**How to Download and Tabulate**

**the CPC Global Temperature Dataset**

February 17, 2020

**Data Overview**

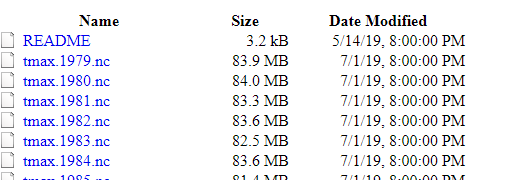
The CPC global temperature dataset provides gridded daily data for TMIN and TMAX, with 0.5 degree grid spacing. TMIN is the daily minimum temperature and TMAX is the daily maximum temperature. The data provides global geographic coverage but excludes areas over oceans. Data is provided from 1979 to the present. There is also a parallel dataset for precipitation that I have not yet examined.

**Downloading the Data**

The CPC temperature dataset can be downloaded from this URL:

[ftp.cdc.noaa.gov/Datasets/cpc\_global\_temp/](ftp://ftp.cdc.noaa.gov/Datasets/cpc_global_temp/)

At this location, you will find a separate TMIN and TMAX file for each year from 1979 through 2020:



It is cumbersome to download these files one at a time. I developed a more efficient approach that involves using Window PowerShell, as follows:

1. Edit the “download\_cpc\_data.ps1” batch file that I have provided such that it points to a local directory on your computer.
2. Place the batch file in the local directory where you wish to store the CPC data.
3. Open Windows PowerShell from the Windows command line and navigate to your designated local directory.
4. Run the following PowerShell command:

**powershell –ExecutionPolicy ByPass –File download\_cpc\_data.ps1**

Step 4 will download the files one at a time. This process may take 10 to 20 minutes.

**Tabulating the Data**

The data are in the form of NetCDF files. This is a common format for weather data. There are various ways to convert the files to CSV format, but the resulting files are quite large. In NetCDF format, each annual CPC file is about 80 megs. Converted to CSV, each file will be about 350 megs. Since we have 22 years of data, and a separate set of files for TMIN and TMAX, the dataset is over 15 gigs in CSV format. This isn’t huge compared to datasets such as ERA5 which provides hourly data (whereas CPC provides daily data), but it is large enough to be cumbersome.

As an illustration on how one might run tabulations on the CPC daily data, I’ve included an R program that can read the data directly in NetCDF form, without first converting it to CSV. The R program loops through the data, one year at a time, computing monthly TMIN and TMAX averages for each latitude/longitude pairs. After looping through all years, it outputs a CSV file that is about 180 megs. So the tabulations convert daily data into monthly averages.

Of course, there are many other metrics that we are interested in other than simply a monthly average, but my R code is intended just as an example of how one can tabulate a large weather database.

I am not an R expert, so please keep that in mind while reading my code. There may be better ways to “skin the cat” than the approach that I have used.

I will share many code snippets over the coming months in the hope that they might be useful to you, but please don’t feel like you are under an obligation to study these snippets. That is, this isn’t “homework”. But for those of you without R experience, or without experience in a different language that can be applied to large dataset, perhaps my code will be of some help.

In addition to R, I also use VBA and C++ to tabulate weather datasets. While I like C++ because it is fast, I think it probably makes sense for me to focus on R and VBA for our work together because the learning curve for C++ is a bit steep, and because speed will generally not be critical to our work.

VBA is built into Excel -- so if you have Excel, you already have VBA.

R can be downloaded for free. However, you may need your IT department to enter an admin password in order to install it.

I have learned R by using YouTube tutorials as well as by studying code written by other people. When I run into trouble coding R, and can’t figure out how to do something, I have found “www.Stackoverflow.com” to be a great resource. Stackoverflow is a website where programmers share problems and solutions with each other. Nine times out of ten, a particular coding question already has a solution presented on Stackoverflow.

**Running my R Code**

My R code for tabulating the CPC temperature data consists of two files: (1) cpc\_tabulate.r and (2) cpc\_tabulate\_support.r. Place both of these files in the same subdirectory. Next, in cpc\_tabulate.r, edit the following lines so that they point to your local subdirectories rather than mine:

Location\_R\_Code <- "D:/JPW/data\_CPC\_Global\_Temp" # location of your R program

Location\_Data <- "D:/JPW/data\_CPC\_Global\_Temp/data\_raw" # location of CPC netCDF data files

Then highlight the code in cpc\_tabulate.r and press RUN. The code will cycle through the CPC data, processing one year of data at a time. It will output a compact dataset of monthly means.