# Station configuration file.

Steps to produce the station configuration file

Current GHCND mingle-list.txt available from (<ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/>)

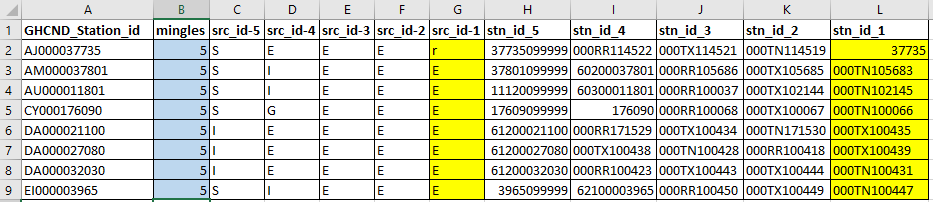
The mingle-list.txt file contains details of the number of stations that are merged to produce one integrated station with all variables available. There are between 1-13 stations used in the mingle/merge process for the GHCND stations.

Step1: Separate the GHCND mingle lists into .xls files by number of station mingles (merges) this is done by Python code ( See Github repository).

Step 2: Open the first mingle .csv file and rename sheet 1 “mingle” . Then Add ghcnd-stations.txt list (available from <ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/>) to a new sheet the .csv mingle file and rename sheet “stations” and save file as (.xlsx) file.

Figure 1 shows an example of a mingle (.xlsx) file. The first column shows the GHCND station ID that has been merged using 5 stations from various sources. The yellow highlighted columns in G (re-named src\_id\_1) show the highest prioritised source ID letter for the merge and column L (re-named stn\_id\_1) shows the station ID for this source. The merge will start with src\_id-5 (S) and then overwrite or merge information from each source from left to right. The highest prioritised source (src\_id\_1) will overwrite the information of a lesser prioritised source in the merge process. The order of merge station IDs matches up with order of the source letters.

Figure 1 Shows an example of a separated mingle (.xlsx) file for a merged GHCNd station.



Step 3: Produce a GHCND source file and match the GHCND source ID with the C3S311a source ID’s (Table 2) Because there are lower- and upper-case letters for the GHCND sources these need to be changed to: (a=ab, b=bc, m=mx, r=rs, s=st, u=uv and z=za). The add the source match information to a sheet in the .xls mingle files and rename sheet sources.

Table 2 Details of the GHCND source ID match with the C3S311a source ID

|  |  |  |
| --- | --- | --- |
| *GHCNd\_source\_Id* | *Source name* | *C3S311a\_Source\_Id* |
| Blank | No source (i.e., data value missing) | blank |
| 0 | U.S. Cooperative Summary of the Day (NCDC DSI-3200) | 161 |
| 6 | CDMP Cooperative Summary of the Day (NCDC DSI-3206) | 162 |
| 7 | U.S. Cooperative Summary of the Day -- Transmitted via WxCoder3 (NCDC DSI-3207) | 120 |
| A | U.S. Automated Surface Observing System (ASOS) real-time data (since January 1, 2006) | 224 |
| ab | Australian data from the Australian Bureau of Meteorology | 225 |
| B | U.S. ASOS data for October 2000-December 2005 (NCDC DSI-3211) | 159 |
| bc | Belarus update | 226 |
| C | Environment Canada | 227 |
| D | Short time delay US National Weather Service CF6 daily summaries provided by the High Plains Regional Climate Center | 228 |
| E | European Climate Assessment and Dataset (Klein Tank et al., 2002) | 229 |
| F | U.S. Fort data | 230 |
| G | Official Global Climate Observing System (GCOS) or other government-supplied data | 231 |
| H | High Plains Regional Climate Center real-time data | 160 |
| I | International collection (non U.S. data received through personal contacts) | 232 |
| K | U.S. Cooperative Summary of the Day data digitized from paper observer forms (from 2011 to present) | 233 |
| M | Monthly METAR Extract (additional ASOS data) | 234 |
| N | Community Collaborative Rain, Hail,and Snow (CoCoRaHS) | 235 |
| Q | Data from several African countries that had been "quarantined", that is, withheld from public release until permission was granted from the respective meteorological services | 236 |
| R | NCEI Reference Network Database (Climate Reference Network and Regional Climate Reference Network) | 237 |
| rs | All-Russian Research Institute of Hydrometeorological Information-World Data Center | 238 |
| S | Global Summary of the Day (NCDC DSI-9618) NOTE: "S" values are derived from hourly synoptic reports exchanged on the Global Telecommunications System (GTS). Daily values derived in this fashion may differ significantly from "true" daily data, particularly for precipitation (i.e., use with caution). | 166 |
| st | China Meteorological Administration/National Meteorological Information Center/ Climatic Data Center (http://cdc.cma.gov.cn) | 239 |
| T | SNOwpack TELemtry (SNOTEL) data obtained from the U.S. Department of Agriculture's Natural Resources Conservation Service | 240 |
| U | Remote Automatic Weather Station (RAWS) data obtained from the Western Regional Climate Center | 241 |
| uv | Ukraine update | 242 |
| W | WBAN/ASOS Summary of the Day from NCDC's Integrated Surface Data (ISD). | 163 |
| X | U.S. First-Order Summary of the Day (NCDC DSI-3210) | 164 |
| Z | Datzilla official additions or replacements | 165 |
| za | Uzbekistan update | 243 |
| mx | Mexican National Water Commision (Comision National del Agua -- CONAGUA) | 196 |

