An Overview of the Global Historical Climatology Network Temperature Database

Thomas C. Peterson* and Russell S. Vose+

ABSTRACT

The Global Historical Climatology Network version 2 temperature database was released in May 1997. This century-scale dataset consists of monthly surface observations from ~7000 stations from around the world. This archive breaks considerable new ground in the field of global climate databases. The enhancements include 1) data for additional stations to improve regional-scale analyses, particularly in previously data-sparse areas; 2) the addition of maximum–minimum temperature data to provide climate information not available in mean temperature data alone; 3) detailed assessments of data quality to increase the confidence in research results; 4) rigorous and objective homogeneity adjustments to decrease the effect of nonclimatic factors on the time series; 5) detailed metadata (e.g., population, vegetation, topography) that allow more detailed analyses to be conducted; and 6) an infrastructure for updating the archive at regular intervals so that current climatic conditions can constantly be put into historical perspective. This paper describes these enhancements in detail.

1. Introduction

Humanity has long been fascinated by the weather. Instruments that could reliably measure air temperature had been developed by the late seventeenth century. Renowned for his manufacture of precision meteorological instruments, D. G. Fahrenheit invented the mercury thermometer in 1714. Soon, individuals and organizations began to establish networks of meteorological instruments to help quantify and record the weather. There were many reasons to do this, ranging from agriculture to forecasting. The first large-scale monitoring efforts were in western Europe. Over time, the implementation of these instruments diffused into the rest of the world. Currently, most countries operate large networks of weather observing stations.

*Global Climate Laboratory, National Climatic Data Center, Asheville, North Carolina.
+Office of Climatology, Arizona State University, Tempe, Arizona.
Corresponding author address: Thomas C. Peterson, Global Climate Laboratory, National Climatic Data Center, 151 Patton Avenue, Room 120, Asheville, NC 28801.
E-mail: tpeterso@ncdc.noaa.gov
In final form 11 August 1997.

Today, climate research relies heavily on the records from instruments at these near-surface weather stations. There are two reasons for this reliance: instrumental records represent direct samples at exact points in space and time, and they have been collected at over 100 000 locations in the past two centuries (F. Wernstedt 1994, personal communication). While other indicators (e.g., tree rings) also record climate variations, they generally are inferential rather than direct measurements of meteorological conditions and are currently available at far fewer locations than their instrumental counterparts. Thus it is the "instrumental network" that constitutes the most spatially and temporally complete record of land surface climate since the onset of the Industrial Revolution (Jones 1994). Unfortunately, not all available historic data have been digitized. In the digital archives, there are many more station years of monthly data available than daily data with correspondingly much better spatial coverage.

Because most instrumental networks were established to monitor local weather and not the long-term climate, there are practical problems in using these data to study climate change. For instance, the records are often not digitized and/or are not readily available outside of the country in which they were measured. An uneven distribution of stations introduces network biases that have significant effects on estimated temperature trends, particularly at the regional scale (Willmott et al. 1994). Instrumental records also often contain data errors resultant from the data recording and archiving processes. These errors, which take many forms (e.g., outliers, truncations), reduce confidence in the analyses. In addition, instrumental records are subject to inhomogeneities caused by many factors, such as local station moves and the introduction of new thermometers. Such inhomogeneities introduce nonclimatic variation into historical records and thus further cloud temporal trends. In short, each of these forces contributes to a bias embedded in the historical record that complicates the detection of climatic change on any scale.

Many efforts to produce long-term monthly global climate databases have addressed these issues, though most emphasized data collection. One of the first and longest running efforts is the World Weather Records (WWR) initiative, which commenced in 1923 and has resulted in the regular publication of decadal series of global climate records ever since (Clayton 1927). Another fine example is the National Center for Atmospheric Research's annually published World Monthly Surface Station Climatology dataset (WMSSC; Spangler and Jenne 1992), which consists of WWR, miscellaneous acquisitions, and the National Climatic Data Center's (NCDC) Monthly Climatic Data for the World for more recent records. Both the WWR and WMSSC are outstanding databases in their own right; however, owing to simple time, resource, and mission constraints, these sets (and others of their kind) have not yet integrated some newly available datasets (e.g., data from United States–Russia bilateral exchanges). Furthermore, neither database contains detailed station homogeneity assessments, limiting their utility in studies of climate change. This issue has been more commonly addressed to some degree by individual researchers (e.g., Wernstedt 1972; Bradley et al. 1985), who compiled their own global and hemispheric datasets for specific applications. The most famous of these is the Jones dataset (Jones et al. 1986; Jones 1994), which has been used extensively in climate research.

In the early 1990s, climatologists from NCDC and the Carbon Dioxide Information Analysis Center (CDIAC) undertook a new initiative aimed at creating a dataset appropriate for the study of climate change at both global and regional scales. Building upon the fine efforts of its predecessors, this database, known as the Global Historical Climatology Network (GHCN), was released in 1992 (Vose et al. 1992). It contains quality-controlled monthly climatic time series from 6039 land-based temperature stations worldwide. Compared to most datasets of this type (e.g., Jones 1994), this initial release of GHCN was larger and had more detailed spatial coverage. Since its creation, thousands of copies have been provided free of charge to researchers, educators, and students around the world, and requests for both the basic dataset and derived products (e.g., gridded temperature anomalies) currently average over 200 per month from NCDC and CDIAC. More importantly, it has become a popular tool in climate change research (e.g., Brown et al. 1993; Young 1993; Groisman et al. 1994a; Groisman et al. 1994b; Karl et al. 1994; Quereda and Monton 1994, 1996; Balling 1995; Baranyi and Ludmany 1995; Epperson et al. 1995; Gutzler, 1996; Adkison et al. 1996; Tayanc et al. 1997).

Given the popularity of GHCN, researchers at NCDC, CDIAC, and Arizona State University have prepared an enhanced database to serve the ever-increasing demand for these data. This archive, GHCN version 2, breaks considerable new ground in the field of global climate databases. Enhancements include 1) data for additional stations to improve regional-scale analyses, particularly in previously data-sparse areas; 2) the addition of maximum–minimum temperature data to provide important climate information not available in mean temperature data alone (e.g., Karl et al. 1993; Easterling et al. 1997); 3) detailed assessments of data quality to increase the confidence in research results; 4) rigorous and objective homogeneity adjustments to decrease the effect of nonclimatic factors on the time series; 5) detailed metadata (e.g., population, vegetation, topography) that allow more detailed analyses to be conducted; and 6) an infrastructure for updating the archive at regular intervals so that current climatic conditions can constantly be put into historical perspective. This paper describes these enhancements in detail.

