

National Climatic Data Center

DATA DOCUMENTATION

FOR

DATASETS 3506-08

(DSI's-3506, -3507, -3508)

Daily U.S. Integrated Data

March 2, 2004

National Climatic Data Center
151 Patton Ave.
Asheville, NC 28801-5001 USA

⋮
⋮

Table of Contents

Topic	Page Number
1. Abstract.....	3
2. Element Names and Definitions:	4
3. Start Date.....	6
4. Stop Date.....	6
5. Coverage.....	6
6. How to order data.....	7
7. Archiving Data Center.	7
8. Technical Contact.....	7
9. Known Uncorrected Problems.....	7
10. Quality Statement.....	7
11. Essential Companion Data Sets.....	11
12. References.....	11

1. **Abstract:** Recognizing that the original data as received by NCDC may contain errors or suspect values, it is beneficial for users that the quality of the data be evaluated. Data validation is dependent on the kind of information inherent in the observation rather than on the network generating the data. Similar kinds of data (e.g., daily data observed at coop, first order, etc. sites) should and will be treated together in an integrated manner rather than independently. Therefore, it will not be necessary to separately verify, as is now done, sets of data from individual networks. Similar kinds of data will be verified together with the same rules and algorithms.

An iterative, linear system is being developed at NCDC to assess the quality of daily data. Conceptually, the first step in the linear process is to integrate daily data from all sources into one data set. The second step is to validate that the information from each data source conforms to the observing and processing rules that apply to the source data. Third, meteorological and physical consistency is assessed for individual, point observations. Fourth, data values are evaluated in a spatial context. Fifth, the data are evaluated in a temporal context. The last step is the resolution of differing values from multiple sources of an element at a given time and place. The iterative concept involves solving (if possible) problems at each step and then reprocessing the data before proceeding to the next step.

The goals of the development are:

- Providing a baseline, quality assessed, daily data set to users
- Consistency of quality assessment across all surface data types
- Consistency of data provided to users (users will not get two different values for a parameter or product)
- Providing a baseline data set for summarizing over longer time intervals (e.g., monthly) thus leading to temporal consistency among different data sets
- Reducing the chances for errors and inconsistencies between data sets that span multiple observing platforms
- Standard 'platform'/format for servicing software, data summarization, visualization, climate monitoring, etc.
- Reduction in quality assessment development and maintenance costs (fewer systems will be built)

The process is intended to be dynamic, flexible and open-ended so that data updates, changes to the algorithms, inclusion of additional algorithms, elimination of algorithms, ordering of the linear process, and other similar maintenance efforts can be accomplished easily. It was therefore decided that a modular approach would be best; software modules, data tables, etc. are independent but linked. The process is also intended to accept input data and algorithms from sources other than NCDC, such as the Regional Climate Centers, State Climatologists, etc.

Historical data for the U.S. through 2002 from nine separate data sets have been reformatted and merged into an integrated set. These nine separate sets are:

1. Summary of the Day from the Cooperative Observer Network (3200)
2. Preliminary Summary of the Day from the Cooperative Observer Network

:
:

- (3202)
3. Summary of the Day from the Cooperative Observer Network from the Midwest Climate Center (3205)
 4. Summary of the Day from the Cooperative Observer network from the Data Rescue Project (3206)
 5. Summary of the Day from the First Order Network (3210)
 6. Summary of the Day from the from the ASOS Network (3211)
 7. Summary of the Day from Air Force stations
 8. Centennial Data Collection
 9. SNOTEL data (6430)

Documentation of the source data sets with a number in parentheses is available from the NCDC website (search for dataset documentation). The Air Force data were obtained from the U.S. Air Force Combat Climatology Command, and the Centennial data collection was compiled from information sent to the NCDC as part of the 100 year anniversary of weather services.

The data from these nine sets were processed through automated "format" checks, and numerous systematic errors were fixed. They were then processed through the consistency checks. There are now three versions of the integrated data: raw, format checked, and consistency checked. Detailed descriptions of these first three steps in the linear process are described in detail in the following sections.

