An Overview of the Global Historical Climatology Network-Daily Database

MATTHEW J. MENNE, IMKE DURRE, RUSSELL S. VOSE, BYRON E. GLEASON, AND TAMARA G. HOUSTON

National Climatic Data Center, Asheville, North Carolina

(Manuscript received 18 May 2011, in final form 23 February 2012)

ABSTRACT

A database is described that has been designed to fulfill the need for daily climate data over global land areas. The dataset, known as Global Historical Climatology Network (GHCN)-Daily, was developed for a wide variety of potential applications, including climate analysis and monitoring studies that require data at a daily time resolution (e.g., assessments of the frequency of heavy rainfall, heat wave duration, etc.). The dataset contains records from over 80 000 stations in 180 countries and territories, and its processing system produces the official archive for U.S. daily data. Variables commonly include maximum and minimum temperature, total daily precipitation, snowfall, and snow depth; however, about two-thirds of the stations report precipitation only. Quality assurance checks are routinely applied to the full dataset, but the data are not homogenized to account for artifacts associated with the various eras in reporting practice at any particular station (i.e., for changes in systematic bias).

Daily updates are provided for many of the station records in GHCN-Daily. The dataset is also regularly reconstructed, usually once per week, from its 20+ data source components, ensuring that the dataset is broadly synchronized with its growing list of constituent sources. The daily updates and weekly reprocessed versions of GHCN-Daily are assigned a unique version number, and the most recent dataset version is provided on the GHCN-Daily website for free public access. Each version of the dataset is also archived at the NOAA/National Climatic Data Center in perpetuity for future retrieval.

1. Introduction

The analysis of multidecadal climate trends and variability is commonly based on monthly and annual summaries of station-based weather data, and records of this time resolution have been widely available in digital form for decades (e.g., Jones et al. 1985, 1986; Vose et al. 1992). However, monthly means and averages are not sufficient for all climate applications. For example, the analysis of changes in the length of the growing season (Kunkel et al. 2004), changes in the frequency of heavy precipitation (Min et al. 2011), and changes in heat wave frequency and duration (Della Marta et al. 2007) all require data at least at the daily resolution. Unfortunately, daily data are comparatively less accessible than monthly values, in part because of impediments (e.g., mandates for cost recovery) in many countries for releasing daily climate data for widespread public use (Alexander et al. 2006). This relative paucity of daily data hampers climate change analysis and model comparison studies (Trenberth et al. 2007).

Here, a database is described whose aim is to address the need for daily climate data over global land areas. The database, known as the Global Historical Climatology Network (GHCN)-Daily dataset, contains daily data from over 80 000 surface stations worldwide, about two-thirds of which are for precipitation only. Like its counterpart for monthly data (Peterson and Vose 1997; Peterson et al. 1998; Lawrimore et al. 2011), GHCN-Daily is composed of daily weather reports from numerous sources that have been merged and subjected to a common suite of quality assurance (QA) reviews. Below, GHCN-Daily's component data sources, methods for data integration and quality assurance, and the resulting spatial and temporal coverage of the dataset are described. The focus is on the core elements of temperature and precipitation, but the database also contains observations for snowfall, snow depth, and numerous other variables. Coverage of these elements is more limited in space and time.

2. Data sources

During the last several decades, the Global Telecommunication System (GTS), operated under the auspices

Corresponding author address: Matthew J. Menne, National Climatic Data Center, 151 Patton Avenue, Asheville, NC 28801.
E-mail: matthew.menne@noaa.gov

DOI: 10.1175/JTECH-D-11-00103.1

TABLE 1. Sources and contacts for the international collection.

Region/country	Source/contact
Countries in West Africa	Météo-France
Countries in East Africa	Kenyan Meteorological Department/P. Ambenji
South Africa and Namibia	South African Weather Service/R. S. Vose
China	National Climate Center China Meteorological Administration/D. R. Easterling
India, Japan, and Thailand	National Center for Atmospheric Research
Brazil	Agência Nacional de Energia Elétrica (ANEEL)/P. Ya. Groisman
Paraguay, Uruguay, and Venezuela	NOAA/Climate Diagnostics Center
Mexico	National Weather Service of Mexico/A. Douglas
Countries in the former USSR	Bilateral Exchange/P. Ya. Groisman
Europe	European Climate Assessment and Dataset (early versions)/A. Klein Tank

of the World Meteorological Organization (WMO), has allowed National Meteorological and Hydrological Services (NMHSs) to share a wide variety of meteorological data regionally and worldwide. However, there has been no formal mechanism or requirement to share daily data via the GTS and no central repository for daily climate reports from national surface networks. In practice, the transmission of daily climate summaries has been treated as optional even for the network of stations that report temperature, precipitation, surface pressure, etc., at fixed synoptic hours [i.e., every 3 or 6 h for stations in the Regional Basic Synoptic Network (RBSN)]. Similarly, stations in the WMO's Regional Basic Climate Network (largely a subset of the RBSN) are only required to provide a monthly climate summary known as CLIMAT and likewise do not consistently transmit daily summaries within their synoptic messages (WMO 2003).

Given this context, the goal in creating a global land surface daily dataset was to maximize the spatial coverage of daily weather data by acquiring historical data from as many stations in as many national observing networks as possible. Several complementary data acquisition tactics were used. The first was to exploit contacts with representatives from national meteorological and hydrological centers around the world to request contribution of their respective data collections. The earliest of these efforts lead to the development of the Global Daily Climatology Network (GDCN; Gleason et al. 2002) dataset. Not surprisingly, GDCN also contained a large collection of U.S. data. However, since GDCN's release, a number of additional archives at National Oceanic and Atmospheric Administration/National Climatic Data Center (NOAA/NCDC) that contain daily data for the United States and its territories have been integrated comprehensively into GHCN-Daily. The second data collection tactic leveraged bilateral and international initiatives, such as the Global Climate Observing System (GCOS) program, which works to facilitate the free exchange of daily data from GCOS surface stations (Peterson et al. 1997). Bilateral agreements, in particular, have resulted in large contributions (hundreds to thousands of station records) from a number of countries. The last (and least preferred) tactic used the optional daily summaries that get transmitted as part of the GTS synoptic messages.

These varied attempts to acquire daily data can be loosely classified into four broad categories: 1) the international collection; 2) the U.S. collection; 3) government exchange data; and 4) the global summary of the day. A brief synopsis of each of these categories follows.

The international collection contains historical records for approximately 20 000 locations outside the United States (from over 100 different countries) and largely reflects the data collection efforts for GDCN. Well over 200 million values of maximum and minimum temperatures and total daily precipitation are included in this collection. As shown in Table 1, international collection records were generally obtained through personal contacts in various countries. As discussed in section 5, contributions to the international collection have resulted in particularly dense station networks with daily precipitation totals in Brazil, South Africa, and India, although the data from this collection are purely historical and are not updated. Precipitation records end generally in the late 1990s for Brazil and South Africa and in 1970 for India.

The U.S. collection contains daily data from a dozen separate datasets archived at NOAA/NCDC. As shown in Table 2, these archives include some of the earliest observations available for the United States (from the U.S. Forts and Voluntary Observer Program covering much of the nineteenth century; Dupigny-Giroux et al. 2007) as well as the latest measurements from the state-of-the-art climate monitoring stations that make up the U.S. Climate Reference Network (deployed early in the twenty-first century). GHCN-Daily thus contains the most complete collection of U.S. daily data available. Data for the United States are comprehensively updated in GHCN-Daily from a number of real-time and time-delayed data feeds. In addition, beginning with the

TABLE 2. Data sources composing the U.S. collection.