2. Sources

One of the primary goals of GHCN version 2 was to acquire additional data in order to enhance spatial and temporal coverage. There were three reasons for this goal: 1) data for recent months allow one to assess current climatic conditions and place them in historical perspective, 2) denser coverage facilitates the analysis of regional climate change, and 3) certain areas (or certain times in certain areas, such as 1920s Africa) are still undersampled even from a perspective of global analysis. Because numerous institutions operate weather stations and because no single repository archives all of the data for all stations, we employed five acquisition strategies to maximize the available pool of data: 1) contacting data centers, 2) exploiting personal contacts, 3) tapping related projects, 4) conducting literature searches, and 5) distributing miscellaneous requests. In general, most parties were cooperative and enthusiastic about donating their data to the GHCN initiative, particularly since GHCN is a World Meteorological Organization (WMO) Global Baseline Data Set. As a result, GHCN version 2 contains data from 31 diverse sources (Table 1).

We started the data acquisition process by approaching institutions that collect, archive, manage, and/or distribute meteorological data. Approximately a dozen datasets were acquired in this fashion. We also exploited personal contacts by contacting colleagues in the search for potential data sources. For example, scientists who visit or work in conjunction with the authors' respective institutions often either have data themselves or are able to facilitate the acquisition of data from another party (e.g., by putting the authors in contact with potential sources). This was another extremely productive means of acquiring data, which yielded approximately 10 new datasets.

When possible, we tapped related projects for potentially useful data. For example, NCDC recently collected and processed station normals for the period 1961–90 as a contribution to WMO (WMO 1996a). On occasion, a WMO member country supplied year/month sequential data in addition to the 30-yr means and other statistics. Upon receipt of such records, the member country was contacted in regard to contributing the time series data to GHCN. The Colonial Era Archives initiative was also tapped in this regard (Peterson and Griffiths 1996). Started as a GHCN subproject to acquire data in very data sparse regions, this initiative digitized early temperature and precipitation records for stations operated by various European countries in their respective overseas colonies. Data for hundreds of early African stations have been incorporated from this source and the digitizing effort has been expanded to Asia and South America (Peterson and Griffiths 1997).

When articles using the appropriate types of climate records were found, the articles' authors were contacted in an attempt to procure data. In general, this approach resulted in very few acquisitions simply because most of the time the data used in the published research had been previously acquired. Internet searches turned up many versions of datasets previously acquired but little in the way of new data. Posts on climate-related electronic bulletin boards yielded no useful data. Apparently, while many researchers need long-term climate data, few are involved in the exacting acquisition and digitization of historic data.

3. Duplicate elimination

A time series for a given station can frequently be obtained from more than one source. For example, data for Tombouctou, Mali, were available in six different source datasets. When "merging" data from multiple sources, it is important to identify these duplicate time series because 1) the inclusion of multiple versions of the same station creates biases in areally averaged temperature analyses, and 2) the same station may have different periods of record in different datasets; merging the two versions can create longer time series.

The goal of duplicate station elimination is to reduce a large set of *n* time series (many of which are identical) to a much smaller set of *m* groups of time series that are unique. In the case of maximum and minimum temperature, 8000 source dataset time series were reduced to 4964 unique time series. This was accomplished in the following fashion. First, the data for every station were compared with the data for every other station. This naturally started with stations whose metadata indicated they were in approximately the same location. Similarity was assessed by computing the total number of months of identical data as well as the percentage of months of identical data. Maximum–minimum temperature time series were considered duplicates of the same station if they shared the same monthly value at least 90% of the time, with at least 12 months of data being identical and no more than 12 being different. This process identified the duplicates, which were then merged to form time series with longer periods of record after a manual inspection of the metadata (to avoid misconcatenations). This process was then repeated on the merged dataset without the initial metadata considerations so every time series was compared to all the other time series in the database. Similarity of time series in this step was judged by computing the length of the longest run of identical values.

Table 1. Sources of data that went into GHCN version 2 temperature database and the number of stations in each dataset. However, for a dataset with a significant percentage of stations that were not used in GHCN (e.g., because they were derived from synoptic data), the number of stations represent the number of GHCN stations with contributions from that data source.

	Number of stations	
Dataset name/contributor	**Mean**	**Max–Min**
National Center for Atmospheric Research's world monthly surface station climatology	3563	0
National Climatic Data Center's maximum–minimum temperature dataset	3179	3179
Deutscher Wetterdienst's global monthly surface summaries dataset	2559	0
Monthly climatic data for the world	2176	0
Climate Prediction Center's CAMS dataset	2124	0
World Weather Records (1971–80)	1912	0
World Weather Records (1961–70)	1858	0
U.S. Summary of the Day Dataset	1463	1463
U.S. Historical Climatology Network	1221	1221
A climatological database for Northern Hemisphere land areas	920	0
Australian National Climate Center's dataset for Australia and surrounding countries	785	785
North American climate data, NCDC	764	764
Bo-Min's dataset for the People's Republic of China	378	0
USSR network of CLIMAT stations	243	0
Daily temperature and precipitation data for 223 USSR stations (NDP-040)	223	223
Two long-term databases for the People's Republic of China (NDP-039)	205	60
ASEAN climatic atlas	162	162
Pakistan's meteorological and climatological dataset	132	132
Diaz's dataset for high-elevation areas	100	0
Douglas' dataset for Mexico	92	0
Ku-nil's dataset for Korea	71	71
Jacka's dataset for Antarctic locales	70	0
Monthly data for the Pacific Ocean–western Americas	60	0
U.S. Historical Climatology Network (Alaska)	47	47
Muthurajah's dataset for Malaysia	18	18
Hardjawinata's dataset for Indonesia	13	13
Fitzgerald's dataset for Ireland	11	11
Sala's dataset for Spain	3	0
Al-kubaisi's dataset for Qatar	1	1
Al-sane's dataset for Kuwait	1	1
Stekl's dataset for Ireland	1	1

Cases where the time series were determined to be duplicates of the same station but the metadata indicated they were not the same station were examined carefully and a subjective decision was made. This assessment provided additional quality control of station locations and the integrity of their data. For example, a mean temperature time series for Thamud, Yemen, had 25 yr (1956–81) of monthly values that were exactly identical to the mean temperature data from Kuwait International Airport (12° farther north). Needless to say, one of these time series was in error. As with most of these problems, determining which time series was erroneous was fairly easy given the data, metadata, knowledge about the individual data sources, duplicate data, and other climatological information available.

The procedure for duplicate elimination with mean temperature was more complex. The first 10 000 duplicates (out of 30 000+ source time series) were identified using the same methods applied to the maximum and minimum temperature datasets. Unfortunately, because monthly mean temperature has been computed at least 101 different ways (Griffiths 1997), digital comparisons could not be used to identify the remaining duplicates. Indeed, the differences between two different methods of calculating mean temperature at a particular station can be greater than the temperature difference from two neighboring stations. Therefore, an intense scrutiny of associated metadata was conducted. Probable duplicates were assigned the same station number but, unlike the previous cases, not merged because the actual data were not exactly identical (although they were quite similar). As a result, the GHCN version 2 mean temperature dataset contains multiple versions of many stations. For the Tombouctou example, the six source time series were merged to create four different but similar time series for the same station (see Fig. 1).

