Quality Control of Pre-1948 Cooperative Observer Network Data

KENNETH E. KUNKEL

Illinois State Water Survey, Champaign, Illinois

DAVID R. EASTERLING

National Climatic Data Center, Asheville, North Carolina

KENNETH HUBBARD

University of Nebraska at Lincoln, Lincoln, Nebraska

KELLY REDMOND

Desert Research Institute, Reno, Nevada

KAREN ANDSAGER, MICHAEL C. KRUK, AND MICHAEL L. SPINAR

Illinois State Water Survey, Champaign, Illinois

(Manuscript received 7 January 2005, in final form 2 May 2005)

ABSTRACT

A recent comprehensive effort to digitize U.S. daily temperature and precipitation data observed prior to 1948 has resulted in a major enhancement in the computer database of the records of the National Weather Service's cooperative observer network. Previous digitization efforts had been selective, concentrating on state or regional areas. Special quality control procedures were applied to these data to enhance their value for climatological analysis. The procedures involved a two-step process. In the first step, each individual temperature and precipitation data value was evaluated against a set of objective screening criteria to flag outliers. These criteria included extreme limits and spatial comparisons with nearby stations. The following data were automatically flagged: 1) all precipitation values exceeding 254 mm (10 in.) and 2) all temperature values whose anomaly from the monthly mean for that station exceeded five standard deviations. Additional values were flagged based on differences with nearby stations; in this case, metrics were used to rank outliers so that the limited resources were concentrated on those values most likely to be invalid. In the second step, each outlier was manually assessed by climatologists and assigned one of the four following flags: valid, plausible, questionable, or invalid. In excess of 22 400 values were manually assessed, of which about 48% were judged to be invalid. Although additional manual assessment of outliers might further improve the quality of the database, the procedures applied in this study appear to have been successful in identifying the most flagrant errors.

1. Introduction

The National Weather Service's (NWS's) cooperative observer network (COOP) is the core climate network of the United States. In operation since the late nineteenth century, it consists primarily of volunteer observers using standard equipment provided by the NWS. The typical suite of elements observed daily includes precipitation (P), maximum temperature ($T_{\max}$), minimum temperature ($T_{\min}$), snowfall (S_f), and snow depth (S_d). Some stations report only precipitation variables. A few stations observe other variables such as pan evaporation and soil temperature.

Beginning in 1948 [when surplus keypunch machines were obtained from the U.S. Postal Service by the then-existing New Orleans, Louisiana, branch of the National Climatic Data Center (NCDC)], cooperative ob-

Corresponding author address: Dr. Kenneth Kunkel, 2204 Griffith Drive, Champaign, IL 61820-7495.
E-mail: kkunkel@uiuc.edu

servations were routinely stored on machine-readable punch cards. As computers became more widely available, these digitized monthly data were stored on electronic media. Although there have been occasional projects to retroactively digitize selected data, most pre-1948 observations have been available only on paper or microfiche. Recently, the U.S. Congress has provided funding to the NCDC for the Climate Database Modernization Program (CDMP 2001), the goal of which is to convert data that are available only in hard-copy form to computerized formats. The pre-1948 COOP data were some of the first datasets chosen for this conversion.

There are a number of potential sources of errors or quality issues in the digitized dataset, which generally fall into three categories: observer error, station discontinuity, and digitization errors. Observer errors include errors in reading the instruments or in writing the observations on the form, and problems with the equipment, including liquid mercury separation in the thermometers (which a good observer would presumably notice). Station discontinuity issues include potential discontinuities introduced into a station's climate record by changes in instrumentation/shields, observing practices, changes in station location, and exposure. The digitization process may introduce errors into the digitized dataset through errors in properly identifying stations by their station identification numbers, errors in identifying data element types (e.g., snowfall keyed as snow depth), and keying errors in individual values. Keying errors may be increased by poor legibility of the preserved documents.

