THE GLOBAL HISTORICAL CLIMATOLOGY NETWORK: A PREVIEW OF VERSION 2

Russell S. Vose*
Carbon Dioxide Information Analysis Center / Environmental Sciences Division
Oak Ridge National Laboratory

Thomas C. Peterson
Global Climate Laboratory
National Climatic Data Center

Richard L. Schmoyer
Mathematical Sciences Division
Oak Ridge National Laboratory

Jon K. Eischeid
Cooperative Institute for Research in Environmental Sciences
University of Colorado

1. INTRODUCTION

Instruments that could reliably measure temperature, precipitation, and pressure were developed by the late 17th and early 18th centuries. The use of these instruments grew rapidly during the 19th and 20th centuries until virtually all countries operated observing networks consisting of large numbers of stations. The spread, or diffusion, of these instruments as well as their present-day distribution is closely tied to commerce, the growth of urban and agricultural centers, the establishment of aviation networks, etc.—in other words, to those human endeavors that are most affected by the weather.

It has been estimated that weather records have been collected at one to two hundred thousand locations since those first instruments were placed in the field (F. Wernstedt, pers. comm., 1993). Numerous applications, from global change studies to climate impact assessments to general circulation models, make use of such historical records (e.g., Diaz et al., 1989; Jones et al., 1989). These records are an integral part of modern climatic research.

Given their importance, it is unfortunate (but not necessarily surprising) that one cannot approach a single researcher or data center to acquire all of the records for all of the stations, or even a large portion of them. In fact, if one were to review the literature, one would notice the existence of dozens of such sets. To which would one turn? The World Weather Records (NESDIS, 1991), which contains ~2000 stations? The World Monthly Surface Station Climatology (Spangler and Jenne, 1992), which contains ~4000? Perhaps some other data set (e.g., Eischeid, 1991)? Perhaps a combination thereof? And how good is any one of these sources? Few researchers/organizations have the time or resources to answer such questions.

**Corresponding author address*: Russell S. Vose, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6335

In 1990, the Carbon Dioxide Information Analysis Center (CDIAC), the National Climatic Data Center (NCDC), and the World Meteorological Organization (WMO) undertook a collaborative effort aimed at solving this problem. The goals of the initiative are:

1. Data acquisition: identifying/acquiring high-quality digital archives and digitizing manuscript records from data-sparse areas.
2. Quality control: eliminating gross errors in data records and noting values that are possibly erroneous.
3. Data adjustment: creating time series free of artificially induced discontinuities.
4. Dissemination: distributing the data to anyone who wants it, any way they want it, free of charge.

The initiative completed its first data product, known as the Global Historical Climatology Network (GHCN) version 1.0, in 1992 (Vose et al., 1992). This data base contains quality-controlled monthly climatic time series from 6039 temperature, 7533 precipitation, 1883 sea level pressure, and 1873 station pressure stations located on global land areas. Relative to most data sets of its type, GHCN version 1.0 is larger and has better spatial and temporal coverage. Over 500 copies have been distributed since its creation. However, given the volume of data that is known to exist, the data set only begins to address the problem at hand.

This paper describes the data and methods being used to compile GHCN version 2.0, an expanded and improved version of its predecessor. Planned for distribution in early 1995, its enhancements will include (1) data for additional stations—perhaps three times as many as in version 1.0, plus maximum/minimum temperature measurements; (2) detailed assessments of data quality, including nearest-neighbor checks; and (3) adjustments for nonclimatic inhomogeneities, such as station relocations and land use changes.

DISCLAIMER

2. IN SEARCH OF CLIMATE DATA

The primary thrust of GHCN version 2.0 has been the search for additional data. A number of measures have been taken in this regard. These include (in order of effectiveness):

1. contacting data centers,
2. exploiting personal contacts,
3. conducting literature searches,
4. digitizing manuscript records,
5. tapping related projects, and
6. distributing miscellaneous requests.

Although certain researchers and particular countries have been unwilling to contribute their digitized data to GHCN, most parties generally are cooperative and enthusiastic when approached about the project. As a result, GHCN version 2.0 will contain data from ~50 data sets (Table 1). Rather than describe each in detail, the following are examples of how various techniques have resulted in the acquisition of specific sets.

