GHCN-Daily: a treasure trove of climate data awaiting discovery

Jasmine B.D. Jaffrés[1,2]
[1]C&R Consulting, Townsville, Australia
[2]College of Science and Engineering, James Cook University, Townsville, Australia

Last revised: 21 October 2018

Table of Contents

1 Introduction

The ghcnd_access toolbox is utilised to extract data from the Global Historical Climatology Network (GHCN)-Daily database (Menne et al., 2012) and change the .dly file format into a more accessible structure. The toolbox can be run in either MATLAB or open source alternative GNU Octave.

To acknowledge the use of this toolbox, please cite: *Jaffrés, J.B.D. (2019) GHCN-Daily: a treasure trove of climate data awaiting discovery. Computers & Geosciences 122, 35-44*.

The complete GHCN-Daily database can be downloaded from ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/all/. The latest version of the GHCN-Daily inventory file is available from ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-inventory.txt, while ghcnd-stations.txt is accessed via ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt.

1.1 Software compatibility

The script has been written and tested in MATLAB (R2017b) and GNU Octave (v4.2.0; Eaton et al., 2015).

2 Data requirements

2.1 ghcnd_access toolbox source

The functions and data samples are included in the ghcnd_access .zip file. Download the file from GitHub (https://github.com/jjaffres/ghcnd_access) or SourceForge (https://sourceforge.net/projects/ghcnd-access) and unzip the package in your preferred location.

2.2 GHCN-Daily data download

The latest available GHCN-Daily weather station data and associated files can be obtained from the NOAA ftp site ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily. The ghcnd-stations.txt and ghcnd-inventory.txt files are also required, as they contain relevant information on station location and activity. These files should re-downloaded whenever new GHCN-Daily data is accessed from the website to ensure that the station list and inventory is complete. The direct data links are listed below:

all GHCN-Daily data: ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd_all.tar.gz

individual .dly files: ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/all/

ghcnd-inventory.txt: ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-inventory.txt

ghcnd-stations.txt: ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt/

When the complete GHCN-Daily database is downloaded, the individual .dly weather station files are contained in a zipped tarfile (ghcnd_all.tar.gz). Extract these files in your chosen input directory prior to running the ghcnd_access toolbox. Ensure that the inventory and

station text files are located in the same directory (or alter the ghcnd_access.m script accordingly).

Note: The ghcnd_access toolbox is capable to extract data from the entire database (in excess of 100'000 files). However, this will take a few hours, especially when precipitation is selected (the most commonly observed variable).

2.3 List of files

The directory and files structure of the ghcnd_access toolbox are described below.

ghcnd_access/	Base directory containing main script, user's guide, readme.txt and all subdirectories.
ghcnd_access/data/	Directory for the GHCN-Daily sample data, downloaded from ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/all/. The folder also contains older versions of the GHCN-Daily inventory and station files (download the latest versions from ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/).
ghcnd_access/output/	Default output location, including extracted GHCN-Daily data and figures.
ghcnd_access/postproc/	Directory containing the dailypostproc.m post-processing subroutine.
ghcnd_access/subs/	Directory containing all subroutines.

3 Running and modifying the script

3.1 User modifications

The ghcnd_access.m script allows for several user modifications. Some code should be modified according to user preferences before the script is invoked. Additional selections are undertaken while the script is running via the input() function. Note that an up-to-date ghcnd-inventory.txt file (modification date no later than the utilised GHCN-Daily data), along with the ghcnd-stations.txt file, should be located in your input folder.

3.1.1 Pre-run modifications

in_dir	Directory of GHCN-Daily data (line 30). Data can be obtained from ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/all/
out_dir	Directory of output (line 31).
unit_which	Conversion of data units (line 34). By default (unit_which = 'SI'), all units are converted to SI units (°C, mm, m/s). For original units, change 'SI' to any other text (e.g. unit_which = '').
dly_which	.dly file names to be considered (line 37). By default, any .dly files (*.dly) in the in_dir directory will be included. The list can

	be reduced to a country-specific import by modifying the line (e.g. 'AS*.dly' for Australian data).
var_options	List of GHCN-Daily variable (line 41). The default list is limited to precipitation (PRCP), maximum temperature (TMAX) and minimum temperature (TMIN). A fourth position is left empty ('). Specify the fourth or replace a variable using the GHCN-Daily notation (cf. the readme.txt file for GHCN-Daily data: ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt/).
qc_apply	Decide whether to keep or remove quality-flagged data (line 47). By default (qc_apply = 1), all quality-flagged data are removed. If qc_apply is not equal to 1, a cell matrix of quality flags will be compiled. Note that qc_apply is only invoked for daily data extraction. For monthly data, all quality-flagged data will be removed.