Preserving the multiple duplicates provides some distinct benefits. It guarantees no concatenation errors. Adding the recent data from one time series to the end of a different time series can cause discontinuities, unless the mean temperature was calculated the same way for both time series. It also preserves all possible information for the station. When two different values are given for the same station–year–month, it is often impossible for the dataset compiler to determine which is correct. Indeed, both may be correct given the different methods used to calculate mean temperature.

Unfortunately, preserving the duplicates may cause some difficulty for users familiar with only one “correct” mean monthly temperature value at a station. There are many different ways to use data from duplicates. All have advantages and disadvantages. One can use the single duplicate with the most data for the period of interest; use the longest time series and fill in missing points using the duplicates; average all data points for that station–year–month to create a mean time series; or combine the information in more complicated ways, such as averaging the first difference ($FD_{year\ 1} = T_{year\ 2} - T_{year\ 1}$) time series of the duplicates and creating a new time series from the average first difference series. Which technique is the best depends on the type of analysis being performed.

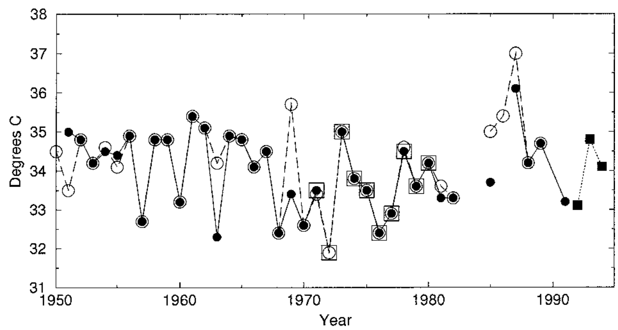

FIG. 1. Tombouctou, Mali, mean temperature data for May from 1950 to 1995. Mean temperature data for Tombouctou were present in six of GHCN’s 31 source datasets, with data starting in 1897. Of the six time series, several of these could be combined, leaving four different Tombouctou mean temperature time series (duplicates). In the graph, each duplicate is indicated by a different symbol. Many of the data points are exactly the same, but the differences between the duplicates were significant enough that the time series could not be combined. The reason why GHCN mean temperature data have duplicates while mean maximum and minimum temperature data do not is because there are over 100 different ways in which daily mean temperature has been calculated by meteorologists.

4. Distribution

GHCN version 2 contains mean temperature data for a network of 7280 stations and maximum–minimum temperature data for 4964 stations. All have at least 10 yr of data. The archive also contains homogeneity-adjusted data for a subset of this network (5206 mean temperature stations and 3647 maximum–minimum temperature stations). The homogeneity-adjusted network is somewhat smaller because at least 20 yr of data were required to compute reliable discontinuity adjustments and the homogeneity of some isolated stations could not be adequately assessed.

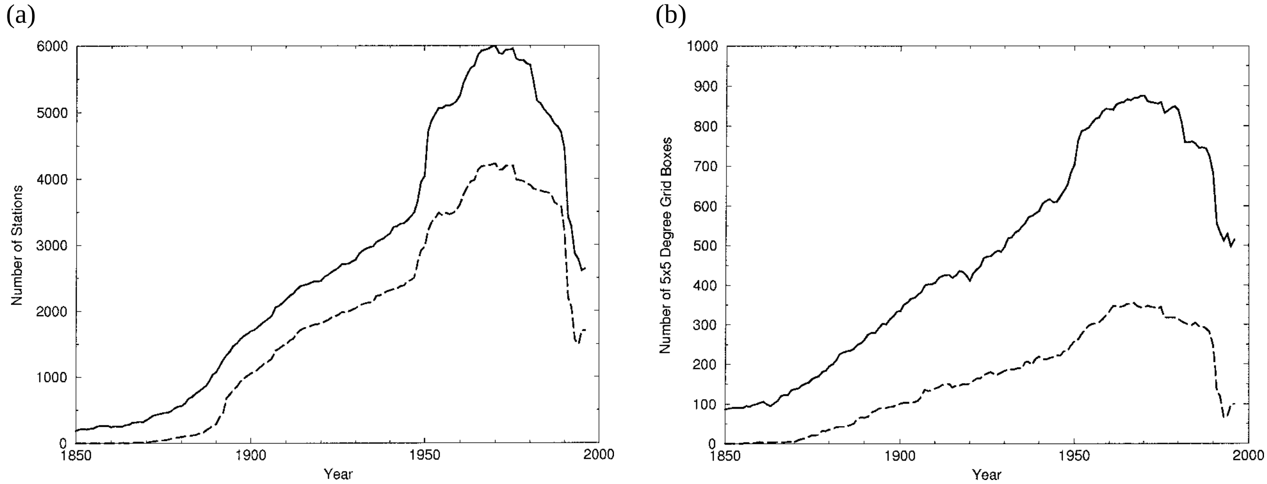

FIG. 2. Time series of the number of stations (a) and the number of 5° × 5° boxes (b) for mean temperature (solid) and maximum and minimum temperature (dashed). The graphs start in 1850, but the earliest mean temperature datum is for January 1701 from Berlin, Germany, and the earliest mean maximum and minimum temperature data in GHCN are for March 1840 from Toronto, Canada. The reasons why the number of stations in GHCN drop off in recent years are because some of GHCN's source datasets are retroactive data compilations (e.g., World Weather Records) and other data sources were created or exchanged years ago. Only three data sources are available in near-real time. The rise in maximum and minimum temperature stations and grid boxes in 1995 and 1996 is due to the World Meteorological Organization's initiation of international exchange of monthly CLIMAT maximum and minimum temperature data over the Global Telecommunications System in November 1994.

Figure 2 shows the variation in the number of stations from 1850 to 1997 and the variation in the global coverage of the stations as defined by 5° × 5° grid boxes. The graphs start in 1850, but the earliest mean temperature datum is for January 1701 from Berlin, Germany, and the earliest mean maximum and minimum temperature data in GHCN are from March 1840 for Toronto, Canada.

With 7280 stations, GHCN is over twice as large as the widely used Jones (1994) 2961-station mean temperature dataset. The spatial distribution of these datasets can be estimated by comparing the number of 5° × 5° grid boxes with station data. GHCN has data in 876 grid boxes, while Jones (1994) has 779. Though the number of grid boxes with data is less in early years, as indicated by Fig. 2, GHCN has more grid boxes with data in the early years as well. For example, in the 1930s GHCN averaged 540 grid boxes with data compared to 425 for Jones (1994); for the decade of the 1900s 375 versus 275; and in the 1870s, 160 versus 125. Some of GHCNs improved spatial coverage in the late nineteenth and early twentieth centuries comes from stations that are no longer operating and lack adequate data for the 1961–90 base period required by Jones (1994).

