National Climatic Data Center

DATA DOCUMENTATION

FOR

DATA SET 9300 (DSI-9300)

Global Historical Climatology Network - Daily, V1.0

October 18, 2004

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

Table of Contents

1. Abstract: The Global Historical Climatology Network (GHCN) - Daily represents a compilation of global daily data into a single and consistent format. This dataset is a continuation of the Global Daily Climatology Network (GDCN, TD9101) which is no longer under development. The data set will serve the needs of researchers, weather-sensitive businesses, agriculture, and policy makers whose are dependent upon complete and accurate analysis of daily temperature and precipitation. Data within the GHCN-Daily have been extensively checked through a series of quality control procedures to ensure erroneous values have been removed and/or identified.

2. Element Names and Definitions:

Access Method for Data Files:

All elements available for a single station are stored in an ASCII format data file. These data files can be accessed using a variety of programming languages and/or software (e.g. spreadsheets). The data format (per one line) is as follows:

Variable	Type	Width	Start Column	End Column
STNID	Character	11	1	11
YEAR	Integer	4	12	15
MONTH	Integer	2	16	17
ELEMENT	Character	4	18	21
DATA	Integer	5	22	26
FLAG1	Character	1	27	27
FLAG2	Character	1	28	28

NOTE: DATA, FLAG1, and FLAG2 repeat 31 times for a single record (e.g. one line). Therefore one record of data represents one month of daily data.

Access Method for Inventory Files:

Station information can be found in the ASCII format inventory file. The format of the inventory file is as follows (in Fortran variable notation):

Variable	Type	Width	Start Column	End Column
STNID	Character	11	1	11
LATITUDE	Real	7	12	18
LONGITUDE	Real	8	19	26
ELEVATION	Integer	5	27	31
ELEVATION ESTIMATE FLAG	Character	1	32	32
DATA SOURCE FLAG	Character	1	34	34
STN_NAME	Character	30	36	65
WMOID	Integer	5	66	70
STNID2	Character	8	71	78
STNID3	Character	8	79	86

Element Names and Definitions for Data Files

STATION ID: The 11 character alphanumeric station identification element is comprised of 2 parts ... a 3-digit country code (see Appendix A) and an 8-digit station code. In some cases the last 5 digits of the 8-digit station code are the 5-digit World Meteorological Organization (WMO) number.

Many countries operate weather stations in foreign lands (e.g. possessions, military bases, research stations, etc.) and therefore it is possible for station IDs to have a 3-digit country code that differs from the country where the station is physically located (e.g. Canadian military base located in Europe).

YEAR: 4-digit year

MONTH: 2-digit month

ELEMENT: 4 character weather element that is one of TMAX (max temperature), TMIN (min temperature), or PRCP (total 24 hour precipitation).

DATA (31): There are 31 daily data values (per record) expressed as a 5-digit integer. The missing value is -9999 and non-missing values are in tenths of a degree Celsius for temperature and tenths of a millimeter for precipitation. Therefore, to obtain the whole measurement unit divide the non-missing integer datum by 10.0 to express value as either whole degree Celsius (to the nearest tenth) or whole millimeters (to the nearest tenth).

Example: (1) TMAX value in 5-digit integer format “00302" would be expressed as 30.2 degrees Celsius. (2) PRCP value in 5-digit integer format “01200" would be expressed as 120.0 millimeters.

DATA MEASUREMENT FLAG(31): There are 31 data measurement flags that are expressed as 1 character. These flags occur immediately following each 5-digit integer datum. Data measurement flags are as follows (Note: flags can apply to Tx = Maximum Temperature, Tn = Minimum Temperature, Pr = Total Precipitation):

‘’ - Blank = no data measurement information (Tx,Tn,Pr)
A - accumulated precipitation value (Pr)
C - begin or continuing precipitation accumulation (Pr)
E - estimated value provided by source (Pr)
J - source provided accumulated precipitation value, however no information on accumulation period was provided (Pr)
T - trace accumulation reported by source (Pr)
R - Daily value derived from the hourly data (i.e. highest hourly or synoptic-reported temperature). (Tx, Tn)