The COOP data represent the highest and lowest temperature values, or precipitation totals, at any time over the 24 h ending at the time of observation, and are ascribed by long-standing convention to the date of the observation (when instruments are reset). In theory, and often in reality, T_{max}, T_{min}, P, and S_f can occur at any time during these 24 h. "Shifting" refers to the assignment of a value (by either the observer or subsequent processing) to a presumed calendar date of actual occurrence, typically the prior day, rather than to the date of the observation (instrument reset), as required by adherence to the formal convention mentioned above. Shifting errors are most common in the T_{max} data for observers with a morning (A.M.) time of observation. Because on most days the actual time of occurrence of T_{max} is in the afternoon of the prior day, some observers have mistakenly believed that they should record the value on the day that it occurred. The presence of shifting creates problems for spatial quality control (QC) when comparing shifted values with observations from neighboring stations that have been correctly recorded. Also, to complicate matters, during certain multiyear intervals prior to 1948, T_{max} values for morning observers were routinely shifted back 1 day by the data processing system before being printed in *Climatological Data*, the official publication for COOP data. This is not mentioned in those publications. Because some past digitization projects (e.g., Kunkel et al. 1998) keyed data from *Climatological Data*, there are additional shifted values in COOP data that were used in the QC of CDMP data, which were keyed from the original forms.

The primary focus of this project was on quality issues affecting individual values, particularly observer and keying errors. In this article, the dataset is described, along with an analysis of the rate of keying errors in individual values. Objective spatial tests for identifying outliers in daily T_{max}, T_{min}, and P were used to flag outliers; these outliers were manually assessed for their validity. The manual assessment process and its results are described in detail, particularly their indications for observer errors in individual values. One focus of the QC effort was on improving the value of the dataset for analysis of extreme temperature and heavy precipitation events, and some of the tests were designed to identify outliers in extreme values.

The continuing concern about climate variability and change ensures that the COOP data will be heavily used for the indefinite future. Therefore, this paper provides considerable detail so that future users can appropriately consider data quality issues in their studies and applications.

2. Dataset description

COOP observations are recorded on paper forms (one sheet per month) and are sent to NCDC at the end of each month. In the 1980s, NCDC copied all paper forms onto microfiche. The keying of these data in CDMP was done from the microfiche images by Image Entry, a private contractor located in London, Kentucky. All monthly data sheets were double-keyed in a two-person process: the second keyer resolved discrepancies between his/her data and the first set of keyed data as he/she keyed the data. Double-keying minimizes the number of keystroke errors, but neither entirely eliminates them, nor eliminates problems due to illegibility of the form. Extremes tests (using state monthly extremes tables; see National Climatic Data Center 2003) were applied during postprocessing to help ensure accurate keying. Values that failed the extremes tests but verified with the source were retained. Estimated values were added to the database to fill in some missing values, in particular, if the daily precipi-

tation values added up to the observer-supplied monthly total for a station, the other days were zero filled. The total number of values keyed for this project exceeded 300 000 000.

This dataset is designated as DSI-3206 by NCDC. The digital COOP data available prior to this are designated as DSI-3200; this includes the routinely keyed COOP data plus the results of various state- and region-based keying projects done through the years. The digitization and quality control processes for these projects varied. A recently developed dataset of keyed COOP data done for 10 U.S. states (Kunkel et al. 1998) was designated as DSI-3205; this set includes data for the 9 Midwestern states of Illinois, Indiana, Iowa, Kentucky, Missouri, Minnesota, Michigan, Ohio, and Wisconsin, in addition to data from New Mexico. The data for DSI-3205 were single keyed from the publication *Climatological Data* (not the original COOP forms) and passed through quality control tests similar to those used on the current data keyed for DSI-3200. These three datasets—DSI-3200, DSI-3205, and DSI-3206—were combined for this project to create the dataset of all keyed COOP data.

