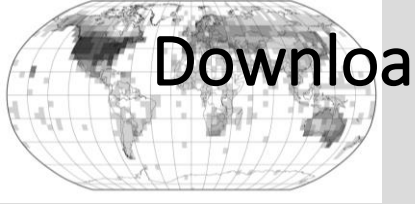


Downloading data from the Global Historical Climate Network (GHCN)-Daily in tidy format using bash

Luis F. Duque, Newcastle University

1. Motivation

The GHCN-daily is a global dataset of land surface stations developed by NOAA. This dataset contains records from over ~100,000 stations in 180 countries, with record lengths ranging from less than a year to more than 175 years. The daily environmental variates available are maximum and minimum temperature, total daily precipitation, snowfall, and snow depth. These records are stored on the NOAA's website and are updated continually. However, the GHCN-daily is messy, and the user has to spend considerable time downloading and structuring the data according to his/her requirements prior to analysis. Therefore, there is a need to automate this prepossessing task. Furthermore, the output of this task should be data structured in a standardized way (e.g., tidy format), so the user can use it in the software of his/her convenience (e.g., bash, R, python, etc.).

2. Original record

- The original record can be downloaded from the following link: https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/
- The records can be obtained: i) by year for all stations (GZ-compressed files), or ii) for all stations for the whole record length (GZIP-compressed tar file)
- Details on the files available and a description of the data format can be checked in the following document https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt

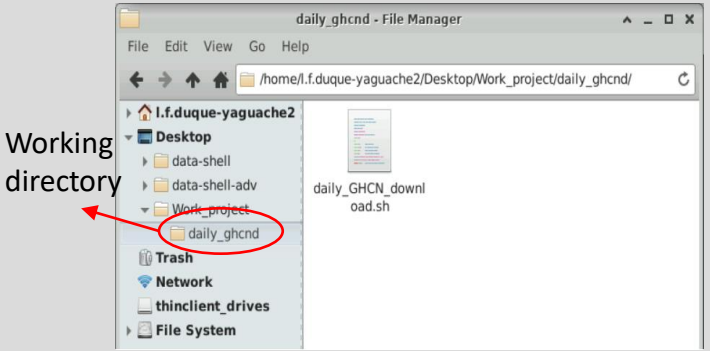
3. The bash script

- I built a script in bash to download records from the GHCN-daily in a tidy format. In this format, each variable must have its own column and each observation its own row. The bash script is available in: https://github.com/lfduquey/GHCN_daily_download/blob/main/GHCN_daily_download.sh
- To run the script, one must install csvkit in the shell.
- The script's inputs are : 1) the initial and final years of the record length, 2) the acronyms of the daily environmental variables, and 3) the code of the stations.

```
l.f.duque-yaguache2@ip-172-31-11-1:~/Desktop/Work_project/daily_ghcnd$ bash daily_GHCN_download.sh 2018 2020
Enter Variables separated by 'space' : TMIN TMAX PRCP
Enter station/s code separated by 'space' : AE000041196 AEM00041218
```

- The acronyms of the daily variables and codes of the stations can be checked in the files "readme.txt" and "ghcnd-stations.txt", respectively, available in https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/
- The outputs are stored in a folder called "Outputs". This folder is created in the directory of the bash script and contains subfolders with the name of the acronyms of the environmental variable downloaded. The tidy data for each station is located within these subfolders and is stored as csv file.

Before running the script



After running the script