Dataset Description Document

Global Summary of the Month/Year Dataset

Version 1.0.1 / March 27, 2017

RESPONSIBILITY

Prepared By:	Jay Lawrimore	Chief, DSB	NOAA/NCEI
	Signature		Date
Reviewed By:	<Name>	<Title>	<Organization>
	Signature		Date
Reviewed By:	<Name>	<Title>	<Organization>
	Signature		Date
Approved By:	<Name>	<Title>	<Organization>
	Signature		Date
Approved By:	<Name>	<Title>	<Organization>
	Signature		Date

REVISION HISTORY

Note: the version number of this document is being maintained to match the version number of the GSOM/GSOY data (vX.Y.Z). If a change to this document is made without a change to the underlying data, a 4th digit is used (e.g., 1.0.0.1 made on 5/5/2016).

Version	Description	Revised Sections	Date
1.0.0	Initial submission	New Document	02/01/2016

1.0.0.1	Updated to reflect changes to element names by DSD/DAB (Brian May and Rich Baldwin). The number of elements was reduced from 55 to 48 because the day of occurrence is now included as an attribute instead of a separate element. Also removed the 3-second wind elements because such an element is not computed.	Section 2.2	05/05/2016
1.0.1	Updated to reflect change of elements DP01, DP10, DP1X. These are now number of days with precipitation >= 0.01”, 0.10”, 1.0”, respectively.	Section 2.2	3/27/2017

Table of Contents

1. Introduction

1.1 Purpose

The purpose of this document is to describe the algorithms, software, and data sets associated with development of the Global Summary of the Month and Global Summary of the Year datasets (hereafter referred to collectively as Global Summary of the Month or GSOM). In the past, NCEI provided summary of the month data for U.S. Cooperative Observer Network (COOP) stations in the DSI-3220 Summary of the Month dataset as the primary means of data stewardship and access (via Annual Climatological Summaries). This was extended in 2011 when the Global Historical Climatology Network-Daily dataset quality control and processing system replaced the COOP legacy processing system. At that time, NCEI's Data Access Branch (DAB) began on-the-fly computations of summary of the month elements, providing customers with summaries based on GHCN-Daily data (via Monthly Climatological Summaries). Although this filled an immediate requirement it had several disadvantages including lack of a permanent archive of the data, no consensus on standards used in computation of the monthly values, incomplete documentation describing the algorithms used for monthly and annual summaries, and some differences in climatological data provided to customers by the two legacy products. Development of this dataset helps resolve such deficiencies and provides NCEI's DAB with data that will meet customer requirements.

1.2 Document Maintenance

This document will be maintained in a manner consistent with version control practices for NCEI. When a new version of the Global Summary of the Month dataset is developed and approved, this document will be reviewed and edited as necessary to ensure that it remains consistent with the current operational version.

2. Data set Description

The Global Summary of the Month dataset (GSOM) consists of 48 climatological variables computed from Summary of the Day observations of the Global Historical Climatology Network-Daily dataset (Menne et al. 2012). Of these, 46 are monthly and annual summary variables and two are season-to-date variables. Each variable is described in section 2.2.

2.1 Quality Control

Quality control (QC) is performed at the summary of the day level through the GHCN-Daily QC processes as described below. Additional quality assurance is performed following computation of the monthly and annual summaries through the use of a validation process involving independent calculations and cross-comparisons to ensure computational accuracy. The schedule for the operational and cross-comparison processes is described in Section 3.1.

GHCN-Daily data are quality controlled using a suite of automated algorithms designed to detect as many errors as possible while maintaining a low probability of falsely identifying true meteorological events as erroneous (Durre et al. 2010). The system consists of 19 tests that detect duplicate data, climatological outliers, and various inconsistencies (internal, temporal, and spatial). Quality control thresholds were determined by manual review of random samples of the values flagged as errors and established so that the false positive rate, or fraction of valid values identified as errors, is minimized. The tests are arranged in a sequence in which the performance of the later checks is enhanced by the error detection capabilities of the earlier tests.

Data which are flagged as part of GHCN-Daily quality control processes are excluded from summary of the month computations. Thresholds were established for the number of missing or flagged values allowed in the computation of a monthly value. For

example, a monthly mean temperature is not computed if more than 5 daily observations or more than 3 consecutive days are missing in a month. The thresholds for each variable are included in section 2.2. Values are included in the datasets to provide the user with information regarding the number of missing or flagged values in the month.

2.2 Data Set Elements

The elements summarized on monthly and annual timescales are defined below.