2. Element Names and Definitions:

Not knowing where the development would lead, it was decided to define a format for the integrated data set that would be simple, flexible, allow for future modification, and allow for inclusion of additional source data sets. The resulting format consists of two parts: 1) identification and 2) data values. The identification part is the first 19 characters of the integrated record format:

<i>Column</i>	<i>Description</i>
1 through 6	Station identifier
7 through 10	Year
11 through 12	Month
13 through 14	Day
15 through 18	Element
19	Identification flag

The second part is the rest of the variable length record:

<i>Column</i>	<i>Description</i>
20 through 21	Number n of data fields (12 characters each) to follow
22+12($n-1$)	Data source code
23+12($n-1$) through 24+12($n-1$)	Observation time
25+12($n-1$) through 30+12($n-1$)	Data value (signed in column 25 if applicable)
31+12($n-1$)	Data measurement flag
32+12($n-1$)	Data quality flag
33+12($n-1$)	Quality assurance flag

All of the columns are defined as character rather than as integer, real, etc.

:
:

thereby allowing for non-numeric values. The source codes correspond the numbered list is section 2 above. The identification and quality assurance flags are initially set to "blank". The element as well as the data measurement and data quality flags are carried forward as they appear in the source data sets, and their values are defined in the documentation for the source data set. One exception is the element name in the SNOTEL data set. For this data set, lower case names are capitalized (conformance with all the other element names), the element "prec" is changed to "PRCP" ("prec" is equivalent to "PRCP" in all of the other data sets, and a blank element name was found to be accumulated precipitation for a water year beginning October 1 and was arbitrarily changed to "PCPY" in the integrated data set.

Some changes were made to the NCDC Summary of the Day from the Cooperative Observer Network (3200) data holdings based on information provided to NCDC by the Midwest Climate Center (Illinois, Indiana, Iowa, Kentucky, Michigan, Minnesota, Missouri, New Mexico, Ohio and Wisconsin data) and the Oklahoma Climate Survey (Oklahoma data). The information provided by these agencies is reflected in replacement values with a data quality flag of "S".

Information in each of the source data sets was converted and merged into the integrated format. Each record in the raw integrated data set therefore contains all the observations from all of the source data sets for a given element, time and place. Note that if a source data set contained both observed and replacement values, then both of these values have been carried forward into the raw data set; flags associated with the source data set identify whether a value is observed or is a replacement.

The integrated data station files are stored within a state subdirectory of the raw main directory.

Using the raw data as input, an initial version of the format checks, as well as several later iterations, indicated that there were non-meteorological problems with the data. As a result, a driver program was written to correct some of the defects prior to performing data checks. These defects are:

- Converting negative temperature ranges to positive values for Data Rescue source data
- Converting the -99 missing data code in the Midwest Climate Center source data to -9999
- Conversion of SNOTEL temperatures from Celsius to Fahrenheit
- Convert wetbulb temperatures beginning on July 1, 1996 from whole degrees to tenths of degrees
- Convert water equivalent data beginning on April 11, 1970 from tenths to hundredths

The driver program was written so if additional systematic problems are found, the correction software can be easily inserted.

A variety of checks are performed through subroutines and tables to allow for easy modification. They are:

- Read errors.
- Proper record lengths.
- Duplicate data fields (duplication of all 12 characters). Duplications were found most often in the Summary of the Day from the First Order

:
:

Network (3210) source data.

- Range checks to ensure that a data value falls within the range specified by the source documentation.
- Valid years, months and days
- Valid source data sets
- Valid elements for a given source data set
- Valid number of data fields for a given source data set
- Valid flags for a given source data set

These checks are not performed when a record contains one data field and the source data set is the Preliminary Summary of the Day from the Cooperative Observer Network (3202). These records are excluded because they are part of an initial, error prone, data set that are the initial input for operational processing at NCDC. Further operational processing puts the quality assured information into the Summary of the Day from the Cooperative Observer Network (3200) data.

Short period records are identified, where short period is defined by a period of record of one month or less or by one month of data at the end of the period of record that is separated by a year from the rest of the data. The short records are likely misidentified station identifiers or times.

