Comprehensive Automated Quality Assurance of Daily Surface Observations

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

National Climatic Data Center, Asheville, North Carolina

(Manuscript received 20 August 2009, in final form 25 January 2010)

ABSTRACT

This paper describes a comprehensive set of fully automated quality assurance (QA) procedures for observations of daily surface temperature, precipitation, snowfall, and snow depth. The QA procedures are being applied operationally to the Global Historical Climatology Network (GHCN)-Daily dataset. Since these data are used for analyzing and monitoring variations in extremes, the QA system is designed to detect as many errors as possible while maintaining a low probability of falsely identifying true meteorological events as erroneous. The system consists of 19 carefully evaluated tests that detect duplicate data, climatological outliers, and various inconsistencies (internal, temporal, and spatial). Manual review of random samples of the values flagged as errors is used to set the threshold for each procedure such that its false-positive rate, or fraction of valid values identified as errors, is minimized. In addition, the tests are arranged in a deliberate sequence in which the performance of the later checks is enhanced by the error detection capabilities of the earlier tests. Based on an assessment of each individual check and a final evaluation for each element, the system identifies 3.6 million (0.24%) of the more than 1.5 billion maximum/minimum temperature, precipitation, snowfall, and snow depth values in GHCN-Daily as errors, has a false-positive rate of 1%−2%, and is effective at detecting both the grossest errors as well as more subtle inconsistencies among elements.

1. Introduction

One of the most active areas of climatological research is the study of changes in the frequency and intensity of extreme events (e.g., Nicholls 1995; Cerveny et al. 2007; Trenberth et al. 2007). In the past, limitations in data availability and computational resources have frequently restricted such studies to particular countries, regions, or individual locations. Thanks to technological advances as well as a greater willingness of countries to share their daily observations, it has become possible to assemble and regularly update a global dataset of historical daily meteorological observations. Such a dataset, the Global Historical Climatology Network (GHCN)-Daily, is being maintained at the National Oceanic and Atmospheric Administration's National Climatic Data Center (NCDC) and has proven useful in many applications requiring daily data (e.g., Alexander et al. 2006; Caesar et al. 2006).

GHCN-Daily consists of more than 1 500 000 000 observations at over 40 000 land-based stations, some of which date back to the mid-1800s. The primary meteorological elements represented include daily maximum and minimum temperature (TMAX and TMIN), 24-h precipitation (PRCP) and snowfall (SNOW) totals, and the snow depth at a certain time of day (SNWD). The data originate from a variety of sources ranging from paper forms completed by volunteer observers to synoptic reports from automated weather stations.

With this diversity of data comes a large variety of measurement, recording, digitization, transmission, and processing problems (Goodison 1978; Robinson 1989, 1990; Wallis et al. 1991; Reek et al. 1992; Nicholls 1995; Kunkel et al. 1998; Brasnett 1999; Kunkel et al. 2005; Daly et al. 2007; Kunkel et al. 2007; Green et al. 2008). Consequently, procedures that can ensure high-quality historical and real-time daily data are critical. Such procedures have been implemented in the GHCN-Daily quality assurance (QA) system.

The core of this system, consisting of 19 outlier, consistency, and other checks on the five primary meteorological elements, is described in this paper. Unlike

Corresponding author address: Imke Durre, National Climatic Data Center, 151 Patton Avenue, Asheville, NC 28801.
E-mail: imke.durre@noaa.gov

DOI: 10.1175/2010JAMC2375.1

many other QA checks, the tests described here are fully automated but have been manually validated using the strategies of Durre et al. (2008a, hereafter DMV08) to ensure satisfactory performance. The paper is therefore intended not only for those interested in documentation of the GHCN-Daily system, but also for readers in search of a QA approach for other sets of meteorological observations.

In section 2, the philosophy used in designing the system is reviewed, and the overall structure of the system is described. Sections 3–7 contain specifics about each group of checks. The performance of the system is discussed in section 8, and concluding remarks are offered in section 9.

2. System design

The QA of daily surface observations is frequently performed in a semiautomatic fashion in which trained validators examine a subset of values flagged by the automatic procedures and override the system's decision whenever they deem a flagged value to be valid (Guttman and Quayle 1990; Schmidlin et al. 1995; Kunkel et al. 1998; Hubbard et al. 2005; Kunkel et al. 2005; Brunet et al. 2006). The primary reason for this approach is that automated procedures tend to yield a significant number of ''false positives,'' that is, valid observations erroneously identified as invalid (Guttman and Quayle 1990; Robinson 1990; Schmidlin et al. 1995; Kunkel et al. 2005). Although effective, the semiautomatic approach is impractical for datasets as large as GHCN-Daily, which must be reprocessed regularly to incorporate additional historical and real-time data. Therefore, the challenge is to design a QA system that, without manual intervention, is effective at detecting a variety of data errors and that has a low false-positive rate, or fraction of flagged values that are false positives. The strategies for developing such a QA system are presented in generic form by DMV08. The following subsections describe how the strategies are specifically applied here and provide an overview of the resulting system.

a. Design considerations

From a user perspective, data values can be erroneous for a variety of reasons. Perhaps the most common attribution is that a value is physically impossible or climatologically implausible for the location and time of year. Other observations might be inconsistent with those on adjacent days or at neighboring stations. In still other cases, values may be repeated for a period of days, or duplicated for months or years, or simply inconsistent with other elements. Consequently, an effective QA system should include not only tests for outliers and spatial inconsistencies, but a comprehensive set of procedures, each of which is designed to detect one of these types of data errors (Reek et al. 1992; Peterson et al. 1998; Brunet et al. 2006; Durre et al. 2008b).

Accordingly, the GHCN-Daily QA system consists of procedures that test for all of the above-mentioned types of data errors. Many of these are replicates or extensions of checks employed by others (Reek et al. 1992; Kunkel et al. 1998; Serreze et al. 1998; Hubbard et al. 2005; Kunkel et al. 2005), while others, such as the frequent-value and megaconsistency checks (see below), utilize concepts that have received little or no attention in the peer-reviewed literature. Since the checks each have specific data requirements that may not be fulfilled at all locations and times, they are designed to operate independently of one another, are arranged in a deliberate sequence, and ignore values flagged by preceding checks in the sequence. Given the large variability in record length and station density in GHCN-Daily, this arrangement maximizes the number of observations that can be checked with at least one QA procedure.

A variety of observing practices also are considered when designing the QA procedures. One of these considerations pertains to the various conventions used for ascribing daily observations to a particular calendar date. Observations for one ''day'' typically summarize conditions during a 24-h period that does not necessarily correspond to a calendar day (i.e., from local midnight to midnight). Depending on convention, they are assigned to the date of either the beginning or end of this period (Janis 2002; Kunkel et al. 2005; Green et al. 2008). The time of day at which the 24-h period begins and ends (i.e., the time of observation) also varies among stations and, in some cases, even among elements at the same station (Schmidlin et al. 1995; Wu et al. 2005). In addition, some data are affected by ''shifting,'' or the attribution of a value (by either the observer or subsequent processing) to a presumed calendar date of occurrence (Reek et al. 1992; Kunkel et al. 2005). These variations in reporting practices must be taken into account by QA procedures testing for inconsistencies among elements or stations (Schmidlin et al. 1995; Wu et al. 2005; You and Hubbard 2006). As described in more detail in subsequent sections, this is accomplished in the GHCN-Daily system by considering observations within a 3-day window centered on the day being tested.

In addition, two noteworthy observing practices affect the design of QA procedures for precipitation-related variables. The first is the reporting of multiday precipitation and snowfall totals by some observers who were unable to make daily measurements on one or more preceding days. When such totals are identified as ''accumulated'' in the raw data, they are excluded from most of the GHCN-Daily QA procedures because they do not

conform to the same limits and rules as the standard 24-h totals. The second precipitation-related practice is the convention followed in some countries during certain periods of time of reporting not only wintertime frozen precipitation, but also summertime hail in the SNOW and SNWD fields. Unfortunately, the electronic data files do not include an indication of whether the reported value is due to snow, hail, or another form of frozen precipitation, often making it virtually impossible for either data users or QA procedures to distinguish reports of hail from the at least equally frequent, erroneous, warm-season nonzero values in the SNOW and SNWD fields. To obtain the highest-quality record of true snowfall observations possible, the QA procedures presented here are designed to identify as errors any nonzero SNOW and SNWD values that are reported at temperatures at which snow is implausible.

b. False-positive rates

Each individual check in the GHCN-Daily QA system is required to have a false-positive rate of no more than 20%; that is, no more than one in five values flagged is allowed to be a false positive. To ensure adherence to this requirement, each test is applied in a preliminary fashion to the entire GHCN-Daily dataset. Random samples of flagged values are then manually inspected for a range of plausible test thresholds, and the fraction of false positives in each sample is determined as in DMV08. The resulting sample false-positive rate, sometimes also referred to as a false-positive ratio, is considered to be an estimate of the QA check's false-positive rate for a given threshold, recognizing that the accuracy of this estimate is a function of a variety of factors, including sampling variability and the subjectivity inherent in manual assessments of any kind (Kunkel et al. 2005; DMV08). In the end, the threshold yielding the highest error detection rate without exceeding the 20% false-positive rate limit is chosen for that QA check. A detailed illustration of how this technique is applied to a specific procedure can be found in DMV08.

In manually assessing the validity of a randomly selected value, we employ a strategy similar to that outlined by Kunkel et al. (2005). Specifically, a flagged value that is judged to be truly invalid is counted as a data error whereas a flagged value deemed to be reasonably plausible is counted as a false positive. A value thought to be questionable counts as half a false positive. The false-positive rate for a QA check is then calculated by dividing the total number of false positives (and half false positives) by the number of values examined.