		Data source code
U.S. Cooperative Summary of the Day (NCDC DSI-3200).	Dataset includes daily observations from over 20 000 stations in the United States and its territories. Although most measurements are taken once per day by volunteer observers as part of the NOAA/National Weather Service (NWS) Cooperative Observer Program (COOP), manual and automated measurements from some "first order" synoptic sites are also included. Some daily records extend back to the late 1800s in this dataset, but most do not begin until 1948 or later. Corrections to fix processing or reporting errors are sometimes made, but the dataset ends in December 2010. The time of observation varies by station.	0
U.S. Cooperative Summary of the Day–CDMP (NCDC DSI-3206).	Dataset includes daily summaries primarily for the years before 1948 from more than 11 000 COOP stations that were keyed as part of NCDC Climate Database Modernization Program (CDMP). Corrections are made occasionally, but new observations are not added. The time of observation varies by station.	6
U.S. First Order Summary of the Day (NCDC DSI-3210).	Dataset contains historical and present-day manual and automated observations from approximately 1600 synoptic stations, including U.S. first-order stations, a selection of Canadian sites, and U.S.-operated stations in other countries. Observations for a specific year and month are added 2–3 months after they were taken, and corrections to historical data may occasionally be applied. These observations are generally with respect to the 24-h period ending at local midnight.	X
U.S. ASOS Summary of the Day, 2000–05 (NCDC DSI-3211).	Dataset contains observations for nearly 900 U.S. Automated Surface Observing System (ASOS) stations between October 2000 and December 2005. The dataset is no longer updated, but corrections are made occasionally. The observations are with respect to the 24-h period ending at local midnight.	B
ASOS Summary of the Day, 2006–present (from NCDC DSI-3505).	Dataset contains reports from U.S.-operated ASOS stations beginning in January 2006. Data are updated daily with observations from the previous day or 2 days before. These observations are with respect to the 24-h period ending at local midnight.	
Surface METAR Monthly Airways Extract (NCDC DSI-6407).	Dataset contains hourly surface weather observations for 1996–2002 at major airports that include a daily summary with respect to the 24-h period ending at local midnight.	M
U.S. Forts and Voluntary Observers	Dataset contains observations from the CDMP's nineteenth-century Forts and Voluntary Observers Database Build Project. Data come from the U.S. Army forts in the early 1800s and from volunteer observer networks managed by the Smithsonian Institution in the mid- and late 1800s. The volunteer networks evolved into the Weather Bureau's Cooperative Observer Network, which continues to operate as the NOAA/NWS COOP. Newly keyed forts and volunteer data are added periodically. Observation times vary.	F
U.S. Climate Reference Network Daily Summary	Daily climate summaries from the U.S. Climate Reference Network. Data begin as early as 2001 and are ongoing. Data are updated at least weekly. Summaries are for the 24-h period ending at local midnight.	R
Real-time Cooperative Summary of the Day updates from the High Plains Regional Climate Center	Provides real-time updates to records for several thousand U.S. COOP stations. Updates are provided by the High Plains Regional Climate Center from observations transmitted by NOAA on a daily basis and represent summaries for the previous 24 h. Observation times vary, but summaries are generally for 24-h periods ending in the morning local time.	H
Latest Cooperative Summary of the Day updates from WxCoder3 (DSI 3207)	Updates for the U.S. Cooperative Summary of the Day data received through the NOAA/NWS WxCoder3 system. Updates are provided on a monthly basis and represent summaries for the previous 24 h. Observation times vary, but are commonly for 24-h periods ending in the morning local time. Begins in 2011.	7
Latest U.S. Cooperative Summary of the Day updates digitized from paper forms	Newly keyed data from COOP forms. Observation times vary.	K
CoCoRaHS	Provides daily rain and snow measurements from CoCoRaHS volunteers. Data begin as early as 1998 and are updated daily. Observation times vary but observers are encouraged to report at 0700 local time.	N

TABLE 3. Sources for government exchange data.

Region/Country	Source/Contact
Canada	Environment Canada/Robert Morris
Australia	Bureau of Meteorology/Cathy Toby
Belarus	Bilateral Exchange/P. Ya Groisman
Ukraine	Bilateral Exchange/P. Ya Groisman
Greater Europe	European Climate Assessment and Dataset (latest updates) (http://eca.knmi.nl/)
Russia	All Russian Research Institute of Hydrometeorological Information–World Data Center (http://www.meteo.ru)
556 GCOS surface stations	Various contacts

2011 data year, dataset index (DSI) 3200 (U.S. cooperative summary of the day) is no longer updated, having been superseded by GHCN-Daily (DSI 9101), which now serves as the official archive for U.S. daily Cooperative Observer Network data.

Government exchange data (Table 3) refer to data collected through official GCOS or bilateral agreements. Under such agreements, Environment Canada and the Australian Bureau of Meteorology, for example, have provided their complete digital, daily database for inclusion in GHCN-Daily (with more than 7500 and 17 000 station records, respectively). Other NMHSs, such as the All-Russian Research Institute of Hydrometeorological Information, have provided large subsets (hundreds of station records) of their digital archives. The European Climate Assessment and Dataset (ECA&D; Klein Tank et al. 2002) project also provides a large collection of government exchange data and currently contains daily data from over 1500 stations in more than 50 countries. Early versions (before 2004) of the ECA&D data were used to form part of the international collection; however, more recently (beginning in 2011), the latest version of the ECA&D is operationally ingested into GHCN-Daily to incorporate monthly updates to the European data. Finally, under the auspices of GCOS, 76 different NMHSs have officially provided daily data for just over half of the 1000+ GCOS Surface Network (GSN) stations that make up the GSN network implementation. Although mechanisms have been set up to regularly update (at least monthly or annually) much of the government exchange data (e.g., Canada, Australia, ECA&D, Uzbekistan, Cypress, Iran, Latvia), such mechanisms have yet to be implemented for most GCOS stations, although new sets of historical data are periodically added.

The global summary of the day contains 24-h summaries encoded in the special "climatological code" group transmitted with SYNOP reports on the GTS. These reports are archived in NCDC's Integrated Surface Dataset (DSI-3505) and the 24-h summary period purportedly ends at midnight (i.e, 2400) UTC. Daily maximum and minimum temperatures from this source are included in GHCN-Daily only when provided as a nominal 24-h climatological summary as indicated in the SYNOP messages, whereas daily precipitation totals are used if they can be summed from two 12-h or four 6-h subtotals (as provided in standard SYNOP code groups). Subdaily summations are identified by associated "measurement" flag codes in the GHCN-Daily data format. Daily summaries from the global summary of the day may differ significantly from climate summaries with 24-h periods ending at local midnight (or at other hours used by convention at a particular NMHS), particularly in the case of precipitation. Nevertheless, data from this GTS source are available for a number of locations that are not contained in any other data archive available to NCDC, and they provide the only source of updates for many stations.