The output from the format checks assigns the identification and quality assessment flags as follows:

Identification Flag

0	Passed all format checks
1	Failed at least one station identifier, year, month, day or element check
S	Short period record

Quality Assessment Flag

0	Passed all data field checks
1	Failed at least one data field check (except for duplication check)
D	Duplicated data field

The integrated data station files are stored within a state subdirectory of the format checked main directory with one exception. The short period records are stored as station files within one subdirectory of the format checked main directory.

3. **Start Date:** 18800101

4. **Stop Date:** 20021231

5. **Coverage:**

- a. Southernmost Latitude: 25.0S
- b. Northernmost Latitude: 50.0N
- c. Westernmost Longitude: -125.0W
- d. Easternmost Longitude: -65.0E

:
:

6. **How to Order Data:**

Ask NCDC's Climate Services about the cost of obtaining this data set.

Phone: 828-271-4800

FAX: 828-271-4876

E-mail: NCDC.Orders@noaa.gov

7. **Archiving Data Center:**

Archive Branch
National Climatic Data Center
151 Patton Avenue
Asheville, NC 28801

8. **Technical Contact:**

National Climatic Data Center
151 Patton Avenue
Asheville, NC 28801

9. **Known Uncorrected Problems:** None.

10. **Quality Statement:** The quality assessment program inputs an entire month of format checked data for a station. Call routines perform the consistency checks. The input matrix is declared to accommodate 28 elements and 39 days (last 4 days of previous month, 31 days of current month, and first 4 days of next month). The matrix is initialized with a missing value code of "-99999.0" and then overwritten with the input data.

The quality assessment program first checks to see if the identification flag is "1". If the flag is "1", the data are not checked but the record is written to the output files. The program next looks at the data fields for a missing value code. If the data value is missing and the data measurement flag is not "S" (included in a subsequent value), then the quality assessment flag is set to an error code since the input data sources are not supposed to contain data values for missing data.

a. Extremes

The data are then checked against a table of extreme values for the following elements:

PRCP	Total Precipitation
F2MN	Faster 2 minute wind speed
TMAX	Maximum Temperature
TMIN	Minimum Temperature
TOBS	Temperature at observation time
TAVG	Average Temperature
F5SC	Fastest 5 second wind speed
AWND	Average Wind Speed
FSMI	Fastest Mile (ddfff)
FSMN	Fastest One-minute Wind (ddfff)
PRES	Station Pressure
RWND	Resultant Wind Speed

:
:

SLVP	Sea Level Pressure
TMPW	Wet Bulb Temperature
FSIN	Fastest Instantaneous Wind (ddfff)
WDMV	24_hour Wind Movement
MNTP	Average Temperature
DPTP	Dew Point Temperature

Any data value that fails the extremes check is assigned a quality assessment flag of "2".

Station extremes are calculated from the format checked data. All of the source data are used to calculate station extremes except those from SNOTEL and the Preliminary Summary of the Day from the Cooperative Observer Network (3202) data, which have numerous data problems. Also, any data with a failed identification flag ("1"), that is the beginning of or end of an accumulation period, is invalid (quality assessment flag of "3"), represents missing data, is a replacement value, or is a short period station are not used in calculating extremes.

If at least 100 values exist for an element-month, station extremes are calculated by fitting the data to a Wakeby probability model. This 5 parameter model yields an excellent fit to the data. The values of zero-bounded elements, such as precipitation, are fit by a mixture

$$P(x) = P(x=0) + P(x>0)$$

where $P(x)$ is the probability of x , $P(x=0)$ is the relative frequency of zero values, and $P(x>0)$ is determined from the Wakeby distribution. Thresholds used for extremes are values corresponding to probabilities of .005 and .995. If data cannot be fit to the model, empirical relative frequency curves are constructed, and thresholds are integer values corresponding to relative frequencies of .005 and .995.

The fallback extreme threshold for instances when the above procedure does not work (e.g., insufficient data) is a statewide extreme. The statewide thresholds are the 10th highest and lowest values of an element in a month that has been observed within the state. These values provide a gross check to flag extraordinary data values.