The manual assessments are made using techniques that proved beneficial in semiautomatic QA (e.g., Guttman and Quayle 1990; Kunkel et al. 2005). For values from the U.S. Cooperative Observer Network, scanned images of the original observer forms are consulted whenever they are available. The value in the data file is considered to be invalid if it differs from that reported on the corresponding observer form or if other characteristics of the form (e.g., a note from the observer) suggest instrument or observer error. Other tools are used when no observer form is available or when an examination of the form does not point to a data problem. For example, when evaluating a value flagged by a spatial consistency check, mapping utilities that display the location of the station with respect to mountains and coastlines can be helpful in assessing the representativeness of the neighboring stations used in the check. In other cases, examination of the data file alone can be sufficient, such as when a temperature value of zero flagged by an outlier check turns out to be repeated on several days in a row.

Using a sample size of 10 values per threshold category, 30–100 values are typically examined before choosing the threshold for any check with a single test threshold. Multiparameter checks often necessitate the inspection of even more values. For logical checks without a test threshold (e.g., a check for nonzero precipitation amounts accompanied by a trace flag), a simple random 10-value sample of all values flagged by the check is used to estimate whether the false-positive rate is below the desired level. In the case of checks for inconsistencies between two or more observations, the evaluation process is also used to determine whether one or all of the values causing the inconsistency should be flagged. In addition, when a procedure could be prone to overflagging in certain regions, as might be the case for a spatial consistency check in complex terrain, for example, the spatial pattern of all flags set by the procedure is examined as illustrated in DMV08.

c. Structure of the QA system

The tests in the GHCN-Daily system can be grouped into five general categories that are executed in the following order: basic integrity checks, outlier tests, internal and temporal consistency checks, spatial consistency checks, and ''megaconsistency'' checks. The basic integrity checks identify cases of data duplication as well as physically implausible values. The outlier checks identify excessive gaps in the distributions of data values as well as observations that deviate excessively from station-specific climatological parameters. The internal, temporal, and spatial consistency checks identify values that deviate significantly from ''adjacent'' observations in time and space. Finally, the megaconsistency checks verify the integrity of all remaining unflagged observations.

Because each procedure ignores previously flagged values, this overall sequence limits ''collateral damage.'' For example, an internal inconsistency check (e.g., for TMAX < TMIN) flags all values composing the inconsistency; consequently, less collateral damage is done if one value can first be flagged by another test (e.g., an outlier test, which might determine that TMIN was erroneous). A second factor determining the order of procedures is the impact of data errors on statistics calculated by the procedures. For example, the spatial consistency checks are preceded by all but the megaconsistency checks in an effort to reduce as much as possible the potential influence of erroneous values at a neighboring station on the spatial comparisons.

The following five sections describe the QA procedures in detail. Several comprehensive tables are included, each listing the tests in the order in which they are applied. While specifics on most tests are provided in the main text, the reader is also referred to the tables for actual test thresholds (and to the appendixes for details on the more complex algorithms).

3. Basic integrity checks

The basic integrity checks are listed in Table 1. Most of these checks [the ''naught'' (zero), duplicate, streak, and frequent-value checks] address various forms of erroneous repetition or duplication of the values. The world record exceedance checks, in contrast, identify values that are larger or smaller than have been recorded at any surface observing station in any nation.

a. Checks for repetition and duplication

The first type of repetition and duplication check addresses the relatively straightforward detection of erroneous zeros, hence the name naught checks. For example, zero is sometimes used incorrectly as a missing value code; consequently, both TMAX and TMIN are flagged if both are 0°C or if both are −17.8°C (i.e., 0°F). Likewise, a precipitation total (PRCP, SNOW, or SNWD) is flagged if the value is greater than zero and the data measurement flag indicates that only a trace of precipitation was recorded.

The next type of check looks for the duplication of sequences of data in different time periods, such as two different years having exactly the same data, or two different months in the same year having exactly the same data (Fig. 1). Such problems typically occur because of keying, transmission, or processing errors (e.g., Kunkel et al. 1998). Note that SNWD is not tested for duplication because there are regions in high latitudes where the same depth can persist for more than one month at a time.

The last type of integrity check focuses on consecutive runs of the same value (i.e., streaks) or frequent occurrences of the same value (i.e., identical values that are closely spaced in time but are not necessarily consecutive). For the streak tests, values are flagged not only if they occur on consecutive days, but also when they continue across days for which no data are available. For the frequent-value checks, observations are flagged if the group overall consists of more than a specified minimum number of observations and if those values exceed a specified climatological percentile (see Table 1 for specifics). In both tests, the minimum number of values that constitute an error and the method for handling values of zero vary by element as illustrated in Table 1. The precipitation time series shown in Fig. 2 is an example of a record containing both streaks and clusters of frequent identical values. In general, the kinds of errors most commonly detected by these checks include cases in which observer, data entry, or data processing errors result in the replication of a particular observation on numerous subsequent days; streaks of temperatures equal to 0° or −17.8°C not already detected by preceding checks; and streaks or clusters of incorrect missing value codes.

b. World record exceedance checks

The world record exceedance checks identify values that cannot be valid under any circumstance, either because they are physically impossible or because they fall outside the range of what has been observed anywhere on the earth. The limits used, shown in Table 1, are those defined and updated by the World Meteorological Organization World Weather/Climate Extremes Archive (Cerveny et al. 2007).

As in Reek et al. (1992) and Feng et al. (2004), the values flagged by these tests typically far exceed the relevant world records and are the result of incorrect units, undocumented or incorrect missing value codes, and other digitization or data coding errors. Although many of these values would also be detected by the climatological outlier checks, their identification at this stage serves two purposes. First, particularly when they appear in groups, the exclusion of such gross errors from the computation of climatological statistics results in a more effective outlier check. Second, at stations with records that are too short for calculating climatological statistics, the world record exceedance checks allow for the detection of at least these outliers.

4. Outlier checks

The outlier checks are listed in Table 2. Generally speaking, an outlier is a value that does not fit the

TABLE 1. Basic integrity checks in the core GHCN-Daily system. Procedures are listed in the order in which they are applied. In addition to the element abbreviations defined in section 1 of the text, (0) and (−1) refer to the current and previous days, respectively. Unless otherwise noted in column 4, only the values that fail to meet the given condition are flagged.

Naught check

Variant	Condition for flagging	Values flagged	Comment
Temperature zeros	TMAX = −17.8°C and TMIN = −17.8°C at U.S.-operated stations; TMAX = 0°C and TMIN = 0°C elsewhere	TMAX and TMIN	Zero is sometimes incorrectly used as a missing value code; U.S. stations report temperatures in °F, which are converted to °C for GHCN-Daily
Trace value	value >0 and data measurement flag = T (trace)	PRCP, SNOW, or SNWD	Trace indicates an amount smaller than the smallest measurable amount, which varies by element and measurement unit

Duplicate check

Variant	Condition for flagging	Values flagged	Comment
Between entire years	All values in one year = all corresponding values in another year	All PRCP or SNOW values in both years	For years with at least three nonzero values
Between different months within the same year	All values in one month = all values in another month	Not SNWD	Compares all days up to the last day of the shorter month; minimum of three nonzero values in the month required for PRCP and SNOW
For the same calendar month in different years	All values in one month = all values in another month	Not SNWD	Compares all days up to the last day of the shorter month; minimum of three nonzero values in the month required for PRCP and SNOW
Between TMAX and TMIN	TMAX = TMIN on 10 or more days within a month	All TMAXs and TMINs in a month	

World record exceedance check

Variant	Condition for flagging	Values flagged	Comment
Temperature	Temperature < −89.4° or >57.7°C	TMAX or TMIN	
Precipitation	PRCP < 0 or >1828.8 mm	PRCP	
Snowfall	SNOW < 0 or >1925 mm	SNOW	
Snow depth	SNWD < 0 or >11460 mm	SNWD	
Snow depth increase	SNWD(0) − SNWD(−1) > 1925 mm	SNWD(0) and SNWD(−1)	

Identical value–streak check

Variant	Condition for flagging	Values flagged	Comment
Temperature	20 or more consecutive identical TMAX or TMIN values; 10 or more consecutive identical nonzero SNOW values; 90 or more consecutive identical nonzero SNWD values	TMAX, TMIN, SNOW, or SNWD	Missing values are skipped, e.g., (32.8, −9999, 32.8) is ''packed'' into (32.8, 32.8)
Precipitation	20 or more consecutive	PRCP	Missing values and zeros are skipped

Identical value–frequent-value check

Variant	Condition for flagging	Values flagged	Comment
	9 or more out of 10 consecutive values are identical and ≥ their respective 30th percentiles; 8 or more out of 10 are identical and ≥50th percentile; 7 or more out of 10 are identical and ≥70th percentile; or 5 or more out of 10 are identical and ≥90th percentile	PRCP	Missing values and zeros are skipped; percentiles computed as for the percentile-based outlier check

presumed frequency distribution at a specific location and time of year. The most common technique for identifying outliers in meteorological data involves normalizing the data values using their mean and standard deviation (STD) over a specified time interval and then flagging those values whose z score exceeds a specified threshold (e.g., Hubbard et al. 2005; Kunkel et al. 2005). This approach, however, has several limitations,

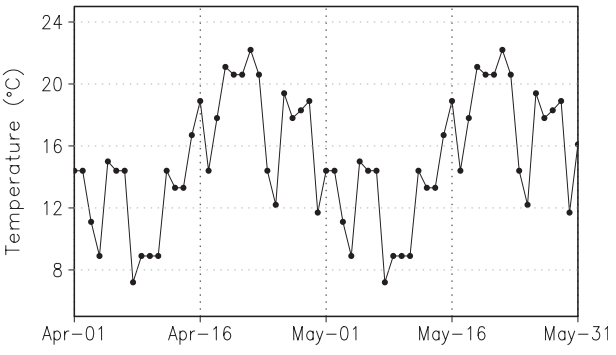

FIG. 1. Daily maximum temperatures during April and May 1967 at Lardeau, Canada (GHCN-Daily station identifier = CA001144580), showing an example of data duplication identified by the duplicate check comparing data from different months within a year (Table 1).