Step 4: You will need to transpose the source IDs (src\_id-5 to src\_id-1 and station IDs (stn\_id\_5 to stn\_ID-1) so that you have two columns of data stn\_ids and src\_ids matching side by side vertically. Create a matrix by highlighting all of cells of the stn\_ids and name matrix1, cerate a matrix for the src\_ids by highlighting them and naming matrix2. Then enter the following code in the frost empty row 1 cell to the right of the data to start the transpose with the lowest priority source (Highlighted yellow in column M Figure 3):

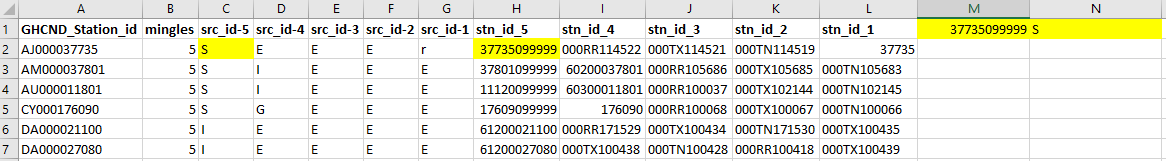
=OFFSET(matrix1,TRUNC((ROW()-ROW($M$1))/COLUMNS(matrix1)),MOD(ROW()-ROW($M$1),COLUMNS(matrix1)),1,1)

In the next cell to right enter the following code to start the transpose with the lowest priority source (Highlighted yellow column N in Figure xxxx)::

=OFFSET(matrix2,TRUNC((ROW()-ROW($N$1))/COLUMNS(matrix2)),MOD(ROW()-ROW($N$1),COLUMNS(matrix2)),1,1)

Then click and drag the formula all the way down until you stop before you see 0 for each result. Note you will have to drag the formula beyond the end of the data the end will vary based on number of mingles multiplied by number of stations in the mingle columns. For example, when we drag the formula for the 5 mingle stations the end of the formula is at row 9715. The number of rows in the st\_id columns = 1943 multiplied by 5 mingles =9715.

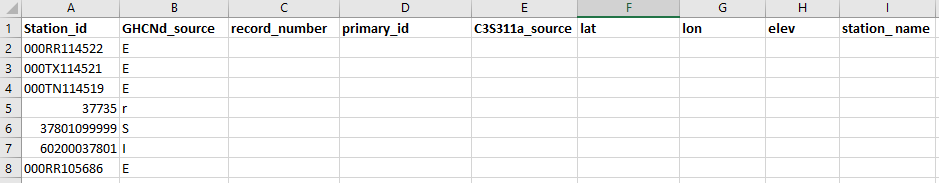
Figure 3 Example showing the highlighted cells in column N and M after entering formula, values match up with values in lowest priority source and station.



Step 5: Copy paste special values only the two columns of transposed data in columns N and M from current sheet to a new sheet in the same (.xlsx) file and rename config.

Step 6: insert headers to you two columns of data: station\_id and GHCND source. And add the following headers shown in Figure 4 to the empty columns in the config sheet.

Figure 4 Example showing the config sheet and headers added.