QUALITY CONTROL FLAG(31): There are 31 quality control flags that are expressed as 1 character. These flags occur immediately following each data measurement flag. The quality control flags are as follows (Note: flags can apply to Tx = Maximum Temperature, Tn = Minimum Temperature, Pr = Total Precipitation):

‘’ - Blank = no quality control information (Tx,Tn,Pr)

D - Duplicate value. (Tx, Tn, Pr)

X - Exceeds known world daily extreme for that particular element (Krause and Flood, 1997). (Tx,Tn,Pr)

I - Internal consistency error. The minimum temperature value is greater than same day maximum temperature value. (Tx,Tn)

K - Streak value

Note: 0.0 mm precipitation values are excluded. (Tx, Tn, Pr)

G - Gap value (see temperature quality control section). (Tx, Tn)

O - Temperature value is greater than or equal to 4.0 bi-weight standard deviations from the period biweight mean. (Tx,Tn)

M - Manually edited value set to missing. (Tx, Tn, Pr)

It is possible that one or more quality control flags may apply to a single datum. Therefore, the above listed flags are listed in order of precedence. Thus, if a value is a streak (e.g. 'K') and also an outlier (e.g. 'O') then the 'K' flag will be set for that datum. Once a value has a non-blank quality control flag, that datum is excluded from subsequent quality control procedures.

Element Names and Definitions for Inventory File

STATION ID: same as for data file (11 character alphanumeric)

LATITUDE: Expressed in decimal degrees to the nearest hundredth of a degree. (7 digits)

LONGITUDE: Expressed in decimal degrees to the nearest hundredth of a degree. (8 digits)

ELEVATION: Expressed in whole meters (5 digits)

ELEVATION_ESTIMATE_FLAG: If a station had incomplete elevation information, attempts were made to estimate the elevation based on longitude and latitude. The Global Land One-km Base Elevation (GLOBE) data set was used to estimate the elevation (Hastings et. al., 1997). If left blank this field indicates the elevation was provided in the supplied metadata. If the elevation was estimated using GLOBE then this flag is set to 'E'. (1 character)

DATA_SOURCE_FLAG: A one-character identifier indicating the source data network from which the data were supplied. In GHCN-Daily 1.0 this flag is either blank or set to 'G' to indicate Global Climate Observing System Network stations (GCOS) (Peterson et. al., 1997) or set to "S" to indicate a Global Climate Observing System Network station that contains synoptic-transmitted data. (1 character)

STATION NAME: Name of place or location of station. (30 characters)

WMO ID: If supplied in the original metadata this 5 digit integer represents the World Meteorological Organization=s 5 digit station identifier. (5 digits)

STATION ID #2: 2nd supplied identifier from original metadata. This often is a cross reference with another station identifier. (8 characters)

STATION_ID_#3: 3rd supplied identifier from original metadata. This often is a cross reference with another station identifier. (8 characters)

3. Start Date (YRMO): 183301

4. Stop Date (YRMO): 200402

5. Coverage: Global

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:

National Climatic Data Center
Federal Building
151 Patton Avenue
Asheville, NC 28801-5001
Phone: (828) 271-4800.

8. Technical Contact:

National Climatic Data Center
Federal Building
151 Patton Avenue
Asheville, NC 28801-5001
Phone: (828) 271-4800.

9. Known Uncorrected Problems:

An undetermined number of stations within Mexico have recorded some of their temperature observations in Fahrenheit. No pattern has yet been identified, so users should pay particular attention to the temperature quality control flags. These flags may or may not be useful in identifying unit conversion issues.

About 400 India stations have a unit-conversion issue for precipitation throughout the period of record.

Please check www.ncdc.noaa.gov/ghcn.html for the latest information on corrections and available supplementary files.

10. Quality Statement: All of the GHCN-Daily data (metadata and data) have been processed through an extensive set of quality control procedures. The quality control consisted of two parts:

(1) Simple datum checks (e.g. correct format, impossible values, out of range values, etc.).

(2) Statistical analysis of sets of observations to locate and identify potential outliers and/or erroneous data.

