My ERA5-Odyssey

Rich Signell - USGS

Pangeo Showcase 10/12/2022

Goals for this talk

- Convey that when creating cloud-optimized datasets, challenges are normal
- Show my approach in the hope that you see something you find useful
- Learn something from you that helps in my workflows! :)

But wait, why this talk?

Q: Didn’t Peter Marsh already demonstrate cloud-optimized access to ERA5 using kerchunk?

A: Yes, but ERA5 on AWS and Planetary Computer contain only *some* of the variables.

USGS Hydrologic Modelers want data from ERA5-Land, like snow_depth and soil moisture.

Data from these vars can be accessed only using the ERA5 API.

How can we use our tools (Dask, Xarray, Fsspec, Intake) to make an effective workflow for creating a cloud-optimized dataset?

The Plan

- Construct an API request to get a NetCDF file about 100mb in size
- Issue API requests in parallel using Dask to get a collection of NetCDF files
- Kerchunk the collection into a single virtual cloud-optimized Zarr dataset
- Explore with hvplot and Panel

Demo on ESIP QHUB

Rendered Notebooks at:

- 00_era5_test_api.ipynb
- 01_era5-land-bitinfo.ipynb
- 02_era5_land_api_dask.ipynb
- 03_era5_kerchunk.ipynb
- 04_era5_land_explorer.ipynb

Repo at: https://github.com/rsignell-usgs/pangeo_showcase_20221012

Issues Encountered

- APIs often have limited ability to scale and may have other limits
- Dask Gateway workers need credentials
- Gateway workers don’t see local filesystem
- Can’t use kerchunk to aggregate NetCDF files that have partial chunks
- Can’t use kerchunk on aggregate NetCDF files using scale_factor and add_offset