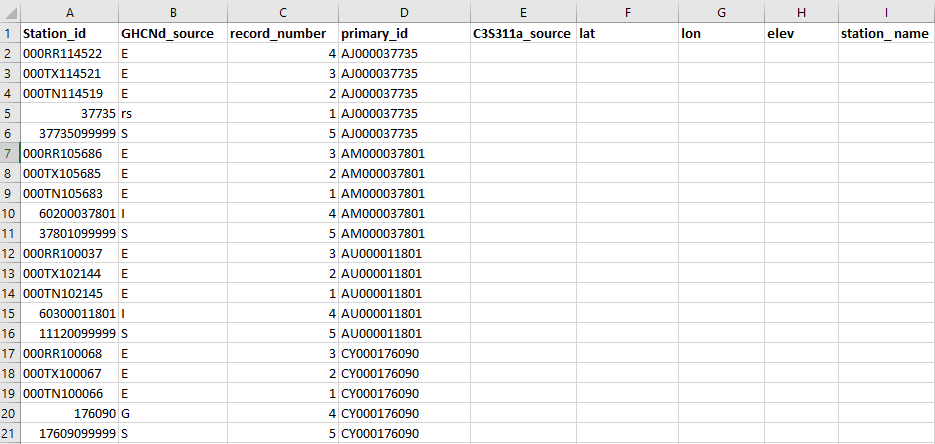
Step 7: Add record numbers to the GHCND\_station\_IDs. See section 2 which presents details on why we use different record numbers for each station. In cell D2 of the config sheet under the record\_number header enter 5, in cell D3 enter 4 , in D4 enter 3, in D5 enter 2, in D6 enter 1. The in D7 enter the formula : =D2 an drag formula down to D11 and click. Then highlight cells D7 to d11 and drag then double click the bottom right of the last cell to copy the formula to all cells in the data range. Then highlight column D and copy paste special values only.

Step 8: You need to match the GHCND primary Id with the station Ids in the mingle sheet. To do this insert a new blank column to the left of Station\_ID column in the config sheet and name it “concatenate”. Do the same in the mingle sheet. Then enter the following formula in cell A2 of the config sheet =CONCATENATE(B2,"-",C2) and double click the bottom right of cell A2 and copy formula to all the cells in the data range. The highlight column A and copy paste special values only. Then in the mingle sheet we need to concatenate each stn\_id with the corresponding src\_id so we can match up the correct GHCN\_Station\_Id with the Station\_Ids in the config sheet. Enter the following formula into cell A2 =CONCATENATE(i2,"-",d2) this will concatenate stn\_id-5 and srcid-5 (e.g 37735099999-S) then double click the bottom right of cell A2 and copy formula to all the cells in the data range.

Step 9: In the config sheet enter the following formula in cell E2 (primary\_id) :

=VLOOKUP(A2,mingle!$A$2:$B$1944,2,FALSE) then double click the bottom right of cell A2 and copy formula to all the cells in the data range. Then double click the bottom right of cell A2 and copy formula to all the cells in the data range. The highlight column primary\_id and copy paste special values only. Highlight column primary\_id and using CTRL F replace #N/A with blank cells, then Highlight all cells and sort values on column primary\_id by a-z so that all blank cells are at the end. In the mingle sheet concatenate the next src and stn using same method as previous. Then go to the first empty cell in the config sheet and enter the same formula as above but change the A2 to the row number of the blank cell. Repeat this process until all the cells in the primary\_id column are populated with GHCND primary station ids. Once completed you can delete the concatenate column in the config sheet (See Figure 5).

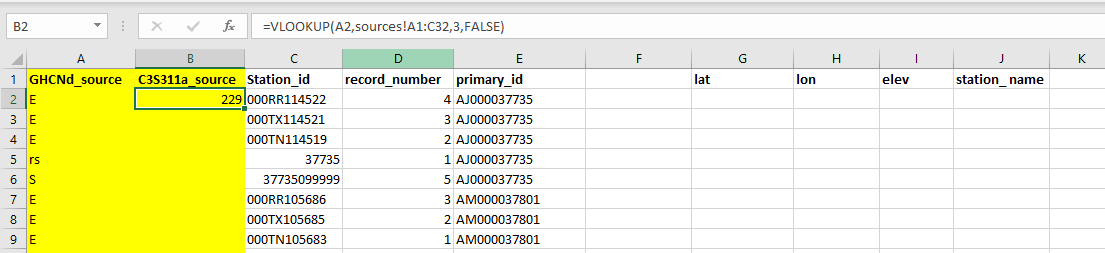
Figure 5 example config sheet with primary\_ids added.



Step 11: Because there are lower- and upper-case letters for the GHCND sources these need to be changed to: (a=ab, b=bc, m=mx, r=rs, s=st, u=uv and z=za) so the vlookup can match with C3S311a source IDs. Highlight all cells and sort by the HHCNd source column. Highlight all cells again and click the data tab top of the excel page and use the filter button. Using the drop-down tab at top of the column select A and/or a then change all the lower-case a to lower case ab, proceed to the next letter and change b to bc repeat for m=mx, r=rs, s=st, u=uv and z=za.