The distribution of mean temperature stations has pronounced spatial variation. As revealed by Fig. 3a, the total station coverage is rather excellent. However, the period of record for these stations is highly variable. For example, some of the station data were digitized by special projects during the 1970s and therefore have no later data. Going back in time to 1900 (Fig. 3b) reveals good coverage in North America, Europe, and parts of Asia and Australia. For the rest of the world the pre-1900 era coverage is spotty. However, the number and distribution is likely sufficient for computing reliable global temperature time series of 100 or more years (Jones 1995; Jones et al. 1997); furthermore, it exceeds that of most other global climate databases (e.g., WMSSC, WWR, Jones). One data source, the Colonial Era Archive project (Peterson and Griffiths 1996), continues to digitize early data from around the world so the pre-1900 data distribution should improve somewhat as these stations get incorporated into GHCN. The distribution of adjusted mean temperature stations is somewhat less dense throughout the record, though the spatial distributions are largely the same as for their unadjusted counterparts with the exception of isolated stations such as St. Helena Island in the tropical South Atlantic.

The distribution of maximum–minimum temperature stations is less complete (Fig. 4). Large spatial gaps are present in maximum and minimum station coverage. The coverage actually is much less complete than analysis of Fig. 4a might indicate because of the highly variable period of record. For example, a significant portion of the maximum and minimum station data in Africa were from the Colonial Era Archive project (Peterson and Griffiths 1996), which digitized preindependence (circa 1960) data. The lack of maximum and minimum temperature records results from the fact that, until recently, few countries exchanged such data on a regular basis. While organizations like WMO and NCDC have made maximum and minimum temperature data exchange a high priority, it will likely be some time before a global coverage of historic maximum and minimum temperature is possible. To date, however, this compilation represents the largest of its kind. The coverage of maximum and minimum temperature in 1900 is very isolated (Fig. 4b).

5. Quality control

GHCN quality control (QC) is a three-stage process. A full description of GHCN QC tests and their justification is given in Peterson et al. (1997b), so the following is a short summary of the QC tests applied to GHCN.

The first stage examines the quality and appropriateness of the source datasets. Thirty-one source datasets contributed temperature data to GHCN while several additional potential sources had to be rejected. The rejections were primarily caused by (a) homogeneity-adjusted data without access to original observations; (b) the monthly data were derived from synoptic reports, which are almost always incomplete, thereby causing unacceptable errors or biases; and (c) significant processing errors that indicated the source dataset was unreliable.

The second stage examined individual station time series. These tests included comparing the stations to gridded climatology (Legates and Willmott 1990) and plotting the stations on Operational Navigation Charts (see section 7 on metadata). Both of these processes uncovered mislocated stations and the former uncovered stations that were digitized 6 months out of phase. Additionally, we tested each time series for significant discontinuities using the cumulative sum test (CUSUM, van Dobben de Bruyn 1968), which looks for changes in the mean. A test called SCUSUM was developed to look for changes in the variance or scale. Finally we looked for runs of three or more months of the same value in the time series.

The third and final stage of GHCN QC evaluated individual data points to determine if they were outliers in time and space. All data points that were determined to be greater than 2.5 biweight standard deviations (Lanzante 1996) from the time series mean were flagged. Each of these flagged data points was then compared to neighboring stations to determine if the extreme value represented an extreme climate event in the region. Over 85% of the previously flagged data

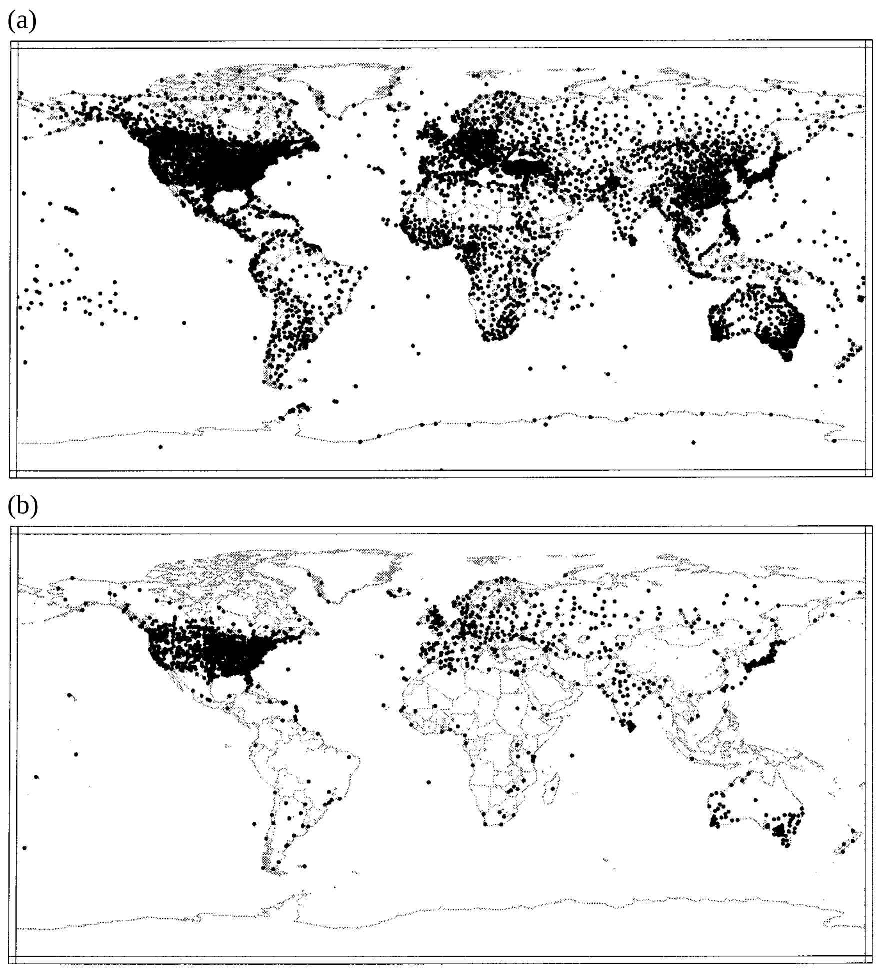

FIG. 3. Maps of GHCN mean temperature station locations: (a) all GHCN mean temperature stations and (b) mean temperature stations with data in 1900. Approximately 1000 GHCN stations have a century or more of mean temperature data. Work is under way to fill in some of the large data-sparse regions shown in (b) by digitizing selected station data from Colonial Era Archives (Peterson and Griffiths 1996).

