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, NC USA

*Journal of Applied Meteorology and Climatology**
**This paper is currently "in press". An early online release version can be found at*

http://journals.ametsoc.org/doi/pdf/10.1175/JTECH-D-11-00103.1

November 15, 2010

Updated March 4, 2011

Updated March 18, 2011

Corresponding Author Address:

Matthew J. Menne
National Climatic Data Center
151 Patton Avenue
Asheville, NC 28801
Phone: (828) 271-4449
Fax: (828) 271-4022
E-mail: Matthew.Menne@noaa.gov

Abstract

A new integrated land surface daily climate dataset is described. The dataset, called the Global Historical Climatology Network (GHCN) Daily dataset, was developed to meet the needs of 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 75000 stations in 179 countries and territories. Numerous daily variables are provided, including maximum and minimum temperature, total daily precipitation, snowfall, and snow depth; however, about two thirds of the stations report precipitation only. A set of quality assurance checks is routinely applied to the full dataset and, where possible, values are updated each day. The dataset is also regularly reconstructed from its 20-plus data source components, ensuring that GHCN-Daily is generally in sync with its growing list of constituent sources. Each update and reprocessed copy of GHCN-Daily is assigned a unique version number and is archived at NCDC for future retrieval. GHCN-Daily and its processing system also serve as the official archive and processing system for U.S. daily data.

1. Introduction

In situ records of temperature and precipitation are essential to studies of climate variability and change. The analysis of multi-decadal climate trends and variability are commonly based on monthly and annual time series of station-based climate observations, 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). Even so, 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 the reluctance in many countries to release daily climate summaries for widespread public use (Alexander et al. 2006). This relative paucity of daily data is an ongoing impediment to climate change analysis and model comparison studies (Trenberth et al. 2007).

Here we describe a database whose aim is to address the need for historical daily temperature and precipitation records over global land areas. The database, known as the Global Historical Climatology Network (GHCN) – Daily dataset, contains daily summary data from over 75000 stations worldwide, about two thirds of which are for precipitation measurement only. Like its counterpart for monthly climate summaries (Peterson and Vose 1997; Peterson et al. 1998), GHCN-Daily is comprised of climate records from numerous sources that have been merged and subjected to a common suite of quality assurance reviews (Durre et al. 2010). In the following, we briefly describe GHCN-Daily's component data sources, methods for data integration and quality assurance, and the resulting spatial and temporal coverage of the dataset. Our focus is on the core elements of temperature and precipitation. Although the database also contains observations for snowfall, snow depth, as well as 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 of the World Meteorological Organization has allowed for rapid data sharing to meet needs of the meteorological and hydrological communities. Nevertheless, while most National Meteorological and Hydrological Centers (NMHCs) operate at least one network of surface observing stations from which daily summaries for temperature and precipitation are produced, there has been no formal mechanism for sharing daily data worldwide and no central repository for the daily summaries from these global observing systems. In practice, the transmission of daily climate summaries has been treated as optional even for the network of stations that report temperature and precipitation observations at fixed synoptic hours (i.e., every three or six hours for stations in the Regional Basic Synoptic Network or 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 daily dataset was to maximize the spatial coverage of daily climate summaries by acquiring historical records 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. GDCN also contained a large contribution of U.S. data, but since GDCN's release, a number of additional archives at the National Oceanic and Atmospheric Administration's (NOAA) National Climatic Data Center (NCDC) that contain daily data for the United States and its Territories have been integrated comprehensively into GHCN-Daily.

The second data collection tactic entailed leveraging off of 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). Bi-lateral agreements, in particular, have resulted in large contributions of daily data from a number of countries. In the absence of such agreements or personal contacts, the third tactic is to incorporate the optional daily summaries that do get transmitted as part of the GTS synoptic messages.

In GHCN-Daily, we assign the results of these varied attempts to acquire daily data 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 20000 locations outside the United States (from over 100 different countries) and largely reflects the data collection efforts that led to the release of 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 summaries 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 provide some of the earliest observations available for the United States (from the U.S. Forts and Voluntary Observer Program covering much of the 19th century; Dupigny-Giroux 2007) to 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 21st century). GHCN-Daily thus contains the most complete collection of U.S. daily climate summaries available, which are comprehensively updated from a number of real-time data feeds. For this reason, GHCN-Daily serves as the official archive for U.S. daily data.

Government Exchange Data (Table 3) refers to data collected through official GCOS or bilateral agreements. In the best case scenario, an NMHC may offer its complete digital, daily climate database for inclusion in GHCN-Daily, which is the case for Canada (with over 7500 station records provided) and Australia (with over 17000 station records). In other cases, NMHC's have provided daily data only for the GCOS Surface Network stations (Peterson et al. 1997) under their jurisdiction. To date, 76 different countries have officially provided daily data for just over half of the 1025 GCOS Surface Network stations that make up the GSN network implementation for the year 2009 (GCOS, 2009). Apart from Canada and Australia, *Government Exchange Data* are regularly updated for only a handful of stations, but new historical contributions are periodically added to this collection.

The *Global Summary of the Day* contains 24-hour 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-hour summary period purportedly ends at midnight UTC (i.e., 24 *Z*). Daily maximum and minimum temperatures from this source are included in GHCN-Daily only when provided as a nominal 24-hour climatological summary as indicated in the SYNOP messages, whereas daily precipitation totals are also included when they must be summed from two 12-hour or four 6-hour sub-totals (as provided in standard SYNOP code groups). Sub-daily summations are identified by associated “measurement” flag codes in the GHCN-Daily data format. In any case, daily summaries from the *Global Summary of the Day* may differ significantly from climate summaries with 24-hour periods ending at local midnight, 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

The process of integrating data from multiple sources into the GHCN-Daily dataset takes place in three steps: 1) screening the 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) mingling 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 mingling of data is part of an automated process that fully re-creates GHCN-Daily on a regular basis using the latest versions of all sources. We expand upon these three steps 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 "inter-station" duplicate check, which compares records from all stations within a source dataset. When more than 50% of a station's record is identical to the data from another station, the longer of the two records is retained for inclusion in GHCN-Daily provided that the metadata indicates that the two sites are in close proximity (i.e., within 40 km). However, if two stations with matching records are more than 40 km apart, neither record is incorporated into the dataset.

The next 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 are matched on the basis of network affiliation and station identification number (ID). If no such ID match can be made, existing cross-referenced lists that identify the correspondence of station ID numbers from different networks are consulted. For example, data for Alabaster Shelby County Airport, Alabama, United States are indexed by Cooperative Observer ID 010116 in NCDC's 3200 and 3206 datasets (among others). The two souces likely can be combined into one GHCN-Daily record based on that ID. In NCDC DSI-3210 (Table 2) and the various other sources for ASOS stations, however, the data for this location are stored under WBAN ID 53864, which must be matched with the corresponding Cooperative station ID using NCDC's Master Station History Record.