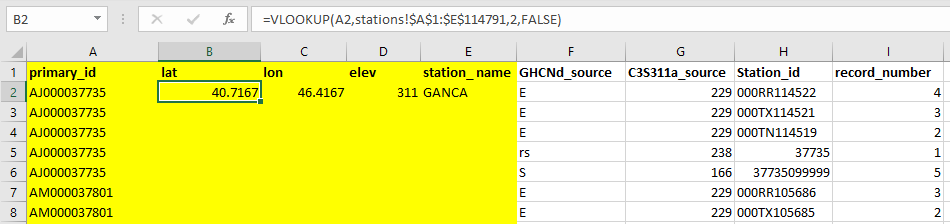
Once this is done arrange the columns A and B as seen in Figure 6 and type the following formula into cell B2 =VLOOKUP(A2,sources!A1:C32,3,FALSE) double click the bottom right of cell A2 and copy formula to all the cells in the data range match up the C3S311a source allocated ID with the GHCND source ID (See Figure 6).The highlight all cells on column B and copy paste special values only.

Figure 6 Shows the columns arranged to look up the C3S311a source ID.



Step 12 involves populating the station information columns in the config sheet from the stations sheet. Arrange the columns in the config sheet a seen in Figure 7. Then in cell b2 enter the formula: =VLOOKUP(A2,stations!$A$1:$E$114791,2,FALSE), in cell C2 enter : =VLOOKUP(A2,stations!$A$1:$E$114791,3,FALSE), in cell D2 enter : =VLOOKUP(A2,stations!$A$1:$E$114791,4,FALSE), in cell E2 enter: =VLOOKUP(A2,stations!$A$1:$E$114791,5,FALSE) and then copy all four column formulas down for all cells in the data range of these columns. Once completed the latitude, longitude, elevation and station name will be matched with each primary Id taken form the stations sheet.

Figure 7 Shows the columns arranged to look up the station information.



When all of these steps have been completed for the first mingle file you must repeat for all mingle files. Then combine all of the config sheets from each mingle file into a one file and this can be used to populate the station configuration file. Additional information on data start and end years, variables available, WMO region and country ID are derived fom the original data files when the they are converted to CDM compliant format. See conversion code on the GitHub repository.

# Optional data Table

Many stations in GHCN-D are made up of data from the same station but from multiple sources. In addition, some of the sub-daily stations have multiple independent location and or elevation information in their associated reports over their operational life time. Many of the stations in the USAF sub-daily data have multiple geo-coordinates (latitude/longitude) and/or varying elevation values entered for reports over the operational life time of an individual station. These different locations could be actual station moves or they could be just data inputting errors. However, this is important metadata information and needs to be retained for the users to interpret the records correctly. If there is ambiguity over a station location due to multiple concurrent metadata records then these can be managed through the optional data table, with the most likely location, arising from the USAF station chronology file, reported in the station\_configuration table and then additional information in the station\_configuration\_optional table – once for each distinct location being reported. Therefore, we have set up multiple configurations in the station configuration table for each sub-daily station where varying location information exists. Practically, this means that if there are 5 distinct sets of geolocation information for a station within the set of reports there will be 5 versions of the station with different record numbers in the station configuration file. Each version will have primary geolocation information that are identical so that on default use the data arise from one single location. However, for 4 versions of the station there will be associated original geolocation information that differs from the primary coordinates given. This information needs to be retained and made available to users. The following section describes how the CDM provides this information.

Figure 8 shows a snapshot of the station\_configuration table showing a sub-daily station that has varying station location metadata over the operational period. In this case we have entered the primary\_id and allocated a record\_number to each station occurrence associated with a different location/elevation. The most likely station latitude/longitude/elevation information derived from the USAF master metadata chronology file is allocated to all the station occurrences with the same primary id. A value of 1 in the optional\_data field indicates that there is additional station metadata information that may be important to users, such as varying historical location information (station moves). This indicates that a user can go to the station\_configuration\_optional table to view the additional metadata for each station configuration (see Figure 9). Where the primary\_id is the link to station for which this entry corresponds, the record\_number links to the station for which this entry corresponds, the kind field indicates the enumerated data type (e.g. 0=integer, 1=numeric, 2=varchar and 3=timestamp with timezone).The field entry code describes the additional data entered into the value field (e.g. 26=alternative latitude, 25=alternative longitude and 29=alternative elevation). The comments field can contain additional information about the entry (see Figure 9).