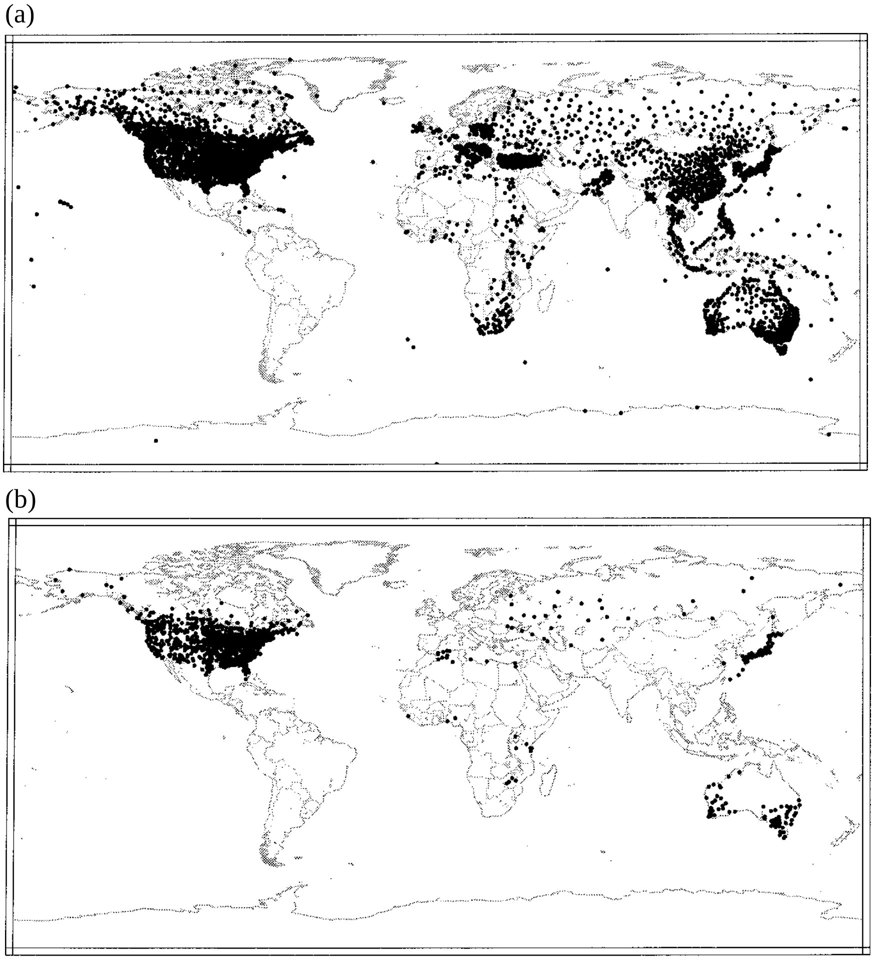

FIG. 4. Maps of GHCN maximum and minimum temperature station locations: (a) all GHCN maximum and minimum temperature stations and (b) GHCN temperature stations with maximum and minimum data in 1900. Because mean monthly maximum and minimum temperatures have not been regularly exchanged until recently, there are large gaps in GHCN's maximum and minimum temperature coverage. These gaps will be slowly filled with the incorporation of new sources of data.

points were determined to be valid using the spatial QC test. Those data points that failed both of these tests were removed from the main GHCN data file but included in a separate file for possible use by researchers possessing additional, potentially corroborating, information.

6. Homogeneity

Most long-term climate stations have undergone changes that make a time series of their observations inhomogeneous. There are many causes for the discontinuities, including changes in instruments, shelters, the environment around the shelter, the location of the station, the time of observation, and the method used to calculate mean temperature. Often several of these occur at the same time, as is often the case with the introduction of automatic weather stations that is occurring in many parts of the world. Before one can reliably use such climate data for analysis of long-term climate change, adjustments are needed to compensate for the nonclimatic discontinuities. GHCN temperature data include two different datasets: the original data and a homogeneity-adjusted dataset. All homogeneity testing was done on annual time series. The homogeneity-adjustment technique used two steps.

The first step was creating a homogeneous reference series for each station (Peterson and Easterling 1994). Building a completely homogeneous reference series using data with unknown inhomogeneities may be impossible, but we used several techniques to minimize any potential inhomogeneities in the reference series. The first of these sought the most highly correlated neighboring station, from which a correlation analysis was performed on the first difference series: $FD_1 = (T_2 - T_1)$. A change in thermometers would alter only 1 yr of data in a first difference series, whereas with the original data such a change alters all following years.

The second minimizing technique was building a first-difference reference series from which the correlations for each year were calculated without including the target year's data. Therefore, if a first-difference year was excessively warm due to a discontinuity, the determination of that year's first-difference reference series data point would not be impacted at all by the discontinuity. In creating each year's first difference reference series, we used the five most highly correlated neighboring stations that had enough data to accurately model the candidate station. From this modeling, the probability of this similarity being due to chance was less than 0.01 as determined by a Multivariate Randomized Block Permutation (MRBP) test using Euclidean distance (Mielke 1984, 1986, 1991).

The final technique we used to minimize inhomogeneities in the reference series used the mean of the central three values (of the five neighboring station values) to create the first difference reference series. In doing so, it was assumed that if there was a significant discontinuity in one of the five stations that year, that station would most likely have the highest or lowest value. The final step in creating the reference series was to turn the first difference reference series into a station time series ($T_1 = 0$; $T_2 = T_1 + \mathrm{FD}_1$) and adjust the values so the final year's value of the reference series equaled the final year's temperature from the candidate series.

With the reference series created, the second step for detecting the inhomogeneities examined the difference series between a station and its reference series (Easterling and Peterson 1995a). It was assumed that the reference series accurately reflected the climate of the region so that any significant departures from climatology could be directly associated with discontinuities in the station data. To look for such a change point, a simple linear regression was fitted to the part of the difference series before the year being tested and another after the year being tested. This test is repeated for all years of the time series (with a minimum of 5 yr in each section), and the year with the lowest residual sum of the squares was considered the year with a potential discontinuity. A residual sum of the squares from a single regression through the entire time series was also calculated. The significance of the two-phase fit was tested with a likelihood ratio statistic using the two residual sum of the squares and the difference in the means of the difference series before and after the discontinuity was evaluated using Student's t-test.

If the discontinuity was determined to be significant, the time series was subdivided into two at that year. Each of these smaller sections were similarly tested. This subdividing process continues until no significant discontinuities were found or the time series was too short to test (< 10 yr). Each of the discontinuities that have been identified was further tested using a Multiresponse Permutation Procedure (MRPP; Mielke 1991). The MRPP test is nonparametric and compares the Euclidean distances between members within each group with the distances between all members from both groups, then returns a probability that two groups more different could occur by random chance alone. The two groups were the 12-yr windows on either side of the discontinuity, though the window is truncated at a second potential discontinuity. If the discontinuity was significant at the 95% level (a probability of 0.05), it was considered a true discontinuity. The adjustment that was applied to all data points prior to the discontinuity was the difference in the means of the (station minus reference) difference series' two windows.