2.1 Contacting Data Centers

This is perhaps the easiest and most productive short-term way of acquiring data. It simply entails contacting the distribution department of the particular center/organization, obtaining a list of all potentially relevant data sets, and acquiring those that seem promising. Roughly a dozen sets, ranging from the U.S. Historical Climatology Network (Karl et al., 1990) to The Comprehensive Pacific Rainfall Data Set (Morrissey et al., 1994) have been acquired in this manner.

2.2 Exploiting Personal Contacts

Personal contact goes a long way toward discovering additional sources of data and making these data available. For example, scientists who visit or work in conjunction with the authors' respective institutions (CDIAC, NCDC, or CIRES) often either directly supply data (i.e., by bringing it with them or sending it personally) or they "smooth" the path for acquiring data from some other party (i.e., by persuading the appropriate individual/organization to be receptive to the idea). One of the best examples of data acquired via personal contact is an archive compiled by Penn State Professor Emeritus Frederick L. Wernstedt. For decades Prof. Wernstedt has gone to great lengths to search out data and to digitize data from nonindustrialized countries. Early in the GHCN project he arrived at NCDC with a suitcase of floppy disks containing records for over 6000 stations. GHCN still has a synergistic relationship with Prof. Wernstedt, financing his travels (most recently to Brazil) in return for shared use of his valuable, hard-sought data.

Other personal contacts emphasize the more political and sometimes tragic side of things. For example, after a workshop in Europe in the spring of 1994, one of the authors learned that the chief of Rwanda's Division of Meteorology, Isaac Rusangiza, might be able to supply climate data for GHCN. After exchanging a letter and a couple of faxes, Mr.

Table 1. List of data sets included in GHCN version 2.0.

Data Base Name	Stations
A Clim. Data Bank for N. Hem. Land Areas (NDP-012)	1120
A Compr. Precip. Data Set for Global Land Areas (TR051)	5562
African Historical Precip. Data (TD-9799)	1169
Al-kubaisi's data set for Qatar	1
Al-sane's data set for Kuwait	1
Assoc. of Southeast Asian Nation's Clim. Atlas	880
Bo-min's data set for the PRC	378
Brewster's data set for Australia	2371
CAC's Clim. Anomaly Monitoring System data set	6078
Canadian Clim. Data	6701
Daily Temp. and Precip. Data for 223 USSR Stations	223
D. Wetterdienst's Global Monthly Surface Summaries	2597
Diaz's data set for high elevation sites	100
Dlamini's data set for Swaziland	33
Douglas' data set for Mexico	92
Forland's data set for Norway	20
Fitzgerald's data set for Ireland	11
GCPS Maximum/Minimum Temp. Data Set	3180
Griffith's data set for Australia	191
Griffith's data set for Brazil	428
Griffith's colonial archive data set	~150
Groisman's data set for the USSR	610
Hardjawinata's data set for Indonesia	13
Heino's data set for Finland	20
Hughes' data set for South Africa	90
Hulme's CRUWLD precip. data set	8215
Jacka's data set for Antarctic sites	70
Koch's data set for South Africa	~450
Ku-nil's data set for Korea	71
Monthly Climatic Data for the World (TD-3500)	2176
Monthly Data for the Pacific Ocean/Western Americas	71
Muthurajah's data set for Malaysia	18
NCAR's data sets for India (DS480.1 & 575.0)	4096
NCAR's/Harnack's data set for S. America (DS572.0)	678
NCAR's/Nicholson's data set for Africa (DS571.0)	1087
NCAR's WMSSC (DS570.0)	~4000
Non-African Historical Precip. Data (TD-9799)	1164
Oladipo's data set for Nigeria	13
Pakistan's Meteorological and Climatological Data Set	125
Sala's data set for Spain	3
The Comprehensive Pacific Rainfall Data Set	249
Two Long-Term Data Bases for the PRC (NDP-039)	205
U.S. first-order station network (TD-3280)	1571
U.S. Historical Climatology Network (NDP-019/R1)	1221
USSR network of CLIMAT stations	243
Wernstedt's data set for global land areas	~6000
World Weather Records, 1961-1970 and 1971-1980	~2000

Rusangiza sent a fax on April 1 stating that if he was supplied with 50 1.44-megabyte floppy diskettes and return postage, he would copy all of the Rwandan data and send them to NCDC. Even though the diskettes were mailed out that same day, it was too late. Rwanda erupted into chaos and the floppies were turned back at the border. There has been no further contact with Mr. Rusangiza since.