If cross reference lists are not available, we can attempt to match stations 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. Nevertheless, 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*.

Lastly, 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 is added as an additional data source to the relevant GHCN-Daily station record already present in the dataset.

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 integrating, or “mingling”, the data from the various sources to create GHCN-Daily. Mingling 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 are mixed. While precipitation, snowfall, and snow depth are mingled separately, 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 U.S. and the *Global Summary of the Day* data, which tend to have observations that apply to 24-hour 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 before being integrated into GHCN-Daily 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 since they contribute the largest amount of data. For stations outside the U.S., 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 quality assurance (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, as well as 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, we employ a multi-tiered QA approach consisting primarily of routine, fully automated procedures as well as 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). 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 one of 19 quality tests. Described in greater detail in Durre et al. (2010), these tests identify a variety of data problems, including the excessive 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 nearly 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 (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).

In essence, 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. Thanks to 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 (i.e., valid observations erroneously flagged as bad) of automated QA procedures (Schmidlin 1995; Kunkel et al. 2005; You and Hubbard 2006). The checks are therefore applied without manual intervention or verification during the frequent and routine reprocessing of the data while at the same time yielding a consistent, reproducible set of quality decisions.

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;
- for large, systematic jumps in the annual mean of a record (such as might be caused by a shift in units); and
- for concentrations of values that fail automated QA procedures.

In addition, two checks have been 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, 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 data set of monthly values (Legates and Willmott 1990a; Legates and Willmott 1990b). In both cases, large station-to-grid differences (or ratios in the case of precipitation) have been examined manually to determine their validity. To date, this technique has helped to identify erroneous coordinates and/or data resulting from incorrect units or from totals reported as zero rather than missing. Where an obvious manual fix to the coordinates was not apparent, the station records in question were "quarantined" and excluded from GHCN-Daily.

We have also employed a semiautomatic method for identifying large jumps and other erratic behavior in time series of annual totals. 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 three-fold increase in precipitation around 1970 at Indian stations and completely dry multi-year periods at locations that normally report abundant precipitation. The affected post-1970 Indian data and the non-Indian stations with large jumps were eliminated from the integrated dataset.

For daily maximum and minimum temperatures, the 13 stations with at least 300 outlier and inconsistency flags were examined after all automated quality control checks had been applied. This analysis revealed two problems that were addressed specifically. First, the pre-1981 records for five stations in Thailand were removed because they contained only maximum temperatures, and these temperatures were around 10°C lower than the maximum temperatures reported during the latter part of the record. Secondly, noting clearly erroneous sections of data at numerous Mexican stations, time series of daily maximum and minimum temperatures from all stations in Mexico were examined visually. Based on this inspection, approximately one third of the Mexican stations were removed from the dataset because their time series exhibited shifts on the order of five to 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 Appendix 1), all temperatures in the month are flagged accordingly.

5. Description of the dataset and processing

With over 75000 station records, GHCN-Daily is likely the most comprehensive global collection of in situ daily climate summaries available. The total number of values for all elements in the dataset is over 2 billion (2000000000), which includes more than 275 million maximum and minimum temperatures, 778 million daily precipitation totals, 240 million observations of daily snowfall, and about 204 million daily snow depths. The remaining values include measurements of additional elements 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.

The interval covered by GHCN-Daily station records varies from less than one year up to 177 years, with the average temperature record spanning 36.7 years and the average precipitation record lasting 33.1 years. Currently, the earliest observation in the dataset is a daily precipitation total recorded on January 1, 1832 at Parramatta, Australia (GHCN-Daily ID = ASN00066046; total = 0.0 mm). The earliest maximum and minimum temperature measurements are from January 2, 1833 at Uccle, Belgium (GHCN-Daily ID = BE000006447; maximum temperature = -1.4°C and minimum temperature = -4.8°C).

Uccle, a GCOS Surface Network station, is still active and provides the longest daily station record in the dataset.

Figures 1 and 2 depict the locations of stations that have at least 10 years of records during successive 30-year 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 of stations with data in 2010 show the distribution of stations that provide updates to GHCN-Daily.

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. Nevertheless, 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.

Nearly all snowfall and snow depth stations are in the Northern Hemisphere, and snowfall is commonly measured only in North America. The precipitation network, in contrast, declines in size abruptly in the late 1960s and continues its decline in numbers until the mid-2000s when the rapid development of the Community, Cooperative, Rain, Hail and Snow (CoCoRaHS; www.cocorahs.org) Network in the United States contributes to a rebound in precipitation station numbers. Throughout the record, the vast majority of temperature stations are 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 20th century.

GHCN-Daily is updated each day using a number of near real-time data streams such that recent observations are added within one or two days of their availability at many thousands of stations. Thus, for climate monitoring or other near real time uses, the latest daily summaries from data update streams should be only one or two days behind the calendar date. It should be noted, however, that while some 20000 stations in GHCN-Daily can be regularly updated, most participating countries have provided historical daily station records only once, and the sole potential for updates to such records is through the daily synoptic summaries archived in NCDC's *Global Summary of the Day*.

In addition to the near-real time updates, GHCN-Daily is fully reassembled and reprocessed on a regular basis (usually once per week). In this reprocessing, the dataset is literally reconstructed from its component sources from start to finish. In particular, the most recent version of each source dataset is re-integrated to form the comprehensive ("mingled") 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 in sync 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’s 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 version number, and every version of GHCN-Daily is archived in its entirety as a separate dataset (along with the latest processing source code). There are two components to the version number. The first component is incremented only when there are changes to the processing algorithms and/or major additions to the database itself. The second component is a timestamp that indicates the time that the dataset was last updated to incorporate newly available data as well as the date of the period of record reprocess to which the updates have been added. To illustrate, the descriptive statistics in this section were generated from GHCN-Daily version 2.50 “por2010100813” (i.e., the period of record reprocessed version of October 08, 2010 at 13 UTC). This version can be retrieved from NOAA/NCDC by requesting Data Set Index (DSI) 9101 version “por.2010100813. Authors are requested to cite the relevant version number and time stamp when GHCN-Daily is used for analysis.

6. Summary and Conclusions

The origins of GHCN-Daily can be traced back a number of years, its predecessor--the GDCN--having been released in 2002. Unlike GDCN, however, GHCN-Daily contains numerous data streams for updated data that enhance the latency of the dataset through rapid and frequent updates. Relative to GDCN, GHCN-Daily also contains a much more comprehensive set of QA checks as well as a more expansive set of historical data sources. GHCN-Daily also now serves as the official archive for daily records from the GCOS

Surface Network (GSN), and merged GSN station records from all possible data sources are provided as a distinct subset of stations for ease of access to the GSN archive. In addition, daily records from stations in the widely used U.S. Historical Climatology Network (USHCN) are 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 summaries for the dataset 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 to GHCN-Monthly.