Annual averages are computed from equally weighted months; i.e., no weighting of months by number of days. This is to remain consistent with the National Data Stewardship Committee's recommendation established in 2015. Annual values are set to missing if one (1) or more Months are missing.

1. TMAX

Monthly Mean Maximum Temperature – Average of daily maximum temperature; computed to hundredths degree Celsius. Values are set to missing if more than 5 daily values are missing or flagged or if more than 3 daily values in a row (consecutive) in a given month are missing or flagged.

The Annual temperature is the average of the monthly temperatures.

DaysMissing (Numeric value): The number of days (from 1 to 5) missing or flagged is provided

For Australia, which includes measurement of multi-day temperature, the GHCN-D elements MDTX and DATX are used. The MDTX measured value is treated as a single day maximum temperature. The other days in the multi-day period defined by DATX are set to missing. Multi-day temperatures that cross a month are ignored – i.e., set to missing any multi-day periods crossing the month. This is done because it is not possible to determine in which month the multi-day maximum temperature occurred.

2. TMIN

Monthly Mean Minimum Temperature – Average of daily minimum temperature; computed to hundredths degree Celsius. Values are set to missing if more than 5 daily values are missing or flagged or if more than 3 daily values in a row (consecutive) in a given month are missing or flagged.

The Annual temperature is the average of the monthly temperatures.

DaysMissing (Numeric value): The number of days (from 1 to 5) missing or flagged is provided. .

For Australia, which includes measurement of multi-day temperature, the GHCN-D elements MDTN and DATN are used. The MDTN measured value is treated as a single day minimum temperature. The other days in the multi-day period defined by DATN are set to missing. Multi-day temperatures that cross a month are ignored – i.e., set to missing any multi-day periods crossing the month. This is done because it is not possible to determine in which month the multi-day minimum temperature occurred.

3. TAVG

Average Monthly Temperature - computed by adding the unrounded monthly mean TMAX (average of the daily maximum temperatures) and TMIN temps (average of the daily minimum temperatures) and dividing by 2; then round to hundredths degree Celsius. Values are set to missing if either the monthly mean TMAX or TMIN temperature is missing.

The Annual temperature is the average of the monthly temperatures.

DaysMissing (Numeric value): The number of days (from 1 to 5) missing or flagged is provided.

Use the same Criteria for all of the following temperature elements; No more than 5 days missing or flagged in the month – and no more than 3 consecutive days missing or flagged.

4. EMXT

Extreme maximum temperature for the month (year) is the highest daily maximum temperature value for specified month (year). In tenths of degree Celsius. (Day of the EMXT value for current month (year) is included as an attribute of this element. This attribute may be missing while EMXT has a valid value, if the extreme occurred during a multi-day period.)

5. EMNT

Extreme minimum temperature for the month (year) is the lowest daily minimum temperature value for specified month (year). In tenths of degree Celsius. (Day of the EMNT value for specified month (year) is included as an attribute of this element. This attribute may be missing while EMNT has a valid value, if the extreme occurred during a multi-day period.)

6. DX90

Number of days with maximum temperature >= 32.2°C/90°F.

7. DX70

Number of days with maximum temperature >= 21.1°C/70°F.

8. DX32

Number of days with maximum temperature <= 0°C/32°F.

9. DT32

Number of days with minimum temperature <= 0°C/32°F.

10. DT00

Number of days with minimum temperature <= -17.8°C/0°F.

11. HTDD

Heating Degree Days - computed when daily average temperature is less than 18.3°C/65°F. HDD = 18.3° – mean daily temperature to tenths degree Celsius. Each day is summed to produce a monthly total.

Each monthly degree day total is summed to produce a yearly total. This is a July through June Annual total for Northern Hemisphere stations (Year is for the ending month; June); January-December for Southern Hemisphere stations.

12. CLDD

Cooling Degree Days - computed when daily average temperature is above 18.3°C/65°F. CDD = mean daily temperature -18.3° to tenths degree Celsius. Each day is summed to produce a monthly total.

Each monthly degree day total is summed to produce a yearly total. This is a January-December Annual total for Northern Hemisphere stations. July-June for Southern Hemisphere stations (Year is for the ending month; June).

13. PRCP

Total Monthly (Annual) precipitation. Precipitation totals are based on daily or multi-day (if daily is missing) precipitation report, in millimeters to tenths.

The value is set to missing if more than 5 daily values are missing or flagged and there is an additional stipulation that there can be no more than 5 consecutive days of accumulation in a month (accumulations that cross a month are ignored, i.e., accumulated values are set to missing). This is to ensure consistency with the newest GHCN-Monthly data set (version 3).