2.3 Conducting Literature Searches

Literature searches can result in significant data acquisitions. For instance, a recent publication by Hulme

(1992) describes an 8215-station rainfall data set for global land areas archived at the University of East Anglia. Upon request, the author (Professor Michael Hulme) graciously agreed to contribute the set to GHCN, resulting in one of the single largest and most valuable acquisitions to date.

Unfortunately, not all efforts bring such good rewards. For example, the WMO document called INFOCLIMA lists data sets available from WMO member countries. After carefully reviewing this document for data sets containing land surface station climate records, letters were distributed to a dozen potential sources. Of the three replies which were received, only one led to the acquisition of data (a 33-station data set for Swaziland from Mr. E. Dlamini).

2.4 Digitizing Manuscript Records

Most industrialized countries were able to begin digitizing their climate data some time ago. In contrast, many less-developed countries lacked (and still lack) the resources to undertake such a project. Yet data for many of these countries exist in libraries in the U.S. and Europe because, in the late 1800s and early 1900s, European countries assembled climate data publications for many of their colonies (particularly those in Africa). Under the direction of Texas A&M Professor John Griffiths, data from some of these publications are currently being digitized for inclusion in GHCN version 2.0. Records for hundreds of stations in data-sparse regions of Brazil are also being digitized by Profs. Griffiths and Wernstedt.

2.5 Tapping Related Projects

GHCN has also benefitted from other projects. For example, NCDC is currently collecting 30-year station means (normals) for the period 1961–1990 as a service to WMO. On occasion, certain WMO member countries (e.g., Qatar) supply year/month sequential data in addition to their 30-year normals. Upon receipt of such records, the member country is contacted in regard to contributing the sequential records to GHCN. Generally, the response to such requests is favorable.

2.6 Distributing Miscellaneous Requests

From time to time, new acquisition ideas are explored. For example, a request for data was posted to CLIMLIST earlier this year. There was little response.

3. QUALITY CONTROL

GHCN version 2.0 will be primarily constructed from what Guttman (1991) calls "secondary data sources ... that have been compiled, adjusted, or summarized by anyone other than the researcher using the data." When using such data, it is imperative that data "validity" and "accuracy" be assessed. "Validity" refers to whether the data fit the particular application; "accuracy" refers to the reliability of the data. To address both concerns, an extensive quality control procedure was developed. In theory, a quality control procedure is intended to ensure that data meet certain standards of excellence. In practice, quality control implies looking for gross data errors. These "errors" take many forms, ranging from inappropriate data to magnetic media problems to formatting errors to data gaps to outliers to time series inhomogeneities. The procedure employed in GHCN version 2.0 is designed to catch the most pervasive and/or egregious problems that were discussed at the *International Workshop on Quality Control of Monthly Climatic Data* (Peterson, 1994). The procedure consists of four parts:

1. selection of data sets,
2. preprocessing of data files,
3. duplicate station elimination, and
4. outlier checks.

3.1 The Right Stuff

Data set "validity" was determined by reviewing any documentation and publications related to a particular data set. A number of data sets (or portions thereof) will not be included in GHCN version 2.0 because they do not meet the "validity" criterion; that is, something inherent in the way they were compiled rendered them inappropriate for this application.

Among those not included are any sets in which monthly means/totals were derived from synoptic reports. Comparisons between monthly temperature and precipitation data which are created from synoptic reports and CLIMAT data, which are monthly data created by the country taking the observations, indicate that synoptically derived values are not up to the quality that are required for GHCN (C. Ropelewski, pers. comm., 1993).