Finally, while GHCN-Daily has already found applications in climate monitoring and assessments (e.g., Alexander et al. 2006; Caeser 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 for inclusion in GHCN-Daily. These contributions can be made as part a new initiative to create a more comprehensive global surface temperature databank (Thorne et al. submitted). In cases where routine updates of such national data contributions are not possible via web services or other 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.

Acknowledgment

The authors wish to thank all of the many National Meteorological and Hydrological Services and their representatives for their 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 also thank Peter Thorne, Anthony Arguez and Tom Peterson for helpful comments on this manuscript. Partial support for this work was provided by the Office of Biological and Environmental Research, U.S. Department of Energy (Grant number DE-AI02-96ER62276); and the NOAA Office of Global Programs, Climate Change Data and Detection Element.

References

Alexander, L.V., X. Zhang, T.C. Peterson, J. Caesar, B. Gleason, A.M.G. Klein Tank, M. Haylock, D. Collins, B. Trewin, F. Rahimzdeh, A. Tagipour, R. Kumar Kolli, J.V. Revadekar, G. Griffiths, L. Vincent, D.B. Stephenson, J. Burn, E. Aguilar, M. Brunet, M. Taylor, M. New, P. Zhai, M. Rusticucci, J.L. Vazquez Aguirre, 2006: Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research – Atmospheres,* **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 co-authors, 2009: Lessons learned from IPCC AR4. *Bull. Amer. Meteor. Soc.*, **90**, 497-513.

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

Durre, I., M.J. Menne, B.E. Gleason, T. G. Houston, and R.S. Vose, 2010: Robust automated quality control of daily surface observations. *J. Appl. Meteorol. Climatol.*, **49**, 1615-1633.

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

Gleason, B.E., T.C. Peterson, P.Y. Groisman, D.R. Easterling, R.S. Vose, and D.S. Ezell, 2002: A new global daily temperature and precipitation dataset. *AMS 13th Symposium on Global Change and Climate Variations*, 4-17 Jan. 2002, Orlando, FL.

GLOBE Task Team and others (Hastings, David A., Paula K. Dunbar, Gerald M. Elphingstone, Mark Bootz, Hiroshi Murakami, Hiroshi Maruyama, Hiroshi Masaharu, Peter Holland, John Payne, Nevin A. Bryant, Thomas L. Logan, J.-P. Muller, Gunter Schreier, and John S. MacDonald), eds., 1999. The Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Model, Version 1.0. National Oceanic and Atmospheric Administration, National Geophysical Data

Center, 325 Broadway, Boulder, Colorado 80305-3328, U.S.A. Digital data base on the World Wide Web (URL: http://www.ngdc.noaa.gov/mgg/topo/globe.html) and CD-ROMs.

Guttman, N.B., and R.G. Quayle, 1990: A review of cooperative temperature data validation. *J. Atmos. Oceanic Tech.*, **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., S.C.B. Raper, B.D. Santer, B.S.G. Cherry, C.M. Goodess, P.M. Kelly, T.M.L. Wigley, R.S. Bradley, and H.F. Diaz, 1985: A Grid Point Surface Air Temperature Data Set for the Northern Hemisphere, U.S. Dept. of Energy, Carbon Dioxide Research Division, Technical Report TRO22, 251 pp.

Jones, P.D., 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. Dept. of Energy, Carbon Dioxide Research Division, Technical Report TR027, 73 pp.

Karl, T. R., V. E. Derr, D. R. Easterling, C. K. Folland, D. J. Hoffman, S. Levitus, N. Nicholls, D. E. Parker, and G. W. Withee, 1995: Critical issues for long-term climate monitoring. *Clim. Change*, **31**, 185–221, doi:10.1007/BF01095146.

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.

Kunkel, K.E., D.R. Easterling, 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.

Legates, D. R., 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.

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

Legates, D.R., and C.J. Willmott. 1990: Mean seasonal and spatial variability in global surface air temperature. *Theoretical and Applied Climatology*, **41**, 11-21.

Menne, M.J., C.N. Williams Jr., and R.S. Vose, 2009: The United States 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.

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

Peterson, T.C., 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., K.M. Willett, R.J. Allan, S. Bojinski, J.R. Christy, N. Fox, S. Gilbert, I. Jolliffe, J.J. Kennedy, E. Kent, A.Klein-Tank, J. Lawrimore, D.E. Parker, N. Rayner, A. Simmons, L. Song, P.A. Stott, B. Trewin, 2011: Guiding the creation of a comprehensive surface temperature resource for 21st century climate science. *Bull. Amer. Met. Soc.*, submitted.

Trenberth, K.E., P.D. Jones, P. Ambenje, R. Bojariu, D. Easterling, A. Klein Tank, D. Parker, F. Rahimzadeh, J.A. Renwick, M. Rusticucci, B. Soden, and P. Zhai, 2007: Observations: Surface and Atmospheric Climate Change. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

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. ORNL/CDIAC-53, NDP-041, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 315 pp.

WMO, 2003: Manual on the Global Observing System. WMO-No. 544. World Meteorological Organization, Geneva.

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

Appendix 1 – Testing for discrepancies in the timing of daily maximum temperature

While 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-hour daily summary period ends in the local morning hours. Because the maximum temperature attained during the 24 hours that preceed 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-hour 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.

While this practice of "shifting" the maximum backwards 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-hour 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-hour 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

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

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 $\bar{y}$ 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 a) there is at least one surrogate series available for comparison, and b) 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 backwards 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 evaluation of potential thresholds as described in Durre et al. (2008).

List of Figures

List of Tables

Table 1. Sources and contacts for the International Collection

Region/Country	Source/Contact
Countries in West Africa	MeteoFrance
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, Thailand	National Center for Atmospheric Research
Brazil	ANEEL (Agencia Nacional De Energia Electrica)/P.Ya. Groisman
Paraguay, Uruguay, Venezuela	NOAA's 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 (http://eca.knmi.nl/)

Table 2. Data Sources comprising the U.S. Collection

U.S. Cooperative Summary of the Day (NCDC DSI-3200).	Dataset includes daily observations from over 20000 stations in the United States and its territories. While most measurements are taken once per day by volunteer observers as part of the NOAA National Weather Service's 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 in data from previous months are made frequently and the dataset is current. Updates are made 4 to 5 months after the close of the observational month. The time of observation varies by station.
U.S. Cooperative Summary of the Day--CDMP (NCDC DSI-3206).	Dataset includes daily summaries primarily for the years before 1948 from more than 11000 COOP stations that were keyed as part of NCDC's Climate Data Modernization Program (CDMP). Corrections are made occasionally, but new observations are not added. The time of observation varies by station.
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 two to three months after they were taken, and corrections to historical data may occasionally be applied. These observations are generally with respect to the 24-hour period ending at local midnight.
U.S. ASOS Summary of the Day, 2000-2005 (NCDC DSI-3211).	Dataset contains observations for nearly 900 U.S. ASOS stations between October 2000 and December 2005. The data set is no longer updated, but corrections are made occasionally. The observations are with respect to the 24-hour period ending at local midnight.
ASOS Summary of the Day, 2006-present (from NCDC DSI-3505).	Dataset contains reports from U.S.-operated Automated Surface Observing System (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-hour 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-hour period ending at local midnight.