3.1.2 Input() modifications

The code for the subsequent inputs is located in the ghcnd_vars.m subroutine:

var_target	Selection of target variable (line 19). By default, the choice is between three main variables - precipitation (PRCP), maximum temperature (TMAX) and minimum temperature (TMIN) – plus a 4th user-defined variable (cf. var_options, line 38 of ghcnd_access.m).
yr_options	This parameter allows the selection of a subset of data based on the first and last year of interest (line 45). Options include 1) the full temporal range, 2) a reduced range or 3) one year only. If option 2 or 3 is selected, further input will be requested for the target years (yr1 and y2, lines 135, 136,139).
yr1	Selection of the first or only year of interest (lines 135,139). Only called if a reduced temporal range or individual year was selected for yr_options.
yr2	Selection of the final year of interest (line 136). Only called if a reduced temporal range was selected for yr_options.
mode_options	This parameter allows the choice of either keeping the data in the daily format or converting the daily data into a monthly value (lines 54,65,67,72). The choice of monthly value (total or average) is variable-dependent. For precipitation (PRCP), options are daily values or monthly totals. For temperature variables (e.g. TMAX), the selection is between daily values and monthly average. Daily total sunshine (TSUN) can also be

provided as total or average monthly values, while wind direction variables can only be extracted in daily format.

3.1.3 Error response - ghcnd-stations.txt

The ghcnd_access.m script was written in the assumption that both ghcnd-stations.txt and ghcnd-inventory.txt have a fixed text width that remains constant over time. That assumption is not always valid. For instance, several stations with prefix FRE* had an additional five columns when the ghcnd-stations.txt file was downloaded in March 2018. Consequently, an error was produced when restructuring the data using the *fread* and *repmat* functions:

Fix the inconsistency in the ghcnd-stations.txt file (excess columns)!

To successfully run the ghcnd_access.m script, these excess columns need to be removed. The following freeware is recommended: Notepad++ (https://notepad-plus-plus.org/) by Don Ho. The following steps are suggested for the deletion of the excess columns within Notepad++:

1) Open ghcnd-stations.txt, then press Alt and select column 86 in row 1;
2) Press Alt + Shift and select a column number sufficiently to the right (e.g. column 100) in the last row. Wait a few seconds to allow the program to select the large amount of area to be removed;
3) Right-click + select Delete;
4) Save file (use original name and folder location).

Once the ghcnd-stations.txt formatting has been corrected, ghcnd_access.m can be re-run (cf. Section 3.1.2).

3.2 MATLAB vs GNU Octave

GNU Octave requires the loading of two packages (financial and nan). The ghcnd_access toolbox automatically loads these packages when the script is run on GNU Octave. MATLAB requires the Statistics and Machine Learning Toolbox™.

3.3 Script running

The following message should appear once the optional inputs have been set and processed:

User input is now complete and your selected GHCN-Daily data - VARS - will now be collated.

where VARS refers to the chosen variable (e.g. TMIN).

The time required to complete the data extraction will depend on the number of weather station (.dly) files from which data has to be obtained and on the computer hardware. Once the above message is shown, the run time ranges from less than a second to several hours (2.5-4 hours for all available precipitation – PRCP - data).

4 Data compilation

4.1 Missing and quality-flagged data

NaN (not a number) values were assigned to any day (per weather station) for which no data was registered. NaN values were also allocated to any values that were flagged during the internal GHCN-Daily database quality checks. These NaN values are ignored for the calculation of the monthly values (nanmean and nansum functions). For monthly output, the ghcnd_access toolbox creates a sparse matrix (count_mth_obs) to count the number of days per month and weather station with data.

Note: While the GHCN-Daily database includes internal quality checks, these are not as extensive as quality checks performed by some of the source agencies (e.g. Australia's Bureau of Meteorology). As the original quality flags from the source agencies are not included in the database, it is likely that some of the extracted data will be of poor quality. Unexpected values should therefore be treated with caution.

4.2 Data units

Several variables in the GHCN-Daily database are provided in non-standard units (e.g. tenths of mm, tenths of °C, tenths of m/s), including precipitation and temperature values. By default, the toolkit converts these variables into traditional units (mm for precipitation, °C for temperature, hours for total sunshine, etc.). Alternatively, the original GHCN-Daily units can be kept (cf. unit_which, line 34 in ghcnd_access.m). Data unit information is provided within the output file (cf. data_unit).