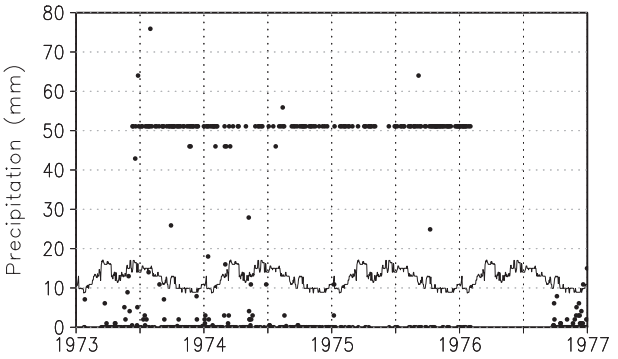

FIG. 2. Time series of daily precipitation totals (dots) during 1973–76 at Balmaceda, Chile (GHCN-Daily station CI000085874), containing 162 values of 51.1 mm that are flagged by the frequent-value check (Table 1). As the check proceeds through the time series (section 3a), the values are first flagged when 5 of them exceeding the 90th percentile (solid line) appear in the sequence of 10 nonzero totals stretching from 10 Jun through 9 Jul 1973.

particularly if used as the only outlier check. First, it can only be applied to time series that are sufficiently long for computing the requisite climatological statistics. Second, a z score based outlier check is prone to overflagging since even a slight skewness in the distribution can result in a larger fraction of true data values exceeding a certain z score than would be expected from the assumption of normality (Wolter 1997; Harmel et al. 2002). Furthermore, the approach is not at all suitable for variables such as PRCP, whose distributions are zero-bounded and highly skewed. Therefore, the GHCN-Daily QA system uses three types of outlier checks: ''gap'' checks (for all elements except SNOW), the traditional z-score check (for TMAX and TMIN), and a percentile-based outlier check (for PRCP). (For SNOW, none of these checks was implemented because no threshold yielded a satisfactory false-positive rate.)

a. Gap checks

The gap checks examine the frequency distributions of observations for individual elements and calendar months. They flag values that compose the distribution's tail when the tail is unrealistically separated from the remaining values. The gap threshold, or maximum allowable separation of the tail from the remainder of the distribution, differs among elements, but is independent of location and time.

The algorithm first sorts all of an element's values observed in a particular calendar month throughout a station's period of record from smallest to largest. Differences between consecutive sorted values are then calculated. If a value is separated by more than the gap threshold from the next largest value, all of the element's values on the far side of the gap (i.e., in the tail of the distribution) are flagged. The definition of the tail depends on the element being analyzed. For the zero-bounded, positively skewed distribution of PRCP, the search for a gap begins at the smallest nonzero value and moves upward (Fig. 3a). When applied to TMAX, TMIN, and SNWD, on the other hand, the procedure analyzes the top and bottom halves of the respective distributions separately, proceeding upward and downward from the median. Although it is zero-bounded like PRCP, SNWD is subjected to the two-tailed rather than the one-tailed test because its median can exceed zero by an amount that is greater than the gap threshold, thus allowing for the occurrence of excessively large gaps at the low end of the distribution (Fig. 3b) as well as the high end. Some typical data problems detected by the gap checks include incorrect missing value codes (Fig. 3a), incorrect units, and SNWD values of zero when SNWD should be missing (Fig. 3b).

b. Climatological outlier checks

The climatological outlier checks compare each observation to parameters computed from all observations at the given location and time of year. For TMAX and TMIN, which generally follow the normal distribution, the z score–based approach described above is used, and the requisite biweight means and STDs (Lanzante 1996) are calculated from data within a 15-day window centered on each day of the year. A temperature is then flagged if it is more than six STDs from the respective mean. For example, the mean and STD of TMIN for 12 June at Thule, Greenland, are −1.2° and 2.43°C, respectively, based on all TMIN observations between 5 and 19 June during the station's 1951–75 record. The temperature of −19.4°C found on that day in 1975 fails this outlier check because it has a z score of −7.5.

TABLE 2. As in Table 1, but for outlier checks. The following abbreviation is used in addition to those defined in the text: z = z score.

Gap check			
Variant	Condition for flagging	Values flagged	Comment
Temperature	Gap in distribution of TMAX or TMIN for the station/calendar month ≥10°C	TMAX or TMIN values on tail side of gap	Two-tailed check (see text)
Precipitation	Gap in nonzero PRCP distribution for station/calendar month ≥300 mm	PRCP values above gap	One-tailed check
Snow depth	Gap in SNWD distribution for station/calendar month ≥350 mm	SNWD values on tail side of gap	Two-tailed check

Climatological outlier check			
Variant	Condition for flagging	Values flagged	Comment
Conventional	$\|z\| \geq 6.0$	TMAX or TMIN	Requires a minimum of 100 values for the period of record in the 15-day window
Percentile based, generic	PRCP ≥ 9 × 95th percentile	PRCP*	Requires a minimum of 20 nonzero values for the period of record in the 29-day window
Percentile based, below freezing	PRCP ≥ 5 × 95th percentile and 0.5(TMAX + TMIN) < 0°C	PRCP	

* An attempt was made to also apply the percentile-based outlier check to SNOW and SNWD. However, particularly in locations where snowfall is rare but can be heavy, this check is associated with an unacceptably high false-positive rate even for the largest ratio thresholds. The same was true when the reference value was changed from the 95th to the 90th or 99th percentile.

The percentile-based outlier check flags precipitation totals that exceed a specified multiple of the corresponding climatological 95th percentile for the calendar day on which the total was observed (Table 2). The percentile is computed from nonzero daily values observed during all available years and within a 29-day window centered on the calendar day of the observation. (A larger window is required than for temperature because only nonzero values are used.) When the day's mean temperature is above freezing or temperature is not available, a multiple of nine is used. For example, at Gold Hill, Utah, the 20 August 1982 total of 238.8 mm shown in Fig. 4 is flagged because it is more than 9 times larger than the corresponding 95th percentile of 19.9 mm. A smaller multiple of five is employed when the day's mean temperature is less than or equal to 0°C. This temperature dependence is implemented because precipitation totals tend to be less extreme at below-freezing temperatures than under warmer conditions.

5. Internal and temporal consistency checks

The internal and temporal consistency checks are listed in Table 3. Internal consistency checks test for violations of logical or physical relationships between two or more elements (e.g., TMAX < TMIN). Temporal consistency checks, on the other hand, compare one element's values on consecutive days, usually to identify unrealistic spikes or dips (Reek et al. 1992; Kunkel et al. 1998). The GHCN-Daily QA system contains six procedures testing for a variety of internal and temporal inconsistencies, three affecting temperature and three affecting the precipitation variables.

a. Temperature

The inherent relationships among daily temperature observations are particularly suitable for internal consistency checks. For example, TMAX and TMIN must be internally consistent not only on the same observational day, but also on adjacent days (Reek et al. 1992; Kunkel et al. 1998; Easterling et al. 1999; Janis 2002). For example, TMAX for a given observational day (i.e., 24-h period ending at the time of observation) cannot be below TMIN reported for the 24-h periods immediately preceding and following that observational day. The power of checks based on these relationships is further enhanced if the temperature at the observation time (TOBS) is available because this temperature, by definition, should lie between TMAX and TMIN for both the previous and subsequent 24 h (Guttman and Quayle 1990).

In the GHCN-Daily system, all possible relationships among TMAX, TMIN, and TOBS are aggregated into one iterative procedure that takes a holistic view of the entire temperature time series rather than considering individual 2- or 3-day periods at a time. This check is described in more detail in appendix A.

Two other checks are included to test for unrealistically large swings in temperature, to the extent that this is possible without excessively flagging actual rapid changes of this kind. The first check, termed the spike/dip test, identifies temperatures that are at least 25°C warmer or colder than the previous and following days.

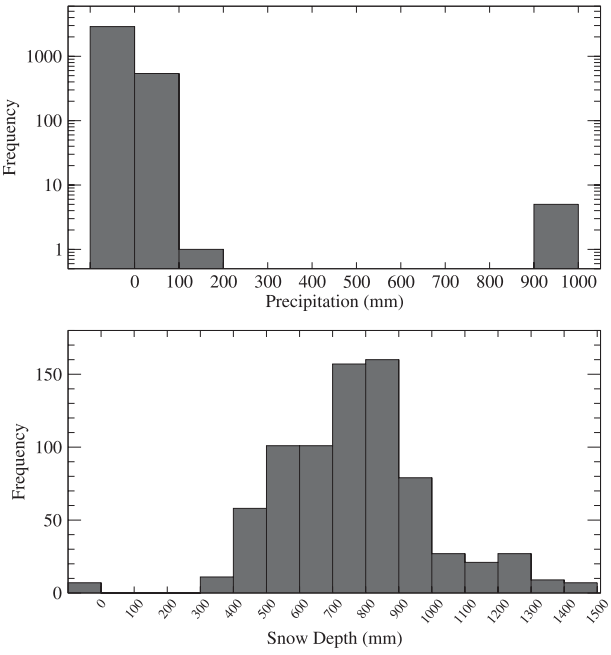

FIG. 3. Examples of values flagged by the gap check (Table 2). (a) Histogram of all daily precipitation totals observed in August during the period of record (1881–2004) at Tiblisy, Georgia (GHCN-Daily station GG000037549), where values of 999 mm (reported on some days in August 1996 and 1997) are flagged because they differ from the next highest precipitation total ever reported in that calendar month by more than the threshold of 300 mm. (b) Histogram of all daily snow depths observed in March during the period of record (1975–2008) at Paxson, AK (GHCN-Daily station USC00507097), where the values of zero (reported in March 1982, 2004, and 2007) are flagged because they differ from the next lowest value ever reported in that calendar month by more than the threshold of 350 mm. In both (a) and (b), the bin size is 100 mm, and each label identifies the inclusive upper boundary of one bin and the exclusive lower boundary of the next bin, such that the first bin includes values equal to 0, the second bin includes values >0 and ≤100 mm, etc.

