORATORS

[Logos of various organizations including NSF, EarthCube, NASA, Google Cloud Platform, AWS, Lamont-Doherty Earth Observatory, NCAR, Met Office, USGS, BIDS, Rhodium Group, Johns Hopkins University, ECMWF, and others]
PANGEO ARCHITECTURE

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

Parallel computing system allows users deploy clusters of compute nodes for data processing. Dask tells the nodes what to do.

Cloud / HPC

Distributed storage

Jupyter for interactive access remote systems

Xarray provides data structures and intuitive interface for interacting with datasets

Xarray

Dask

Jupyter

web browser

end user
## Build Your Own Pangeo

<table>
<thead>
<tr>
<th>Storage Formats</th>
<th>HDF</th>
<th>OPeNDAP</th>
<th>Cloud Optimized COG/Zarr/Parquet/etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND-Arrays</td>
<td>NumPy</td>
<td>DASK</td>
<td>More coming…</td>
</tr>
<tr>
<td>Data Models</td>
<td>xarray</td>
<td>Iris</td>
<td></td>
</tr>
<tr>
<td>Processing Mode</td>
<td>Jupyter</td>
<td>Batch</td>
<td>Serverless</td>
</tr>
<tr>
<td>Compute Platform</td>
<td>HPC</td>
<td>Cloud</td>
<td>Local</td>
</tr>
</tbody>
</table>
PANGEO DEPLOYMENTS

HTTP://PANGEO.IO/DEPLOYMENTS.HTML

PANGEO.PYDATA.ORG
BINDER.PANGEO.IO

Over 1000 unique users since March!

NCAR Cheyenne
(Scale using job queue system)

NASA Pleiades

Google Cloud Platform

Microsoft Azure

aws

(Scale using Kubernetes)
What's in a typical Pangeo?

- JupyterHub interface
- Tools to deploy dask clusters
- Customizable software/hardware environment

Current effort to federate pangeo deployments for problem specific uses (e.g. `cds.pangeo.io`?)

Custom deployments:

- `polar.pangeo.io`
- `solar.pangeo.io`
- `ocean.pangeo.io`
- `hydroshare.pangeo.io`

And more coming…
• **BinderHub**
  - Highly customizable Jupyter environment
  - Automates Git repo -> docker image -> Jupyter notebook
  - Automates deployment of Dask clusters
• Easiest way to share Pangeo workflows
PANGEO IN A NUTSHELL

• **Scientific Python ecosystem**
  ‣ flexible, open-source, community driven

• **Interoperable**
  ‣ integrates with existing/developing tools used by science community

• **Analysis ready data formats**
  ‣ cloud optimized data (e.g. zarr)

• **Intuitive self-describing data models**
  ‣ e.g. xarray, Iris

• **Scalable**
  ‣ e.g. Dask, Kubernetes

• **Interactive**
  ‣ e.g. Jupyter, JupyterHub, BinderHub

• **Cross platform**
  ‣ HPC, Cloud, local computing
WHAT’S COMING FOR PANGEO

- Governance (https://github.com/pangeo-data/governance)
- Funding (new projects from NASA and NSF)
- AWS Open Datasets Program and Pangeo compute resources
- Science focus on remote sensing datasets
- Looking for new community partners
HOW TO GET INVOLVED

HTTP://PANGEO.IO

• Access and existing Pangeo deployment on an HPC cluster, or cloud resources (eg. binder.pangeo.io)

• Adapt Pangeo elements to meet your projects needs (data portals, etc.) and give feedback via GitHub: github.com/pangeo-data/pangeo

• Participate in open-source software development!
SHARING DATA IN THE CLOUD

Traditional Approach: A Data Access Portal

Data Granules (netCDF files)

- file.0001.nc
- file.0002.nc
- file.0003.nc
- file.0004.nc

Data Center

Data Access Server

Client

Internet
ON-DEMAND ANALYSIS-READY DATA

- **Too big to move**: assume data is to be used but not copied
- **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
Sharing Data in the Cloud

Direct Access to Cloud Object Storage

Data Granules
(netCDF files or something new)

Cloud Object Storage

chunk.0.0.0
chunk.0.0.1
chunk.0.0.2
chunk.0.0.3

Catalog

Cloud Compute Instances

Client

Client

Client

Cloud Data Center