Archives in which climatic values have undergone major "adjustments" (e.g., corrections for discontinuities) will also be excluded. Although it is sometimes difficult to find data that have not been sanitized in any manner, the intermingling of "pure" and "adjusted" records can create some acute spatial problems. For example, data for Carlsbad, New Mexico, were available from two data sets that were used to compile GHCN version 1.0. One series only contained raw data, whereas the other contained data in which discontinuity adjustments were applied. Because the annual temperature between the two correlated at an r^2 of less than 0.01 and their metadata were dissimilar, both stations were inadvertently included in the data set.

3.2 Preprocessing of Data Files

Data received are rarely in a condition that would permit immediate use, regardless of the source or data set documentation. To guarantee data of the highest possible quality, extensive preprocessing reviews are conducted. The reviews involve checking data sets for completeness, reasonableness, and consistency. Although they have common objectives, the reviews are tailored to each data set, often requiring extensive programming efforts and days (or even weeks) of manpower. Some of the routine preprocessing checks applied to each new data include the following:

1. determining if the physical characteristics of each data file (e.g., number of lines, record length) agree with supplied documentation (if any);
2. determining if variable storage locations (i.e., the columns in which variables are located) and variable types (i.e., integer, real, or character) are consistent throughout each file and agree with supplied documentation (if any);
3. determining the number of unique stations in each file, whether the file contains any duplicate stations and (by comparison with documentation) if any stations are missing or if any undocumented stations are present;
4. determining if all date variables have reasonable values, if each file is chronologically sorted, if the period of record for each station agrees with supplied documentation (if any), and if duplicate station/date entries are present;
5. determining if the units of each climate variable agree with supplied documentation (if any) or determining the units if documentation is lacking; searching for specially defined values (e.g., missing value codes or trace rainfall codes) and undocumented, physically meaningless values; checking the frequency of occurrence of all possible data values;
6. searching for the presence of various flag codes and whether they have documented meanings; searching for contradictions between flag codes and data values and between flag codes and other flag codes; and
7. determining if the units of each metadata variable (e.g., coordinates or elevation) agree with the supplied documentation (if any) or determining the units if documentation is lacking; searching for the presence of specially defined metadata values or undocumented, meaningless values.

When problems are detected, the original data set compiler is contacted for additional information or revised files, if possible.

3.2 Too Much of a Good Thing

A time series for a given station can frequently be obtained from more than one source data set. For example, precipitation data for Beijing, China, are available in both Eischeid et al. (1991) and Shiyan et al. (1991). When "merging" data from multiple sources, it is important to identify these duplicate station time series because (1) the inclusion of multiple time series for a particular location can create acute spatial problems, as in the case of Carlsbad, New Mexico; (2) each may have a different period of record that, when "mingled," results in a single series that has a longer or more complete record than either of the originals; and (3) the intercomparison of comparable year/month values in each can help identify other data quality problems.

Unfortunately, a number of factors complicate the identification of duplicate stations. These include:

1. nonidentical or inaccurate station numbers, names, coordinates, and elevations (i.e., metadata); and
2. data which are the same for some years and not others (this can be caused by poor data processing practices or inhomogeneity adjustments having been applied to one or both series).

The duplicate station elimination procedure used in GHCN version 2.0 will take such problems into account.

3.3 The Needle in the Haystack

GHCN version 2.0 will contain a massive quantity of data (at least 10,000 temperature and precipitation stations). Considering the available resources, it will not be possible to search for extreme values/outliers in the same manner employed when compiling version 1.0 (i.e., by plotting and visually inspecting every time series). As a result, it has been necessary to develop an inventory of the most common and egregious errors detected when compiling version 1.0 and then to devise tests that can detect those problems. In general, the tests fall into three categories: serial checks, spatial checks, and intervariable checks.