3. Data integration

As shown in Table 4, the process of integrating data from multiple sources into the GHCN-Daily dataset takes place in three steps: 1) eliminating source data for stations whose location is unknown or questionable; 2) classifying each station in a source dataset either as one that is already represented in GHCN-Daily or as a new site; and 3) combining the data from the different source datasets to form comprehensive station records. The first two of these steps are performed whenever a new source dataset or additional stations become available. The combining of data is part of an automated process that fully regenerates GHCN-Daily on a regular basis (usually once per week) using the latest versions of all sources. These three steps are explained further below.

In the initial step, a station's record from a particular source dataset is considered for inclusion in GHCN-Daily provided it meets the following conditions: First, it must be identified with a location name, latitude, and longitude using the metadata associated with the source dataset or from other standard station history information. Second, its period of record must contain 100 or more daily values for at least one of five core GHCN-Daily elements (maximum temperature, minimum temperature, precipitation, snowfall, or snow depth). Third, the record must not fail the "intra-source" duplicate check, which compares records from all stations within a source dataset. If more than 50% of a station's record is identical to the data from another station in the same source, the longer of the two records is retained for inclusion in GHCN-Daily, provided that the metadata indicate that the two sites are in close proximity (i.e., within

TABLE 4. Basic procedure for adding new source data to GHCN-Daily.

Step 1: Eliminate source data for stations whose location is unknown or questionable.	1.1	Associate station records with a location name, latitude, and longitude using the metadata provided with the source dataset or from other available station history information.
	1.2	Eliminate from consideration records with fewer than 100 values for all core elements.
	1.3	Check for duplicated station records within the new source.
Step 2. Classify each station in source data either as one that is already represented in GHCN-Daily or as a new site.	2 [alternative (a)]	Cross reference the source station ID with source IDs already combined in GHCN-Daily, or
	2 [alternative (b)]	Compare the similarity of the new source station records to stations records already contained in GHCN-Daily during their overlap period, or
	2 [alternative (c)]	Compare the coordinates and name of the new source stations to the station names and coordinates of stations already in GHCN-Daily.
Step 3. Combine the data from the different source datasets to form comprehensive station records.	3.1	Add station record as a source to an existing GHCN-Daily station record if a match is found in step 2 or as a new station if there is no match.
	3.2	Recombine all existing station records using the new source plus previously available sources according to the hierarchy of source precedence.

40 km). However, if two stations with matching records are more than 40 km apart, neither is incorporated into the dataset.

The second step is to determine whether, thanks to a different source, data for the same location are already contained in GHCN-Daily or whether the location of the station record is new to GHCN-Daily. Whenever possible, station records from a new source are matched to records already in GHCN-Daily via the station identification numbers (IDs). However, it is common for a single meteorological station to have multiple network affiliations, which means that different source datasets may index the same station data to different IDs. Station lists (e.g., as supplied by an NMHS) are sometimes available that cross-reference IDs used by different organizations. To illustrate, data for Alabaster Shelby County Airport, Alabama, are indexed by Cooperative Observer Network ID 010116 in NCDC's 3200 and 3206 datasets (among others). Given the common ID, data from these two sources should likely be combined into one GHCN-Daily record. In NCDC DSI-3210 (Table 2), however, and in the various other sources for airport observations, data for this location are stored under WBAN ID 53864, which must be matched with the corresponding cooperative station ID using NCDC's Multinetwork Metadata System.

If cross-reference lists are not available, a new source of data for a particular station may be compared to station records already contained in GHCN-Daily. If data from the new source match the data for a station already added to GHCN-Daily at a rate of at least 50% for all elements during their common overlap period and the new station and the preexisting GHCN-Daily station are identified to be within 40 km of one another (based on their respective coordinates), then the new station data are added as an additional data source to the relevant GHCN-Daily station record already present in the dataset.

Finally, stations may be matched on the basis of their names and location alone. This strategy is more difficult to automate than the other two approaches because multiple stations within the same city or town may be identified with the same name and small differences in coordinates can be the result of either differences in accuracy or the existence of multiple stations in close proximity to each other. This type of matching was conducted for stations outside the United States whose data from the global summary of the day needed to be matched with data from the international collection.

The implementation of the above classification strategies yields a list of GHCN-Daily stations and an inventory of the source datasets to be integrated for each station. These lists form the basis for step 3, the integrating (or combining) of the data from the various sources to create GHCN-Daily. Combining takes place according to a hierarchy of data source precedence and in a manner that attempts to maximize the amount of data included while also minimizing the degree to which data from sources with different characteristics such as times of observation are mixed. Although precipitation, snowfall, and snow depth are allowed to come from separate sources during a particular month, maximum and minimum temperatures are considered together in

order to ensure that the temperatures for a particular station and day always originate from the same source. This is important, for example, in the case of the real-time data feeds for the United States and the global summary of the day data, which tend to have observations that apply to 24-h summary periods that differ from those reported by other sources. For this reason, these sources are used only if no observations are available from any other source for that station, month, and element. Among the other sources, each day is considered individually; if an observation for a particular station and day is available from more than one source, the observation from the most preferred source available is used in GHCN-Daily. The hierarchy of data sources used in cases of overlap is based on several criteria. In general, data that have received the greatest amount of scrutiny are chosen over fully automated, real-time data streams. At stations operated by the United States, sources providing a cooperative summary of the day are given preference over other data streams because they contribute the largest amount of data. For stations outside the United States, the official governmental exchange data are preferred over the international collection when summaries from these two sources are available for the same station, element, and day.

4. Quality assurance

The QA approach to GHCN-Daily is based on several basic design considerations. First, given the large number of station records, a growing number of meteorological elements, and frequent additions of both historical and real-time data, it is impractical to rely on network-wide manual verification of the outcome of quality assurance algorithms as is commonly done in many existing QA systems (e.g., Guttman and Quayle 1990; Hubbard et al. 2005; Kunkel et al. 2005). Rather, a fully automated QA system is necessary for GHCN-Daily that is reliable enough to run "unsupervised." Automated systems also have the advantage of providing traceable and reproducible results, which is a necessary component to tracking the provenance of climate data. At the same time, integration of new station records can introduce data problems that may go undetected by routine, automated QA checks. Such problems include undocumented changes to units of measure and the assignment of data records to incorrect station identifiers (Peterson et al. 1998). Consequently, the occasional application of additional automatic and semiautomatic fundamental data integrity checks is also necessary. Because of these design considerations, a multi-tiered QA approach was used. This approach consists primarily of routine, fully automated procedures with some additional overall data record integrity checks that are implemented occasionally (e.g., when a significant amount of historical data are added to the dataset) and that require some manual evaluation. Each of these procedures is described briefly below.

To begin, during routine processing, the data are first passed through a "format checking program" that looks for problems such as nonexistent months or days, invalid characters in data fields, and so forth. This routine sets offending records to missing. The primary purpose of this program is to ensure that our integration methods do not either introduce or retain records that violate the intended and documented GHCN-Daily data format. Next, a comprehensive sequence of fully automated QA procedures identifies daily values that violate 1 of 19 quality tests. Described in greater detail in Durre et al. (2010), these tests identify a variety of data problems, including the duplication of data records; exceedance of physical, absolute, and climatological limits; excessive temporal persistence; excessively large gaps in the distributions of values; internal inconsistencies among elements; and inconsistencies with observations at neighboring stations. This system flags approximately 0.3% of over 2 billion data values, and it has been estimated that 98%–99% of the values flagged are true data errors and only 1%–2% are false positives (i.e., valid observations erroneously flagged as bad; Durre et al. 2010). This level of performance was achieved through careful selection and evaluation of procedures and test thresholds using the techniques described by Durre et al. (2008).