The following flags will be used:

- No Flag: All days with precipitation in the month (No missing, None Flagged, and No Accumulations).
- Measurement Flags (a or T)
    - This does not include the use of “I” (incomplete month) because it cannot be consistently used; for example when a Trace will need to be placed in the measurement flag location. Instead, the number of days missing will be shown in the days_miss flag.
    - a: Any Accumulation within the month
        - Trace flag trumps an accumulation flag.
    - T for Trace Amount.
- Source Flag: The source flag from GHCN-Daily. (If there are multiple sources within the same month the source present most often in the month is used – sometimes a Z; Datzilla flag is mixed in a month.
- Days Miss Flag: Numeric; Include number of days (from 1 to 5) missing or flagged.

Use the same criteria for all Precipitation and Snow elements (No more than 5 days missing or flagged in the month and no more than 5 days in a row of accumulation.)

For Elements 16 through 25, Multi-day precipitation totals are excluded from the calculations – i.e., all days in the multi-day accumulation period are treated as missing for this element.

14. EMXP

Highest daily total of precipitation in the month (year) in tenths of millimeters (Non-Accumulation). The day that EMXP occurred for the month (year) is included as an attribute of this element.

15. DP01

Number of days with >= 0.01 inch/0.254 millimeter in the month (year). (Non-Accumulation) Note: values originally recorded in inches as 0.01” are stored as 0.3

millimeters in GHCN-Daily; technically this test is for values greater than or equal to 0.3 mm.

16. DP10

Number of days with >= 0.1 inch/2.54 millimeter in the month (year). (Non-Accumulation) Note: values originally recorded in inches as 0.10” are stored as 2.5 millimeters in GHCN-Daily; technically this test is for values greater than or equal to 2.5 mm.

17. DP1X

Number of days with >= 1.0 inch (25.4mm) precipitation in the month (year). (Non-Accumulation)

18. SNOW

Total Monthly (Annual) Snowfall in millimeters.

19. EMSN

Highest daily snowfall in the month (year) in millimeters. The day EMSN occurred for the month (year) is included as an attribute of this element.

20. DSNW

Number of days with snowfall >= 1 inch (25 mm). Snowfall is provided in mm in GHCN-D so 25 mm is used instead of 25.4 mm.

21. DSND

Number of days with snow depth >= 1 inch (25 mm). Snow depth is provided in mm in GHCN-D so 25 mm is used instead of 25.4 mm.

22. EMSD

Highest daily Snow Depth in the month (year) in millimeters. The day EMSD occurred for the month (year) is included as an attribute of this element.

23. EVAP

Total Monthly Evaporation to tenths of millimeters. Precipitation missing/flagged criteria is used.

Use Temperature missing/flagged criteria for the following elements.

24. MNPN

Monthly (Annual) mean minimum temperature of evaporation pan water in hundredths degree Celsius.

25. MXPN

Monthly (Annual) mean maximum temperature of evaporation pan water in hundredths degree Celsius.

26. WDMV

Total monthly (annual) wind movement over evaporation pan in kilometers. Precipitation missing/flagged criteria is used.

27. TSUN

Daily total sunshine (minutes). This element contains historical data only, except for four (4) stations in the U.S. that continue to measure total sunshine (Miami, FL; Buffalo, NY; Pocatello, ID; Grand Rapids, MI).

Total sunshine for the month is computed using Temperature missing/flagged criteria (5/3).

28. PSUN

Average of the daily percent of possible sunshine. Daily percentages are reported, not computed. For monthly percent of possible, the daily values are averaged using

Temperature missing and flagged criteria (no more than 5 missing days in the month, no more than 3 in a row missing).

For the following wind elements the Temperature missing/flagged criteria are used (no more than 5 missing days in the month, no more than 3 in a row missing)

29. AWND

Monthly (Annual) average wind speed. Average the Daily AWND values in GHCN-D to get monthly and annual averages. (tenths of meters per second).

30. WSFM

Maximum Wind Speed - Fastest mile. Maximum wind speed for the month (year) reported as Fastest Mile.

31. WDFM

Wind Direction for Maximum Wind Speed – Fastest Mile.

32. WSF2

Maximum Wind Speed - Fastest 2-minute. Maximum wind speed for the month (year) reported as Fastest 2-minute.

33. WDF2

Wind Direction for Maximum Wind Speed – Fastest 2-minute wind.