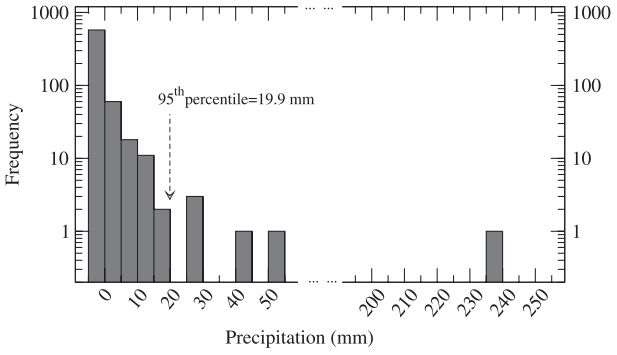

FIG. 4. Histogram of daily precipitation totals reported between 6 Aug and 3 Sep throughout the 1966–90 period of record at Gold Hill (GHCN-Daily station USC00423260) showing an outlier flagged by the percentile-based climatological outlier check (Table 2). The bin size is 5 mm, and every other bin is labeled. Each label identifies the inclusive upper boundary of one bin and the exclusive lower boundary of the next bin, as in Fig. 3. The total of 238.8 mm is flagged because it is more than 9 times larger than the corresponding 95th percentile of 19.9 mm.

The second check, termed the lagged range test, looks for differences in excess of 40°C (i) between TMAX and the warmest TMIN reported on the previous, same, and following days and (ii) between TMIN and the coldest TMAX in the 3-day window centered on the day of the TMIN. The inclusion of the lagged comparisons in the range check avoids the flagging of dynamically induced temperature changes, such as those that can occur in conjunction with frontal passages in interior North America. (Note that there are actually six variations of the lagged range test when TOBS is also included.)

b. Precipitation

Internal consistency checks for precipitation elements typically include comparisons between SNOW and SNWD, SNOW and liquid-equivalent PRCP, and SNOW and TMIN (i.e., nonzero snowfall at temperatures considered too warm for snow) (Reek et al. 1992; Serreze et al. 1998; Brasnett 1999). Although generally effective, these algorithms have two occasional shortcomings. First, they are often based on empirical relationships that neglect the many physical factors and measurement practices affecting the accuracy of both precipitation and snowfall measurements (Goodison 1978; Robinson 1989; Groisman and Legates 1994; Roeber et al. 2003). Second, they generally do not take into account differences in the times at which various elements are reported, a factor that can lead to inconsistencies that are not the result of a data error (Schmidlin et al. 1995).

Consequently, the precipitation consistency checks employ the three previously listed relationships with test thresholds that have been vetted by the threshold selection process described in section 2. To take into account differences in observation time among elements, the checks consider, in one way or another, the 3-day window centered on the day in question. In the simplest of these checks, a nonzero snowfall amount or increasing snow depth is flagged when TMIN exceeds 7°C on the same, previous, and subsequent day. The situation is slightly more complex when comparing two precipitation-related variables because an event total can be split between two observational days in one variable, but not in the other. In the SNOW–SNWD consistency check, for example, an increase in snow depth is considered excessively large when compared with snowfall only when it exceeds the snowfall sums for the previous plus current and current plus subsequent days. Although quite conservative, this approach avoids the systematic flagging of inconsistencies that result from differences in observation time alone.

TABLE 3. As in Table 1, but for internal and temporal consistency checks on temperature. The numbers in parentheses refer to ranges of days: 0 = the day on which the check is centered; −1 = the previous day; 1 = the subsequent day; −1:1 = the prior, current, and subsequent days.

Iterative temperature consistency check		
Variant	Condition for flagging	Values flagged
	Inconsistencies among TMAX, TMIN, and TOBS	In each iteration, values with the most inconsistencies
Spike/dip check		
Variant	Condition for flagging	Values flagged
	Value(0) is at least 25°C larger or at least 25°C smaller than value(−1) and value(1)	TMAX(0) or TMIN(0)
Lagged temperature range check		
Variant	Condition for flagging	Values flagged
TMAX/TMIN	TMAX(0) ≥ max[TMIN(−1:1)] + 40°C	TMAX(0) and TMIN(−1:1)
TMAX/TOBS	TMAX(0) ≥ max[TOBS(−1:1)] + 40°C	TMAX(0) and TOBS(−1:1)
TMIN/TMAX	TMIN(0) ≤ min[TMAX(−1:1)] − 40°C	TMIN(0) and TMAX(−1:1)
TMIN/TOBS	TMIN(0) ≤ min[TOBS(−1:1)] − 40°C	TMIN(0) and TOBS(−1:1)
TOBS/TMAX	TOBS(0) ≤ min[TMAX(−1:1)] − 40°C	TOBS(0) and TMAX(−1:1)
TOBS/TMIN	TOBS(0) ≥ max[TMIN(−1:1)] + 40°C	TOBS(0) and TMIN(−1:1)
Snow–temperature consistency check		
Variant	Condition for flagging	Values flagged
SNOW	SNOW(0) > 0 and min[TMIN(−1:1)] ≥ 7°C	SNOW(0)
SNWD	SNWD(0) − SNWD(−1) > 0 and min[TMIN(−1:1)] ≥ 7°C	SNWD(0) and SNWD(−1)
Snowfall–snow depth consistency check		
Variant	Condition for flagging	Values flagged
	SNWD(0) − SNWD(−1) > SNOW(0) + SNOW(−1) + 25 mm and SNWD(0) − SNWD(−1) > SNOW(0) + SNOW(1) + 25 mm	SNOW(0), SNWD(0), and SNWD(−1)
Snow–precipitation consistency check		
Variant	Condition for flagging	Values flagged
SNOW with 0 PRCP	SNOW(0) ≥ 100 mm and max[PRCP(−1:1)] = 0	SNOW(0) and PRCP(0)
SNOW/PRCP ratio	SNOW(0) ≥ 200 mm and SNOW(0) ≥ 100[PRCP(0) + PRCP(−1)] and SNOW(0) ≥ 100[PRCP(0) + PRCP(1)]	SNOW(0) and PRCP(0)
SNWD increase with 0 PRCP	SNWD(0) − SNWD(−1) ≥ 100 mm and max[PRCP(−1:1)] = 0	SNWD(0), SNWD(−1), and PRCP(0)
SNWD/PRCP ratio	SNWD(0) − SNWD(−1) ≥ 200 mm and SNWD(0) − SNWD(−1) ≥ 100[PRCP(0) + PRCP(−1)] and SNWD(0) − SNWD(−1) ≥ 100[PRCP(0) + PRCP(1)]	SNWD(0), SNWD(−1), and PRCP(0)

In general, as in Reek et al. (1992), the consistency checks are set to flag all values involved in an inconsistency because the evaluation process revealed no justification for systematically incriminating only one of the elements. The one exception is the snow–temperature consistency check; it does not flag temperature since in 90% of the cases evaluated, it was the snowfall or snow depth that was found to be in error.

6. Spatial consistency checks

The spatial consistency checks are listed in Table 4. Generally speaking, these tests involve comparing an observation with concurrent observations at surrounding sites, or "neighbors." The approach usually employs a statistical technique (such as regression or interpolation) to generate an estimate at the "target" station and then to flag those target values that deviate excessively from the neighbor-based estimates (Eischeid et al. 1995; Hubbard et al. 2005; Kunkel et al. 2005; Hubbard et al. 2007). For temperature in particular, estimates derived from spatial regression are generally more accurate than those produced using other methods (Eischeid et al. 2000; Hubbard and You 2005; Hubbard et al. 2007).

A disadvantage of estimation-based techniques, however, is that large differences between the estimated and observed values may result from the failure of the estimation method to accurately depict complex spatial

TABLE 4. As in Table 3, but for spatial consistency checks. See the text for details on the selection of neighbors and for explanations of the conditions for flagging.

Regression check		
Variant	Condition for flagging	Values flagged
	Residual ≥ 8°C or ≤ −8°C and normalized residual ≥ 4.0 or ≤−4.0	TMAX or TMIN

Spatial corroboration check		
Variant	Condition for flagging	Values flagged
Anomaly based	Anomaly at target differs by 10°C from all day 0, −1, and 1 anomalies at first 3–7 neighbors	TMAX(0) or TMIN(0) at target
Percentile based	Data value at target differs from all day 0, −1, and 1 values at first 3–7 neighbors by an amount dependent on the corresponding percent rank difference	PRCP(0) at target

Spatial snow–temperature consistency check		
Variant	Condition for flagging	Values flagged
SNOW	SNOW(0) at target > 0 and min[TMIN(−1:1)] at first 3–7 neighbors ≥7°C	SNOW(0) at target
SNWD	SNWD(0) − SNWD(−1) > 0 and min[TMIN(−1:1)] at first 3–7 neighbors ≥7°C	SNWD(0) and SNWD(−1) at target

relationships, such as in areas with high topographic variability or during frontal passages (Kunkel et al. 2005; Hubbard et al. 2007). Furthermore, accurate neighbor-based estimates appear to be difficult to obtain for variables with high spatial variability, such as daily precipitation (Hubbard et al. 2005; Kunkel et al. 2005). An alternative is to perform pairwise comparisons between each target observation and concurrent nearby observations, flagging target values that are not corroborated by any neighbor value (Peterson et al. 1998; Higgins et al. 2000; Kunkel et al. 2005).