The addition of stations in DSI-3206, compared to what had been previously available in DSI-3200 and DSI-3205 (Figs. 1–3), represents a substantial improvement to the digital record. For temperature (Fig. 1), the number of additional stations is from about 1000 in the late 1800s to about 2000 in the 1940s. For precipitation (Fig. 2), the increase is from more than 1000 in the late 1800s to about 4000 in the 1940s. (The peak between 1948 and 1951 is due to the temporary inclusion of stations from the Hydroclimatic Network of the U.S. Army Corps of Engineers.) The spatial distribution of additional long-term (defined here as those with less than 10% missing data for the period of 1895–2000) temperature stations (Fig. 3) indicates substantial increases in density along the East, Gulf, and West Coasts. Less significant increases occurred in parts of the intermountain West (where fewer COOP stations were operational in the pre-1948 era) and in the upper Midwest [where most data were already keyed in for

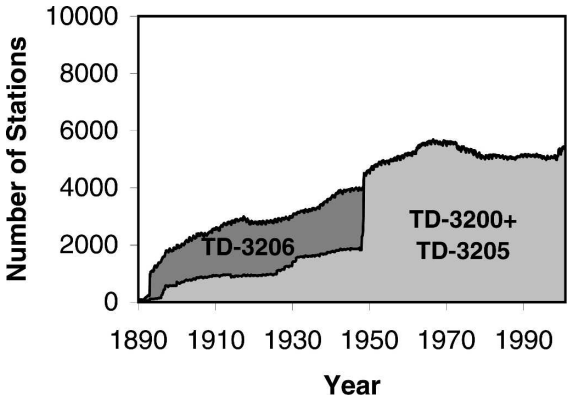

FIG. 1. Time series of the number of stations in the combination of DSI-3200 and DSI-3205 compared to DSI-3206 for the period of 1890–2000 for temperature.

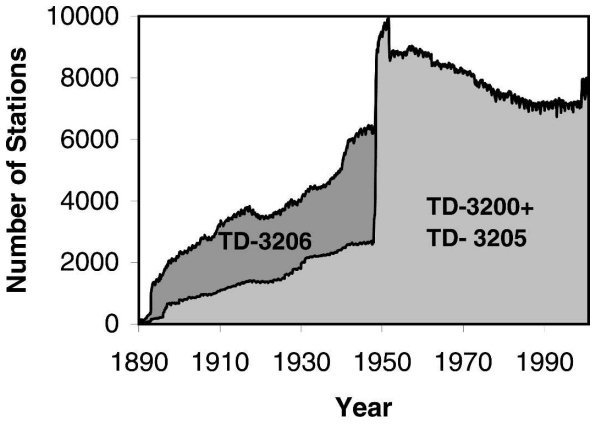

FIG. 2. Same as in Fig. 1, but for precipitation.

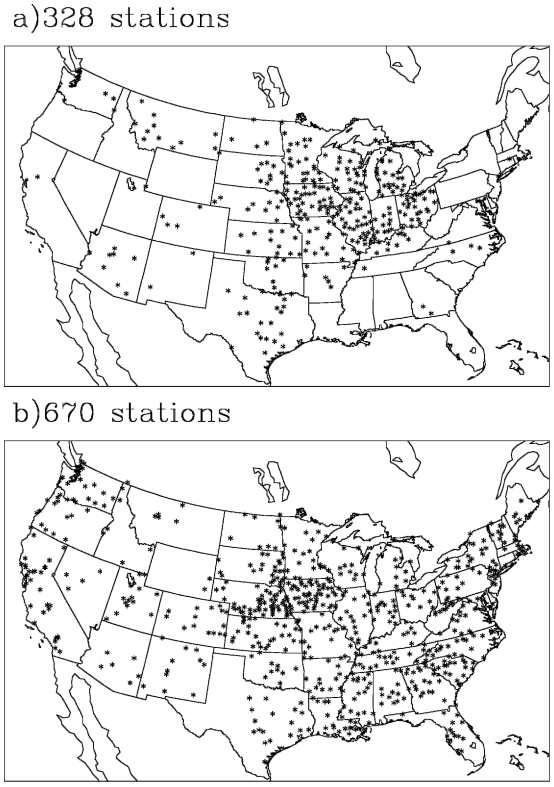

FIG. 3. Map of the location of long-term stations (less than 10% missing data for the period of 1895–2000) for (a) DSI-3200 and DSI-3205 and (b) stations added by including DSI-3206.