All the homogeneity testing was done with annual time series because annual reference series are more robust than monthly series. However, the effects of most discontinuities vary with the season. Therefore, monthly reference series were created and differences in the difference series for each month were calculated both before and after the discontinuity. These potential monthly adjustments were then smoothed with a nine-point binomial filter and all the months were adjusted slightly so the mean of all the months equaled the adjustment determined by the annual analysis.

Our approach to adjusting historical data is to make them homogeneous with present-day observations, so that new data points can easily be added to homogeneity-adjusted time series. Since the primary purpose of homogeneity-adjusted data is long-term climate analysis, we only adjusted time series that had at least 20 yr of data. Also, not all stations could be adjusted. Remote stations for which we could not produce an adequate reference series (the correlation between first-difference station time series and its reference time series must be 0.80 or greater) were not adjusted. The homogeneity-adjusted version of GHCN includes only those stations that were deemed homogeneous and those stations we could reliably adjust to make them homogeneous. Therefore, the homogeneity-adjusted GHCN dataset is smaller than the original data version and the earliest data in the homogeneity-adjusted time series is 1850.

One thousand two hundred twenty-one homogeneity-adjusted stations in the United States were computed using a different technique. These are high quality rural stations taken directly from the U.S. Historical Climatology Network (U.S. HCN; Easterling et al. 1996a), a sister project to GHCN. These data were adjusted using a metadata approach as part of the creation of the U.S. HCN and their adjusted time series were directly incorporated into GHCN. For climate analysis confined to the United States, the U.S. HCN is the preferred dataset because its stations are well distributed, mostly rural stations that were selected based upon their location and their station history metadata indicating that they were the best stations available in the United States for long-term climate analysis.

A great deal of effort went into the homogeneity adjustments. Yet the effects of the homogeneity adjustments on global average temperature trends are minor (Easterling and Peterson 1995b). However, on scales of half a continent or smaller, the homogeneity adjustments can have an impact. On an individual time series, the effects of the adjustments can be enormous. These adjustments are the best we could do given the paucity of historical station history metadata on a global scale. But using an approach based on a reference series created from surrounding stations means that the adjusted station's data is more indicative of regional climate change and less representative of local microclimatic change than an individual station not needing adjustments. Therefore, the best use for homogeneity-adjusted data is *regional* analyses of long-term climate trends (Easterling et al. 1996b). Though the homogeneity-adjusted data are more reliable for long-term trend analysis, the original data are also available in GHCN and may be preferred for most other uses given the higher density of the network.

7. Metadata

For long-term climate stations, there are two types of metadata. The first type is historical metadata that indicate changes with time. Many countries maintain detailed station history files that document relevant station attributes such as the type of thermometer used and when the instruments changed. Such metadata are very difficult if not impossible to acquire on a global basis. Therefore, historical metadata are not available for GHCN. The second type of metadata is information about the stations and their present environments. We have compiled a variety of this type of metadata that will facilitate research applications using GHCN.

Like most station databases, these metadata start off with station name, latitude, longitude, and elevation. Wherever possible, these were obtained from the current WMO station listings (WMO 1996b). Some stations in GHCN do not have elevation metadata. To provide all stations with some elevation information, an elevation value was interpolated to the station location from a 5-min gridded elevation database (Row and Hastings 1994) and this elevation is provided in addition to official station elevations. In areas with significant orography, these interpolated metadata will have limited specific accuracy. But they can provide useful information about the station's elevation.

Each station in GHCN was located on Operational Navigation Charts (ONC). With a scale of 1:1 000 000 (1 cm on the map covers 10 km on the earth), ONC were created by the U.S. Department of Defense. Available through the National Oceanic and Atmospheric Administration (NOAA), these charts are used by pilots all over the world. ONC have elevation contours, outlines of urban areas, locations of airports and towns, and for most of the world, a simple vegetation classification. We located every GHCN station on ONC to both quality control station locations and to derive five types of metadata.

1) *Population.* Examining the station location on an ONC would determine whether the station was in a rural or urban area. If it was an urban area, the population of the city was determined from a variety of sources. We have three population classifications: rural, not associated with a town larger than 10 000 people; small town, located in a town with 10 000 to 50 000 inhabitants; and urban, a city of more than 50 000. In addition to this general classification, for small towns and cities, the approximate population is provided.

   These population metadata represent a valuable tool for climate analysis; however, the user must bear in mind the limitations of these metadata. While we used the most recent ONC available, in some cases the charts or the information used to create the charts were compiled a decade ago or even earlier. In such cases the urban boundaries in rapidly growing areas were no longer accurate. The same is true for the urban populations. Wherever possible, we used population data from the then-current United Nations Demographic Yearbook (United Nations 1993). Unfortunately, only cities of 100 000 or more inhabitants were listed in the yearbook. For smaller cities we used population data from several recent atlases. Again, although the atlases were recent, we do not know the date of source of the data that went into creating the atlases. Additionally, this represents only one moment in time; an urban station of today may have been on a farm 50 years ago, though it is probably valid to assume that if a station is designated rural now, it was most likely rural 50 years ago. Knowing the importance of avoiding the effect of urban warming by preferring rural stations in climate analysis, these population metadata have been used as one of the criteria in the initial selection of the Global Climate Observing System (GCOS) Surface Network (Peterson et al. 1997a).

2) *Airport locations.* Airports are, of course, clearly marked on ONC charts. If a station is located at an airport, this information along with the distance from its associated city or small town (if present) are included as part of GHCN metadata.

3) *Topography.* ONC make detailed orography available to pilots. We used this information to classify the topography around the station as flat, hilly, or mountainous. Additionally we differentiated between mountain valley stations and the few mountaintop stations that can provide unique insights into the climate of their regions.

4) *Coastal locations.* Oceanic influence on climate can be significant, so these metadata include (a) if the station is located on an island of less than 100 km^2 or less than 10 km in width at the station location, (b) if the station is located within 30 km of the coast it is labeled as coastal and the distance to the coast is provided, and (c) if the station is adjacent to a large (greater than 25 km^2) lake, that too is noted because it can have an influence on a station's climate.

5) *Vegetation.* If the station is rural, the vegetation for that location is documented. The classifications used on the ONC are forested, clear or open, marsh, ice, and desert. Not all ONC had complete vegetation data, so these metadata are not available for all stations. An additional source of vegetation data is included in GHCN metadata: the vegetation listed at the nearest grid point to each station in a $0.5° \times 0.5°$ gridded vegetation dataset (Olson et al. 1983). This vegetation database creates a global vegetation map of 44 different land ecosystem complexes comprising seven broad groups. These metadata do not indicate the exact vegetation type at the station location, but they do provide useful information. In particular, an ecosystem classification can be used to some degree as a surrogate for climate regions since vegetation classes depend, to a large extent, on climate.