The last column of the station extremes tables identifies status codes for the extremes. 0 indicates that the Wakeby model was successful, 1-4 indicates unsuccessful modeling and 9 indicates statewide extremes were used.

b. Consistency Checks

The following consistency checks are performed (failure relationships):

- Spike check for all temperature elements except temperature range:
 $|\text{temp}(\text{day } 1) - \text{temp}(\text{day } 2)|$ and $|\text{temp}(\text{day } 2) - \text{temp}(\text{day } 3)| \geq 9$
- Flat line:
 $\text{temp}(\text{day } 1) = \text{temp}(\text{day } 2) = \text{temp}(\text{day } 3) = \text{temp}(\text{day } 4) = \text{temp}(\text{day } 5)$
- Maximum temperature:
 $\text{TMAX} \leq \text{TAVG}$ (average)
 $\text{TMAX} \leq \text{MNTP}$ (mean)

:
:

$TMAX \leq TOBS$ (observation time temperature)
 $TMAX \leq OT07$ (observation time temperature 0700)
 $TMAX \leq OT14$ (observation time temperature 1400)
 $TMAX \leq OT21$ (observation time temperature 2100)
 $TMAX \leq DPTP$ (dewpoint)
 $TMAX \leq TMPW$ (wet bulb)
 $TMIN$ (day 1) $\geq TMAX$ (day)

- Minimum temperature:
 - $TMIN \geq TAVG$ (average)
 - $TMIN \geq MNTP$ (mean)
 - $TMIN \geq TOBS$ (observation time temperature)
 - $TMIN \geq OT07$ (observation time temperature 0700)
 - $TMIN \geq OT14$ (observation time temperature 1400)
 - $TMIN \geq OT21$ (observation time temperature 2100)
 - $TMIN \geq DPTP$ (dewpoint)
 - $TMIN \geq TMPW$ (wet bulb)
- Mean Temperature:
 - $MNTP \leq TMPW$ (wet bulb)
 - $MNTP - \text{integer}[(TMAX + TMIN)/2] \neq 0$
 - $TAVG - \text{integer}[(TMAX + TMIN)/2] \neq 0$
- Dew point:
 - $DPTP \geq TMPW$ (wet bulb)
 - $DPTP \geq MNTP$ (mean)
- Temperature Range:
 - $TRNG - (TMAX - TMIN) \neq 0$
- Degree Days:
 - $HTDG \neq 65 - MNTP$ (mean) if $MNTP < 65$
 - $HTDG \neq 0$ if $MNTP > 64$
 - $CLDG \neq MNTP$ (mean) - 65 if $MNTP > 65$
 - $CLDG \neq 0$ if $MNTP < 66$
- Precipitation:
 - $PRCP = 0$ or T (trace) and $SNOW \neq 0$
 - $PRCP = 0$ and $TMAX = TMIN$
- Water Equivalent:

:
 :

WTEQ > 0 and SNWD (snowdepth) < 2
WTEQ = 0 and SNWD (snowdepth) ≥ 2

- Wind Speed:
 - FSIN (instantaneous) ≤ F5SC (5 second)
 - FSIN (instantaneous) ≤ FSMN (1 minute)
 - FSIN (instantaneous) ≤ F2MN (2 minute)
 - FSIN (instantaneous) ≤ AWND (average scalar)
 - FSIN (instantaneous) ≤ RWND (average vector)
 - F5SC (5 second) ≤ FSMN (1 minute)
 - F5SC (5 second) ≤ F2MN (2 minute)
 - F5SC (5 second) ≤ AWND (average scalar)
 - F5SC (5 second) ≤ RWND (average vector)
 - FSMN (1 minute) ≤ F2MN (2 minute)
 - FSMN (1 minute) ≤ AWND (average scalar)
 - FSMN (1 minute) ≤ RWND (average vector)
 - F2MN (2 minute) ≤ AWND (average scalar)
 - F2MN (2 minute) ≤ RWND (average vector)
 - AWND (average scalar) ≤ RWND (average vector)
 - FSMI (fastest mile) ≤ AWND (average scalar)
 - FSMI (fastest mile) ≤ RWND (average vector)

The checks examine all observed data as well as all combinations of observed and replacement data (if they exist).