In the example given in Figures 8 and 9 the interpretation would be that there are 3 unique locations associated with station identifier WMO \_01003 within the individual reports. The first location and that directly associated with all occurrences of an observation is 77 degrees North, 15.5 degrees east. However, only that set of observations associated with record\_number 1 truly arise at that location. The second location differs in that the longitude is 15.55 degrees East (varying by less than the 0.2 degrees threshold). All observations associated with record\_number 2 have this value in the original associated report. The final version differs in both latitude (77.0017) and longitude (15.5348), again both variations lying well within the stated tolerance for inclusion in the present release. All occurrences with record\_number 3 share this set of coordinates in the original reports. All three versions agree that the elevation is 12m a.s.l. Note that the occurrences of locations sometimes overlap (record number 1 and 3) and sometimes are distinct (2 occurs in a block distinct from 1 and 3). This may point toward real relocations or simply corrections / adjustments to locations used in message transmission. Much further work is required still to disentangle issues, and this is why the selection criteria outlined in Section 8.1.1 were applied.

Figure 8. Snapshot of station configuration table for WMO USAF sub daily stations.

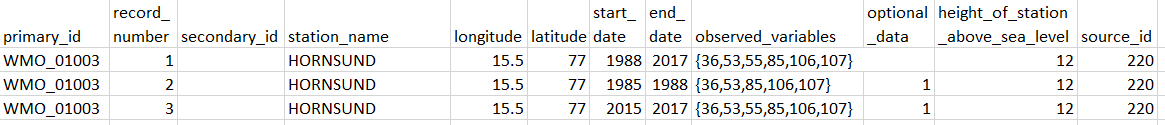


Figure 9. Snapshot of station configuration optional table

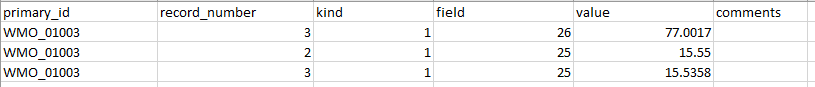
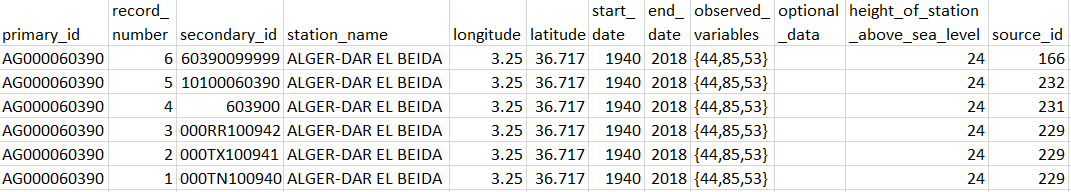


Figure 10 shows a snapshot of the station\_configuration table for a daily station that consists of station data merged from distinct primary sources. In this case we have entered the primary identifier (primary\_id) and allocated a (record\_number) to each different station source used in the data merge. The primary station location information is used for all the different station record numbers regardless of whether there are differences between sources. The different record numbers correspond to the different station configurations and subsequent data sources. This allows for the different data sources for each station to be defined in the source\_id field. We have also retained the original identifier (secondary\_id) which was allocated by each data source. If additional metadata was available for any station a value of 1 would have been entered in the optional\_data field. A user could look up the station\_configuration\_optional table to view the additional metadata as seen in Figure 9. A user can thus associate each observation with the primary source and if that primary source metadata differs from the applied geolocation ascertain the nature of the differences.

Figure 10. Snapshot of station configuration table showing the different configurations for a daily station that consist of merges from multiple sources.



Some GHCND merge decisions involve merges of stations within a source. In such a case (most prevalently GSOD) there will exist multiple versions of the station arising from the single source in question. These are differentiated by the secondary\_id which points to the mingling of two records from within a single source. In addition, for the ECA&D source (and uniquely for that source presently) each element is associated with a unique identifier for historical reasons such that for a ECA&D station reporting for example maximum and minimum temperatures and precipitation it will have three identifiers in that source. To retain provenance we have provided in such cases n identifiers (where n is equal to the number of ECA&D identifiers) and each element is mapped onto its ECA&D identifier. Thus there will be multiple ocurrences of source identifier 229 (ECA&D) for several daily stations and the secondary\_id (Figure 10 final three entries) is necessary to use in addition to the source\_id.