Manual review of random samples of flagged values was used to set the test threshold of each procedure such that its false-positive rate is minimized. In addition, the tests are arranged in a deliberate sequence in which the performance of the later checks is thought to be enhanced by the error detection capabilities of the earlier ones. As a result of this comprehensive manual assessment during the QA development phase, the algorithms are effective at detecting the grossest errors as well as more subtle inconsistencies among elements without the typically higher rate of false positives of automated QA procedures (Schmidlin et al. 1995; Kunkel et al. 2005; You and Hubbard 2006).

The second tier of quality assurance includes record integrity checks, which are implemented only occasionally. These consist of checks for

- climatological means that are inconsistent with a station's location;
- large, systematic jumps in the annual mean of a record (such as might be caused by a shift in reporting units); and
- concentrations of values that fail automated QA procedures.

In addition, two checks have been manually performed to identify stations with grossly incorrect coordinates: 1) a comparison of each station's elevation to the Global One-Kilometer Base Elevation (GLOBE) dataset (Globe Task Team et al. 1999) and 2) a comparison of the long-term monthly station averages of maximum temperature, minimum temperature, and total precipitation to an independently constructed gridded dataset of monthly values (Legates and Willmott 1990a,b). This technique has helped identify cases of erroneous coordinates and data with incorrect reporting units or totals reported as zero rather than missing. Where an obvious manual correction to the coordinates was not apparent, the station records were "quarantined" and excluded from GHCN-Daily.

A semiautomatic method for identifying large jumps and other erratic behavior in time series of annual totals was also applied. Gross shifts in precipitation time series were identified by means of the standard normal homogeneity test (Alexandersson 1986) applied to station time series of annual precipitation totals computed from the daily data. The two major problems revealed by this test included a two- to threefold increase in precipitation (likely caused by a transition in reporting units) and completely dry multiyear periods at locations that normally report abundant precipitation. The affected stations with large jumps were eliminated from the integrated dataset.

A manual examination of station records where maximum and minimum temperatures failed the outlier and/or inconsistency checks at least 300 times revealed two problems. In one case, five stations reported only maximum temperatures before 1981 and these temperatures were around 10°C lower than the maximum temperatures reported during the latter part of the record. Second, a set of about 100 stations was removed from the dataset because their time series exhibited shifts on the order of 5°–10°C during some portion of their records or failed to follow an annual cycle where one would be expected.

In the last record integrity check, U.S. temperature records for which the time of observation has been documented are tested for inconsistencies between the reported observation time and the reported temperatures. Such inconsistencies are known to be present in the data as a result of various observing and digitization practices and errors (e.g., Reek et al. 1992; Kunkel et al. 2005). Such errors are best identified by means of comparison with hourly temperature observations at neighboring synoptic stations (Janis 2002). Whenever the daily maximum temperatures within a month are judged to be inconsistent with corresponding maximum temperatures derived from the hourly data (see the appendix), all temperatures in the month are flagged accordingly.

5. Description of the dataset and processing

Figures 1 and 2 depict the locations of stations that have at least 10 years of records during successive 30-yr intervals starting in 1861. Like its monthly counterpart, the concentration of stations with observations of temperature or precipitation in GHCN-Daily is denser over North America and Eurasia than over Africa, Antarctica, and South America. In the case of GHCN-Daily, however, the densest historical station networks come from the United States, Canada, and Australia, a reflection of the comprehensive contributions from these countries. Nevertheless, Brazil, India, and South Africa have also contributed records from very dense national precipitation networks. The maps for the year 2010 provide an indication of the density of stations that can be updated in GHCN-Daily.

With over 80 000 station records from 180 countries and territories (Table 5), GHCN-Daily is likely the most comprehensive global collection of in situ land surface daily climate summaries available. The total number of values for all elements in the dataset is well over 2 billion, including nearly 300 million maximum and minimum temperatures and more than 800 million daily precipitation totals (as well as 240 million observations of daily snowfall and about 220 million daily snow depths). Additional elements are available at select U.S. stations, most notably temperature at observation time, snow water equivalent, pan evaporation, and the occurrence of various weather phenomena. About 70% of all values come from North American stations.

Figure 3 depicts the temporal evolution of the station network. Daily summaries are available from a relatively small number of stations before 1890 when the number of stations reporting maximum and minimum temperature (precipitation) is about 2.5% (8.9%) of the peak number. The total number, spatial distribution, and temporal completeness generally increase through time for all variables, although both the temperature and precipitation networks attain their maximum density in the 1960s. The interval covered by GHCN-Daily station reports varies from less than 1 year up to 245 years, with the average temperature record spanning 36.7 years and the average precipitation record lasting 33.1 years.

The number of temperature stations as well as the total number of snowfall and snow depth stations remains roughly the same at near-peak levels through the present. The precipitation network, in contrast, declines in size abruptly in the late 1960s, largely because a source for thousands of Indian precipitation records ends around 1970. The decline in the number of available precipitation reports continues until the mid-2000s, when the rapid development of the Community Collaborative Rain, Hail and Snow Network (CoCoRaHS; http://www.cocorahs.org) in the United States contributes to a rebound in