The GHCN-Daily system contains both estimation and corroboration tests that are tailored to each variable. Specifically, TMAX and TMIN are subjected to a spatial regression check and to a corroboration check that tests whether the temperature anomaly at the target lies significantly outside the range of the anomalies at selected neighbors. The primary benefit of the latter is that it is applicable in areas where high spatial variability or incompleteness of the data prevents the development of a suitable regression relationship. In contrast, PRCP is only evaluated with a modified form of the same corroboration check. Finally, nonzero SNOW and increases in SNWD are evaluated by testing whether daily minimum temperatures at neighboring stations are too high to make snowfall at the target location plausible.

a. Spatial regression check

The term ''spatial regression'' implies that a regression relationship is developed for a specific window in time in which the temperature at the target location functions as the dependent variable, and temperatures at selected nearby stations serve as the independent variables (Eischeid et al. 1995; Hubbard et al. 2005). In the spatial regression check employed here, TMAX and TMIN are analyzed separately for each station and year/month. Each day's estimate is an average of neighbor observations selected from a 3-day window centered on the day in question, weighted by the simple linear regression coefficient and index of agreement for each target–neighbor pair. The appropriate neighbors are chosen, and the corresponding regression coefficients and indices of agreement are computed, for each year/month separately, using the approach described in appendix B. To determine whether a target observation should be flagged, both the residual and the corresponding standardized residual (i.e., the residual normalized by the mean and STD of all residuals within the regression window) must exceed their respective thresholds (Fig. 5).

While this check is similar to previous implementations for estimating or assuring quality of daily temperatures (cf. Eischeid et al. 2000; Hubbard et al. 2005), it has been refined in several ways to minimize the risk of false positives. First, rather than using the correlation or root-mean-square error as a criterion for selecting and weighting neighbors, the index of agreement (d) is used here because it provides a measure of both the covariation and the absolute differences between the target and neighbor series (Legates and McCabe 1999). Therefore, the selection of neighbors with high d values should reduce the risk of outliers in the residual that are caused by errors in the estimate rather than in the observation. Second, the 3-day window is used to reduce estimation errors that can be caused by interstation differences in either the timing of meteorological events or the time of observation (Wu et al. 2005; You and Hubbard 2006). Finally, the use of the absolute threshold in addition to the standardized threshold reduces the risk of overflagging when the STD of the residual is small.

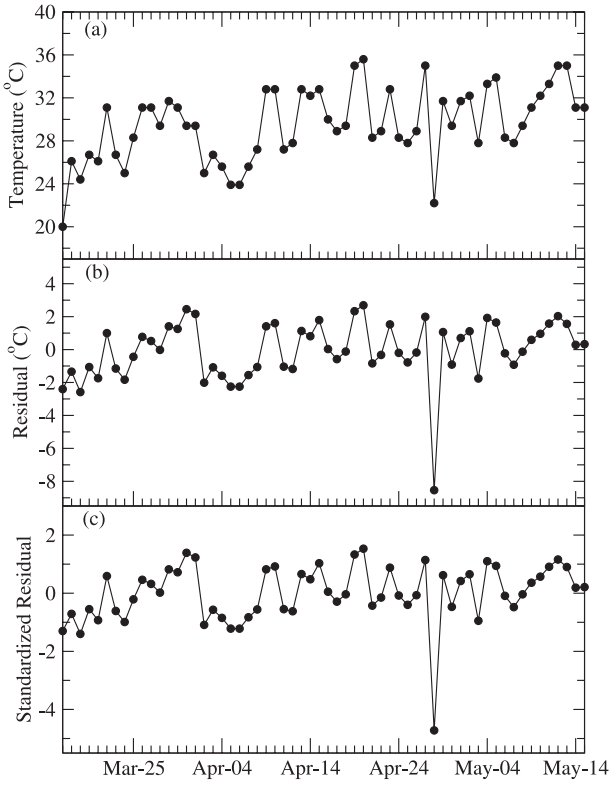

FIG. 5. Time series containing a temperature flagged by the spatial regression check (Table 4). (a) Daily maximum temperatures at Bracketville, TX (GHCN-Daily station USC00411007), between 17 Mar and 15 May 1991; (b) the corresponding residual time series; and (c) the time series of standardized residuals. The temperature of 22.2°C on 28 Apr is flagged because the residual and standardized residual on that day are greater than 8°C and 4.0 standardized units, respectively. See section 6a for additional information.

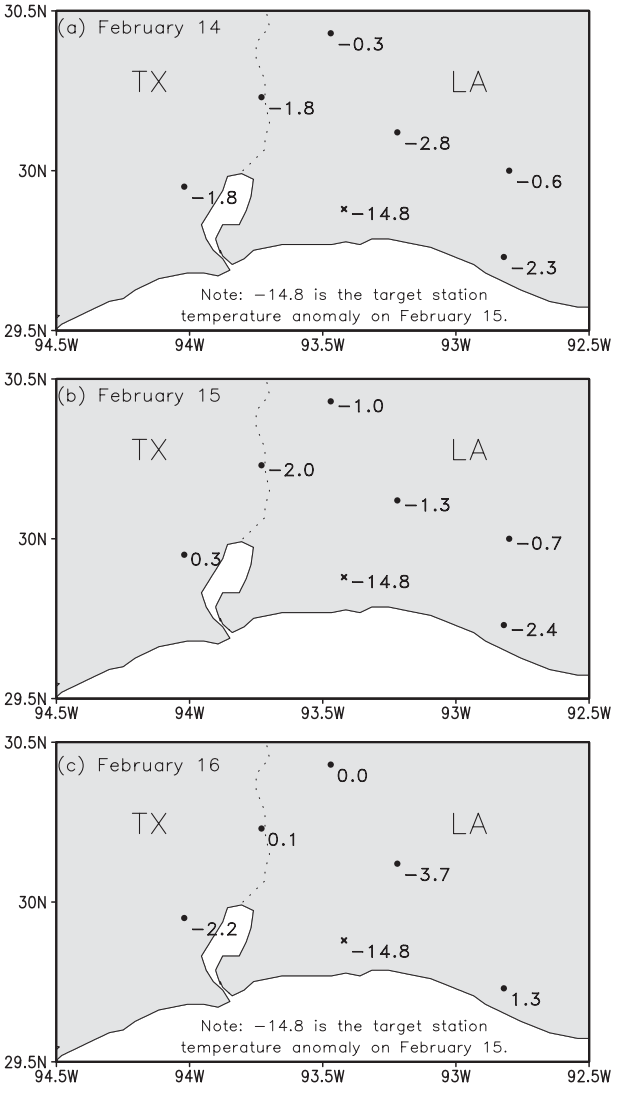

FIG. 6. Maps illustrating the spatial corroboration check on temperature (Table 4). Shown are the daily minimum temperature anomaly at Hackberry, LA (GHCN-Daily station USC00163979), on 15 Feb 2002 and the daily minimum temperature anomalies to which this "target value" is compared: (a) the six available neighbor anomalies on day −1 (14 Feb); (b) the six neighbor anomalies available on day 0 (15 Feb); and (c) the five neighbor anomalies available on day +1 (16 Feb). The target value is indicated by an X symbol, the neighbor values by filled circles. The target anomaly of −14.8°C is flagged because it is 11.1°C lower than the coldest temperature anomaly among the neighbor values within the 3-day window.

b. Spatial corroboration checks

The spatial corroboration checks determine whether the value in question falls significantly outside the range of values reported at neighboring stations. As in the spatial regression check, neighbors must be located within 75 km of the target, and each target observation is compared with neighboring values at lags of −1, 0, and 1 day. In this case, however, neighbors are selected solely based on their physical proximity to the target station and data availability. The test is applied only if, on each of the days of the 3-day window, at least three neighbors are available; if more than seven neighbors are present, only the nearest seven are used.

For TMAX and TMIN, the check is performed on anomalies from the long-term mean. The long-term mean is calculated in the same manner as for the *z*-score-based outlier check. A temperature is identified as an error if the corresponding anomaly differs from all neighbor anomalies by at least 10°C (Fig. 6); that is, an error flag is set only if none of the temperature anomalies at the selected neighbors and within the 3-day window "corroborate" the temperature anomaly at the target location. By virtue of differences in data requirements and neighbor selection, the temperature corroboration and spatial regression checks complement each other in an

important way. Since there is no need to calculate a measure of agreement outside of the 3-day window containing the value being tested, the corroboration check can be applied at times and locations where the spatial regression check is not applicable [e.g., where the 2-month window required for the regression is insufficiently complete or the correlation between the fit and target series within that window is too low (appendix B)]. At the same time, however, the spatial regression check, when applicable, is capable of detecting spatial inconsistencies that are smaller in magnitude than those detectable with the simple corroboration approach. As a result, each of the two procedures detects errors not identified by the other.

In the case of precipitation, the corroboration test is applied to raw daily totals (Fig. 7). If the target observation falls outside the range of the neighbor values, the difference between the target value and the next highest or lowest neighbor value must exceed a threshold that is inversely related to the difference between the climatological percent ranks of the corresponding target and neighbor totals (appendix C). The dependence of the test threshold on percent rank differences implies that considerably larger target–neighbor differences are tolerated when the difference in percent rank is small (e.g., when heavy precipitation is observed throughout the region) than when the difference in probabilities is large (e.g., for isolated heavy totals at the target location). In this respect, the procedure is similar to that of Kunkel et al. (2005).

c. Spatial snow–temperature consistency check

In the spatial snow–temperature consistency check, nonzero SNOW and an increase in SNWD are flagged when TMIN at the nearest three–seven neighbors is at or above 7°C on the preceding, current, and subsequent days. (TMIN at the target location is not considered.) The procedure thus augments the snow–temperature internal consistency check by detecting nonzero SNOW totals and SNWD increases under implausibly warm conditions even when temperatures are not reported at the target station.