5 Output

Once the ghcnd_access toolbox has been successfully run with at least one .dly file, a MAT-file (.mat) will be generated. The output file content and name will depend on 1) whether daily or monthly output was requested and 2) the selected target variable (e.g. precipitation or maximum temperature). The file name is structured according to the date source (GHCND), period (_day or _mth) and variable (e.g. _PRCP).

5.1 Daily data

The file name of the generated MAT-file will vary depending on the chosen variable. The file name is structured according to the date source (GHCND), daily value (_day) and chosen variable (e.g. _PRCP). Examples include:

GHCND_day_PRCP.mat	daily precipitation (mm)
GHCND_day_TMIN.mat	daily minimum temperature (°C)

5.1.1 Daily data file content

data_unit	Contains extracted data unit information (e.g. °C, m/s, tenths of mm, etc.).

date_vec	Date vector for the entire selected temporal range (obtained using the datenum function). To obtain the date in [year, month, day] format, use the datevec function.
ghcnd_data	The content of this matrix depends on qc_apply (Section 3.1.1). If qc_apply ≠ 1, this matrix contains all daily data of the selected variable. When qc_apply = 1, only the daily values that passed quality checks are included. Each row represents a different weather station. Each row cell dimension corresponds to the equivalent row cell dimension in ghcnd_date_indiv. Each row number corresponds to the equivalent row number in ghcnd_gauge_info. Note: When the ghcnd_data matrix size exceeds the 2^{31} limit, the matrix is split into 2-5 matrices (ghcnd_data1, ghcnd_data2, etc.) depending on ghcnd_data matrix size.
ghcnd_date_indiv	Cell matrix of weather-station-specific date vectors within the range of date_vec. Every row cell dimension corresponds to the equivalent row cell dimension in ghcnd_data. Note: When the ghcnd_data matrix size exceeds the 2^{31} limit, the ghcnd_date_indiv matrix is split into 2-5 matrices (ghcnd_date_indiv1, ghcnd_date_indiv2, etc.) to match with ghcnd_data. Each ghcnd_date_indiv# matrix contains data for the corresponding weather stations in ghcnd_date_indiv#.
ghcnd_Qflag	The content of this matrix depends on qc_apply (Section 3.1.1). If qc_apply ≠ 1, then ghcnd_Qflag is a cell matrix of quality flags, with index_Qflag providing the index vectors corresponding to ghcnd_data. If qc_apply = 1 (all quality-flagged data were removed), then qc_apply contains the string '*All quality-flagged data were removed*' instead.
index_Qflag	Matrix of indices referring to quality-flagged data in ghcnd_data. The content of this matrix depends on qc_apply (Section 3.1.1). If qc_apply ≠ 1, then index_Qflag contains index vectors corresponding to ghcnd_data. If index_Qflag = 1 (all quality-flagged data were removed), then index_Qflag is empty.
ghcnd_gauge_info	A cell matrix containing six columns (Station ID, Latitude, Longitude, Elevation, Date Start, Date End)^. Each row number corresponds to the equivalent row number in ghcnd_data. Note: When the ghcnd_data matrix size exceeds the 2^{31} limit, the ghcnd_gauge_info matrix is split into 2-5 matrices (ghcnd_gauge_info1, ghcnd_gauge_info2, etc.) to match with

	ghcnd_data. Each ghcnd_gauge_info# matrix contains data for the corresponding weather stations in ghcnd_data#.
	^Note: When the target variable contains multiple names (i.e. SN**, SX**, WT** or WV**), a 7th column (Variable Name) is recorded in ghcnd_gauge_info to specify the variable sub-category (e.g. SN31).
header_gauge_info	Header for ghcnd_gauge_info, description of column content.
year_range	A string listing your data range (e.g. '*1952-2016*').

When a large number of .dly files is accessed and data compiled over the full temporal range, a large amount of irrelevant data would be generated for weather stations that were operational for a relatively brief period. Hence, individual date vectors are provided for every weather station to reduce data size and memory issues (cf. ghcnd_date_indiv). The date vector for the full temporal range (cf. date_vec) is provided separately to allow for time-specific data extraction.

5.1.2 Post-processing - daily data extraction to Excel

An ancillary file (dailypostproc.m) is available to convert the output data from .mat into .xlsx (Excel) format. This ancillary file will extract the information contained in header_gauge_info, ghcnd_gauge_info, ghcnd_data and ghcnd_date_indiv. The extracted .xlsx file is saved in the output folder. The file name is structured according to the date source (GHCND), the selected variable (e.g. _PRCP) and a fixed suffix (_collated).

