Efficient Access to MERRA-2 Reanalysis Data from the Cloud Archives

Poster #: IN31E-2046
AGU 2024Fall Meeting
Washington, D.C.
December 9-13, 2024
iPoster QR code

NASA GODDARD EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

Suhung Shen[1,2] Christopher M. Battisto[1,3], Jennifer M. Adams[1,3], Brianna R. Pagan[1,3], Lena F Iredell[1,3] and Mahabal Hegde[1]

suhung.shen@nasa.gov (1)NASA GES DISC (2)George Mason University (3)ADNET Sys. Inc.

Abstract

The Modern-Era Retrospective analysis for Research and Applications, Version (MERRA-2) dataset is now available in the AWS West Region-2 cloud. The free data downloading policy remains unchanged. The NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) offers cloud services such as OPeNDP subsetting, and time series services. This presentation will showcase use cases that testing the procedures and find the most efficiency workflows to access and process the data for finding climate extremes. The use cases include a) accessing and processing data directly in the cloud S3 bucket without downloading it, and b) downloading subset data to process on the user's local system.

Procedures to Find Extremes

There are several methods to calculate the daily climatology data. In this study, we use the moving average method for a defined window. As the "window" moves forward, the average is recalculated by dropping the oldest value and adding the next one.

Data Collection: MERRA-2 tavg1_2d_slv_Nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Single-Level Diagnostics (M2T1NXSLV_5.12.4)

Variable: temperature at 2-meter (T2M)

Base period: 1991-2020 (30 years)

Window size: 5

Daily statistics: daily mean, min, and max

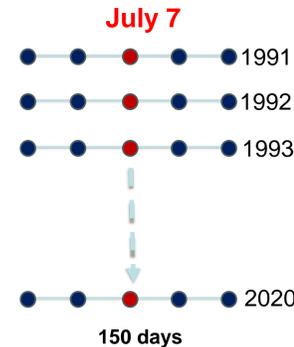**Climatology types:** mean, percentiles

Figure 1: Illustration of days used to calculate one day of the daily climatology.

Figure 2: Flowchart to calculate the daily climate, anomalies, and extremes.

Study Case: Extreme Hot Weather on July 7, 2024 over Western US

In early July 2024, large regions of California and some Nevada had experienced heatwaves. According to Los Angles Times, the Furnace Creek Visitors Center (N 36°27.70, W 116°52.00) in Death Valley National Park has reached 55 C (132 F) on July 7, 2024. It is about 10 -12 C above normal for the hottest part of the year. In the same day, in Grand Canyon National Park, as temperatures rose into the mid-90s F, a hiker was found dead, according to the National Park Service. The high temperatures were also blamed for a massive fish die-off in Lake Elizabeth in Northern California on July 5 2024, reported by according to Fremont city officials.