A potential risk of the snow–temperature spatial consistency check is the misidentification of valid observations as errors when the neighboring stations are located at considerably lower elevations compared to the target location. At least in the case of GHCN-Daily, however, our evaluation does not indicate a systematic occurrence of such false positives. Nevertheless, if this procedure is applied to data in which isolated mountain stations are common, it may be necessary to choose only those neighbors with elevations similar to that of the target station.

7. Megaconsistency checks

The megaconsistency checks are listed in Table 5. The principle behind these checks is to ensure that, after all other QA procedures have been applied, certain relationships hold for each station's entire record of unflagged values. There are three such checks: the extremes megaconsistency check, the snow–temperature megaconsistency check, and the snow season reality check.

Applied to each station and calendar month separately (e.g., January at Jan Mayen, Norway), the extremes megaconsistency check looks for two types of inconsistencies: TMINs that are higher than the highest unflagged TMAX for the station and calendar month and TMAXs that are lower than the lowest unflagged TMIN (Fig. 8). The test requires that the period of record for the station and month contains at least 140 values of the element whose extreme is used in the test. Note that this test is only necessary because missing values occasionally prevent the application of the corresponding internal consistency check.

Analogously, the snow–temperature megaconsistency check flags nonzero values of SNOW and SNWD when even the lowest TMIN ever recorded for the station and calendar month is greater than or equal to 7°C, provided that at least 140 TMINs are available for determining the lowest TMIN. It thus helps to identify reports of snow at locations and times of year when temperatures have never been cold enough to support such reports.

Last, the snow season reality check tests for nonzero reports of SNOW during the warm half of the year at locations where the cold half of the year has always been snow-free. As such, it facilitates the detection of some obviously erroneous nonzero snowfall and snow depth values that are not detected by any of the tests involving snow and temperature. This is the case, for example, in areas where only precipitation-related measurements are available.

8. System performance

The overall performance of the QA system was assessed by applying the entire set of procedures to the full GHCN-Daily dataset and analyzing the results in two ways. First, the resulting flag rates and estimated false-positive rates were summarized for each group of procedures and for the system as a whole. Second, the validity of 50 randomly chosen values (half flagged, half unflagged) was assessed for each element. Although the quantitative results from this final assessment are presented, emphasis is placed on the qualitative information gained since the samples are too small to be statistically representative of the entire dataset. (A much larger sample size would have been preferable, but it also would

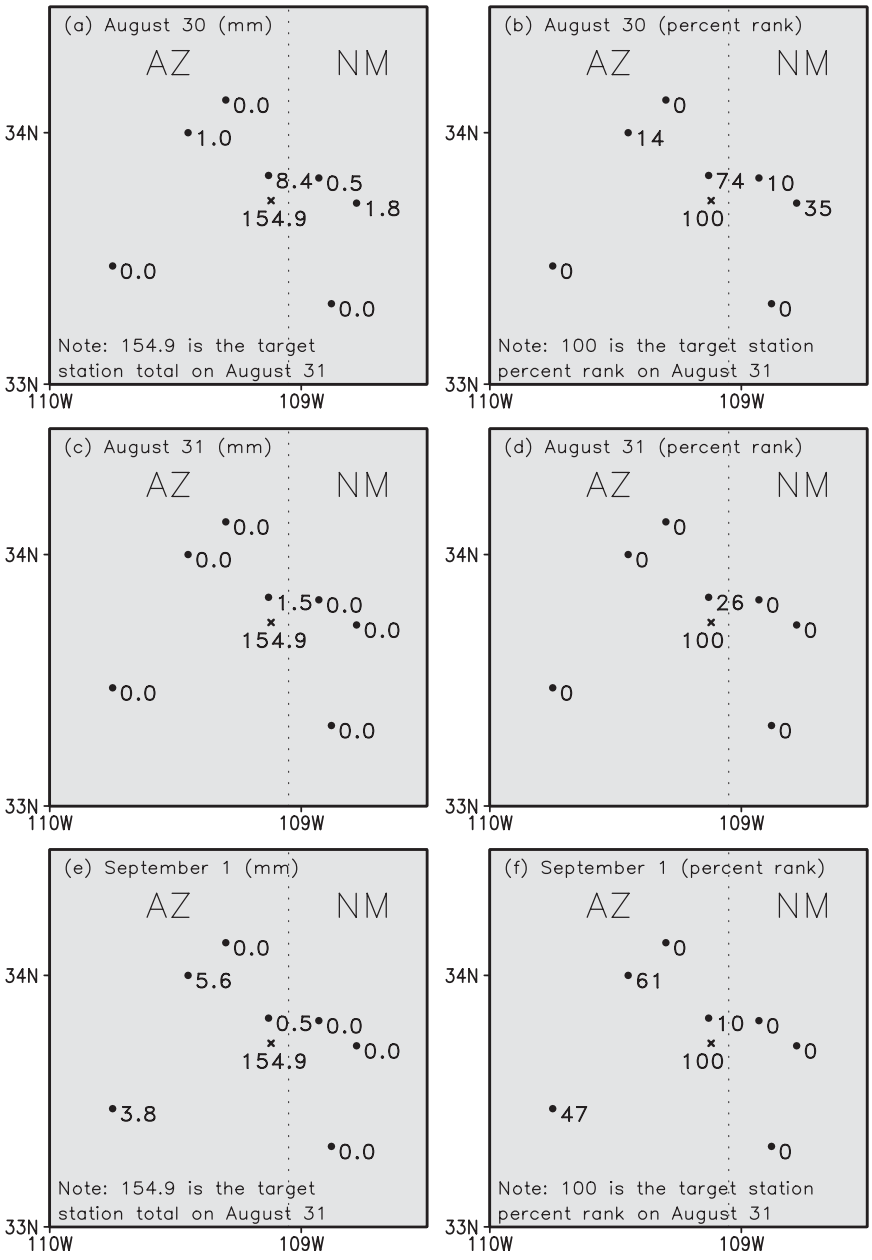

FIG. 7. Maps illustrating the spatial corroboration check (Table 4) applied to a 154.9-mm precipitation total at Alpine, AZ (GHCN-Daily station USC00020174), on 31 Aug 1996. In addition to this target total or its percent rank (X symbol), the maps show all neighbor information (filled circles) used in the check: (a) neighbor totals on day −1; (b) neighbor percent ranks on day −1; (c) neighbor precipitation totals on day 0; (d) neighbor percent ranks on day 0; (e) neighbor totals on day +1; and (f) neighbor percent ranks on day +1. The minimum absolute target–neighbor percent rank difference is 26, yielding a test threshold of 120.3 mm (appendix C). The target value is flagged because the minimum absolute target–neighbor difference among totals is 146.5 mm and therefore exceeds the threshold.

have been too large for a manual assessment to be practical.)

Overall, 0.24% of all GHCN-Daily observations are flagged (Table 6). The internal and temporal consistency checks set the vast majority of these flags, roughly 0.21% of the dataset. The basic integrity checks, flagging 0.02% of the data, account for many of the remaining flags. Note that these flag rates are consistent with those of

TABLE 5. As in Table 1, but for megaconsistency checks.

Extremes megaconsistency check		
Variant	Condition for flagging	Values flagged
TMAX	TMAX < lowest TMIN for the station and calendar month	TMAX
TMIN	TMIN > highest TMAX for the station and calendar month	TMIN
Snow–temperature megaconsistency check		
Variant	Condition for flagging	Values flagged
	SNOW > 0 or SNWD > 0 when lowest TMIN for station and calendar month >7°C	SNOW or SNWD
Snow season reality check		
Variant	Condition for flagging	Values flagged
Northern Hemisphere	SNOW > 0 (or SNWD > 0) for all days in May–October and SNOW = 0 (or SNWD = 0) in November–April	SNOW or SNWD
Southern Hemisphere	SNOW > 0 (or SNWD > 0) for all days in November–April and SNOW = 0 (or SNWD = 0) in May–October	SNOW or SNWD

comparable procedures designed to detect gross data errors and limit the number of false positives (Reek et al. 1992; Kunkel et al. 1998; Feng et al. 2004; Brunet et al. 2006).

One method for approximating the overall false-positive rate is to sum the estimated number of false positives from all procedures and then divide by the total number of flagged values. When this method is applied to the results in Table 6, approximately 1% of the flags are estimated to be false positives. Similarly, a 2% false-positive rate is obtained when the same method is applied to the results in Table 7 (i.e., the randomly selected data).

Overall, it appears that no group of procedures has a false-positive rate greater than 15% (Table 6), and the estimated false-positive rate for each element is less than 5% (Table 7). The highest false-positive rates are associated with the outlier checks and spatial consistency checks (10% and 13%, respectively), although they each flag less than 0.01% of the data. The only other procedures for which false positives were identified are the duplicate year/month check on SNOW, the frequent-value check, the spike/dip check, the lagged range check, and the internal consistency check between temperature and SNOW/SNWD.