Note: Only apply this code if the number of extracted weather stations are within the Excel data limits (e.g. 16,384 columns in Excel 2010).

Known GNU Octave bug: Calling xlswrite in GNU Octave v4.2.0 may produce an error if you use io-2.4.5 or earlier. Ensure that the code in line 128 of xlswrite.m (cf. C:\Octave\...\octave\packages\io...) says: *rstatus = r_extnd = 0;* (instead of: *rstatus = 0;*). See also http://hg.code.sf.net/p/octave/io/rev/05ccd8106966.

5.2 Monthly data

The file name of the generated MAT-file will vary depending on the chosen variable and daily data conversion. Data processing currently includes monthly average (avg) and monthly total (tot). The file name is structured according to the date source (GHCND), monthly value (_mth_avg or _mth_tot) and chosen variable (e.g. _PRCP). Examples include:

GHCND_mth_tot_PRCP.mat	monthly total precipitation (mm)
GHCND_mth_avg_TMAX.mat	monthly average maximum temperature (°C)

5.2.1 Monthly data file content

count_mth_obs	Number of days with data per month and weather station. The dimension of the sparse count_mth_obs matrix is identical to ghcnd_data.

	Note: When the ghcnd_data matrix size exceeds the 2^{31} limit, the count_mth_obs matrix is split into two (count_mth_obs1 and count_mth_obs2) to match with ghcnd_data. Each matrix contains data for the entire chosen temporal range but half of the weather stations (columns/2, rows).
data_unit	Contains extracted data unit information (e.g. °C, m/s, tenths of mm, etc.).
date_vec	Date vector (obtained using the datenum function). To obtain the date in [year, month, day] format, use the datevec function.
ghcnd_data	All extracted data, converted into a monthly value. Each column represents a different weather station. Each row number corresponds to the equivalent row number in count_mth_obs and date_vec. Each column number corresponds to the equivalent row number in ghcnd_gauge_info. Note: When the ghcnd_data matrix size exceeds the 2^{31} limit, the matrix is split into two (ghcnd_data1 and ghcnd_data2). Each matrix contains data for the entire chosen temporal range but half of the weather stations (columns/2, rows).
ghcnd_gauge_info	A cell matrix containing six columns (Station ID, Latitude, Longitude, Elevation, Date Start, Date End)^. Each row number corresponds to the equivalent column number in ghcnd_data. Note: When the ghcnd_data matrix size exceeds the 2^{31} limit, the ghcnd_gauge_info matrix is split into two (ghcnd_gauge_info1 and ghcnd_gauge_info2) to match with ghcnd_data. Each matrix contains data for half of the weather stations (columns, rows/2). ^Note: When the target variable contains multiple names (i.e. SN**, SX**, WT** or WV**), a 7th column (Variable Name) is recorded in ghcnd_gauge_info to specify the variable sub-category (e.g. SN31).
header_gauge_info	header for ghcnd_gauge_info, description of column content
year_range	A string listing your data range (e.g. '*1952-2016*').

When monthly data are extracted (e.g. average temperature or total precipitation), an additional matrix (count_mth_obs) is created that lists the number of days with data per month and weather station. The matrix count_mth_obs permits the elimination of scarcely sampled months at a later stage, without re-extracting the data from the .dly files.

5.3 Matrix size limit in MAT-files

GNU Octave does not permit the saving of matrices that exceed 2^{31} bytes (matrices exceeding that limit are skipped when saving the MAT-file). MATLAB allows saving of larger matrices

(with the '-v7.3' switch). However, the resulting MAT-file size is significantly larger than a MAT-file with identical data stored in sufficiently small matrices. Further, MAT-files saved with the '-v7.3' switch tend to have a longer loading time. Hence, the ghcnd_access toolbox creates submatrices when ghcnd_data exceeds 2^{31} bytes, with each submatrix containing fewer than 2^{31} bytes.

The splitting of matrices is applied to all matrices and vectors that contain weather station-specific information. These include:

ghcnd_data#	daily and monthly data
ghcnd_gauge_info#	daily and monthly data
ghcnd_date_indiv#	daily data only
count_mth_obs#	monthly data only

Note that MAT-files created with GNU Octave tend to be significantly larger than the equivalent MAT-files saved by MATLAB (daily PRCP data: 14 GB vs 2 GB; monthly PRCP data: 2.6 GB vs 0.13 GB).

6 References

Eaton, J.W., Bateman, D., Hauberg, S. and Wehbring, R., 2015. GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computation.

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E. and Houston, T.G., 2012. An overview of the global historical climatology network-daily database. Journal of Atmospheric and Oceanic Technology, 29(7): 897-910.