Help us, Zarr. You're our only hope!

Migrating multi-dimensional arrays to the cloud for data-heavy workflows

Christine Smit [1,2], Brianna Pagán [1,3]

[1] NASA, GES DISC, [2] Telophase,[3] Adnet

Who are we?

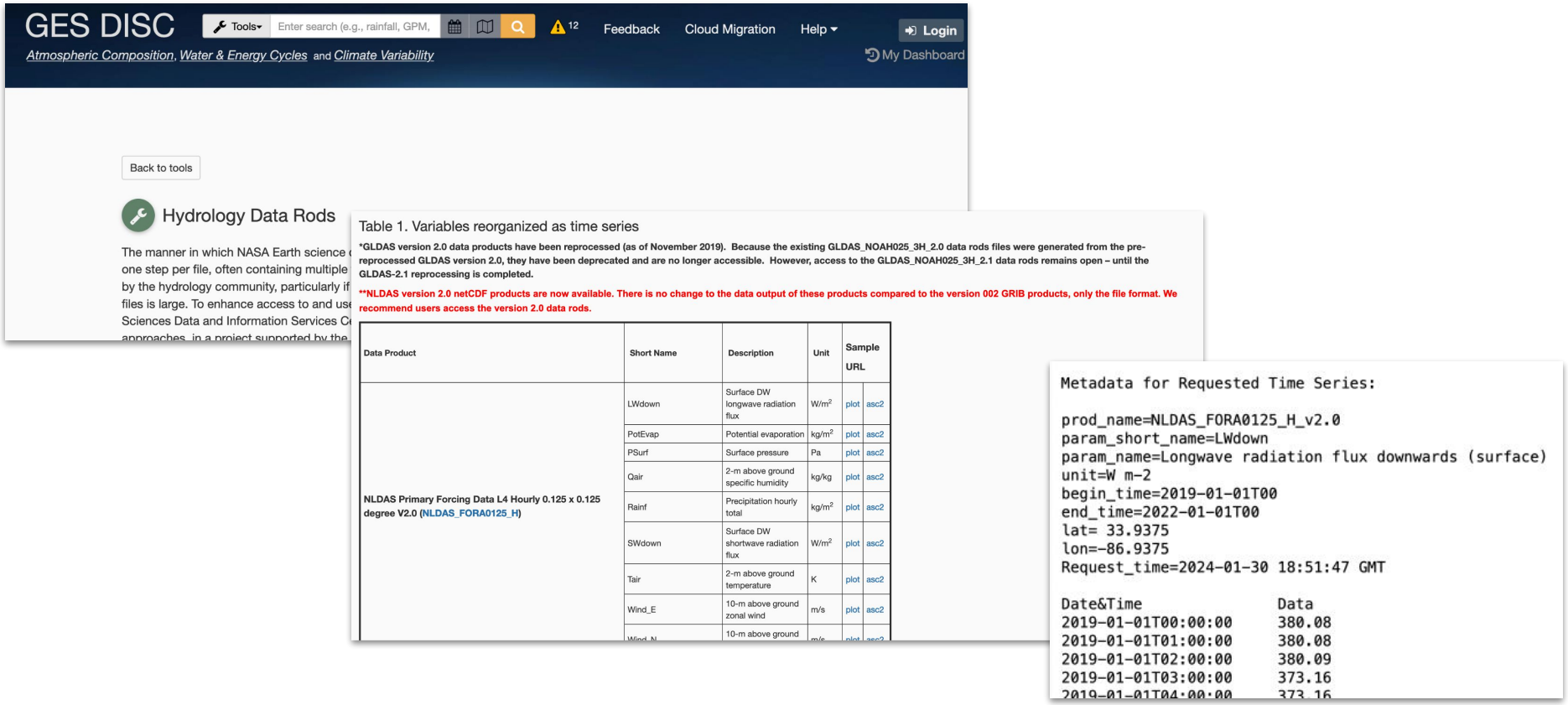
NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC)

We

- archive and curate Earth science data
- create services to make that data easier to use

https://disc.gsfc.nasa.gov/information/documents?title=Who%20We%20Are

The task: migrate a time series service



https://disc.gsfc.nasa.gov/information/tools?title=Hydrology%20Data%20Rods

The task

- Previous implementation:
  → netCDF files stored locally on a single server's disk and chunked for reasonably fast access along any dimension, including the time dimension

- New implementation:
  → zarr stores stored in s3, chunked for reasonably fast access along any dimension, including the time dimension

Example

1 year of precipitation data over NASA Goddard in Greenbelt, MD:

https://api.giovanni.earthdata.nasa.gov/proxy-timeseries?data=GPM_3IMERGHH_06_precipitationCal&location=%5B38.995%2C-76.885%5D&time=2020-01-01T00%3A00%3A00%2F2020-12-31T23%3A59%3A59

Behind the scenes, this zarr store is:

- Global, 0.1 degree
- 23 years, half-hourly

Work in progress

- Making zarr stores in data cache public