The relative inefficiency of the popular outlier and spatial consistency checks is consistent with the results of other studies that have manually assessed the validity of flags set by their QA procedures. For example, Wolter (1997) demonstrates that the process of normalizing values by the STD can lead to overflagging when the distribution of measured values is skewed relative to the normal distribution. Furthermore, Kunkel et al. (2005) found that even for carefully designed outlier and spatial consistency checks of various kinds, the false-positive rate increases rapidly when the test threshold is lowered, even when the threshold is in the extreme tail of the distribution of the test parameter (e.g., z score > 5.0) where only a small fraction of the values reside (Kunkel et al. 2005). Nevertheless, the outlier and spatial consistency checks are applied here with the thresholds shown in Tables 2 and 4 because most of the errors

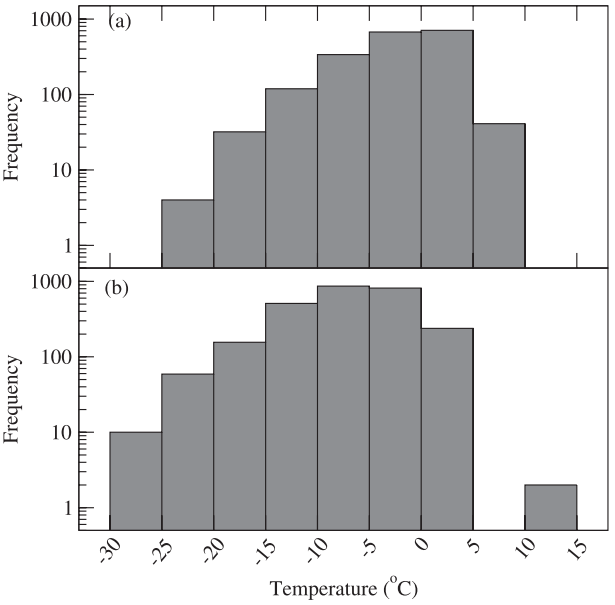

FIG. 8. Histograms of all January (a) daily maximum temperatures and (b) daily minimum temperatures reported at Jan Mayen (GHCN-Daily station JN000099950) illustrating the extremes megaconsistency check (Table 5). The bin size is 5°C, and every bin is labeled. Each label identifies the inclusive upper boundary of one bin and the exclusive lower boundary of the next bin, as in Fig. 3. The 10.3° and 10.6°C TMINs (both reported in January 1929) are flagged by the check because they exceed the highest unflagged January TMAX (9.5°C) reported during the station's 1921–2009 record. See section 7 and Table 5 for additional details.

TABLE 6. Summary statistics about each group of checks in the core GHCN-Daily system. Results are based on GHCN-Daily data through 2008. The flag rate for a group represents the sum of the flags set by each procedure in the group, divided by the total number of values in the dataset. A group's corresponding false-positive rate is estimated as follows: (i) for each component check and element, the false-positive rate obtained during the manual threshold selection process is multiplied by the number of values flagged by the check; then (ii) the resulting numbers of false positives are summed and divided by the total number of values flagged by all of the procedures.

Group of procedures	Flag rate	Estimated false-positive rate
Basic integrity checks	0.0209%	3%
Outlier checks	0.0055%	12%
Internal and temporal inconsistency checks	0.2068%	0%
Spatial inconsistency checks	0.0094%	9%
Megaconsistency checks	0.0001%	0%
All procedures	0.2427%	1%

TABLE 7. Results from assessing the validity of 25 flagged and 25 unflagged values of TMAX, TMIN, PRCP, SNOW, and SNWD in GHCN-Daily after the QA system has been applied. ''Questionable'' values are counted as half valid. The percentage of valid values is the sum of the number of values not considered invalid and 0.5 times the number of questionable values, divided by the number of values evaluated. For flagged values, this is equivalent to the false-positive rate. Values that are part of an inconsistency are considered to be invalid as long as at least one value causing the inconsistency is judged to be a data error.

Element	Type of value	Percent valid
TMAX	Flagged	0%
	Unflagged	100%
TMIN	Flagged	4%
	Unflagged	98%
PRCP	Flagged	4%
	Unflagged	100%
SNOW	Flagged	0%
	Unflagged	100%
SNWD	Flagged	4%
	Unflagged	92%
All	Flagged	2%
	Unflagged	99%

detected by them would otherwise not be identified, and an examination of the spatial distribution of the flags set by these procedures revealed no significant bias that could be the result of overflagging conditions in a particular region. The benefit of flagging these errors is therefore assumed to outweigh the impact of the accompanying 10%–15% false-positive rate.

It is clear from our manual assessments of individual procedures that whenever a test threshold is chosen, a number of invalid values are left unflagged. Although some of these are ultimately detected by other procedures, other (typically less extreme) errors manage to pass all of the tests. For example, multiday precipitation totals exist that are neither identified as such accumulations in the data nor sufficiently unusual in magnitude to be detected by one of the QA procedures. In addition, there are situations in which low temporal or spatial resolution prevents the detection of a particular error, as was the case for the two snow depth values judged to be invalid during the overall evaluation (Table 7). Should the impact of such lingering errors later be judged to be excessive, an attempt could be made to develop additional procedures that address the specific data problems rather than relax the thresholds of existing checks (DMV08).

9. Concluding remarks

A QA system is presented here that effectively quality assures large datasets of daily observations of five primary temperature and precipitation elements. Although fully automated, the system has a false-positive rate of roughly 1%–2%, a rate much lower than that of typical semiautomated QA procedures. This false-positive rate is achieved by employing a set of basic integrity, outlier, and consistency tests with complementary error detection capabilities whose design is informed by manual assessments of the validity of samples of values they flag. Considering both the evaluations performed during the design of each individual check and the final overall evaluation, a total of approximately 2000 values was assessed prior to the deployment of the system.

When applied to NCDC's GHCN-Daily dataset, the system flags 0.24% of the observations, a flag rate comparable to that achieved by other fully automated systems designed to incur a small number of false positives. The system's flag rate is a function of the choice of procedures and test thresholds, of the percentage of the data affected by errors that these procedures can detect, and of the spatial and temporal completeness of the data that dictates which procedures can be applied to a particular data value. Considering the data completeness requirements of the various types of procedures, the system's capability to detect data errors is maximized at stations where all five primary elements are reported consistently for at least 7 yr, and where at least three neighbors with similarly complete overlapping periods of record are available within a 75-km radius. One clear advantage of this fully automated system is that the entire dataset is processed in a uniform and reproducible manner. In addition, the dataset can be reprocessed whenever additional data or QA algorithms become available or whenever existing processing procedures are enhanced.

The above description of the GHCN-Daily QA system is intended first and foremost as documentation of the methodology behind the QA flags in the GHCN-Daily dataset (http://www.ncdc.noaa.gov/oa/climate/ghcn-daily/). Combined with the presentation of design and evaluation strategies in DMV08, however, it also serves to illustrate a framework for designing and implementing comprehensive automated QA systems that may be extended to other datasets. Users interested in applying our specific programs (written in FORTRAN 95 on a Linux platform) to their own datasets may visit the GHCN-Daily Web page for the appropriate point of contact.

Acknowledgments. Partial support for this work was provided by the Office of Biological and Environmental Research, U.S. Department of Energy (Grant DE-AI02-96ER62276). We thank Xungang Yin for assistance in the creation of figures and the reviewers for their constructive comments.

APPENDIX A

Description of the Iterative Internal Consistency Check on Temperature

The iterative internal consistency check evaluates each temperature in a station's record for violations of expected relationships with other temperatures on the same and adjacent days. It then uses complex logic to decide which values are in error and repeats the test until no additional violations are found. The procedure consists of four steps:

1) Each running pair of consecutive days is tested for the following seven conditions, counting the number of violations found for each element and day:
   - TMAX(0) < TMIN(0) − 1°C
   - TOBS(0) > TMAX(0) + 1°C
   - TOBS(0) < TMIN(0) − 1°C
   - TMAX(0) < TMIN(1) − 1°C
   - TMIN(0) > TMAX(1) + 1°C
   - TMAX(1) < TOBS(0) − 1°C
   - TMIN(1) > TOBS(0) + 1°C

   Here, (0) and (1) refer to the current and next days, respectively. Note that by moving through a time series in this fashion, comparisons are, by implication, also made between the current and previous days. The 1°C tolerance allows for minor inconsistencies resulting from variations in response characteristics among thermometers, which are relevant when multiple sensors are used to measure different types of temperature (Guttman and Quayle 1990).

TABLE A1. Sample temperature observations containing internal inconsistencies. The example is taken from Clay City, IL (GHCN-Daily station USC001 12687). The GHCN-Daily internal consistency check flags TMIN on 1 Mar (boldface) because it is inconsistent with both TMAX and TOBS on 28 Feb (italics). See the text for further details.

Date	TMAX (°C)	TMIN (°C)	TOBS (°C)
27 Feb 1985	13.3	−3.9	−3.3
28 Feb 1985	*−1.7*	−7.2	*−6.7*
1 Mar 1985	8.3	**3.9**	3.9
2 Mar 1985	10.0	0.6	0.6

2) Having proceeded through the entire time series, the values with the largest number of violations are flagged.
3) Steps 1 and 2 are repeated, ignoring the values flagged in previous iterations, until no additional inconsistencies are found.
4) Should any days remain on which TMAX < TMIN, both TMAX and TMIN are flagged on those days. In other words, all cases with TMAX < TMIN on the same day are flagged regardless of the magnitude of the difference.

For illustrative purposes, consider the example shown in Table A1. On all days listed, TMIN is less than the corresponding TMAX, and the same day's TOBS lies in between. However, TMIN on 1 March is less than both TMAX and TOBS on 28 February. Accordingly, during the first pass of the check, TMIN on 1 March accumulates two violations, while TMAX and TOBS on the preceding day each accumulate one violation. TMIN on 1 March is then excluded from the check during the second pass, resulting in the elimination of all violations. Consequently, only TMIN on 1 March is flagged.

APPENDIX B

Computations for the Spatial Regression Check

The coefficients and indices of agreement required for the spatial regression check are calculated from observations within a window stretching from 15 days before the beginning of the month to 15 days after the end of the month. For example, the regression relationships for April 1991 (Fig. 5) are based on temperatures reported between 17 March and 15 May of that year. Within each regression window, daily estimates of TMAX and TMIN at each station are calculated as described below.