The NASA MERRA-2 reanalysis datasets provide over 40 years of hourly data. The extreme hot weather over the western US was revealed in the MERRA-2 data.
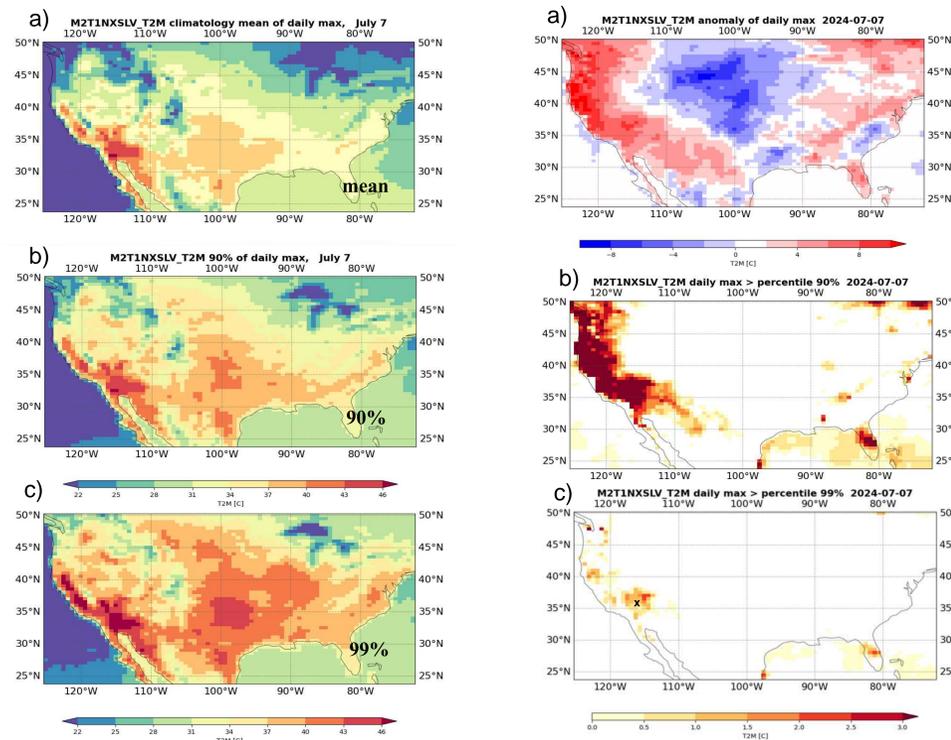

Figure 3: Daily climatology of MERRA-2 daily maximum temperature at 2-meters on July 7 over US: a) climatology mean, b) 90 percentiles, and c) 99 percentiles

Figure 4: Anomalous daily maximum temperature at 2-meters over US for July 7, 2024: a) difference from the climatology mean; b) areas with temperatures larger than 90% ; c) areas with temperatures larger than 99%
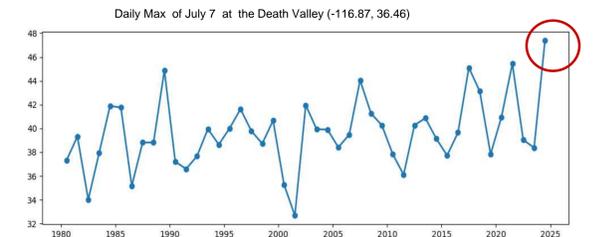

Figure 5: Time series of the daily max temperature on July 7 from 1980 to 2024 at the Death Valley, US (116.87 W, 36.46N, near x mark in Fig4c), indicating that 2024 has the highest temperature in the model history.
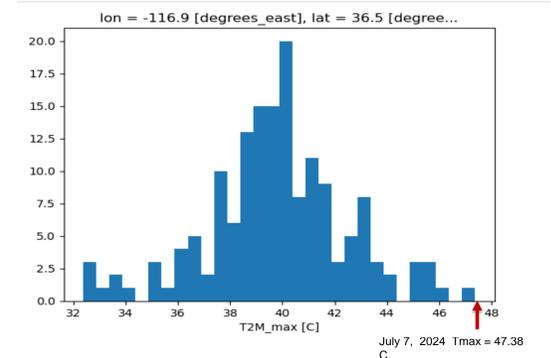

Figure 6: Probability distribution of the daily max temperature on July 7 (with a 5-day window) from 1991 to 2020 at the same location as in Fig. The arrow indicate the extreme high temperature on July 7, 2024.

Examples to Search and Access Data with using Earthaccess

Direct access data in S3

```
import earthaccess
import xarray as xr
prod = 'M2T1NXSLV'
version = '5.12.4'
var = 'T2M'
bdate = 1991-07-05
edate = 1991-07-09

granule = earthaccess.granule_query().short_name(prod).
          temporal(bdate, edate).get(5)

if len(granule)>0:
    ##open granules in S3
    ds = xr.open_mfdataset(earthaccess.open(granule), chunks={})
    x = ds[var]
```