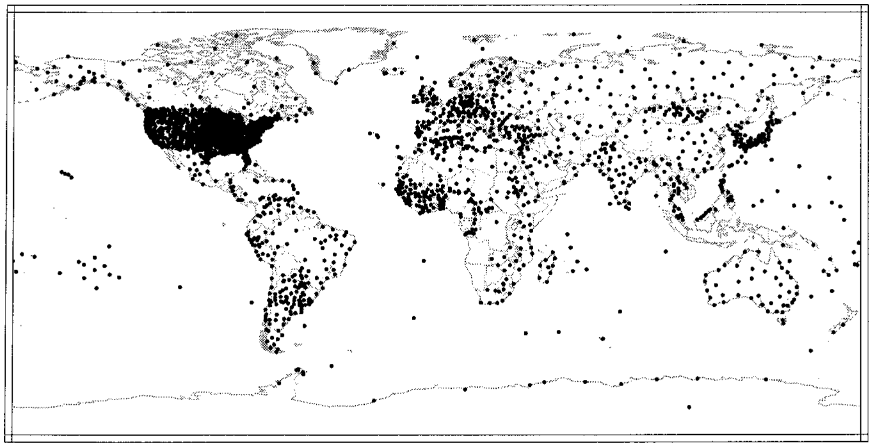

FIG. 5. GHCN mean temperature stations that can be regularly updated. Many of these stations will be updated with maximum and minimum temperature data as well. The three sources of data for updating are the U.S. Historical Climatology Network, a subset of the U.S. First Order stations, and monthly CLIMAT reports transmitted over the Global Telecommunications System.

8. Updates

Thirty-one different sources contributed temperature data to GHCN. Many of these were acquired through second-hand contacts and some were digitized by special projects that have now ended. Therefore, not all GHCN stations will be able to be updated on a regular basis. Of the 31 sources, we are able to perform regular monthly updates with only three of them (Fig. 5). These are 1) the U.S. HCN, 1221 high quality, long-term, mostly rural stations in the United States; 2) a 371-station subset of the U.S. First Order station network (mostly airport stations in the United States and U.S. territories such as the Marshall and Caroline Islands in the western Pacific); and 3) 1502 Monthly Climatic Data for the World stations (subset of those stations around the world that report CLIMAT monthly code over the Global Telecommunications System and/or mail reports to NCDC). Other stations will be updated or added to GHCN when additional data become available, but this will be on a highly irregular basis.

9. Concluding remarks

In creating GHCN version 2, the goal was to produce a high quality global climate database suitable for the widest possible usages. This required breaking considerable new ground: adjusting the data for inhomogeneities using an approach that did not rely on sparse station history information; providing the original data in addition to homogeneity adjusted; expanding the database to include maximum and minimum as well as mean temperature data; increasing the number of stations, which facilitates regional climate analysis, by incorporating over 30 source datasets and digitizing selected stations; eliminating duplicate stations with an approach that both preserves duplicates if they provide additional information and guarantees that no discontinuities are created by inappropriate concatenation of time series; applying a multifaceted quality control approach to ferret out a wide variety of potential problems in the data; providing expanded station metadata, ranging from population to orography; and instigating regular updating of the data. The resultant product, GHCN version 2, is a mean monthly maximum, minimum, and mean temperature dataset that is available to researchers and others free of charge via the World Wide Web.

There are 4.7 million station months of temperature data in GHCN, starting in 1701 and continuing to the present. Derived from 300 million individual readings of thermometers, GHCN embodies the systematic observations of our environment by tens of thousands of individuals over centuries of human history. We feel honored to be a part of this process and gratefully acknowledge the debt we owe to the largely selfless work of individual weather observers. In this time of concern about our global climate, these data are becoming increasingly important and the contributions conscientious individual weather observers made over the past decades and centuries promise to help the climate research community answer questions about the decades to come.

10. Availability

GHCN version 2 is available free of charge from the National Climatic Data Center's Web site: http://www.ncdc.noaa.gov/ol/climate/research/ghcn/ghcn.html. You may want to double-check your Web site entry to make sure it brings up GHCN's home page (ol stands for online).

Acknowledgments. Creating GHCN was far more than a two-person endeavor. We gratefully acknowledge the contributions to various aspects of GHCN that were made by David Easterling, Richard Schmoyer, Thomas Karl, Vyachevslav Razuvaëv, Tom Boden, Paul Jamason, Dale Kaiser, Rob Quayle, John Griffiths, Mike Crowe, Robert Cushman, Tim Owen, Catherine Godfrey, and Red Ezell. We would also like to thank Philip Jones for his review of this article and for insightful discussions over the last several years. GHCN is jointly produced by the National Climatic Data Center/NESDIS/NOAA, the Carbon Dioxide Information Analysis Center/Oak Ridge National Laboratory/DOE, and Arizona State University. Funding for GHCN has been provided by DOE under Interagency Agreement DE-AI05-900R21956 and is currently being provided by NOAA Climate and Global Change Climate Change Data and Detection program.

References

Adkison, M. D., R. M. Peterman, M. F. LaPointe, D. M. Gillis, and J. Korman, 1996: Alternative models of climatic effects on sockeye-salmon. *Fish. Oceanogr.,* **5,** 137–152.

Balling, R. C., 1995: Analysis of German climatic variations during the period of instrumental record. *Geophys. Res. Lett.,* **22,** 223–226.

Baranyi, T., and A. Ludmany, 1995: Role of the solar main magnetic dipole field in the solar-tropospheric relations. 1. Semiannual fluctuations in Europe. *Ann. Geophys.* **13,** 427–436.

Bradley, R. S., P. M. Kelly, P. D. Jones, C. M. Goodess, and H. F. Diaz, 1985: A climatic data bank for northern hemisphere land areas, 1851–1980. DPE Tech. Rep. TR017, 335 pp. [Available from Carbon Dioxide Research Division, Office of Basic Energy Services, Office of Energy Research, U.S. Department of Energy, Washington, DC 20545.]

Brown, J. F., T. R. Loveland, J. W. Merchant, B. C. Reed, and D. O. Ohlen, 1993: Using multisource data in global land-cover characterization. *Photogr. Eng. Remote Sensi.,* **59,** 977–978.

Clayton, H. H., 1927: *World Weather Records*. Vol. 79, Smithsonian Miscellaneous Collection, 1196 pp.

Easterling, D. R., and T. C. Peterson, 1995a: A new method for detecting and adjusting for undocumented discontinuities in climatological time series. *Int. J. Climatol.,* **15,** 369–377.

——, and ——, 1995b: The effect of artificial discontinuities on recent trends in minimum and maximum temperatures. *Atmos. Res.,* **37,** 19–26.