The user is referred to Appendix C for a complete list of all quality control procedures used in step 1 (above). More extensive explanation for step 2 is provided below for both temperature (maximum and minimum) and precipitation.

Temperature QC:

The data for TMAX and TMIN elements are run through a variety of quality control procedures (in order):

1) Simple Format checks (see Appendix C).

2) Duplicate Checks: The three duplicate checks that are done are as follows:

A) Check for different years that have the same data.
B) Check for different months within the same year that have the same data.

C) Check for the same months within different years that have the same data.

Duplicate day quality control flag = “D”

3) Bounds Checks. (See Appendix C).

4) Streak Check. (See Appendix C).

5) Gap Check: For a given station-element (TMAX or TMIN) frequency distribution, the median of the distribution is found, and then moving both right and left from the median, an algorithm identifies any “gaps” in the frequency distribution of 10.0 degrees Celsius or more that separate one or more values from the main distribution. The idea is that such a gap would represent outlier values of interest to dataset users for potential exclusion.

6) Outlier Check: Users are referred to the technique used in the Global Daily Climatology Network (GDCN) Version 1.0 (TD9101, from National Climatic Data Center) and to Lanzante (1996) for an overview of biweight mean and biweight standard deviation calculations.

Specifically for the Global Historical Climatology Network – Daily Version 1.0, TMAX and TMIN values were flagged as outliers (“O”) if the corresponding datum was >= 4.0 biweight standard deviations from the “window” biweight mean. The window biweight mean is defined as all values included within the 15 day window centered on the date of the datum, for all years within the station record. The biweight standard deviation is also calculated using this same window.

7) Manually Edited values: For data that is determined to have known problems, but for which no automated algorithm has been developed to identify the erroneous data, an “M” flag is given to the data and each datum is set to the dataset missing value.

Precipitation QC:

The precipitation quality control is the same as the temperature quality control with the following exceptions:

A) Appendix C outlines the Bounds and Streak Check differences.

B) No “Gap” and “Outlier” check is performed on the precipitation data.

Duplicate Stations:

With millions of days of daily data, direct comparison of all station data to all other station data was not feasible due to computational time limitations. However, a station to station duplicate identification check was developed that compares the 15th day (for all years) of each month to the 15th day of each corresponding month. The comparison was only done between the same element (e.g. compare STN A TMIN to STN B TMIN). The purpose of this test is to identify potential station pairs that have 100% of the same data on the 15th day of each month for all non-missing days. The test requires that at least 60 comparisons (12 months x 5 years) of data are present before a duplicate can be declared. Appendix F lists the station pairs and the element that are potentially duplicated.

11. Essential Companion Datasets: None.

12. References:

Hastings, David A., and P. K. Dunbar, 1997. The development of global digital elevation data. Proceedings, 18th Conference on Remote Sensing, Kuala Lumpur. pp. J=S-3-1 to J=S-3-6.

Krause, Paul F. and Kathleen Flood. (1997). Weather and Climate Extremes (TEN-0099). Fort Belvoir, VA: U.S. Army Engineer Topographic Engineering Center).

Lanzante, John R. 1996. Resistant, robust, and nonparametric techniques for the analysis of climate data. Theory and examples, including applications to historical radiosonde station data. Int. J Climatol., 16, 1197-1226.

Peterson, Thomas, Harald Daan, Philip Jones, 1997: Initial Selection of a GCOS Surface Network. Bulletin of the American Meteorological Society: Vol. 78, No. 10, pp. 2145B2152.

Appendix A

3-digit country codes (used as the first 3 digits of the 11 character station ID):

101 – Algeria
102 – Angola
103 – Benin
104 – Botswana
105 – Burkina Faso
107 – Cameroon
110 – Chad
112 – Congo
113 – Cote d'Ivoire
115 – Egypt
116 – Eritrea
117 – Ethiopia
118 – Gabon
122 – Kenya
124 – Libya
125 – Madagascar
126 – Malawi
127 – Mali
128 – Mauritania
129 – Mauritius
130 – Morocco
131 – Mozambique
133 – Niger
137 – Senegal
138 – Seychelles
139 – Sierra Leone
141 - South Africa
148 – Sudan
149 – Tanzania
152 – Tunisia
155 – Zambia
156 – Zimbabwe
201 – Afghanistan
202 – Bahrain
205 - China
207 – India
208 – Iran
209 - Iraq
210 - Japan
213 - Kyrgyzstan
215 – Mongolia
218 – Oman
219 - Pakistan
221 – Republic of Korea
222 - Russia
223 - Saudi Arabia
224 – Sri Lanka
227 - Tajikistan
228 - Thailand
229 – Turkmenistan
230 – United Arab Emirates
231 – Uzbekistan
232 - Vietnam
301 – Argentina