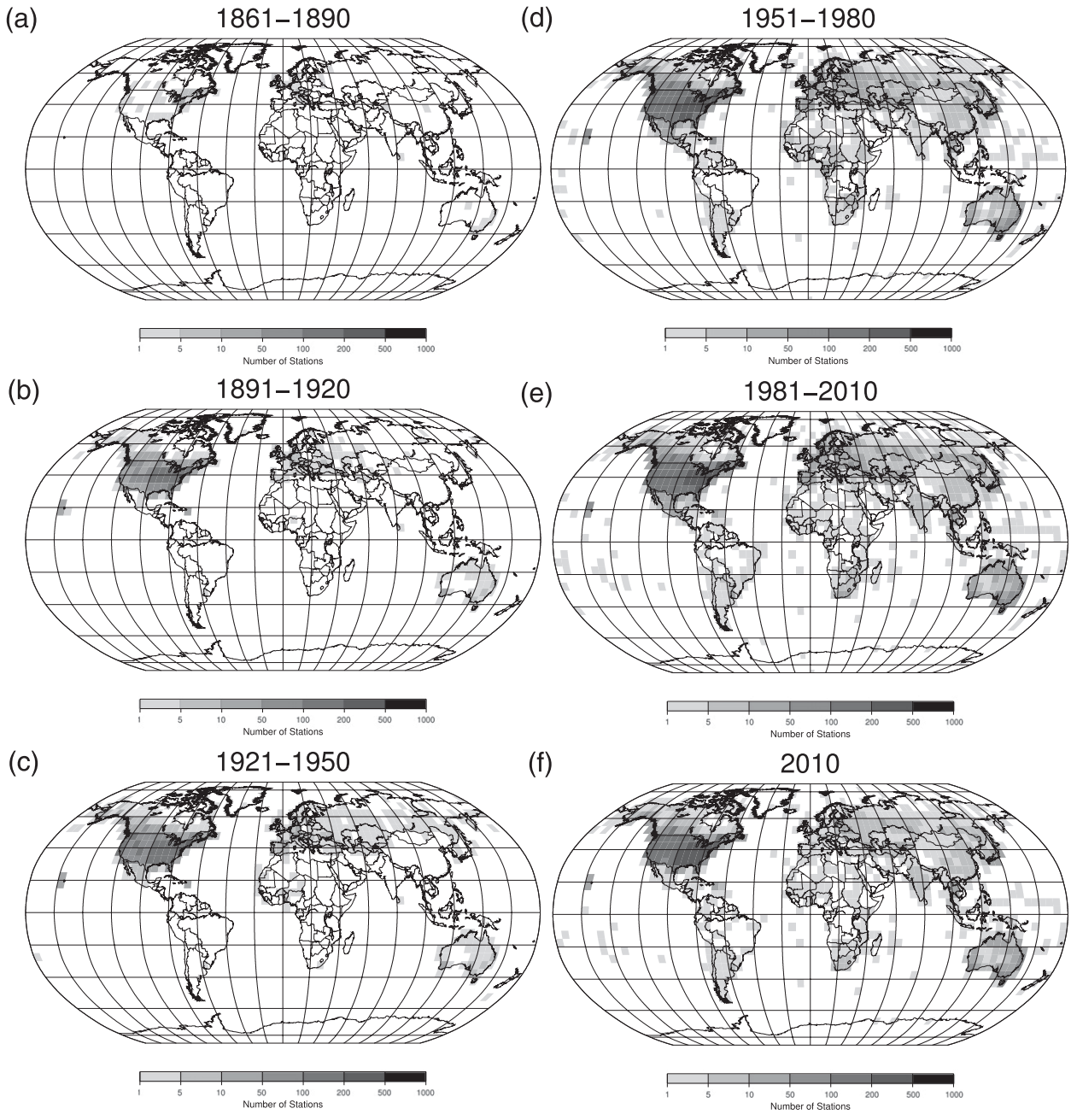

FIG. 1. (a)–(f) Density of GHCN-Daily stations with daily maximum and minimum temperature.

precipitation station numbers. Throughout the record, the vast majority of temperature stations is in North America because of the comprehensiveness of the U.S. and Canadian contributions, whereas the number of precipitation stations is more evenly split between North America and the rest of the world for most of the twentieth century. Nearly all snowfall and snow depth stations are from the Northern Hemisphere, and snowfall is commonly measured only in North America.

GHCN-Daily is updated each day using a number of near-real-time data streams such that recent observations are added within 1 or 2 days of their availability at many thousands of stations. In these cases, the latest daily climate summaries should be only 1 or 2 days behind the calendar date. In the case of time-delayed updates, values are generally delayed by 1 or 2 months. However, it should be noted that, although more than 20 000 stations in GHCN-Daily can be regularly updated

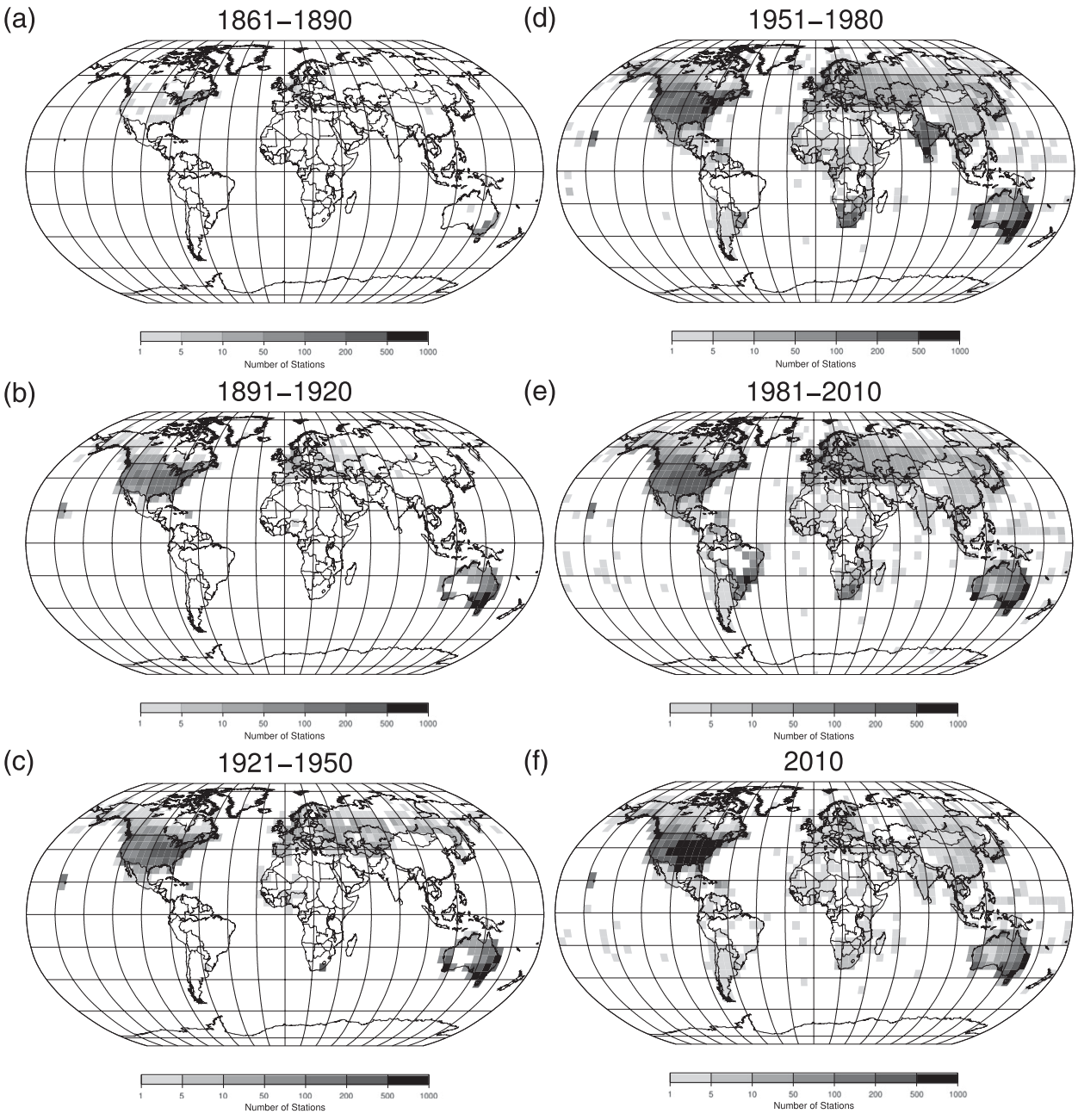

FIG. 2. (a)–(f) Density of GHCN-Daily stations with daily precipitation.

(and more than 30 000 stations contain values within the past year), most participating countries have provided historical daily station records only once. Some of these stations are not necessarily currently active but, in the absence of ongoing formal exchange mechanisms, the sole potential for updates for many GHCN-Daily stations is through the daily synoptic summaries archived in NCDC's global summary of the day. Potential updates from the global summary of the day have yet to be fully exploited in GHCN-Daily; however, values from this source tend to be incomplete and have the timing issues mentioned in section 2. For this reason, other mechanisms for data sharing are encouraged, as outlined in the summary and conclusions. Nevertheless, new and existing bilateral agreements for routine data sharing will continue to enhance the database.

In addition to the near-real-time and time-delayed updates, GHCN-Daily is fully reprocessed on a regular basis (usually once per week), which entails reconstructing the dataset from its component sources from start to finish.

TABLE 5. List of countries and territories with data in GHCN-Daily and their corresponding Federal Information Processing Standard (FIPS) codes. (Source: ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-countries.txt.)