——, T. R. Karl, E. H. Mason, P. Y. Hughes, and D. P. Bowman, 1996a: United States Historical Climatology Network (U.S. HCN) Monthly Temperature and Precipitation Data. ORNL/CDIAC-87, NDP-019/R3, 280 pp. [Available from Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831.]

——, T. C. Peterson, and T. R. Karl, 1996b: On the development and use of homogenized climate data sets. *J. Climate,* **9,** 1429–1434.

——, and Coauthors, 1997: Maximum and minimum temperature trends for the globe. *Science,* **277,** 364–367.

Epperson, D. L., J. M. Davis, P. Bloomfield, T. R. Karl, A. L. McNab, and K. P. Gallo, 1995: Estimating the urban bias of surface shelter temperatures using upper-air and satellite data. Part I: Development of models predicting surface shelter temperatures. *J. Appl. Meteor.,* **34,** 340–357.

Griffiths, J. F., 1997: Some problems of regionality in application of climate change. *J. Int. Soc. Biomet.,* in press.

Groisman, P. Ya., T. R. Karl, and R. W. Knight, 1994a: Observed impact of snow cover on the heat-balance and the rise of continental spring temperatures. *Science,* **263,** 198–200.

——, ——, ——, and G. L. Stenchikov, 1994b: Changes of snow cover, temperature, and radiative heat-balance over the Northern Hemisphere. *J. Climate,* **7,** 1633–1656.

Gutzler, D. S., 1996: Low-frequency ocean–atmosphere variability across the tropical western Pacific. *J. Atmos. Sci.,* **53,** 2773–2785.

Jones, P. D., 1994: Hemispheric surface air temperature variations: A reanalysis and an update to 1993. *J. Climate,* **7,** 1794–1802.

——, 1995: Land surface temperatures—Is the network good enough? *Climate Change,* **31,** 545–558.

——, R. Raper, R. S. Bradley, H. F. Diaz, P. M. Kelly, and T. M. L. Wigley, 1986: Northern Hemisphere surface air temperature variations: 1851–1984. *J. Climate Appl. Meteor.,* **25,** 161–179.

——, T. J. Osborn, and K. R. Briffa, 1997: Estimating sampling errors in large-scale temperature averages. *J. Climate,* **10,** 2548–2568.

Karl, T. R., and Coauthors, 1993: A new perspective on recent global warming: Asymmetric trends of daily maximum and minimum temperatures. *Bull. Amer. Meteor. Soc.,* **74,** 1007–1023.

——, R. W. Knight, and J. R. Christy, 1994: Global and hemispheric temperature trends: Uncertainties related to inadequate spatial sampling. *J. Climate,* **7,** 1114–1163.

Lanzante, J. 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.

Legates, D. R., and C. J. Willmott, 1990: Mean seasonal and spatial variability in global surface air temperature. *Theor. Appl. Climatol.,* **41,** 11–21.

Mielke, P. W., 1984: Meteorological applications of permutation techniques based on distance functions. *Handbook of Statistics,* P. R. Krishnaiah and P. K. Sen, Eds., Vol. 4, North Holland, 813–830.

——, 1986: Non-metric statistical analysis: Some metric alternatives. *J. Statist. Plann. Inference,* **13,** 377–387.

——, 1991: The application of multivariate permutation methods based on distance functions in the earth sciences. *Earth-Sci. Rev.,* **31,** 55–71.

Olson, J. S., J. A. Watts, and L. J. Allison, 1983: Carbon in live vegetation of major world ecosystems. ORNL-5862, Environmental Science Division Publ. 1997, 164 pp. [Available from Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831.]

Peterson, T. C., and D. R. Easterling, 1994: Creation of homogeneous composite climatological reference series. *Int. J. Climatol.,* **14,** 671–679.

——, and J. F. Griffiths, 1996: Colonial era archive data project. *Earth Sys. Monitor,* **6,** 8–16.

——, and ——, 1997: Historical African data. *Bull. Amer. Meteor. Soc.,* **78,** 2869–2872.

——, H. Daan, and P. Jones, 1997a: Initial selection of a GCOS surface network. *Bull. Amer. Meteor. Soc.,* **78,** 2145–2152.

——, R. S. Vose, R. Schmoyer, and V. Razuvaëv, 1997b: Quality control of monthly climate data: The GHCN experience. *Int. J. Climatol.,* in press.

Quereda, J., and E. Monton, 1994: *¿Hacia un Cambio Climático*? Jaime I University, 27 pp.

——, and ——, 1996: L'Elevation de la temperature en Espagne mediterraneenne: Tendance naturalle ou effet de lurbanization? *Publications de l'Association Internationale de Climatologie,* **9,** 487–495.

Row, L. W., III, and D. A. Hastings, 1994: *TerrainBase Worldwide Digital Terrain Data, Documentation and CD-ROM,* National Geophysical Data Center, CD-ROM. [Available from National Geophysical Data Center, 325 Broadway, Boulder, CO 80303.]

Spangler, W. M. L., and R. L. Jenne, 1990: World monthly surface station climatology (and associated data sets). National Center for Atmospheric Research Dataset DS570.0, 20 pp. [Available from NCAR, P.O. Box 3000, Boulder, CO 80307.]

Tayanc, M., M. Karaca, and O. Yenigun, 1997: Annual and seasonal air temperature trend patterns of climate change and urbanization effects in relation to air pollutants in Turkey. *J. Geophys. Res.,* **102,** 1909–1919.

United Nations, 1993: *Demographic Yearbook 1993.* Department for Economic and Social Information and Policy Analysis, United Nations, 1032 pp.

van Dobben de Bruyn, C. S., 1968: *Cumulative Sum Tests: Theory and Practice. Griffin's Statistical Monographs and Courses,* No. 24, Griffin, 82 pp.

Vose, R. S., R. L. Schmoyer, P. M. Steurer, T. C. Peterson, R. Heim, T. R. Karl, and J. Eischeid, 1992: The Global Historical Climatology Network: Long-term monthly temperature, precipitation, sea level pressure, and station pressure data. ORNL/CDIAC-53, NDP-041, 325 pp. [Available from Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831.]

Wernstedt, F. L., 1972: *World Climatic Data.* Climatic Data Press.

Willmott, C. J., S. M. Robeson, and J. J. Feddema, 1994: Estimating continental and terrestrial precipitation averages from rainguage networks. *Int. J. Climatol.,* **14,** 403–414.

WMO, 1996a: Climatological Normas (CLINO) for the Period 1961–1990. Publ. 847, 768 pp.

——, 1996b: Weather Reporting, Observing Stations. Publ. 9. [Available from World Meteorological Organization, Case Postale No. 2300, CH-1211, Geneva 2, Switzerland.]

Young, K. C., 1993: Detecting and removing inhomogeneities from long-term monthly sea-level pressure time series. *J. Climate,* **6,** 1205–1220.