Serial checks involve comparing an observation in a time series with some other observation in the same time series and then deciding, based upon some threshold, whether the observation is problematic. Examples of problematic values from a serial perspective include the following:

1. cases in which the same data value occurs for several consecutive months in one or more years;
2. cases in which consecutive years have the same data;
3. cases in which the same data value occurs in the same month of several consecutive years;
4. cases in which a month with missing data was "left out" of the file rather than being set to missing, causing all subsequent data values to be placed in the preceding month (e.g., October's data value would be listed under September);
5. cases in which an observation appears extreme when compared with other values in the same month;
6. cases in which an observation appears extreme when compared with values in adjacent months; and
7. cases in which there is a severe change in mean or variance.

Spatial checks involve comparing data at one station to data at another and then deciding, based upon some threshold, whether the data are problematic. Examples of problematic values from a spatial perspective include the following:

1. cases in which an observation appears extreme when compared with values in adjacent stations; and
2. cases in which the long-term mean at a station appears extreme when compared with the long-term mean at adjacent stations.

Intervariable checks are quite straightforward. They include the following:

1. cases which violate the relationship: minimum temperature ≤ mean temperature ≤ maximum temperature; and
2. cases which violate the relationship: station pressure ≤ sea level pressure (unless the station is below sea level).

It should be noted that erroneous values will not be corrected; rather, they will flagged as problematic, stored in a separate file, or set to missing.

4. DATA ADJUSTMENTS

The importance of using homogeneous climatological time series in climate research has recently received much attention. A homogeneous time series is defined as one where variations are caused only by variations in weather and climate (Conrad and Pollak, 1962). Using climatological time series containing artificially induced variations can lead to inconsistent conclusions. For example, Hansen and Lebedeff (1987) used data that had not been adequately examined for inhomogeneities and provided analyses of temperature trends for different regions of the globe. One station, St. Helena Island, was used to represent a large area in the tropical south Atlantic, and indicated a considerable amount of warming over the past century. However, most of the warming apparent in the temperature record occurred when the elevation of the station decreased approximately 200 meters because of a move in 1972, causing an abrupt warming of approximately 2°C (P. Michaels, pers. comm., 1992).

A wide variety of factors can create an inhomogeneous time series. Sometimes inhomogeneities in time series occur as a gradual, artificially induced trend (e.g., urban warming, drift in instrument calibration). Others are manifest as abrupt changes in scale or variance (e.g., station relocations, instrumentation replacements). Because numerous applications require homogeneous time series, GHCN version 2.0 will include temperature and precipitation time series in which discontinuities have been removed.

The technique to detect discontinuities in station temperature time series consists of four parts. First, a homogeneous reference time series for each station is created from nearby stations (Peterson and Easterling, 1994). Second, the station time series is subtracted from its reference series to create a difference series. Third, the technique tests the difference series for changes in statistical characteristics indicating a discontinuity (Easterling and Peterson, 1994). And fourth, the magnitude of the inhomogeneity is determined and an adjustment is made in the station data to account for the discontinuity. This technique will be applied to all GHCN temperature data to create a data set where much of the effects of inhomogeneities such as station moves, changes in instrumentation, and changes in station environment have been systematically and objectively removed.

Adjustments to precipitation data will be primarily based on metadata, with an emphasis on accounting for changes in the catch of frozen precipitation caused by changes in instrumentation, especially the installation of wind shields (Groisman, 1991). In general, stations where metadata indicate no change in observing practices or instrumentation will comprise the bulk of the homogeneous precipitation data set.

5. AVAILABILITY

The data base is available, free of charge, as a Numeric Data Package (NDP-041/R1) from CDIAC at Oak Ridge National Laboratory. It can be ordered from CDIAC by telephone, fax, or electronic mail and is distributed via anonymous FTP, 9-track magnetic tape, and 8-mm (e.g., exabyte) tape. For additional information contact:

Address: Carbon Dioxide Information Analysis Center
Oak Ridge National Laboratory
Post Office Box 2008
Oak Ridge, Tennessee 37831-6335
Phone: (615) 574-0390
Fax: (615) 574-2232
E-mail: cdp@ornlstc (Bitnet)
cdp@ornl.gov (Internet)
cdiac (Omnet)
FTP: cdiac.esd.ornl.gov (128.219.24.36)
enter "anonymous" as the userid
enter your e-mail address as the password
cd /pub/ndp041r1

