Comparison of Satellite Reanalysis Data and Weather Station Data

Merra-2

Merra 2 is a data set provided by NASA that describes weather conditions globally since 1980.

File format

The file structure of the data for Merra-2 is described here. To understand this document you first need to understand how the data for Merra-2 is hosted. The data for Merra-2 is hosted on several disks in some facility operated by NASA. When you fetch data from the database, you have to specify which database you are wanting the data from. The document linked above lists the naming specifications for each database in section 5.2 on page 12. For example `M2T1NXRAD` is the name of the database for radiation data. The document also lists all the attributes that are stored in each database. Note that attribute IDs are *not* unique across databases. Make sure that you are downloading from the correct database. There are multiple databases for different time calculations, eg. 1 hour time frame, 3 hour time frame ect.

Downloading Data

To download the data you will need to register to get an account with GEO DISC. It is free and fast.

There are a few ways of downloading data. The Merra-2 databases hold the data for the entire world. So to download just one date of data for one database can be an extremely large amount of data. It can be as high as hundreds of MBs per day. This makes downloading data without restrictions completely impractical. You can download directly from the Goddard Earth Sciences Data and Information Services Center (GEO DISC) by searching for the database that you want. However, this does not allow for downloading subsets directly. Be wary of the size of data that you are downloading.

GEO DISC has a web api service that allows you to use http requests to fetch data from them. The fastest way to download the data is to go onto GEO DISC, search for the database you are wanting to use, select "subset/get data", then choose "Get File Subsets using the GES DISC Subsetter". Then input the dates, variables, bounding box of area that covers the data you are needing, and other prompted information. Then click "get data". Then it will generate a text file filled with http request links. To test a link you can paste a link into any browser, and you will download the corresponding netcat file.

To download the data from the compiled list of links, you can follow directions from here. Personally I used curl to download the links by following this tutorial.

This process is much easier to do on a linux machine, or on a Windows machine with WSL installed. Note that WSL is installed on some government machines, but requires admin access to use. I personally used an installed version of WSL on my personal machine.

This process will likely download one netcat (.nc) file per day. You will have to run the program once for every database that you query.

There is a GitHub repo with a merra-data scraper online but this tool has limited usefulness. For some reason when using the tool it can run into issues where a http request will get an 503 (internal server issue) error often. This left substantial gaps in the data. It also took exponentially longer to use. I would recommend using the official documented ways of downloading the data.

Missing Data

I have had problems with all methods of downloading data. I would verify that all the days have been downloaded, check that the file size is appropriate (> 1KB) and other precautions. When using the webscraper I ran into several 503 http responses, and when using curl there were large chunks of data (months) that were not properly downloaded.

Merra-2 Data Downloaded

All attributes are downloaded for the years 2018 to 2022, inclusive

Database	Attribute Short Name	Attribute Long Name	Units
M2T1NXSLV	T2M	Temperature 2 meters	Kelvin
M2T1NXSLV	V10M and U10M	Vectorized wind components	m/s
M2T1NXSLV	QV2M	2 meter specific humidity	%
M2T1NXLND	TSOIL1, TSOIL2, TSOIL3, and TSOIL4	Soil temperature, level 1-4	Kelvin
M2T1NXLND	PRECTOTLAND	Total precipitation land; bias corrected	Kg/(m^2 s)
M2T1NXLND	GWETPROF	Ave prof soil moisture	%
M2T1NXLND	GWETROOT	Root zone soil wetness	%
M2T1NXLND	GWETTOP	Surface soil wetness	%
M2T1NXRAD	SWGDN	Incident shortwave land	W/m^2

Note that for TSOIL1-4, GWETPROF, GWETROOT, and GWETTOP there is a static data set for each location that describes the depth of the attributes there. This can be found in this document if you search for the corresponding attribute.

ERA-5 Land

ERA-5 Land is a reanalysis of weather data ran by Copernicus, which is based in the EU.

File Format

The ERA-5 Land data can be downloaded as a grib file or as a zipped netcat file. There are limitations to the size of a single piece of data that you can retrieve from their servers, but you are able to customize almost everything. There is only one place that you have to query.

The official documentation can be found here.

Downloading the Data

official documentation

Notice that technical development of the ERA-5 services are all done on linux. This makes it so that a lot of the tools used to download the tools will only work on a linux machine. This includes some python packages that they use.

To download the data you have to first get an account. It is an easy process. You can then go to this link to download the data. Fill in the parameters that you want and then at the end click `Show API request` . This is the pythonic way of downloading the data. There is a python script here that can get the data from the request through their official downloading method called ERA5_webscraper.py. That program will send a request to their servers which will prepare it. The preparation process can take some time. After the program will download it. The program is designed to download 1 year of data for 1 attribute at a time. Program is taken from this page

To run the program there is some preliminary steps that you have to take described on the webpage and in the header comments in the file.