U.S. Forts and Voluntary Observers	Dataset contains observations from the Climate Database Modernization Program's 19th 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 Cooperative Observer Program (COOP). Newly keyed Forts and volunteer data are added periodically. Observation times vary.
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-hour period ending at local midnight.
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. Data fed into GHCN-Daily extends back to 2004. 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 hours. Observation times vary, but summaries are generally for 24-hour periods ending in the morning local time.
Real time Cooperative Summary of the Day updates from NOAAPort (DSI 3201)	Real time updates for the U.S. Cooperative Summary of the Day data. Updates are provided on a daily basis and represent summaries for the previous 24 hours. Observation times vary, but summaries are generally for 24-hour periods ending in the morning local time.
Latest U.S. Cooperative Summary of the Day updates digitized from paper forms (DSI 3202)	Newly keyed data from Cooperative Observer forms. Observation times vary as in DSI-3200.
Community Collaborative Rain, Hail and Snow (CoCoRaHS) Network	Provides daily rain and snow measurements from CoCoRaHS volunteers. Data begin as early as 1998 and are updated daily.

Table 3. Sources for Government Exchange Data

Region/Country	Source/Contact
Canada	Environment Canada/Robert Morris
Australia	Bureau of Meteorogology/Cathy Toby
Belarus	Bilateral Exchange/P.Ya Groisman
Ukraine	Bilateral Exchange/P.Ya Groisman
556 GCOS Surface Stations	Various contacts

Table 4. List of countries and territories with data in GHCN-Daily and their corresponding FIPS codes.

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
AM	Armenia	LQ	Palmyra Atoll [United States]
AO	Angola	LT	Lesotho
AQ	American Samoa [United States]	LU	Luxembourg
AR	Argentina	LY	Libya
AS	Australia	MA	Madagascar
AU	Austria	MD	Moldova
AY	Antarctica	MG	Mongolia
BA	Bahrain	MH	Montserrat [United Kingdom]
BB	Barbados	MI	Malawi
BC	Botswana	MK	Macedonia
BD	Bermuda [United Kingdom]	ML	Mali
BE	Belgium	MO	Morocco
BF	Bahamas, The	MP	Mauritius
BK	Bosnia and Herzegovina	MQ	Midway Islands [United States}
BL	Bolivia	MR	Mauritania
BN	Benin	MT	Malta
BO	Belarus	MU	Oman
BP	Solomon Islands	MV	Maldives
BR	Brazil	MX	Mexico
BY	Burundi	MY	Malaysia
CA	Canada	MZ	Mozambique
CD	Chad	NC	New Caledonia [France]
CE	Sri Lanka	NG	Niger
CF	Congo (Brazzaville)	NH	Vanuatu
CH	China	NL	Netherlands
CI	Chile	NO	Norway
CJ	Cayman Islands [United Kingdom]	NP	Nepal
CK	Cocos (Keeling) Islands [Australia]	NU	Nicaragua
CM	Cameroon	NZ	New Zealand
CO	Colombia	PA	Paraguay
CQ	Northern Mariana Islands [United States]	PC	Pitcairn Islands [United Kingdom]
CS	Costa Rica	PE	Peru
CT	Central African Republic	PK	Pakistan
CU	Cuba	PL	Poland
CY	Cyprus	PM	Panama
DA	Denmark	PO	Portugal
DR	Dominican Republic	PP	Papua New Guinea
EC	Ecuador	PS	Palau
EG	Egypt	RM	Marshall Islands
EI	Ireland	RO	Romania
EN	Estonia	RP	Philippines
ER	Eritrea	RQ	Puerto Rico [United States]
ES	El Salvador	RS	Russia
ET	Ethiopia	SA	Saudi Arabia
EZ	Czech Republic	SE	Seychelles
FG	French Guiana [France]	SF	South Africa

FI	Finland	SG	Senegal
FJ	Fiji	SH	Saint Helena [United Kingdom]
FM	Federated States of Micronesia	SI	Slovenia
FP	French Polynesia	SL	Sierra Leone
FR	France	SP	Spain
FS	French Southern and Antarctic Lands [France]	ST	Saint Lucia
GB	Gabon	SU	Sudan
GG	Georgia	SV	Svalbard [Norway]
GL	Greenland [Denmark]	SW	Sweden
GM	Germany	SY	Syria
GP	Guadeloupe [France]	SZ	Switzerland
GQ	Guam [United States]	TD	Trinidad and Tobago
GR	Greece	TH	Thailand
GT	Guatemala	TI	Tajikistan
GV	Guinea	TK	Turks and Caicos Islands [United Kingdom]
GY	Guyana	TL	Tokelau [New Zealand]
HO	Honduras	TN	Tonga
HR	Croatia	TO	Togo
HU	Hungary	TS	Tunisia
IC	Iceland	TU	Turkey
ID	Indonesia	TV	Tuvalu
IN	India	TX	Turkmenistan
IO	British Indian Ocean Territory [United Kingdom]	TZ	Tanzania
IR	Iran	UG	Uganda
IS	Israel	UK	United Kingdom
IT	Italy	UP	Ukraine
IV	Cote D'Ivoire	US	United States
IZ	Iraq	UV	Burkina Faso
JA	Japan	UY	Uruguay
JM	Jamaica	UZ	Uzbekistan
JN	Jan Mayen [Norway]	VE	Venezuela
JQ	Johnston Atoll [United States]	VM	Vietnam
KE	Kenya	VQ	Virgin Islands [United States]
KG	Kyrgyzstan	WA	Namibia
KN	Korea, North	WF	Wallis and Futuna [France]
KR	Kiribati	WZ	Swaziland
KS	Korea, South	ZA	Zambia
KT	Christmas Island [Australia]	ZI	Zimbabwe
KU	Kuwait		
Total	179		