302 - Bolivia
303 – Brazil
304 – Chile
305 - Colombia
306 - Ecuador
308 – Paraguay
309 - Peru
313 - Uruguay
314 – Venezuela
315 – French Guiana
401 - Barbados
403 - Canada
405 - Costa Rica
408 - El Salvador
410 - Guatemala
412 - Honduras
413 - Jamaica
414 - Mexico
415 - Nicaragua
416 – Panama
419 – Stations Operated by France
423 - Bahamas
425 - United States of America
426 - Greenland
501 – Australia
502 – Fiji
503 – Indonesia
504 - Kiribati
505 – Malaysia
507 – New Zealand
508 – Papua New Guinea
509 – Phillipines
512 – Soloman Islands
514 - French Polynesia
517 – Tonga
518 – Tuvalu
520 – Vanuatu
531 - New Caledonia
541 - Samoa
602 - Armenia
603 - Austria
604 - Azerbaijan
605 - Belarus
606 - Belgium
607 – Bosnia
611 – Czech Republic
612 - Denmark
613 - Estonia
614 - Finland
615 - France
616 - Georgia
617 - Germany
618 – Greece
619 - Hungary
620 - Iceland
621 - Ireland
622 - Israel
623 - Italy
625 - Kazakhstan
626 - Latvia
628 - Lithuania

629 – Luxembourg
630 - Malta
631 - Moldova
633 - Netherlands
634 - Norway
635 - Poland
636 – Portugal
637 - Romania
641 - Slovakia
642 - Slovenia
643 - Spain
645 - Sweden
646 - Switzerland
647 – Syria
648 – Former Yugoslav Republic of Macedonia
649 - Turkey
650 – Ukraine
651 - United Kingdom

Appendix B

The National Climatic Data Center staff listed below were responsible for obtaining GHCN-Daily data:

Dr. Pavel Ya. Groisman
Visiting UCAR Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Dr. Thomas C. Peterson
Physical Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Dr. David R. Easterling
Principal Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Mr. Russell S. Vose
Chief, Climate Analysis Branch
National Climatic Data Center
Asheville, NC 28801-5001

Appendix C (Format and Bounds Checks)

Data and station information format quality control:

Note: the usage below of “invalid” means checks for improper characters (e.g. TMAX value of “34!*.3”)

1. Check for invalid station latitude

2. Check for invalid station longitude

3. Check for station -90.0 <= Latitude <= 90.0
                     -180.0 <= Longitude <= 180.0

4. Check for invalid elevation

5. Check for -408 meters <= Elevation <= 8850 meters

6. Check if the 4-digit year element within the station data file is: 1830 <= YEAR <= 2010.

7. Check that month field is 1 <= MONTH <= 12.

8. Check for invalid elevation estimation flag.

9. Check for invalid data source flag

10. Check for invalid WMO ID

11. Check whether corresponding data file exists

12. Check for invalid year

13. Check for invalid month

14. Check for invalid element

15. Check for invalid datum

16. Check for invalid data measurement flag

17. Check for invalid quality control flag

14. Check for precipitation < 0.0 mm.

14. Check for maximum temperature < minimum temperature for the same day (quality control flag = “I”).

16. Check for daily element value exceeding known world extreme according to the U.S. Army Corps of Engineers World Extremes (Krause and Flood, 1997). Quality control flag = “X”.