34. WSF1

Maximum Wind Speed – Fastest 1-minute. Maximum wind speed for the month (year) reported as Fastest 1-minute.

35. WDF1

Wind Direction for Maximum Wind Speed – Fastest 1-minute wind.

36. WSFG

Peak Wind Gust Speed – FG. Maximum wind gust speed for the month (year). Note: It is permissible to have a wind gust speed without a direction. Gust speeds are sometimes measured without a report of direction.

37. WDFG

Wind Direction for Peak Wind Gust Speed.

38. WSF5

Peak Wind Gust Speed – Fastest 5-second wind. Maximum wind gust speed for the month (year) reported as Fastest 5-second wind.

39. WDF5

Wind Direction for Peak 5-second Wind Gust Speed.

Temperature missing criteria are used for all soil temperature elements. No more than 5 missing days in the month, no more than 3 in a row missing.

For the following six elements (40-45) yy can equal 01 to 08 indicating up to eight possible combinations of soil cover and depth for any particular month-year. These are listed below. The soil cover and depth are annotated in attributes of each element. Note that the soil cover and depth attributed to an element such as MX03 is specific to a particular month-year. The element MX03 may have attributes of a different soil cover and depth in another month-year.

Soil cover

1 Grass
2 Fallow
3 Bare ground
4 Brome grass
5 Sod

6 Straw mulch
7 Grass muck
8 Bare muck
0 Unknown

Soil Depth
Inches (cm)
2 (5)
4 (10)
8 (20)
20 (50)
40 (100)
60 (150)
72 (180)
unknown

40. MXyy

Monthly (Annual) mean of daily maximum soil temperature.

41. MNyy

Monthly (Annual) mean of daily minimum soil temperature.

42. HXyy

Highest maximum soil temperature for the month (year).

43. HNyy

Highest minimum soil temperature for the month (year).

44. LXyy

Lowest maximum soil temperature for the month (year).

45. LNyy

Lowest minimum soil temperature for the month (year).

Additional Heating/Cooling Degree Day elements added to support LCD products.

46. HDSD

Heating Degree Day **(This is a season-to-date element.)**

A running total of monthly HDD through the end of the most recent month. Each month of HTDD is summed to produce a season-to-date total. Season starts in July for Northern Hemisphere stations (Year is for the ending month); starts in January for Southern Hemisphere stations.

47. CDSD

Cooling Degree Day **(This is a season-to-date element.)**

A running total of monthly CDD through the end of the most recent month. Each month of CLDD is summed to produce a season-to-date total. Season starts in January for Northern Hemisphere stations; starts in July for Southern Hemisphere stations (Year is for the ending month).

48. FZFx (x = 0 through 9)

First/Last Freeze Days **(This is an annual element.)**

x=

0 First freeze of the year that is less than or equal to 32F (0C).
1 First freeze of the year that is less than or equal to 28F (-2.2C).
2 First freeze of the year that is less than or equal to 24F (-4.4C).
3 First freeze of the year that is less than or equal to 20F (-6.7C).
4 First freeze of the year that is less than or equal to 16F (-8.9C).
5 Last freeze of the year that is less than or equal to 32F (0C).
6 Last freeze of the year that is less than or equal to 28F (-2.2C).

7 Last freeze of the year that is less than or equal to 24F (-4.4C).
8 Last freeze of the year that is less than or equal to 20F (-6.7C).
9 Last freeze of the year that is less than or equal to 16F (-8.9C).

This element requires information in the daily records that is not contained in the monthly summaries.

Missing Criteria – Data for all 12 months are required. For each month no more than 5 days missing in the month/no more than 3 consecutive days missing. – Else FZFx is missing.

The YEAR (dividing date) for this element is 1 August (this is consistent with the NCEI 1981-2010 Normals project). This is the only element with a ‘year’ that begins on 1 August.

For countries with multi-day temperature (Australia), the multi-day temperatures are included in the computation in the following manner - the first/last freeze days are the last day of the multi-day period (maximum length of multi-day period allowed is 4 days because no more than three consecutive missing or flagged days of temperature observations are allowed).

Note that if the minimum temperature for the month is within a multi-day period, the temperature is reported but not the date. Thus it is possible a first/last freeze day could have a date (last in the multi-day period) but the minimum temperature would not.