First, suitable neighbors are chosen on the basis of data completeness within the regression window, distance from the target location, and their index of agreement with the target observations in the window. A station is

considered a potential neighbor if it lies within 75 km of the target, and at least 40 days within the window contain observations at both the target and the neighbor. If more than three such neighbors are available, they are sorted according to their index of agreement with the target during the regression window, and the neighbors with the seven (or fewer if seven are not available) highest indices of agreement are chosen.

Following Legates and McCabe (1999), the index of agreement is defined as

$$d = 1.0 - \frac{\sum_{i=1}^{m} |y(i) - x(i)|}{\sum_{i=1}^{m} [|x(i) - \bar{y}| + |y(i) - \bar{y}|]}, \quad \text{(B1)}$$

where d is the index of agreement, m is the number of days in the window, $x(i)$ and $y(i)$ are the observations at the target and neighbor on day i, and y denotes an average over all observations in the time window. Thus, high values of d are an indication of both high correlation and small absolute differences between x and y.

Each day's estimated TMAX or TMIN at the target station is then calculated using the following formula:

$$\widehat{y(i)} = \frac{\sum_{k=1}^{n} [b(k) + a(k)x'(i,k)]d(k)}{\sum_{k=1}^{n} d(k)}, \quad \text{(B2)}$$

where $\widehat{y(i)}$ is the estimate on day i, n is the number of neighbors, $a(k)$ and $b(k)$ are the slope and intercept for neighbor k, respectively, and $x'(i, k)$ is the observation at neighbor k within the 3-day window centered on day i that is closest in magnitude to the target observation on day i. For example, if the target observation is 23.9°C, and the observations at a neighbor are 18.3°, 10.6°, and 7.2°C on the preceding, concurrent, and subsequent days, respectively, the 18.3°C temperature would be used in the calculation of the estimate. Once all estimates in the regression window have been computed, the algorithm proceeds to the actual regression check only if the estimates are correlated with the observed time series at a level of 0.8 or higher, that is, when the fit adequately describes the variations in the target series.

The size of the regression window and radius within which neighbors are selected are consistent with the sensitivity studies of Hubbard and You (2005). Our choice to use between 3 and 7 neighboring stations represents a compromise among Hubbard and You's (2005) suggestion to use as many as 10 neighbors, Eischeid et al.'s (2000) recommendation to avoid overfitting the target series by using a maximum of 4 neighbors, and station density throughout the record.

APPENDIX C

Steps in the Spatial Corroboration Check on Precipitation

The following is a description of the steps involved in the corroboration check on precipitation. An illustrative example is shown in Fig. 7.

Each precipitation total with sufficient nearby observations is tested as follows:

1) A "minimum absolute target–neighbor difference" is obtained from the pairwise differences between the precipitation total being evaluated and each neighbor total within the 3-day window. If the target observation exceeds the largest neighbor value or is less than the smallest neighbor value, the minimum target–neighbor difference is set to the absolute value of the pairwise difference that is the smallest in magnitude; otherwise, it is set to zero.
2) The minimum absolute target–neighbor difference is then also determined from the climatological percent ranks of the respective totals. Each percent rank is the rank (in percent) of the total among all nonzero values observed throughout the station's period of record during a 29-day window centered on the relevant day, provided that at least 20 nonzero values are present within the window during all years combined.
3) If sufficient data are available to calculate the minimum absolute percent rank difference, the target total is flagged if the minimum absolute difference among totals exceeds the following test threshold:

$$\text{Threshold} = -45.72 \ln \Delta_{\text{rank}} + 269.24, \quad \text{(C1)}$$

where the threshold is expressed in mm, and Δ_{rank} is the minimum absolute target–neighbor difference based on percent ranks. The threshold function was derived by performing a threshold selection evaluation (DMV08) in three categories of Δ_{rank} (0%–5%, 40%–60%, and 90%–99%), choosing a threshold in each category that yielded a 20% false-positive rate, and fitting a logarithmic function to the resulting three sets of (Δ_{rank} threshold) pairs.
4) If percent ranks are not available for either the target total or a sufficient number of neighbors, as might be the case for very short records or in extremely dry regions, the test threshold is set to the maximum of the above function, or 269.24 mm. Although quite

crude, a comparison of the minimum absolute target–neighbor difference to this threshold nevertheless allows for the detection of extremely egregious spatial inconsistencies.

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.

Brasnett, B., 1999: A global analysis of snow depth for numerical weather prediction. *J. Appl. Meteor.,* **38,** 726–740.

Brunet, M., and Coauthors, 2006: The development of a new data set of Spanish Daily Adjusted Temperature Series (SDATS) (1850–2003). *Int. J. Climatol.,* **26,** 1777–1802.

Caesar, J., L. Alexander, and R. 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.

Cerveny, R. S., J. Lawrimore, R. Edwards, and C. Landsea, 2007: Extreme weather records: Compilation, adjudication, and publication. *Bull. Amer. Meteor. Soc.,* **88,** 853–860.

Daly, C., W. P. Gibson, G. H. Taylor, M. K. Doggett, and J. I. Smith, 2007: Observer bias in daily precipitation measurements at U.S. Cooperative Network stations. *Bull. Amer. Meteor. Soc.,* **88,** 899–912.

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

——, R. S. Vose, and D. B. Wuertz, 2008b: Robust automated quality assurance of radiosonde temperatures. *J. Appl. Meteor. Climatol.,* **47,** 2081–2095.

Easterling, D. R., T. R. Karl, J. H. Lawrimore, and S. A. Del Greco, 1999: United States Historical Climatology Network daily temperature, precipitation, and snow data for 1871–1997. ORNL/CDIAC-118, NDP070, Carbon Dioxide Information Analysis Center, Environmental Sciences Division Publ. 4887, Oak Ridge National Laboratory, 84 pp. [Available from National Technical Information Service, 5285 Port Royal Rd., Springfield, VA 22161, and online at http://www.ornl.gov/~webworks/cppr/y2002/rpt/101454.pdf.]

Eischeid, J. K., C. B. Baker, T. Karl, and H. F. Diaz, 1995: The quality control of long-term climatological data using objective data analysis. *J. Appl. Meteor.,* **34,** 2787–2795.

——, P. A. Pasteris, H. F. Diaz, M. S. Plantico, and N. J. Lott, 2000: Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *J. Appl. Meteor.,* **39,** 1580–1591.

Feng, S., Q. Hu, and W. Qian, 2004: Quality control of daily meteorological data in China, 1951–2000: A new data set. *Int. J. Climatol.,* **24,** 853–870.

Goodison, B. E., 1978: Accuracy of Canadian snow gauge measurements. *J. Appl. Meteor.,* **17,** 1542–1548.

Green, J. S., M. Klatt, M. Morrissey, and S. Postawko, 2008: The Comprehensive Pacific Rainfall Database. *J. Atmos. Oceanic Technol.,* **25,** 71–82.

Groisman, P. Y., and D. R. Legates, 1994: The accuracy of U.S. precipitation data. *Bull. Amer. Meteor. Soc.,* **75,** 215–227.

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

Harmel, R. D., C. W. Richardson, C. L. Hanson, and G. L. Johnson, 2002: Evaluating the adequacy of simulating maximum and minimum daily air temperature with the normal distribution. *J. Appl. Meteor.,* **41,** 744–753.

Higgins, R. W., W. Shi, E. Yarosh, and R. Joyce, 2000: *Improved United States Precipitation Quality Control System and Analysis*. NCEP/Climate Prediction Center Atlas 7, 40 pp. [Available online at http://www.cpc.ncep.noaa.gov/research_papers/ncep_cpc_atlas/7/index.html.]

Hubbard, K. G., and J. You, 2005: Sensitivity analysis of quality assurance using spatial regression approach—A case study of the maximum/minimum air temperature. *J. Atmos. Oceanic Technol.,* **22,** 1520–1530.

——, 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.

——, N. B. Guttman, J. You, and Z. Chen, 2007: An improved QC process for temperature in the daily cooperative weather observations. *J. Atmos. Oceanic Technol.,* **24,** 201–213.

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

Kunkel, K. E., and Coauthors, 1998: An expanded digital daily database for climatic resources applications in the Midwestern United States. *Bull. Amer. Meteor. Soc.,* **79,** 1357–1366.

——, 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.

——, M. A. Palecki, K. G. Hubbard, D. A. Robinson, K. T. Redmond, and D. R. Easterling, 2007: Trend identification in twentieth-century U.S. snowfall: The challenges. *J. Atmos. Oceanic Technol.,* **24,** 64–73.

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

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.

Nicholls, N., 1995: Long-term climate monitoring and extreme events. *Climatic Change,* **31,** 231–245.

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.

Robinson, D. A., 1989: Evaluation of collection, archiving, and publication of daily snow data in the United States. *Phys. Geogr.,* **10,** 120–130.

——, 1990: The United States Cooperative climate-observing systems: Reflections and recommendations. *Bull. Amer. Meteor. Soc.,* **71,** 826–831.

Roeber, P. J., S. L. Bruening, D. M. Schultz, and J. V. Cortinas Jr., 2003: Improving snowfall forecasting by diagnosing snow density. *Wea. Forecasting,* **18,** 264–287.

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.

Serreze, M. C., M. P. Clark, and D. I. McGinnis, 1998: Characteristics of snowfall over the eastern half of the United States and relationships with principal modes of low-frequency atmospheric variability. *J. Climate,* **11,** 234–250.

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.

Wallis, J. R., D. L. Lettenmaier, and E. F. Wood, 1991: Daily hydroclimatological data set for the continental United States. *Water Resour. Res.,* **27,** 1657–1663.

Wolter, K., 1997: Trimming problems and remedies in COADS. *J. Climate,* **10,** 1980–1997.

Wu, H., K. G. Hubbard, and J. You, 2005: Some concerns when using data from the Cooperative Weather Station Network: A Nebraska case study. *J. Atmos. Oceanic Technol.,* **22,** 592–602.

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