Source: ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-countries.txt

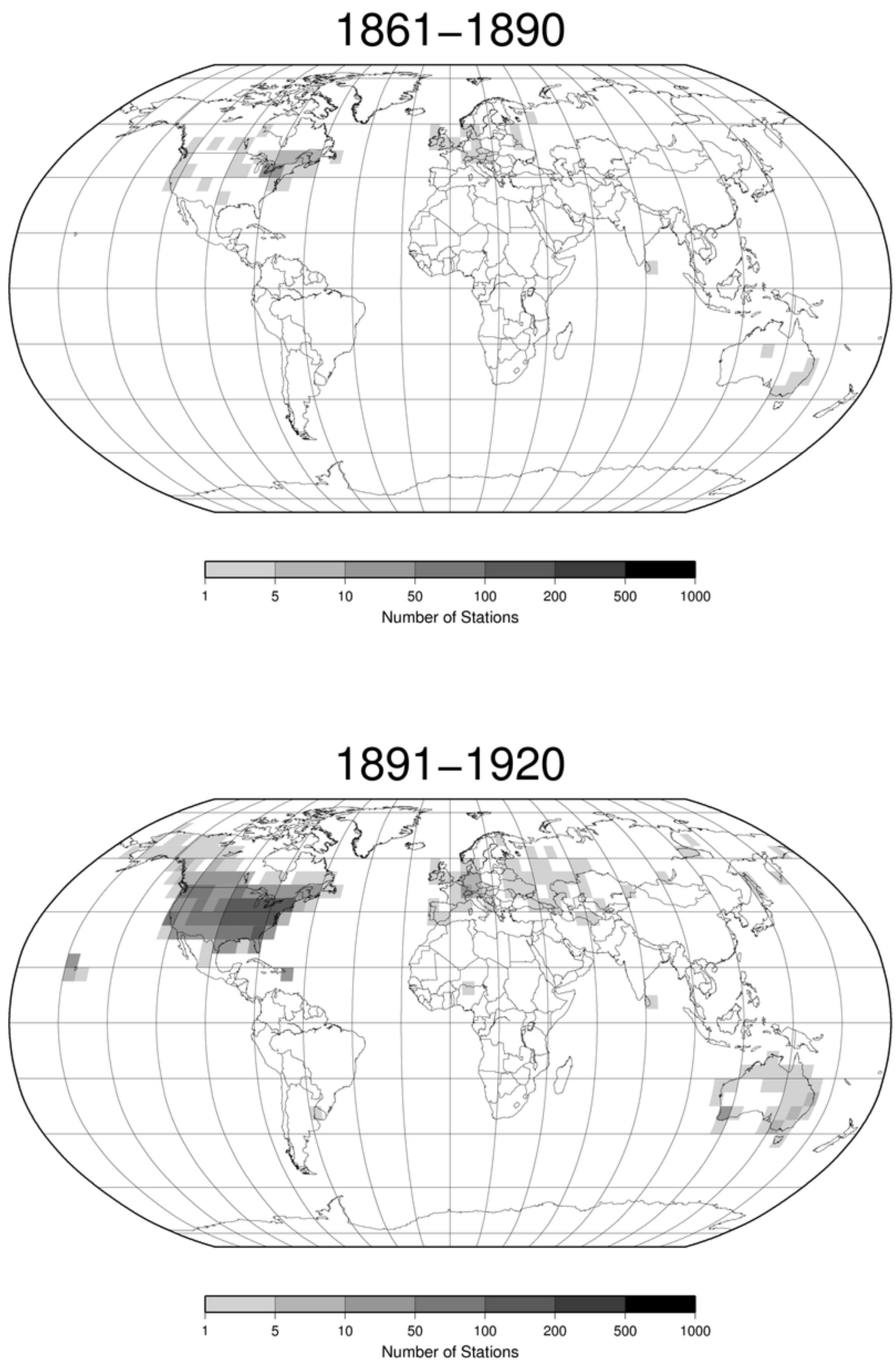
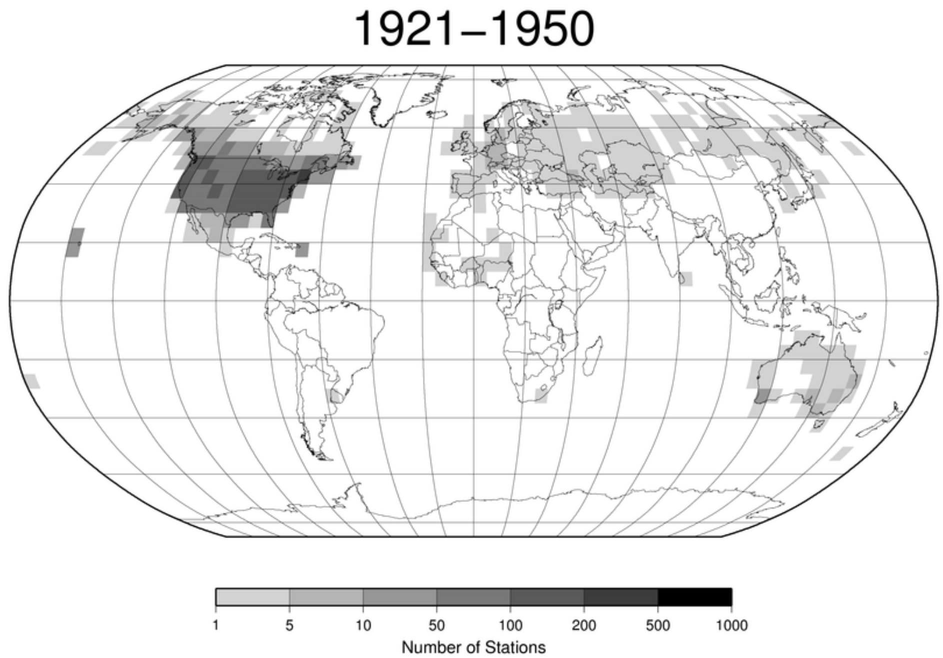


Figure 1. Locations of GHCN-Daily stations with daily maximum and minimum temperature.