The data base can also be acquired from the Research Customer Service Group at NCDC. It can be ordered from NCDC by telephone, fax, or electronic mail. For additional information contact:

Address: Research Customer Service Group
National Climatic Data Center
Federal Building
Asheville, North Carolina 28801
Phone: (704) 271-4994
Fax: (704) 271-4876
E-mail: tross@ncdc.noaa.gov
Mosaic: http://www.ncdc.noaa.gov

Acknowledgements: CDIAC is sponsored by the U.S. Department of Energy, Office of Health and Environmental Research, Global Change Research Program. Oak Ridge National Laboratory is managed by Martin Marietta Energy Systems, Inc. for the United States Department of Energy under contract DE-AC05-84OR21400. Publication No. 4320, Environmental Sciences Division, ORNL. Russell S. Vose is a staff member of the Energy, Environment, and Resources Center, University of Tennessee, Knoxville.

6. REFERENCES

Conrad, V. and C. Pollak. 1962. *Methods in Climatology*, Harvard University Press, Cambridge, Massachusetts.

Diaz, H. F., R. S. Bradley, and J. K. Eischeid. 1989. Precipitation fluctuations over global land areas since the late 1800s. *Journal of Geophysical Research* 94:1195–1210.

Easterling, D. R. and T. C. Peterson. 1994: A new technique for detecting and adjusting undocumented artificial discontinuities in climatic time series. *International Journal of Climatology*, in press.

Eischeid, J. K., H. F. Diaz, R. S. Bradley, and P. D. Jones. 1991. A comprehensive precipitation data set for global land areas. DOE/ER-6901T-H1. Carbon Dioxide Research Division, U.S. Department of Energy, Washington, D.C.

Groisman, P., Koknaeva, V. V., Belokrylova, T. A., and Karl, T. R. 1991. Overcoming biases of precipitation measurement: A history of the USSR experience. *Bulletin of the American Meteorological Society* 72:1725–1733.

Guttman, N. B. 1991. Using secondary data sources. *Climatic Change* 18:95–105.

Hansen, J. and S. Lebedeff. 1988. Global surface air temperatures: Update through 1987. *Geophysical Research Letters* 15:323–326.

Hulme, M. 1992. A 1951–80 global land precipitation climatology for the evaluation of general circulation models. *Climate Dynamics* 7:57–72.

Jones, P. D., P. M. Kelly, C. M. Goodess, and T. R. Karl. 1989. The effect of urban warming on the Northern Hemisphere temperature average. *Journal of Climate* 2:285–290.

Karl, T. R., C. N. Williams, and F. T. Quinlan. 1990. United States Historical Climatology Network (HCN) serial temperature and precipitation data. NDP-019/R1. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Morrissey, M. L., M. A. Shafer, S. Postawko, and B. Gibson. 1993. The Comprehensive Pacific Rainfall Data Base. Oklahoma Climatological Survey Contribution Number 93-003, Norman, Oklahoma.

National Environmental Satellite Data and Information Service (NESDIS). 1987–1991. World Weather Records, 1971–1980 (Vols. 1–4). U.S. Department of Commerce, Washington, D.C.

Peterson, T. C. 1994. Report of the International Workshop on Quality Control of Monthly Climatic Data. Environmental Sciences Division Publication No. 4223, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Peterson, T. C. and D. R. Easterling. 1994. Creation of homogeneous composite climatological reference series. *International Journal of Climatology* 14:671–679.

Shiyan, T., Fu Congbin, Z. Zhaomei, and Z. Qingyun. 1991. Two long-term instrumental climatic data bases of the People's Republic of China. NDP-039. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Spangler, W. M. and R. L. Jenne. 1990. World Monthly Surface Station Climatology (and associated datasets). National Center for Atmospheric Research, Boulder, Colorado.

Vose, R. S., R. L. Schmoyer, P. M. Steurer, T. C. Peterson, R. Heim, T. R. Karl, and J. K. Eischeid. 1992. The Global Historical Climatology Network: Long-term monthly temperature, precipitation, sea level pressure, and station pressure data. NDP-041. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.