FIPS code	Country	FIPS code	Country
AC	Antigua and Barbuda	KZ	Kazakhstan
AE	United Arab Emirates	LA	Laos
AF	Afghanistan	LG	Latvia
AG	Algeria	LH	Lithuania
AJ	Azerbaijan	LO	Slovakia
AL	Albania	LQ	Palmyra Atoll (United States)
AM	Armenia	LT	Lesotho
AO	Angola	LU	Luxembourg
AQ	American Samoa (United States)	LY	Libya
AR	Argentina	MA	Madagascar
AS	Australia	MD	Moldova
AU	Austria	MG	Mongolia
AY	Antarctica	MI	Malawi
BA	Bahrain	MK	Macedonia
BB	Barbados	ML	Mali
BC	Botswana	MO	Morocco
BD	Bermuda (United Kingdom)	MP	Mauritius
BE	Belgium	MQ	Midway Islands (United States)
BF	Bahamas, The	MR	Mauritania
BK	Bosnia and Herzegovina	MT	Malta
BL	Bolivia	MU	Oman
BN	Benin	MV	Maldives
BO	Belarus	MX	Mexico
BP	Solomon Islands	MY	Malaysia
BR	Brazil	MZ	Mozambique
BY	Burundi	NC	New Caledonia (France)
CA	Canada	NG	Niger
CD	Chad	NH	Vanuatu
CE	Sri Lanka	NL	Netherlands
CF	Congo (Brazzaville)	NO	Norway
CH	China	NP	Nepal
CI	Chile	NU	Nicaragua
CJ	Cayman Islands (United Kingdom)	NZ	New Zealand
CK	Cocos (Keeling) Islands (Australia)	PA	Paraguay
CM	Cameroon	PC	Pitcairn Islands (United Kingdom)
CO	Colombia	PE	Peru
CQ	Northern Mariana Islands (United States)	PK	Pakistan
CS	Costa Rica	PL	Poland
CT	Central African Republic	PM	Panama
CU	Cuba	PO	Portugal
CY	Cyprus	PP	Papua New Guinea
DA	Denmark	PS	Palau
DR	Dominican Republic	RI	Serbia
EC	Ecuador	RM	Marshall Islands
EG	Egypt	RO	Romania
EI	Ireland	RP	Philippines
EN	Estonia	RQ	Puerto Rico (United States)
ER	Eritrea	RS	Russia
ES	El Salvador	SA	Saudi Arabia
ET	Ethiopia	SE	Seychelles
EZ	Czech Republic	SF	South Africa
FG	French Guiana (France)	SG	Senegal
FI	Finland	SH	Saint Helena (United Kingdom)
FJ	Fiji	SI	Slovenia
FM	Federated States of Micronesia	SL	Sierra Leone
FP	French Polynesia	SP	Spain
FR	France	ST	Saint Lucia

TABLE 5. (*Continued*)

FIPS code	Country	FIPS code	Country
FS	French Southern and Antarctic Lands (France)	SU	Sudan
GB	Gabon	SV	Svalbard (Norway)
GG	Georgia	SW	Sweden
GL	Greenland (Denmark)	SY	Syria
GM	Germany	SZ	Switzerland
GP	Guadeloupe (France)	TD	Trinidad and Tobago
GQ	Guam (United States)	TH	Thailand
GR	Greece	TI	Tajikistan
GT	Guatemala	TL	Tokelau (New Zealand)
GV	Guinea	TN	Tonga
GY	Guyana	TO	Togo
HO	Honduras	TS	Tunisia
HR	Croatia	TU	Turkey
HU	Hungary	TV	Tuvalu
IC	Iceland	TX	Turkmenistan
ID	Indonesia	TZ	Tanzania
IN	India	UG	Uganda
IO	British Indian Ocean Territory (United Kingdom)	UK	United Kingdom
IR	Iran	UP	Ukraine
IS	Israel	US	United States
IT	Italy	UV	Burkina Faso
IV	Cote D'Ivoire	UY	Uruguay
IZ	Iraq	UZ	Uzbekistan
JA	Japan	VE	Venezuela
JM	Jamaica	VM	Vietnam
JN	Jan Mayen (Norway)	VQ	Virgin Islands (United States)
JQ	Johnston Atoll (United States)	WA	Namibia
KE	Kenya	WF	Wallis and Futuna (France)
KG	Kyrgyzstan	WQ	Wake Island (United States)
KN	Korea, North	WZ	Swaziland
KR	Kiribati	ZA	Zambia
KS	Korea, South	ZI	Zimbabwe
KT	Christmas Island (Australia)		
KU	Kuwait		

During the reprocess, the most recent version of each source dataset is reintegrated to form the comprehensive (combined) GHCN-Daily station records, and all period of record values are subjected to the latest suite of QA checks. This type of reprocessing helps to ensure that GHCN-Daily is synchronized with its source archives and that all daily climate records are uniformly subjected to the latest set of QA tests. This approach to dataset construction and maintenance honors the intent of a key research need required to ensure the climate record for climate studies, which was highlighted as a "lesson learned" from the Intergovernmental Panel on Climate Change Fourth Assessment Report (Doherty et al. 2009).

Moreover, to ensure version control and traceability each updated and reprocessed version of GHCN-Daily is assigned a unique three-part version code, and every version of GHCN-Daily is archived in its entirety as a separate dataset (along with the latest processing source code). The first component of the version code is incremented only when there are changes to the processing algorithms and/or major additions to the database itself. The second part indicates whether it is an update or a newly reprocessed version. The third part is a timestamp that indicates when the update or reprocessing was done. To illustrate, the descriptive statistics in this section were generated from GHCN-Daily version 2.90-upd-2011112910, an update from 29 November 2011 that initiated at 1000 UTC (and which was produced by appending recently available data updates to the last fully reprocessed version 2.90-por-2011112514). This version can be retrieved from NOAA/NCDC by requesting DSI 9101 version 2.90-upd-2011112910. Authors are requested to cite the relevant version number and timestamp when GHCN-Daily is used for analysis.

6. Summary and conclusions

GHCN-Daily supersedes the Global Daily Climate Network dataset, which was released in 2002. Compared to the GDCN, GHCN-Daily includes a more expansive