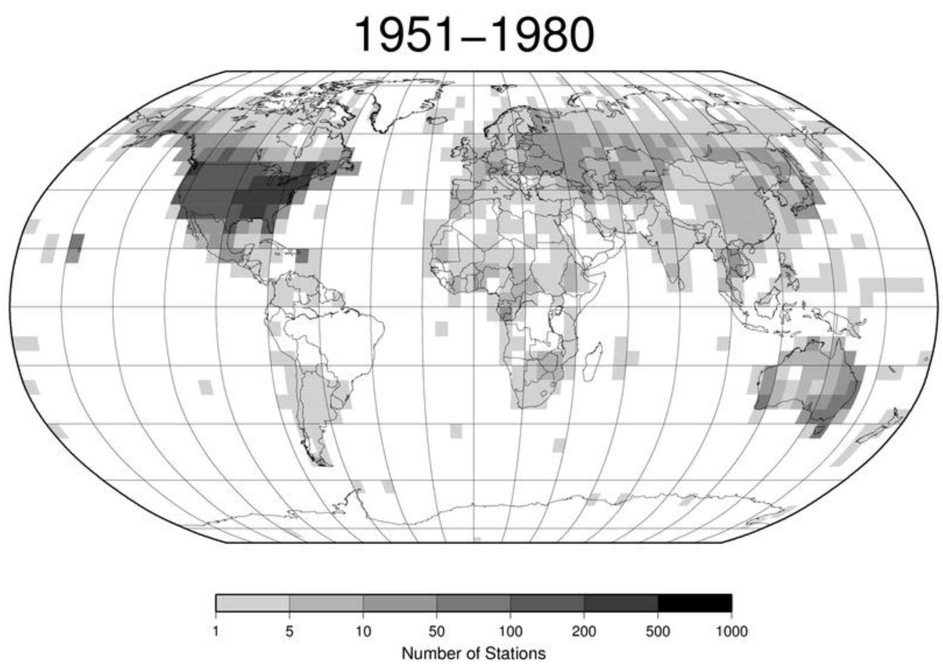
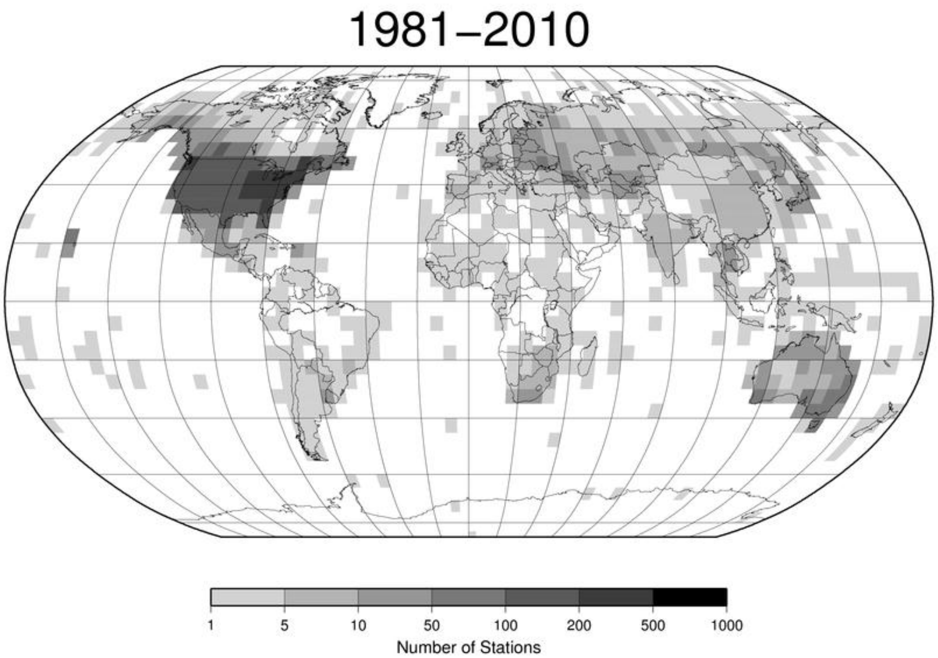


Figure 1. (Cont).



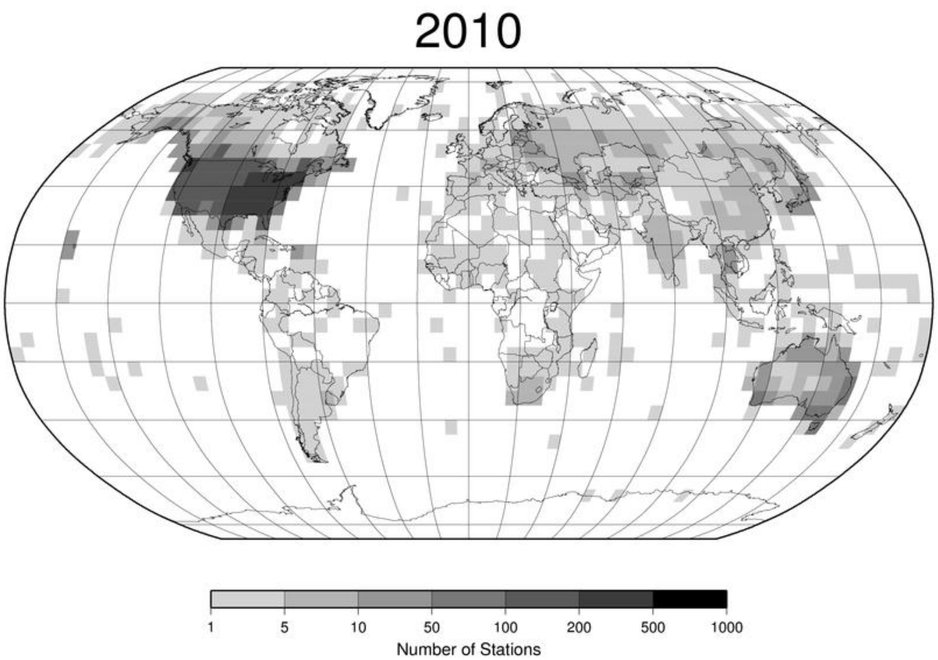


Figure 1. (Cont).