3. Dataset Processing

3.1 Processing Outline

The Level 1 Data Flow diagram (Figure 1) shows a process that relies on two sources of input data; a master station list and GHCN-Daily station data. The process is governed

by a master script (mkGhcnMS.scr) which executes two Java programs in sequence; *createGhcnM.java* computes summary of the month statistics from the GHCN-Daily data. The summary of the month output is subsequently used as input to *createGhcnY.java* which computes the annual summaries. (If either of these programs aborts, a notification e-mail is sent to Data Operations Branch (DOB) personnel responsible for this process; currently Ron Ray.) The individual station files containing the monthly and annual data are then combined into a monthly and annual file (ghcnd-MS.dat and ghcnd-YS.dat, respectively) by *combineMsYsData.java*. These two files are compressed, a checksum is executed using mkChecksum.java and the manifest files are sent to Gulp2 where the data are available for Archive Branch and Data Access Branch to retrieve.

Weekly Processing Timeline

The GSOM data are processed weekly to ensure consistency with the GHCN-Daily period of record update that occurs each week. The GHCN-D period of record update typically completes each Tuesday by 1200L to 1330L. (The completion time can be affected by IT Branch system changes that occur periodically.) Data are available on ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ at the conclusion of the GHCN-D update process. An automated process that runs continuously in cron identifies the new data and executes an ftp from that ftp location to the production server (vapor). This occurs within one hour of the completion of the GHCN-D update. The process that performs the automated load is maintained by DSB personnel (currently David Wuertz).

The GSOM computational process begins at 1700L Wednesday on the production server. This process takes 12 to 24 hours, completing by 1700L each Thursday. Following data verification as described in the section below, data are made available on Gulp2 for subsequent archive and access activities.

Data Validation

Independent validation of the GSOM output is performed in an automated fashion each time the GSOM process completes. Validation begins with computation of monthly precipitation and snow statistics from GHCN-D, executed at the same time as the operational GSOM process. (The computations associated with precipitation accumulations require additional processing not required of the other variables.) After the operational GSOM process completes, cross-comparisons of all variables between the operational output and the independent process takes place (typically Thursday afternoon or evening). Note that before validation begins there is automated confirmation that the GSOM update process completed. The validation process is performed on 30 stations. A log file of output is produced. An e-mail notification is provided at the completion of the validation process indicating the status of the comparisons. If any differences between the operational and the validation process exist, the e–mail provides details regarding the problem. The e-mail is sent to the owner of the GSOM process (currently Ron Ray), the owner of the validation process (Scott Applequist), and the Datzilla Gatekeeper (Bryant Korzeniewski). Although discrepancies are expected to be extremely rare, when they do occur a manual review is required to determine the source of any problem with the goal of completing resolution NLT Friday.

3.2 Data Collection and Integration

All input data are provided via the GHCN-Daily data set. This data set is updated and archived daily. Data are available on ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/.

3.3 Dataset Output and Version Control

The data for all variables and stations are contained in two ASCII files. The filenames are included in the GSOM/GSOY Submission Agreement. The format of these files is described in an xml document which is maintained in subversion (https://conman.ncdc.noaa.gov/svn-repos/cab/ghcndMoYrSummary/trunk/build.xml)

4. Other Documentation

In addition to this document, other documents are provided describing the dataset as it transitions from a research dataset to operations within the NCEI framework. These documents are required for an Operational Readiness Review (ORR). A brief description of each is included below.

4.1 Submission Agreement

A submission agreement is written and reviewed by different sectors of NCEI, including archive, access, customer service, and IT. This agreement is required in order for all data to be archived by NCEI. A Data Set Readiness Review (DSRR) briefing is held before the ORR, and all sectors approve the dataset for archive. In addition, a separate briefing is given to the Customer Engagement Branch (CEB) before the ORR.

4.2 Production Plan

The GHCN-Daily data set is updated on a daily basis to incorporate observations from the current and previous day. Full period-of-record updates are made as part of regular processing that takes place once a week. To ensure any additions or changes are captured in the GSOM product, period of record statistics are recalculated once per week and the output files distributed to Archive Branch where they are available for data access. The Production Plan document provides additional information regarding the update system and timing.

4.3 Maintenance and Reprocessing Plan

This document provides the individual steps needed to perform maintenance on the GSOM processing system. This includes activities needed for completion of full period of record reprocessing if there is a need to either add new elements or modify the algorithm of existing elements to make improvements or correct software bugs. This is completed through release of a new version as described in this document.

4.4 Security Report

The software is made available to IT Branch for a security review and a report provided by IT Branch. Any security concerns are addressed before the software is placed in operations.

4.5 V&V Report

This document highlights Verification and Validation (V&V) that has been completed on the GSOM processing system.