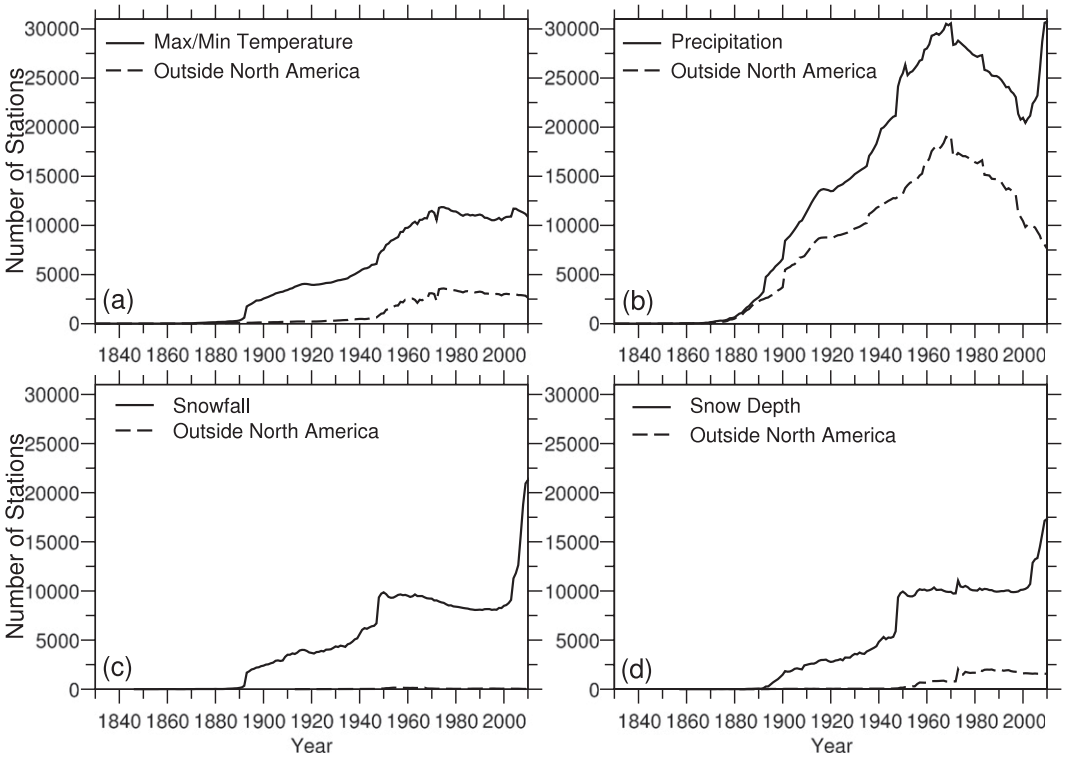

FIG. 3. Time series of the number of stations in GHCN-Daily with (a) maximum and minimum temperature, (b) precipitation, (c) snowfall, and (d) snow depth.

set of historical data sources, numerous data streams that improve the latency of the dataset through frequent updates, and a much more comprehensive set of QA checks. GHCN-Daily also serves as the official archive for daily records from the GSN. The merged GSN station records from all available data sources are provided as a distinct subset of stations for ease of access to the GSN archive. Daily records from stations in the widely used U.S. Historical Climatology Network (USHCN) are also provided as a separate subset of GHCN-Daily and have been used for updating the USHCN version 2 monthly temperatures (Menne et al. 2009) since 2006. In 2011, GHCN-Daily also became the official database for all U.S. daily data.

In spite of the label Global Historical Climatology Network, it is important not to interpret this name to mean that the dataset can be used to quantify all aspects of climate variability and change without any additional processing. Historically (and in general), the stations providing daily data were not managed to meet the desired standards for climate monitoring (e.g., Karl et al. 1995). Rather, the stations were deployed to meet the demands of agriculture, hydrology, weather forecasting, aviation, etc. Notably, GHCN-Daily has not been homogenized to account for artifacts associated with the various eras in reporting practice at any particular station (i.e., for changes in systematic bias). Users, therefore, must consider whether the potential for changes in systematic bias might be important to their application. In addition, GHCN-Daily and GHCN-Monthly are not currently internally consistent (i.e., GHCN-Monthly is not necessarily derived from the data in GHCN-Daily); however, GHCN-Daily is anticipated to be a major source of future updates and enhancements to GHCN-Monthly.

Finally, although GHCN-Daily has already found applications in climate monitoring and assessments (e.g., Alexander et al. 2006; Caesar et al. 2006), its utility could always be enhanced with additional data for regions outside of North America. For this reason, we encourage new data contributions and particularly welcome the addition of complete national daily climate archives. These contributions can be made as part a new initiative to create a more comprehensive global surface temperature databank (Thorne et al. 2011). In cases where routine updates of such national data contributions are not possible via web services or other routine and preferably automated means, the development and exchange of official ''climate quality'' daily messages over the GTS analogous to the monthly CLIMAT messages should be encouraged. In summary, GHCN-Daily

is best viewed as a dynamic, integrated daily dataset to which new data sources and variables will continue to be added. Enhancements to the methods for quality assurance are also likely to be developed over time, with routine homogeneity assessments a likely future addition.

Acknowledgments. The authors wish to thank the many National Meteorological and Hydrological Services and their representatives for contributions to GHCN-Daily. We would also like to acknowledge the many people at NCDC who worked to collect daily data from around the world, especially Tom Peterson and Pasha Groisman. We thank Peter Thorne, Anthony Arguez, Tom Peterson, and two anonymous reviewers for helpful review comments. Partial support for this work was provided by the U.S. Department of Energy Office of Biological and Environmental Research (Grant DE-AI02-96ER62276) and the NOAA/Office of Global Programs, Climate Change Data and Detection Element.

APPENDIX

Testing for Discrepancies in the Timing of Daily Maximum Temperature

Although there are numerous potential causes of discrepancies in the timing of daily maxima and minima, a common discrepancy arises in the United States with observations from Cooperative Observers whose 24-h daily summary period ends in the local morning hours. Because the maximum temperature attained during the 24 h that precede a morning observation time is usually reached sometime during the previous afternoon, a number of volunteer observers who observe in the morning attribute the daily maximum to the previous calendar day when recording the value (Reek et al. 1992). In such cases, the observer usually records the 24-h minimum on the current calendar day (i.e., the day on which the summary period actually ended, which is the desired practice for recording all daily variables, including daily maximum temperature). Moreover, historically, Cooperative Observer paper forms were commonly keyed in a similar way: that is, whereby daily maximum temperatures were systematically assigned to the previous day for morning observers.

Although this practice of ''shifting'' the maximum backward by one day for morning observation times has some logic, it can unfortunately lead to internal inconsistencies within a sequence of daily maxima and minima and often leads to confusion in interpreting daily temperature summaries. For this reason, the purported observation times for U.S. observers are used in conjunction with hourly temperature values from synoptic stations to identify cases in which there appear to be systematic discrepancies between the time of observation at a station and its reported daily maximum temperatures within a particular month. In this check, surrogate daily maximum temperature series are generated from nearby synoptic stations such that the daily summary matches the 24-h period ending at the target station's time of observation. Suitable surrogate ''neighboring'' series are chosen for comparison with the target as a function of the completeness of their hourly data within the month (required to compute a 24-h maximum), distance from the target location, and the index of agreement d between the target and surrogate maximum temperatures within the data month. Specifically, a surrogate series is used in the check if it is from a synoptic station within 75 km of the target, has at least 20 days of generated maximum temperatures in common with the target series, and has an index of agreement [Eq. (A1)] d of at least 0.7 with the target series. If more than three such series are available, they are sorted according to their d value with the target series, and the seven surrogate series (or fewer if seven are not available) with the highest indices of agreement are chosen. Following Legates and McCabe (1999), d is defined as

$$d = 1.0 - \frac{\sum_{i=1}^{m} |y_i - x_i|}{\sum_{i=1}^{m} (|x_i - \overline{y}| + |y_i - \overline{y}|)}, \qquad \text{(A1)}$$

where m is the number of days in the window; x_i and y_i are the observations from the target and surrogate series, respectively, on day i; and $\overline{y}$ denotes an average over all observations in the month for the surrogate series. Thus, high values of d are an indication of both high correlation and small absolute differences between x and y.