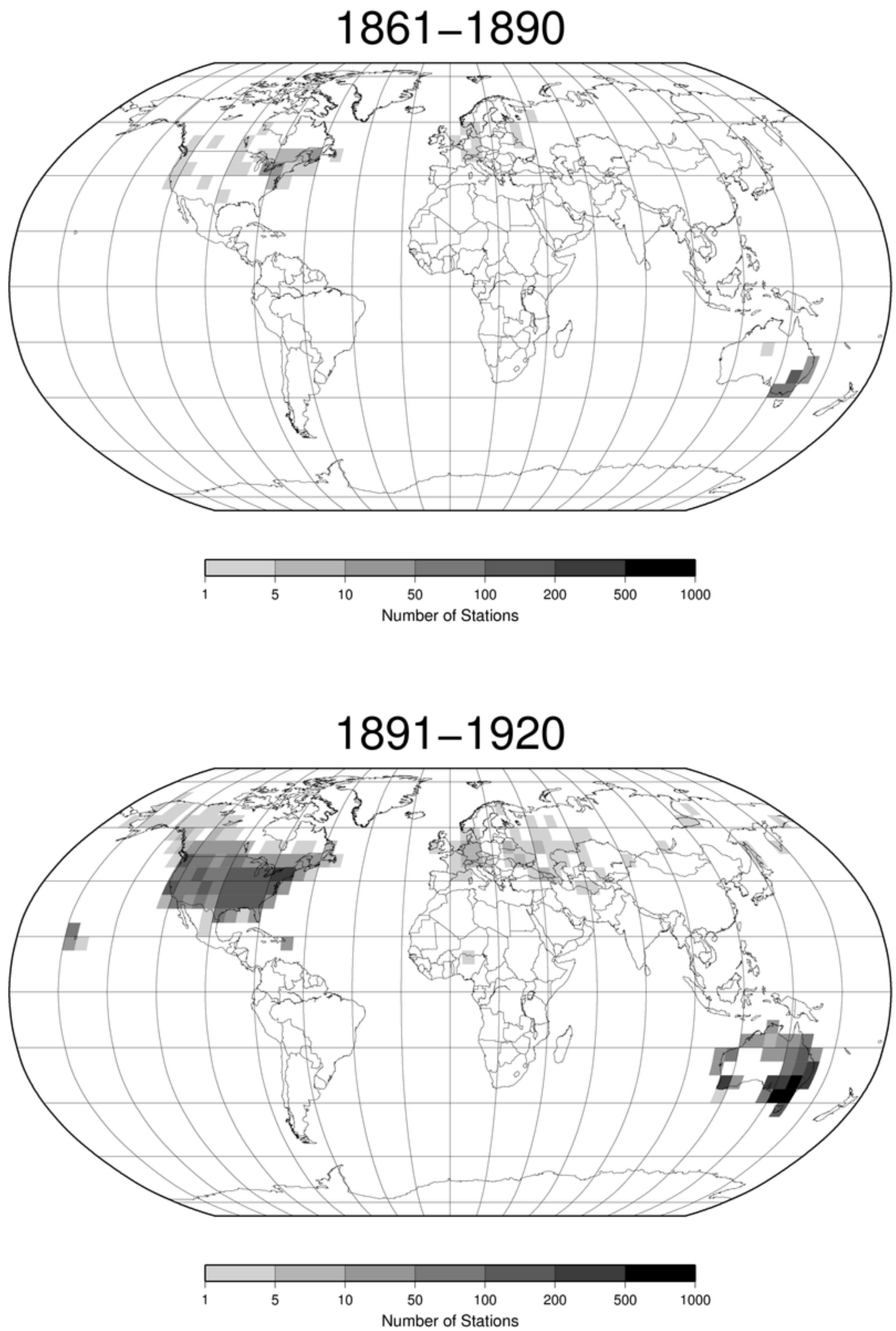
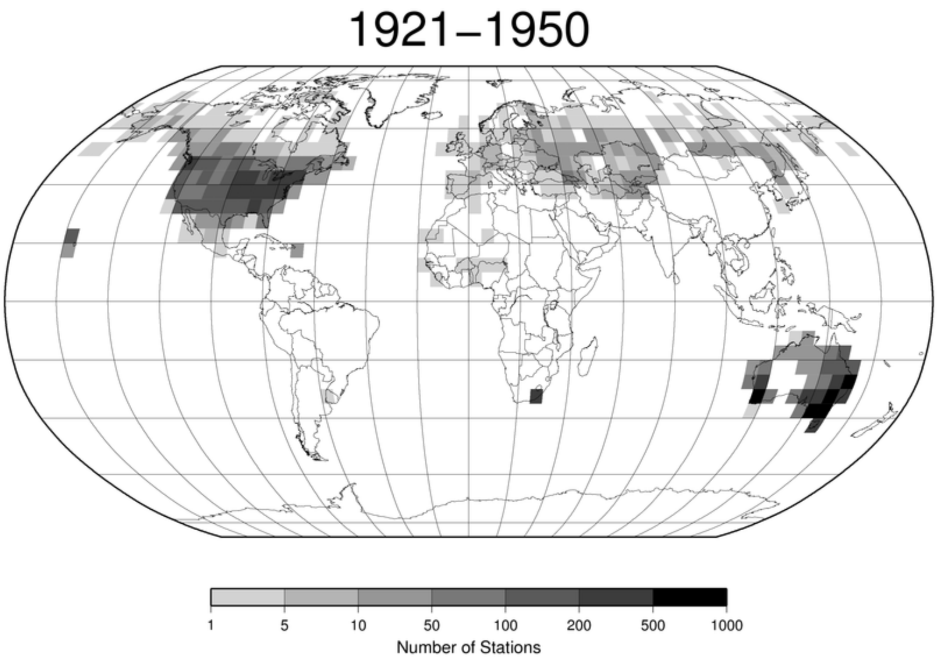


Figure 2. Locations of GHCN-Daily stations with daily precipitation.



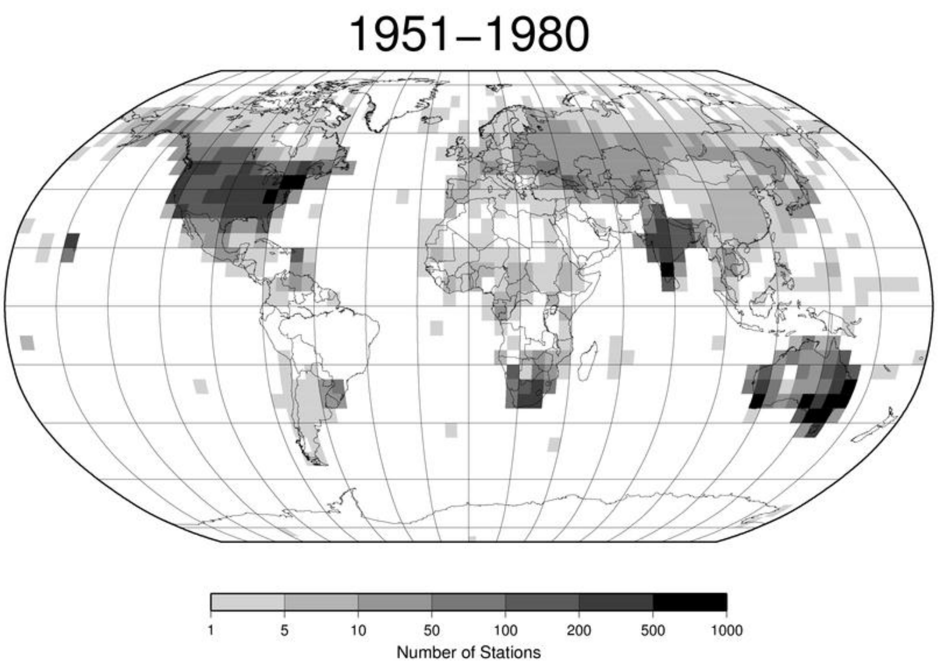
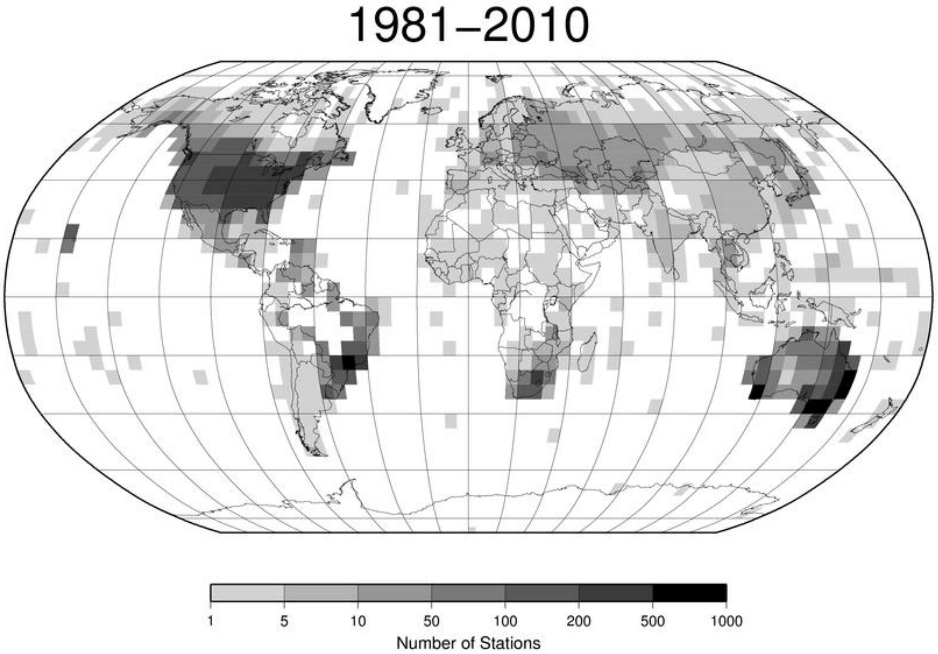


