Pangeo Forge CMIP6 pipeline

links/references:

- New github repo:
  - https://github.com/pangeo-forge/cmip6-pipeline
- Diana's notes on Naomi's procedure:
  - https://paper.dropbox.com/doc/CMIP6-pre-processing-notes--A6deg_32vBRvEOAHvWAyaMCAAg-tqcfV5mr0FtDZ2HCl4OQU
- Naomi's github repo for current collection procedure:
  - https://github.com/naomi-henderson/cmip6collect
  - netcdf (ESGF links) -> xarray (python notebook) -> zarr (Google Cloud)
- Ag's link to some google slides:
  - https://docs.google.com/presentation/d/1sHVkPbIP819JPIwe0m0T4XAhZrfO-kAT7emOBPryNeA/edit?usp=sharing
- CMIP6 API outline:
  - https://paper.dropbox.com/doc/CMIP6-in-the-cloud-API-Outline--A6fDe6qEAQ05K9DA7vb8mI9UAQ-ku7mdtcEVr79Qdiyhz9t6
- Charles' Pangeo catalog repo:
  - https://github.com/pangeo-data/pangeo-datastore

2020-08-27

Attendance:

1. Diana Gergel / Rhodium Group/Climate Impact Lab/ @dgergel
2. Naomi Henderson / LDEO / @naomi-henderson
3. Charles Blackmon-Luca / LDEO / @charlesbluca
4. Kelly McCusker / Rhodium Group / kmccusker@rhg.com
5. Ag Stephens (CEDA) / ag.stephens@stfc.ac.uk

Intros:

- Ag works at CEDA. One of the projects is to build useful tools for the community, run an ESGF node. Bug parallel file system, exploring object store, Pangeo. Test project to load 200Tb of CMIP6 in Caringo object store (https://www.caringo.com/). The ESGF netcdf files are local, so they do not need the ESGF search API, just need to concatenate and store as zarr.
- Charles CMIP6 GCS to AWS, now automating the process. Working on handling

continual updates. Using intake/intake-esm, but now working on stac (especially for online catalog)
- Kelly, Climate data streams - especially downscaling CMIP6 data
- Diana, downscaling CMIP6, need end-to-end open-source, reproducible pipeline and CMIP6 is an important part of this

Agenda:

- Goals (Naomi)
    - Current: refactor into a form usable by others, more general (not just GCS, etc)
    - Future: develop/use pangeo-forge flows/tasks (prefect)
- Meeting business (chaired by Diana)

    Introductions (5-10 min)

    History of CMIP6 Cloud Migration (5-10 mins)

    Current status of workflow (15-20 mins)

        - Overview CMIP6 API Outline

    CEDA/ESGF updates?

    Discussion

- Ag mentioned the memory issue when creating datasets:

    https://github.com/roocs/clisops/issues/27

- Useful discussion of STAC on Gitter:

    https://gitter.im/SpatioTemporal-Asset-Catalog/Lobby