World Extremes used for this data set:

highest temperature: El Azizia, Libya 57.8 deg C, September 13, 1922

lowest temperature: Vostok, Antarctica -89.4 deg C, July 21, 1983

greatest 24-hr rainfall: Foc-Foc, Réunion Island 1828.8 mm, January 7th to 8th,

1966

17. Check whether day exists (e.g. April 31st)

18. Check for streaks, or consecutive runs of the same value, within each element. For TMAX and TMIN, all streaks of 15 or more days (10 or more days for PRCP) are identified by setting the quality control flag to “K” for each day within the streak.

Appendix D

Global Historical Climatology Network - Daily 1.0 Use and Distribution LICENSE

The data and metadata within the Global Historical Climatology Network – Daily Version 1.0 are governed by the following:

"The following data and products have conditions placed on their commercial use. The non-U.S. data cannot be redistributed within or outside of the U.S. for any commercial activities."

Appendix E

Acknowledgments:

Below is a list (not necessarily complete) of people and/or institutions who assisted in the construction of this data set by either directly or indirectly providing data and/or advice.

Mr. Red Ezell
Computer Programmer
National Climatic Data Center
Asheville, NC 28801-5001

Dr. David R. Easterling
Principal Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Mr. Russell S. Vose
Chief, Climate Analysis Branch
National Climatic Data Center
Asheville, NC 28801-5001

Mr. Brant Liebmann
Research Scientist
Climate Diagnostics Center
Boulder, CO 80305-3328

Dr. Mark Morrissey
Environmental Verification and
Analysis Center (EVAC)
Norman, OK 73069

ANEEL (Agência Nacional De
Energia Elétrica) - Brazil
www.aneel.gov.br

Mr. Robert Morris
Manager, Information Svcs. Div.
Meteorological Service of Canada
Environment Canada
Toronto, Ontario
M3H 5T4

Dr. Imke Durre
Physical Scientist
Climte Analysis Branch
National Climatic Data Center
Asheville, NC 28801-5001

Dr. Thomas C. Peterson
Physical Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Dr. Pavel Ya. Groisman
Visiting UCAR Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Mr. Roy Jenne
Manager, SCD Data Support Sec.
NCAR
Boulder, CO 80307-3000

Dr. Arthur V. Douglas
Dept. of Atmospheric Science
Creighton University
Omaha, NE 68178

KNMI (Koninklijk Nederlands
Meteorologisch Instituut)
www.knmi.nl/samenw/eca
The Netherlands

National Climate Center
China Meteorological Admin.
Beijing, China

Mr. Johan Koch
South African Weather Service
Pretoria, South Africa

Appendix F

Potential Duplicate Station Pairs:

14103355200	14103355500	PRCP
22200034928	22200034929	PRCP
23100038450	23100038457	PRCP
30300339028	30328713550	PRCP
30300438011	30328822680	PRCP
30300438021	30328836050	PRCP
30300438022	30328930310	PRCP
30300442010	30327843220	PRCP
30300537019	30338152490	PRCP
30300539037	30338111290	PRCP
30300543010	30337139880	PRCP
30300638007	30338325070	PRCP
30300638014	30338228320	PRCP
30300735009	30338490060	PRCP
30300737006	30338443130	PRCP
30300737023	30338545710	PRCP
30300737027	30338547040	PRCP
30300835068	30338692420	PRCP
30300935012	30338886150	PRCP
30300937013	30338847060	PRCP
30300937018	30338954180	PRCP
30301037049	30348153590	PRCP
30301040002	30347190380	PRCP
30301138002	30348335340	PRCP
30301339000	30348712010	PRCP
30301339044	30348712870	PRCP
30301340003	30347797860	PRCP
30301439039	30348911520	PRCP
30301540004	30357084760	PRCP
41400003074	41400076405	TMAX
41400003074	41400076405	TMIN
41400019052	41400076393	TMAX
41400019052	41400076393	TMIN
42500091334	42500914851	TMAX
42500091334	42500914851	TMIN
42500091376	42500914460	TMIN
42500166244	42500166246	TMAX
42500166244	42500166246	TMIN
42500166659	42500166664	TMAX
42500166659	42500166664	TMIN
61600037189	61600037260	PRCP
61600037375	61600037379	PRCP
62500036705	62500036905	PRCP
62600026422	62600026423	PRCP
65000033355	65000033356	PRCP