Figure 2. (Cont.).



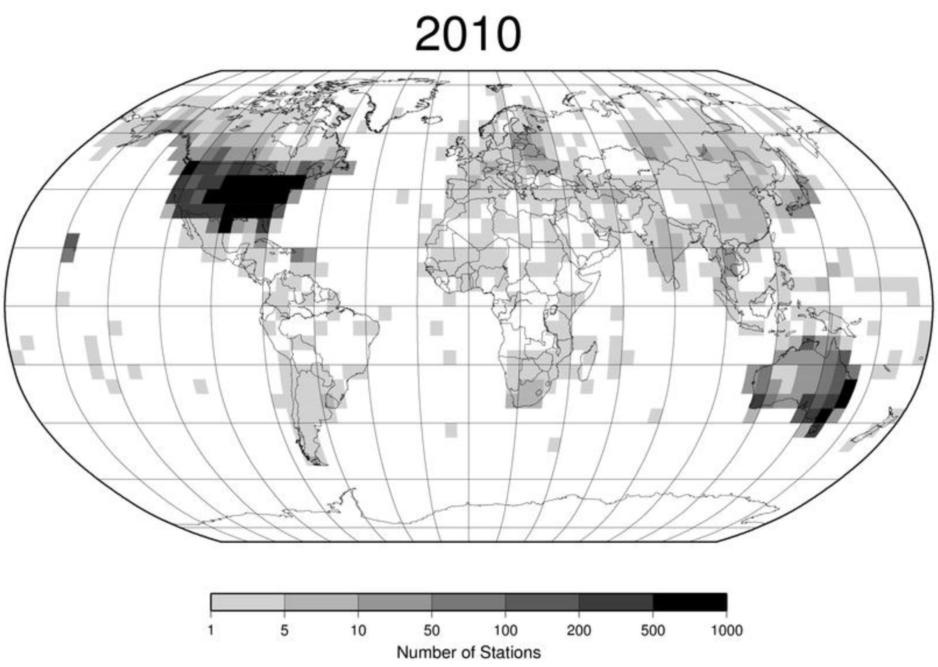


Figure 2. (Cont.).

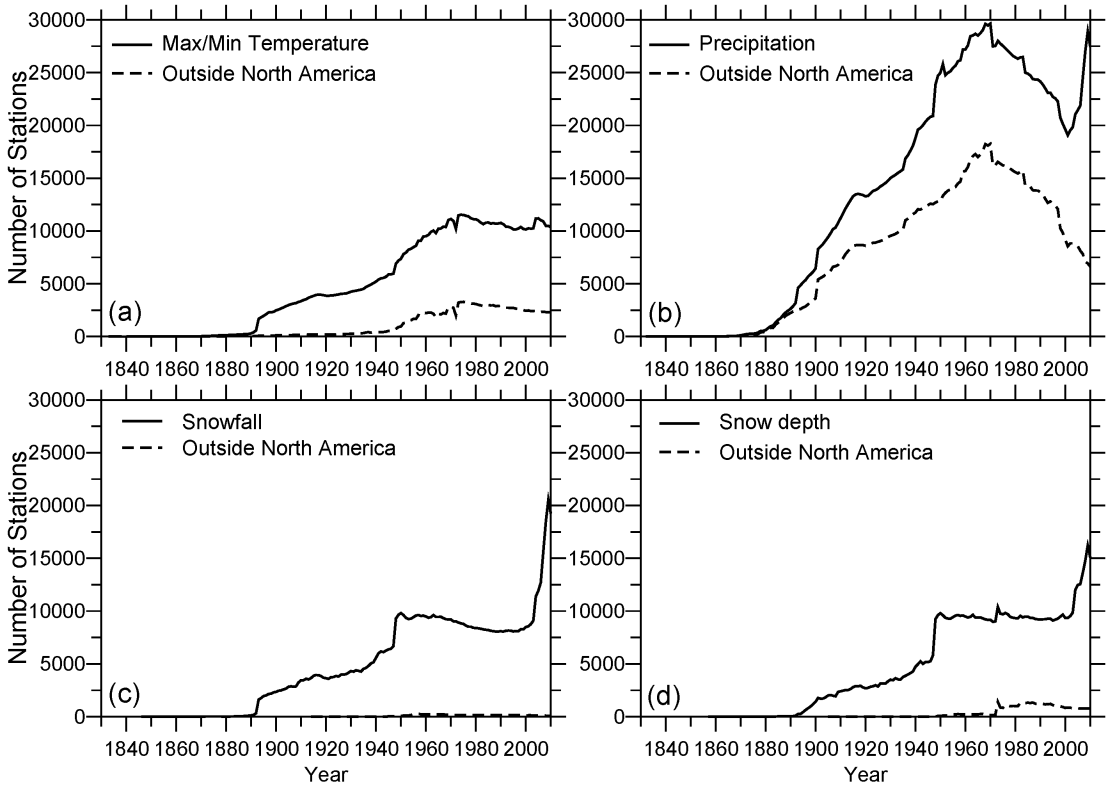


Fig. 3. Time series of the number of stations in GHCN-Daily with (a) maximum and minimum temperature; (b) total precipitation; (c) snowfall; and, (d) snow depth.