Find granules in cloud OPeNDAP

```
import earthaccess
import xarray as xr
prod = 'M2T1NXSLV'
version = '5.12.4'
var = 'T2M'
bdate = 1991-07-05
edate = 1991-07-09
files = []
granule = earthaccess.granule_query().short_name(prod).
          temporal(bdate, edate).get(5)

if len(granule)>0:
    # Parse out OPENDAP links
    for result in granule:
        for url in result['umm']['RelatedUrls']:
            if 'OPENDAP' in url['Description'].upper():
                files.append(url['URL'])

... parallel downloading subsetted data from OPENDAP

##open local files
ds = xr.open_mfdataset(local_files)
```

Figure 7: Wall clock time (in minutes) required to compute one day of daily climate data using the procedures outlined in the flowchart (Fig. 2). In this example, downloading subset data in parallel to a Windows PC with 4 workers performs comparably to processing data in the cloud. Increasing the number of parallel download workers could further reduce the time. Accessing data remotely via OPeNDAP is significantly slower and is not recommended for handling large datasets.
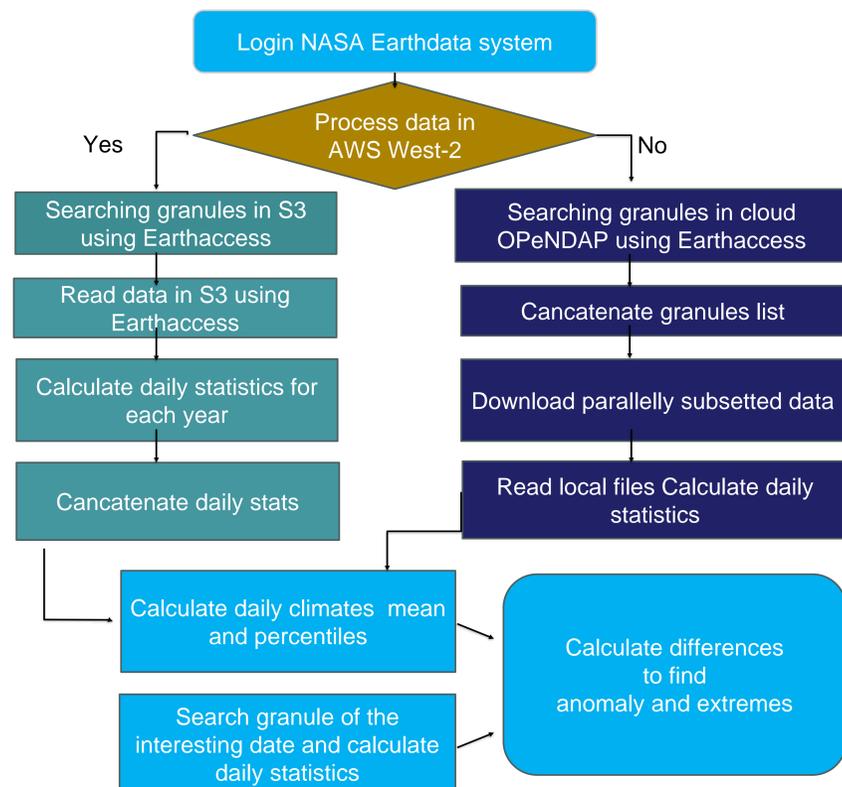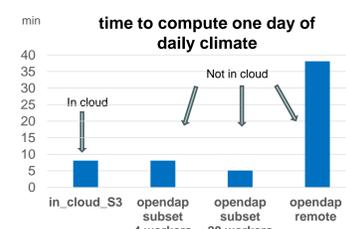

Summary and Resources:

- The cloud OPeNDAP service enables users to download only the interesting data, which helps to save user's storage and the cloud egress cost.
- The python package Earthaccess is powerful and simplified the code for finding data from the NASA Earth data system
- The cloud computing enables the parallel downloading that improves significantly the data downloading and processing speed
- The calculation of climate daily is time consuming. Pre-generating and archive daily climate for popular variables may benefit users to study climate extremes

Acronyms:
GES DISC: Goddard Earth Sciences (GES) Data and Information Services Center (DISC)
MERRA-2: Modern-Era Retrospective analysis for Research and Applications, Version 2