Failures of checks that examine either all observed or all replacement data generate a quality assessment code of "3". However, if a data value fails both the extreme check as well as a consistency check, then the quality assessment is set to "4". It is important to note that if **any** single-day temperature consistency check fails, then **all** temperature and temperature dependent (e.g., degree days) values for the day are flagged. This scheme was adopted because of the interdependencies among all of the temperature elements. Similarly, if any wind check fails, all winds are flagged. Failure of the precipitation/temperature check generates a failure flag for all temperatures, precipitation, and snow. Failure of the water equivalent check generates flags for both water equivalent and snow depth. Multi-day check failures are treated the same way as single-day checks in that values are flagged for all days included in the check. An exception is the failure of the spike check; values are flagged only for the day of the spike.

Failures of checks that examine combinations of observed and replacement data are treated somewhat differently. For these checks, only the replacement values of the elements being checked are flagged when a failure occurs. Note however, that the observed data may be assigned a failure flag because of the scheme described in the previous paragraph even though the combination of observed and replacement values passes a particular check.

:
:

The format checks are a relatively straightforward approach to verifying that the source data conform to the rules described in the documentation. By necessity, the checks are data set dependent rather than generic. From the first iteration to the current iteration of these checks, numerous systematic errors were fixed. The process has shown that the software provides an excellent starting point for data cleanup. Excluding the Preliminary Summary of the Day from the Cooperative Observer Network (3202) and the NCDC SNOTEL data set, the number of error messages from the latest run of the format checks is slightly under 11,000. This number reflects a very small percentage of the total of about 45gB of data. The majority of the errors is station specific and is either improper units resolution or what appear to be improper missing codes. The Preliminary Summary of the Day from the Cooperative Observer Network (3202) data is known to be full of format errors, but these were ignored because the period of record is short and because the information is eventually included in other source data sets. Despite the known poor quality, the data are included mainly to use whatever "good" data are available as confirmation of information from other sources. The NCDC SNOTEL data are generally archived as received. Historical data were obtained from USDA in about 1997. Improving SNOTEL data at USDA is a current priority, and it is expected that the SNOTEL data housed at USDA will eventually replace the SNOTEL data in the integrated data set.

The philosophy behind the quality assessment performed to date is simply to identify suspect data based on meteorological and climatological principles and experience. Since the observers' or automated observing systems' values are the official representation of the weather occurring at a place and time, and because we were not present when the observation was taken, no attempt is made to determine "correct" values. We can only identify what must be wrong or what might be wrong; we cannot determine what is right. A practical consequence of this philosophy is the treatment of the flagging of replacement values. The algorithms look at how the replacements interact with the observed values rather than as treating the replacements as observed values. Another practical consequence is to intentionally identify more suspicious data than may be truly suspicious. An example is the flagging of all temperature elements for a day when one temperature check fails. This overflagging occurs because we expect the analyst to look at the interdependence among elements and to look at all values that could be affected by one suspicious or wrong value. Also consistent with this philosophy is the decision not to quality assess the short period records; if the identifier is uncertain, then the data values have no meaning.

A corollary to this philosophy is the current willingness to incorporate assessment tools and experiences of the climatological community. Sharing of algorithms, ideas, and experiences will lead to improvement of the process and to a much better baseline data set. A dialog among Regional Climate Centers, NCDC, USDA and others is ongoing and hopefully will continue far into the future.

The current quality assessed integrated data set should be considered as an interim product. Spatial and temporal checks need to be included along with additional consistency checks. The current consistency checks are weighted heavily towards temperature elements, so more checks are needed for the other elements. Examples are multi-day relationships among precipitation, snowdepth and snowfall; SNOTEL data consistency; evaporation checks; and soil temperature and depth consistency.

:
:

Format errors should be corrected if possible, the identification section of the short period records should be verified or corrected, and data flagged as suspicious should be analyzed. After modifications are made from this review, the data should be iteratively reprocessed and modified. Additional source data sets, such as the USDA SNOTEL data, should also be integrated into the data set. Update software will therefore be needed to include the additional data and to modify existing records.

11. **Essential Companion Datasets:**

12. **References:**

:
: