



National Park Service

Inventory and Monitoring Division

Processing Climate Data from the Authoritative Data Sources Using Automated Technique

Revision Date: July 17, 2012

Authors: Kirk Sherrill, Brent Frakes, Jeff Stark

Table of Contents:

1. Overview	3
2. Preparation	5
2.1. Obtain Python Code	5
2.2. Obtain Write Permissions for the Database	6
2.3. Confirm SQL-Server Management Studio Is Installed	6
3. Process Climate Data	6
3.1. Edit the Process Script by Datatype	6
3.2. Run Process Scripts	7
3.2.1. Output Directory	7
3.2.2. Review LogFile	9
4. Spot Check for Reformatting Errors	9

List of Figures:

Figure 1: General Workflow for Processing and Loading Station Data	5
Figure 2: Example Process_GHCN.py script	7
Figure 3: Processing Output Directory	7
Figure 4: Example of 01_Acquired GHCN folder	8
Figure 5: Example of 02_Reformatted GHCN folder	8
Figure 6: Example of 03_Combined GHCN folder	9
Figure 7: Logfile.txt	9

1. Overview

This SOP outlines the process of automatically acquiring, reformatting, and combining the climate data with the assumption that the data either (1) has been downloaded from the respective data sources or (2) was submitted as a time series using the SOP for Submitting Data for Upload to the Climate Database (<https://irma.nps.gov/App/Reference/Profile/2166928>).

As part of the NPS climate database support effort (See the Climate Database Protocol (<https://irma.nps.gov/App/Reference/Profile?code=2167254>), the Inventory & Monitoring Division (IMD) will quarterly receive requests to download data from national-level authoritative data sources and upload to the IMD Climate Database (<https://nrinfo.nps.gov/Reference.mvc/Profile?code=2166927>). The data are maintained by several data sources in various databases. Table 1 shows a list of the climatic data sources, the associated climatic variables per data source which are processed by the NPS, and the dates the IMD will regularly synchronize with the data providers.

Authoritative Data Source	Dataset	Variables	Abbreviation	Unit	Synchronization Date
National Weather Service	GHCN/COOP	Daily Precipitation	PRCP	1/100 of an Inch	April 1 – May 1
		Daily Snowfall	SNOW	Inches	
		Snow Depth at Observation Time	SNWD	Inches	
		Daily Maximum Temperature	TMAX	°F	
		Daily Minimum Temperature	TMIN	°F	
		Snow Water Equivalent	WTEQ	1/10 of an inch	
Natural Resource Conservation Service	SNOTEL	Daily Precipitation	PRCP	Inches	April 1 – May 1
		Snow Water Equivalent	SWE	Inches	
		Daily Maximum Temperature	TMAX	°C	
		Daily Minimum Temperature	TMIN	°C	
Natural Resource Conservation Service	SnowCourse	Snow Water Equivalent	SWE	Inches	April 1 – May 1
Western Regional Climate Center	RAWS	Daily Precipitation	PRCP	1/100 of an Inch	April 1 – May 1
		Daily Maximum Temperature	TMAX	°F	
		Daily Minimum Temperature	TMIN	°F	
United States Geological Survey	GAGE	<i>Discharge, CFS</i>	00060	Cubic Feet Per Second	April 1 – May 1

Figure 2. Climatic data sources and associated climatic variables, abbreviations, and units.

While processing the data, IMD will add, when appropriate, additional processing flags for data issues/errors not identified by the source agencies. Common cases include:

- Observations that are not physically possible – Examples include temperature above the maximum observed global temperature and negative accumulated precipitation
- Observations that are illogical – For example, maximum temperature values less than minimum temperatures

It is important to emphasize that the IMD will preserve the data and all data quality and measure flags exactly as they are from the sources, even when errors are identified.

Only in the following cases will there be any modification to the source data:

- Non-numeric observation values – Since our database requires a valid number, we will set this value to NULL and flag the observation accordingly
- Two unexplained and differing observations for a single day – Because we do not know which one to use, we will set the observation value, the data quality flags, and the data measure flags to NULL for that date and add a data processing flag.
- Invalid dates – We will not include any data for the given observation. We will flag the entire version of data indicating that one or more dates are invalid.

When possible, Representation State Transfer (REST) calls are used to request data for each station and then save the data to a specified folder location although manual techniques are employed where necessary.

Error! Reference source not found. provides a general overview of the steps which are further articulated in the remaining sections.

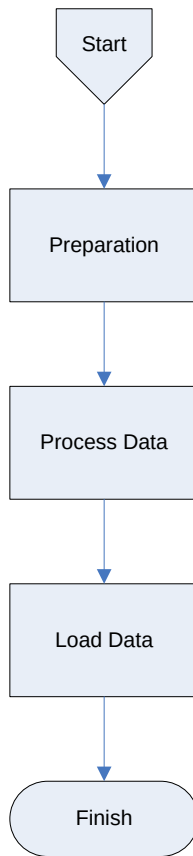


Figure 3. General Workflow for Processing and Loading Station Data

2. Preparation

2.1. Obtain Python Code

Each data source has its own master script (*Process_GHCN.py*, *Process_SNOTEL.py*, *Process_SNOWCOURSE.py*, *Process_RAWS.py*, and *Process_USGS.py*) that are located as a holding at the Reference profile for this SOP:

<https://irma.nps.gov/App/Reference/Profile?code=2170576>

In addition, working local copies are also maintained in the Tortoise SVN code repository at <https://data.nature.nps.gov/svn/GISApps/Climate>. The data source master scripts have an accompanying module *ClimateModule.py*, which is also a holding on this SOP's reference, and can be found in the scripting repository.

The Processing scripts perform the following:

- Programmatically downloads all SNOTEL, Snow Course, GHCN/COOP, RAWS and USGS stations using the REST calls. It is intended to be used on a quarterly basis to update all requested climatic weather station data.
- Reformats source data to tab-delimited text files. Each row in the files represents a unique combination of network, station, parameter, and date.
- Combines all station data and creates a single combined text file by data type.

2.2. Obtain Write Permissions for the Database

You will need to obtain write permissions from the database administrator. Contact Simon Kingston at (970) 225-3551 or Simon_Kingston@nps.gov.

2.3. Confirm SQL-Server Management Studio Is Installed

To load the data into the database, you must have SQL Server Management Studio loaded on your computer. If you do not, initiate an IT HelpDesk Request to get it installed.

3. Process Climate Data

3.1. Edit the Process Script by Datatype

Bulleted below are both parameters which must be entered by the user, and are used to define the weather station data to be processed:

- *rootFolder*- The output data directory is organized by date of download, where downloaded source data(i.e. script output) will be saved to the following folder:

\\inp2300fcspete1\Base_Carto\OIME_Processing\ClimateDatabase\{YYYYMMDD}

- *logFileNameAndPath*- Where the logfile will be stored. This is the same path as the rootFolder.

Both parameters should be reviewed and, where appropriate, modified for all

An example *Process_GHCN.py* script is shown in **Error! Reference source not found.** In this example, the rootFolder is

Y:\\OIME_Processing\\ClimateDataBase\\20120629 where the data and logfile will be stored.

```

dataSource = 'GHCN'                                     #'USGS'|'SNOTEL'|'GHCN'|'SNOWCOURSE'|'RAWS'
rootFolder = 'Y:\\OIME_Processing\\ClimateDataBase\\20120629' #Folder where the acquired files will be written
dBaseConnectionString = 'DRIVER={SQL Server Native Client 10.0};SERVER=INP2300FCSGUM06\\GIS_Working_Data;UID=climate_reader;PWD=clim@te'
logFileNameAndPath = 'Y:\\OIME_Processing\\ClimateDataBase\\20120629\\logfile.txt'
OverwriteExistingStations = False                       #Indicates whether to overwrite any station files that have already been acquired

#####
import ClimateModule
reload (ClimateModule)
from ClimateModule import *

Acquire(dataSource, rootFolder, dBaseConnectionString, logFileNameAndPath, OverwriteExistingStations)

Reformat(dataSource, rootFolder, logFileNameAndPath)

Combine(dataSource, rootFolder, logFileNameAndPath)

```

Figure 4. Example Process_GHCN.py script

3.2. Run Process Scripts

Run the datatype scripts that you wish to load to the database.

3.2.1. Output Directory

Once the data has been processed, three folders will be created in the YYYYMMDD directory: *01_Acquired*, *02_Reformatted*, and *03_Combined*. Within each of these folders, datatype (GHCN, SNOTEL, SNOWCOURSE, RAWS, and USGS) folders will appear.

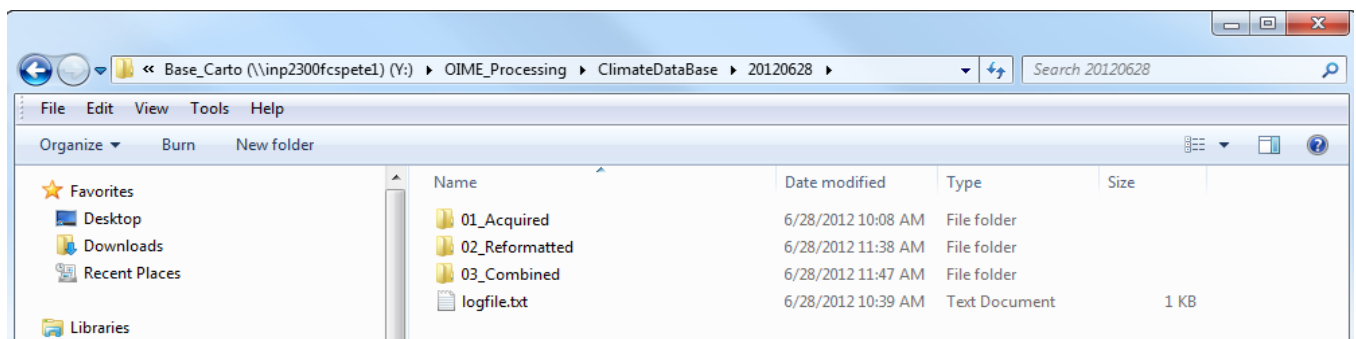


Figure 5. Processing Output Directory

The *01_Acquired* folder will hold all raw station data from the data sources. The file name will contain the Network, StationID, and date of download.

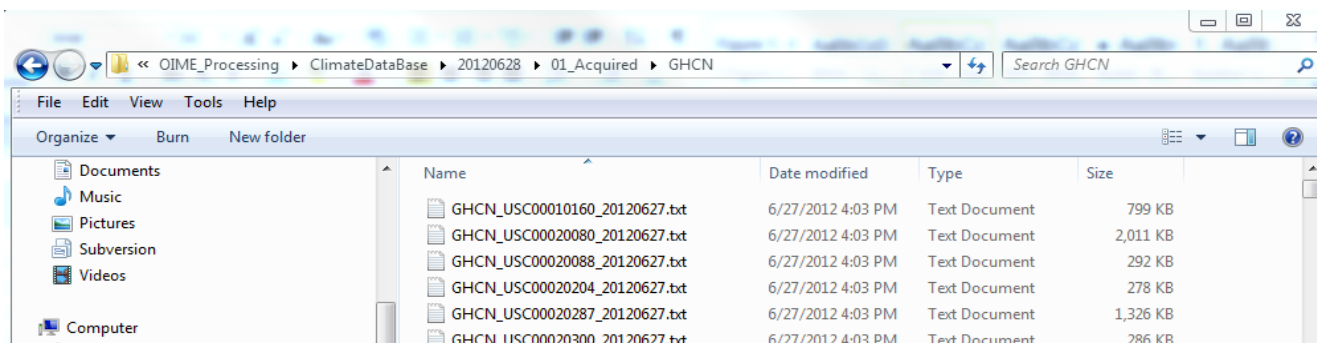


Figure 6. Example of 01_Acquired GHCN folder

Next, the *02_Reformat* folder contains the reformatted, tab-delimited text files with the following columns:

- SourceFile - data file name
- Network - data provider and/or program
- StationID–Station ID used by the data provider
- Date –Observation data
- Parameter - (meteorological/hydrological element)
- MeasureUnit–Unit of measurement
- Value – A number representing the observed value
- DataQualityFlags– Comma-delimited set of data quality flags
- DataMeasureFlags - – Comma-delimited set of data measure flags

Reformatted files are named with the original file name, with a “_Converted.txt” suffix added.

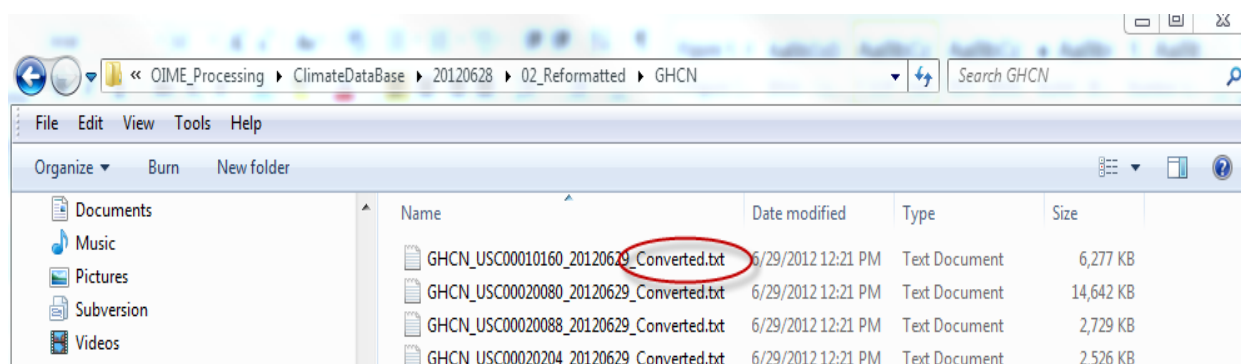


Figure 7. Example of 02_Reformatted GHCN folder

Lastly, the *03_Combined* folder combines all station data from the *02_Reformatted* folder and creates a single combined text file (by datatype).

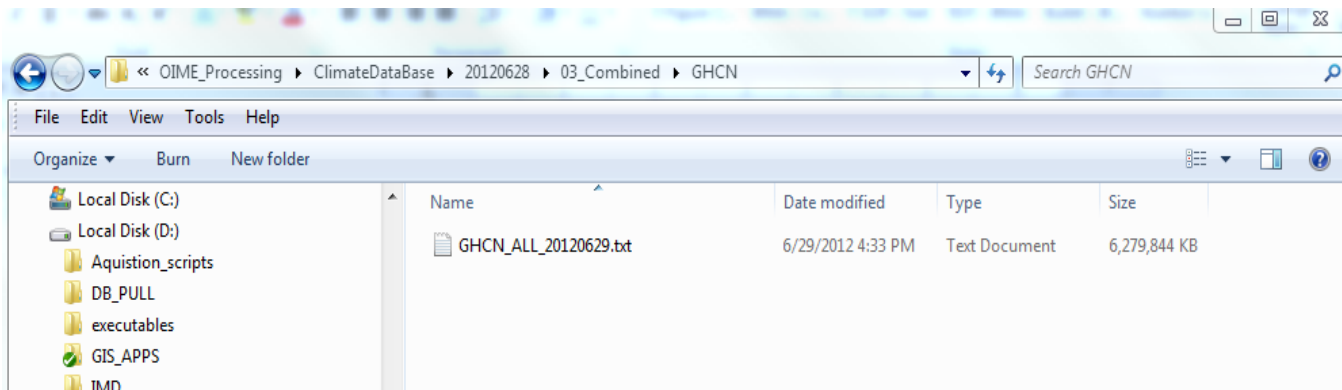


Figure 8. Example of 03_Combined GHCN folder

3.2.2. Review LogFile

A combined log (LogFile.txt) for all sources is created and placed in the YYYYMMDD directory. For each data source the log file lists the date and time of creation, the data source, what function was used, and how many stations were returned.

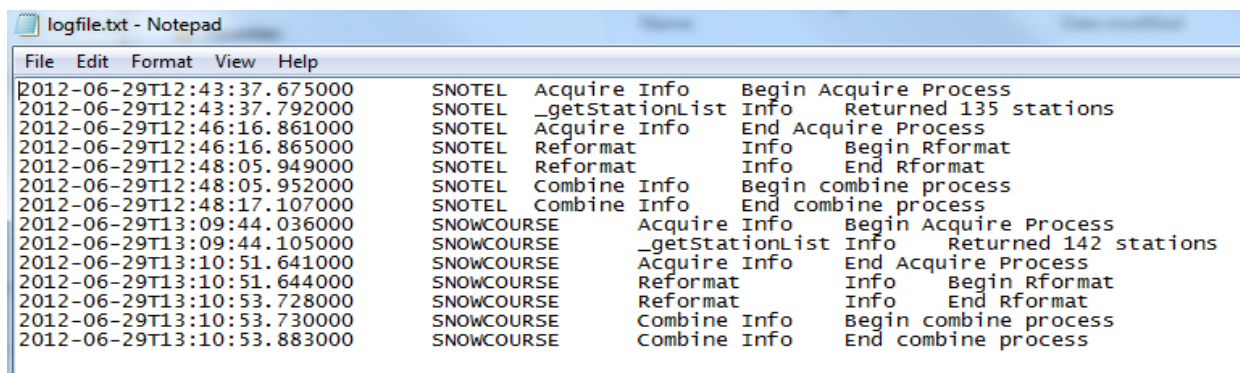


Figure 9. Logfile.txt

3.3. Spot Check for Reformatting Errors

To ensure processing was successful, 10% (or a maximum of 30) of the files for each source will be randomly selected and spot-checked to ensure the processing completed properly.

For a given file, the following will be checked:

- 10% (or a maximum of 30 records) will be randomly selected.
- Start dates and end dates are the same

If any discrepancies are detected between the source and processed data, it will be necessary to diagnose the source of errors prior to subsequent uploading of data to the SQL Climate Database.