This download process can take a while to complete. (~2 hours per 5 years per attribute)

Era5-Land Downloaded Data

Attribute Name	Short Name	Units
2m temperature	T2M	Kelvin
Wind speed vectorized	v10m, u10m	m/s
soil temperature level 1-3	st1, st2, st3	Kelvin
surface solar radiation downwards	rad	J/m^2
total precipitation	totprec	m

Attribute Name	Short Name	Units
volumetric soil water layer 1-3	sw1, sw2, sw3	%

Note the temperature levels and layers are a predetermined depth range. 0-7 cm for level 1, 7-28 cm for level 2, 28-100 cm for level 3.

Documentation for attributes

CaSPAr

CaSPAr is the Canadian Surface Prediction Archive, which is made in conjunction with several Canadian universities and Environment Canada.

Downloading the Data

To download the data you can follow the instructions here. You will have to create at least a couple accounts, one for CaSPAr itself and one for Globus. Once you send your request you will recieve an email with a link to the data in the Globus file sharing system. To download this data directly to your computer you will need to install a version of Globus Connect Personal. It is offered for Windows, Linux, and macOS. Download link. Once the software is downloaded, and you have given access to one of your drives or folders in your drive you go back to the browser Globus file share and once you have both file locations set up you can press `start` to start the file transfer process.

This is something that the executable is unable to be downloaded onto government computers due to firewalls and compartmentalization computer. If you are able to run an executable in a drive with permissions then it is able to operate as per its documentation.

Datasets

We are able to access several reanalysis products through casper. One of which we will do our comparison with is the HRDPS, or the High Resolution Deterministic Prediction System, which provides forecasts for Canada. Direct, unfiltered data can be accessed here. These files are quite large since it covers all attributes and all of Canada. I would recommend the subset generator tool that can be found here.

HRDPS readme

Downloaded HRDPS Data

Note these were chosen somewhat arbitrarily to test the data download process.

Attribute Short Name	Attribute Description	units
HRDPS_P_FB_SFC	Downward solar flux	

Attribute Short Name	Attribute Description	units
HRDPS_P_LA_SFC	Latitude	Decimal Degrees
HRDPS_P_LO_SFC	Longitude	Decimal Degrees
HRDPS_P_PR_SFC	Forecast: Quantity of precipitation	
HRDPS_P_TT_10000	Forecast: Air temperature	
HRDPS_P_UU_10000	Forecast: U component of wind	
HRDPS_P_VV_10000	Forecast: V component of wind	

Station Data

There are several ways to access our own weather station data to use for comparison against Merra and Era datasets. One of the easiest ways to do through is through the use of the historical data links found here. This link gets semi-regularly updated from the mySQL database which holds all the data in a queryable format. These links provide csv formatted data.

To download the data from a query to use for analysis, it is best to write a mySQL query that directly places the output into a csv. This can be by following this syntax:

```
SELECT * FROM data_60
INTO OUTFILE 'output-data.csv'
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
```

The directory that will output the file is `C:\ProgramData\MySQL\MySQL Server 8.0\Data\historical` on the remote server. You are able to add directories in the `\historical` directory to clean up that location.

You can find more information on the weather station attributes and units in the file `Data-Definition-Table.xlsx` .

Station Data Used

Attribute Name	Units
AvgAir_T	Celsius
AvgWS	m/s
RH	%
Soil_TP5_TempC - Soil_TP100_TempC	Celsius

Attribute Name	Units
SolarRad	MJ/m^2
TBRG_Rain or Pluvio_Rain	mm
Soil_TP5_VMC - Soil_TP100_VMC	%

File Format of Compiled Data

It is good to notice that while the human-readable format exists, it is not the most efficient in terms of space and speed. If you are needing to do large amounts of processing work on the data, and you are running into physical limitations, then it may be required to instead use the raw data provided. The xarray python library is able to do operations on netCDF files extremely efficiently. The downside to using the raw data is that it is fragmented, by day for Merra-2, and by attribute and year for ERA5 Land.

The data is compiled into a file with the columns StnID, StationName, Latitude, Longitude, Elevation, StationData, MerraData, EraLandData. There will be 1 row for every hour, for every weather station, for 5 years. The years used were from 2018 to 2022 inclusive. It is important to note that since both the Merra-2 and ERA5-Land data sets are based on a grid, the latitude and longitude coordinates of the weather station is not exactly the center of the grid. This also means that there can be several mappings to one grid if weather stations are sufficiently close together. There will be a separate csv for every attribute that the analysis is done on.