

What's next for Pangeo?

*Tom Nicholas
(on behalf of the Pangeo community)

 Worthy

GORDON AND BETTY
MOORE
FOUNDATION

Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE



*tom@cworthy.org

@TomNicholas

Pangeo has done lots of cool stuff, but **we could go further** 🚀

This talk:

- An opinionated and deliberately provocative list of things we could do next 🧑🏻
 - Mix of gripes 😡 and cool suggestions 😎
- Rapid-fire to stimulate your brain 🧠
- (Credit goes to others for ideas, but any mistakes mine) 🙏
- Call me out in discussion after if I missed something! 🙌

1) User Experience - What's gone well

- Smooth over differences in file access (i.e. netcdf3/4 / HDF5 / Grib / Tiff etc.)
- High-level but still **powerful abstractions** in UI
- Pretty **seamless scaling** beyond a single CPU

1) User Experience - What next? 🚀

- Zero-to-Pangeo is still daunting
- Chunks are annoying to have to think about
- Reliability improvements at scale
- All code should handle **physical units** for me

2) Infrastructure - What's gone well

- Large numbers of geoscience users moved to the Cloud
- Institutional jupyterhubs on HPC too
- Commercial / non-profit providers want to provide us services that fit well with our goals

2) Infrastructure - What next? 🚀

General Pangeo-backed Jupyter/BinderHubs are going away...

- Stop providing Pangeo-centric cloud infra now idea has been demo'ed, sending everyone to Coiled / 2i2c / their institution?
- “Pangeo”-backed Coiled service?
- Public cloud buckets?

3) Software - What's gone well

- **Modular stack** with open standards and interfaces
- Genuinely **domain-agnostic** core libraries
- Software standards generally fairly high

3) Software - What next? 🚀

- **Fewer cookbooks**, more code features
- Truly **arbitrary scaling**, not just fall over if too too big
 - Different distributed array backends
 - **Serverless** distributed arrays? Across GPUs??
- **Rust-ify** to optimize key parts of the stack
 - e.g. fsspec in Rust, Rust reader for Zarr
- Better **ML integration**, especially dataloaders

3) Software - What next? 🚀

- Geospatial
 - Still no standard way to associate geospatial coordinate information (CRS + coordinates) with Xarray data.
 - GeoXarray was supposed to resolve these problems
 - Is it being developed at all? Why not?
 - Partly for this reason, the GeoZarr standard has stalled 😬

3) Software - What next? 🚀

- Xarray flexible indexes' potential has not been realized
 - Geospatial Indexes as above
 - Interval Index
 - Wraparound Index (e.g. longitude)
 - KDTree Index
- Even more open standards!
- Hypothesis testing everywhere
- Query optimization (i.e. dask-expr for xarray)
- Making movies from data
- 3D visualization

4) Data Management - What's gone well

- Data accessibility through the cloud
- Kerchunk as a universal no-copy interface

4) Data Management - What next? 🚀

- **Version-controlled Zarr** stores (see [Earthmover.io](https://earthmover.io))
- **Cataloging** could be so much better
- Thinking in Trees of data
- ETL pipelines
 - Pangeo-Forge is great, but ambitious and future uncertain
- HPC data transfer
 - Skyplane? ✈️

5) Community - What's gone well

- Impact: ~60 Pangeo-related talks at AGU this year!
- Developing for **real user needs** by blurring devs/users
- Educational resources (i.e. Pythia)
- Public discussions
 - On discourse
 - Recording everything on github

5) Community - What next? 🚀

- More adoption in other fields of science
- Resources on how to be a good participant/dev/maintainer
- Live chat - discord? <https://discord.gg/ex5qqEyyTz>
- Diversity
 - Active outreach to underrepresented communities
 - Support for languages other than English
- Integration across Pangeo entities (P. Europe, P. Forge, Pythia...)



6) Funding and Careers - What's gone well

- Won some grants
- Have NumFOCUS as a fiscal sponsor
- Have a well-used “jobs” section on Discourse

6) Funding and Careers - What next? 🚀

- Funding for maintenance is hard unless you're a really big project
- More paid internships and mentoring for early-career scientists
- Need a pipeline for generating maintainers
- Credit is not proportional to effort
- No viable career path for anyone interested in pushing the boundaries on this stuff...

7) Democratization of Science - What's gone well

- Critical Climate Science datasets made truly public
 - (CMIP6 and ERA5)
- Some great outreach / education projects built upon Pangeo
 - ClimateMatch
 - Ghana oceanography summer school - via 2i2c Hub

7) Democratization of Science - What next? 🚀

- Direct knowledge transfer initiatives to non-western world
- LLM-powered natural-language interfaces?
- HPC in the cloud

8) Scientific Publishing - What's gone well

- Tools to make (Zarr) data actually available, not just “Upon Reasonable Request”
- Some great outreach / education projects built upon Pangeo
 - ClimateMatch
 - Ghana oceanography summer school - via 2i2c Hub

8) Scientific Publishing - What next? 🚀

- Cost models for archiving this data
- Web-based visualization of uploaded datasets
- Automated software / dataset citation network
- More nuanced models of credit
- Aspire to better than Jupyter Notebooks as a publication format

Fin. Discussion time!

- Go forth and comment
 - Google doc: <https://tinyurl.com/4dp8wbpc>
- Thanks to various people for ideas both yesterday and over the past few years

