



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

GES DISC Tools Enter search (e.g., rainfall, GPM) Feedback Cloud Migration Help Login My Dashboard

Atmospheric Composition, Water & Energy Cycles and Climate Variability

Back to tools

Hydrology Data Rods

The manner in which NASA Earth science data is organized is one step per file, often containing multiple files. The hydrology community, particularly in the United States, has a large file size. To enhance access to and use of these data, the Earth Sciences Data and Information Services Center is developing new approaches in a project supported by the

Table 1. Variables reorganized as time series

*GLDAS version 2.0 data products have been reprocessed (as of November 2019). Because the existing GLDAS_NOAH025_3H_2.0 data rods files were generated from the pre-reprocessed GLDAS version 2.0, they have been deprecated and are no longer accessible. However, access to the GLDAS_NOAH025_3H_2.1 data rods remains open - until the GLDAS-2.1 reprocessing is completed.

**NLDAS version 2.0 netCDF products are now available. There is no change to the data output of these products compared to the version 002 GRIB products, only the file format. We recommend users access the version 2.0 data rods.

Data Product	Short Name	Description	Unit	plot	asc2
NLDAS Primary Forcing Data L4 Hourly 0.125 x 0.125 degree V2.0 (NLDAS_FORA0125_H)	LWdown	Surface DW longwave radiation flux	W/m ²	plot	asc2
	PotEvap	Potential evaporation	kg/m ²	plot	asc2
	PSurf	Surface pressure	Pa	plot	asc2
	Qair	2-m above ground specific humidity	kg/kg	plot	asc2
	Rainf	Precipitation hourly total	kg/m ²	plot	asc2
	SWdown	Surface DW shortwave radiation flux	W/m ²	plot	asc2
	Tair	2-m above ground temperature	K	plot	asc2
	Wind_E	10-m above ground zonal wind	m/s	plot	asc2
	Wind_N	10-m above ground meridional wind	m/s	plot	asc2

```

Metadata for Requested Time Series:

prod_name=NLDAS_FORA0125_H_v2.0
param_short_name=LWdown
param_name=Longwave radiation flux downwards (surface)
unit=W m-2
begin_time=2019-01-01T00
end_time=2022-01-01T00
lat= 33.9375
lon=-86.9375
Request_time=2024-01-30 18:51:47 GMT

Date&Time          Data
2019-01-01T00:00:00  380.08
2019-01-01T01:00:00  380.08
2019-01-01T02:00:00  380.09
2019-01-01T03:00:00  373.16
2019-01-01T04:00:00  373.16

```

<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