5. References

Durre, I., M. J. Menne, B. E. Gleason, T. G. Houston, and R. S. Vose, 2010: Comprehensive automated quality assurance of daily surface observations. J. Appl. Meteor. Climatol., 49, 1615–1633, doi:10.1175/2010JAMC2375.1.

Menne MJ, Durre I, Vose RS, Gleason BE, Houston TG. 2012. An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology* **29**, 897-910, doi: 10.1175/JTECH-D-11-00103.1.

6. Figures

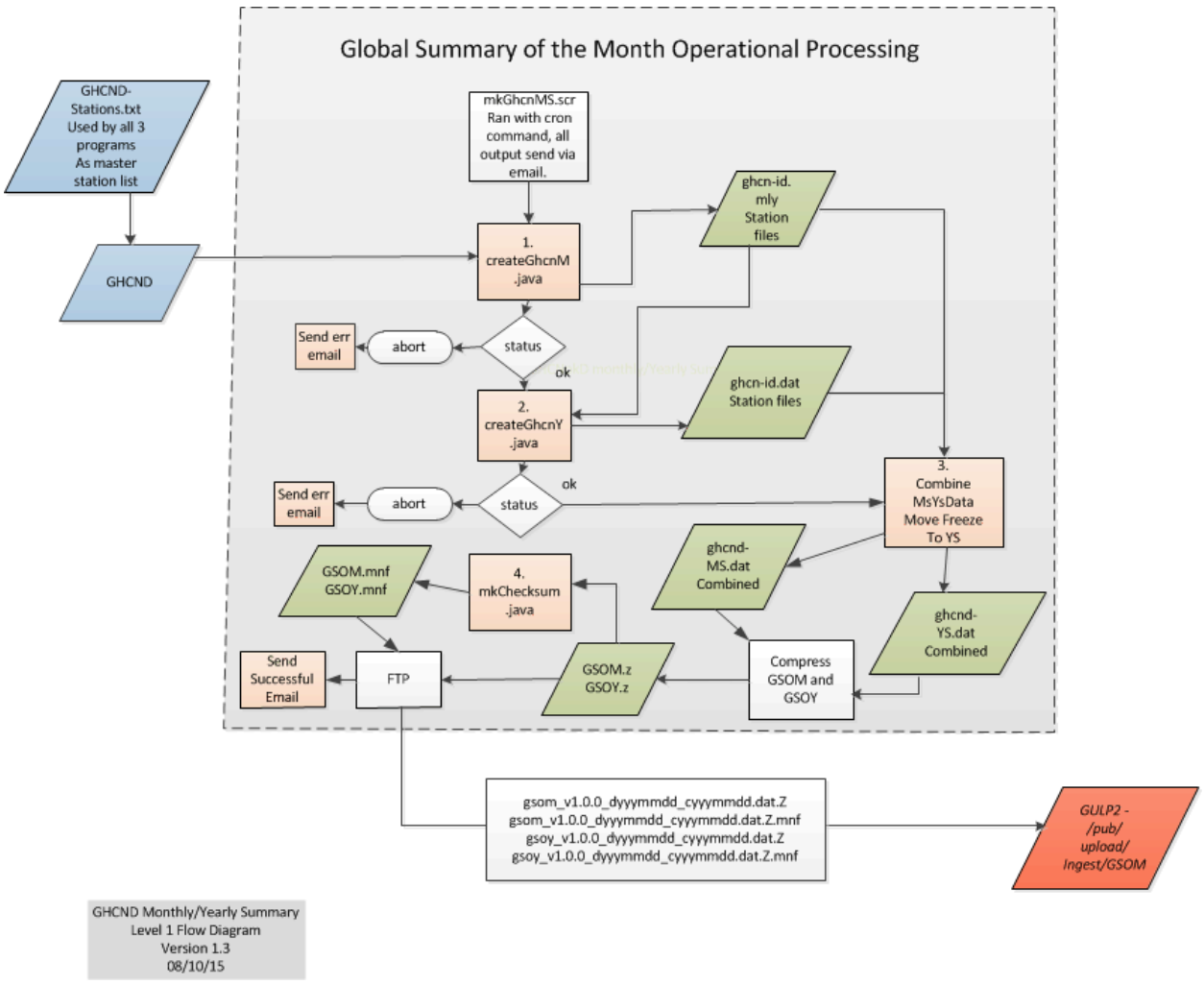

Figure 1. *Level 1 Data Flow Diagram for the Global Summary of the Month/Year data set. (Times of completion are included in Section 3.1.)*