A target series is identified as having an apparent systematic issue with the timing of daily temperatures when (i) there is at least one surrogate series available for comparison and (ii) the index of agreement between the target series and all available surrogate series is higher when the surrogate maximum temperature series are systematically shifted forward or backward by one day. More specifically, the d values between the target and all shifted surrogate series must improve by more than 0.2 relative to the value calculated between the target and unlagged surrogate series. The use of a minimum improvement in d as well as the requirement for d to be at least 0.7 for an unlagged comparison comes from a systematic manual evaluation of potential thresholds as described in Durre et al. (2008).

REFERENCES

Alexander, L. V., and Coauthors, 2006: Global observed changes in daily climate extremes of temperature and precipitation. *J. Geophys. Res.,* **111,** D05109, doi:10.1029/2005JD006290.

Alexandersson, H., 1986: A homogeneity test applied to precipitation data. *Int. J. Climatol.,* **6,** 661–675.

Caesar, J., L. Alexander, and R. S. Vose, 2006: Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. *J. Geophys. Res.,* **111,** D05101, doi:10.1029/2005JD006280.

Della-Marta, P. M., M. R. Haylock, J. Luterbacher, and H. Wanner, 2007: Doubled length of western European summer heat waves since 1880. *J. Geophys. Res.,* **112,** D15103, doi:10.1029/2007JD008510.

Doherty, S. J., and Coauthors, 2009: Lessons learned from IPCC AR4: Scientific developments needed to understand, predict, and respond to climate change. *Bull. Amer. Meteor. Soc.,* **90,** 497–513.

Dupigny-Giroux, L.-A., T. F. Ross, J. D. Elms, R. Truesdell, and S. R. Doty, 2007: NOAA's Climate Database Modernization Program: Rescuing, archiving, and digitizing history. *Bull. Amer. Meteor. Soc.,* **88,** 1015–1017.

Durre, I., M. J. Menne, and R. S. Vose, 2008: Strategies for evaluating quality assurance procedures. *J. Appl. Meteor. Climatol.,* **47,** 1785–1791.

——, ——, B. E. Gleason, T. G. Houston, and R. S. Vose, 2010: Robust automated quality control of daily surface observations. *J. Appl. Meteor. Climatol.,* **49,** 1615–1633.

Gleason, B. E., T. C. Peterson, P. Ya. Groisman, D. R. Easterling, R. S. Vose, and D. S. Ezell, 2002: A new global daily temperature and precipitation data set. Preprints, *13th Symp. on Global Change and Climate Variations,* Orlando, FL, Amer. Meteor. Soc., P1.16. [Available online at https://ams.confex.com/ams/annual2002/webprogram/Paper27803.html.]

GLOBE Task Team, and Coauthors, 1999: The Global Land One-Kilometer Base Elevation (GLOBE) Digital Elevation Model, version 1.0. National Oceanic and Atmospheric Administration National Geophysical Data Centerdigital database. [Available online at http://www.ngdc.noaa.gov/mgg/topo/globe.html.]

Guttman, N. B., and R. G. Quayle, 1990: A review of cooperative temperature data validation. *J. Atmos. Oceanic Technol.,* **7,** 334–339.

Hubbard, K. G., S. Goddard, W. D. Sorensen, N. Wells, and T. T. Osugi, 2005: Performance of quality assurance procedures for an applied climate information system. *J. Atmos. Oceanic Technol.,* **22,** 105–112.

Janis, M. J., 2002: Observation-time-dependent biases and departures for daily minimum and maximum air temperatures. *J. Appl. Meteor.,* **41,** 588–603.

Jones, P. D., and Coauthors, 1985: A grid point surface air temperature data set for the Northern Hemisphere. U.S. Department of Energy Carbon Dioxide Research Division Tech. Rep. TRO22, 251 pp.

——, S. C. B. Raper, B. S. G. Cherry, C. M. Goodess, and T. M. L. Wigley, 1986: A grid point surface air temperature data set for the Southern Hemisphere 1851-1984. U.S. Department of Energy Carbon Dioxide Research Division Tech. Rep. TR027, 73 pp.

Karl, T. R., and Coauthors, 1995: Critical issues for long-term climate monitoring. *Climatic Change,* **31,** 185–221, doi:10.1007/BF01095146.

Klein Tank, A. M. G., and Coauthors, 2002: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *Int. J. Climatol.,* **22,** 1441–1453.

Kunkel, K. E., D. R. Easterling, K. Hubbard, and K. Redmond, 2004: Temporal variations in frost-free season in the United States: 1895–2000. *Geophys. Res. Lett.,* **31,** L03201, doi:10.1029/2003GL018624.

——, ——, K. Redmond, K. Hubbard, K. Andsager, M. Kruk, and M. Spinar, 2005: Quality control of pre-1948 cooperative observer network data. *J. Atmos. Oceanic Technol.,* **22,** 1691–1705.

Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie, 2011: An overview of the Global Historical Climatology Network monthly mean temperature dataset, version 3. *J. Geophys. Res.,* **116,** D19121, doi:10.1029/2011JD016187.

Legates, D. R., and C. J. Willmott, 1990a: Mean seasonal and spatial variability in gauge corrected global precipitation. *Int. J. Climatol.,* **10,** 111–127.

——, and ——, 1990b: Mean seasonal and spatial variability in global surface air temperature. *Theor. Appl. Climatol.,* **41,** 11–21.

——, and G. J. McCabe Jr., 1999: Evaluating the use of ''goodness-of-fit'' measures in hydrologic and hydroclimatic model evaluation. *Water Resour. Res.,* **35,** 233–241.

Menne, M. J., C. N. Williams Jr., and R. S. Vose, 2009: The U.S. Historical Climatology Network monthly temperature data, version 2. *Bull. Amer. Meteor. Soc.,* **90,** 993–1007.

Min, S.-K., X. Zhang, F. W. Zwiers, and G. C. Hegerl, 2011: Human contribution to more-intense precipitation extremes. *Nature,* **470,** 378–381, doi:10.1038/nature09763.

Peterson, T. C., and R. S. Vose, 1997: An overview of the Global Historical Climatology Network temperature database. *Bull. Amer. Meteor. Soc.,* **78,** 2837–2849.

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

——, R. S. Vose, V. N. Razuvaev, and R. L. Schmoyer, 1998: Global Historical Climatology Network (GHCN) quality control of monthly temperature data. *Int. J. Climatol.,* **18,** 1169–1179.

Reek, T., S. R. Doty, and T. W. Owen, 1992: A deterministic approach to the validation of historical daily temperature and precipitation data from the Cooperative Network. *Bull. Amer. Meteor. Soc.,* **73,** 753–765.

Schmidlin, T. W., D. S. Wilks, M. McKay, and R. P. Cember, 1995: Automated quality control procedure for the ''water equivalent of snow on the ground'' measurement. *J. Appl. Meteor.,* **34,** 143–151.

Thorne, P. W., and Coauthors, 2011: Guiding the creation of a comprehensive surface temperature resource for twenty-first-century climate science. *Bull. Amer. Meteor. Soc.,* **92,** ES40–ES47.

Trenberth, K. E., and Coauthors, 2007: Observations: Surface and atmospheric climate change. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 235–336.

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. Oak Ridge National Laboratory Environmental Sciences Division Publ. 3912, 324 pp.

WMO, 2003: Manual on the Global Observing System: Volume 1—Global aspects. World Meteorological Organization Document WMO 544, 58 pp.

You, J., and K. G. Hubbard, 2006: Quality control of weather data during extreme events. *J. Atmos. Oceanic Technol.,* **23,** 184–197.