the project of Kunkel et al. (1998)]. A similar distribution characterizes the additional long-term precipitation stations (not shown).

3. Keying errors in DSI-3206

Ideally, in the keying process for DSI-3206, only data that were not already included in DSI-3200/DSI-3205 would have been keyed. In practice, data for some stations for some periods were rekeyed for DSI-3206, resulting in an inadvertent overlap between the old and new digitized datasets. The number of keying errors in DSI-3206 may be estimated by examining the inadvertent overlap between the digitized data that are newly available in DSI-3206 and the digitized data that were previously available in DSI-3200/DSI-3205, where values appear in both datasets for the same day (here keying error rates will refer to errors in values, not errors in individual digits). Because of the focus of DSI-3205 on the Midwest (plus New Mexico), the degree of overlap between DSI-3206 and DSI-3200/DSI-3205 varies greatly between the Midwest and the rest of the country. Table 1 includes overlap for the Midwest (plus New Mexico), separate from the rest of the country, and expresses overlap as the percentage of all days that appear in both datasets. In general, the overlap is much greater for the Midwest, about 54% for T_{max} and T_{min} and 70% for P. Because DSI-3205 included snowfall and snow depth only for the state of Michigan, the overlap for these element types is much lower, 6% and 13%, respectively. For the rest of the country, the overlap ranges from 2% to 3.5%.

Discrepancies in values for days appearing in both datasets must result from individual keying errors in either dataset or publication errors in *Climatological Data*. Keying errors may be individual keystroke errors or keying errors due to the illegibility of the preserved or published version of the data. The frequency of discrepancies ranges from less than 1% for snowfall to over 8% for T_{max}. Because we do not know of any issues systematically affecting the discrepancy rate for T_{min}, the 2% discrepancy rate is assumed here to be the individual keying error rate in values. The error rate for T_{max} is greatly influenced by shifted values. An analysis indicated that shifting accounts for about two-thirds of the discrepancies and, thus, the keying error rate is less than 3%, or similar to T_{min}.

A number of the discrepancies were examined to try to determine their source. In practice, this process involved a person (subjectively) comparing the two keyed values with the microfiche of the original form. The majority of individual values with discrepancies were hard to read on the microfiche due to legibility issues, which were usually a combination of poor handwriting and/or poor microfiche reproduction of the original records. Due to this legibility issue, in at least 10% of the cases, on examination of the microfiche, the expert climatologist could not read the value clearly enough to say which of the two keyed values was correct. The discrepancy rates shown in Table 1, thus, represent upper limits on the keying error rates in individual values in both digital datasets.

For T_{max} and T_{min}, the great majority (approximately 88%) of the discrepancies, and, therefore, individual keying errors, are 5.6°C (10°F) or less. (All original units in the database are English, and, for this study, were entered and manipulated exclusively in those units.) Due to their small magnitude, they are undetectable by the quality control tests described in the next section. Approximately 6% of the discrepancies, affecting about 0.1% of all temperature values, are 11.1°C (20°F) or more. Approximately 6% of the precipitation discrepancies, affecting less than 0.1% of all precipitation values, are more than 1 in. These larger errors are generally detectable by the quality control tests described in the next section.

4. Quality control process

The primary purpose of the quality control for this project was to identify the largest errors in individual values, particularly those that might affect analyses of extreme temperature and heavy precipitation events. Automated procedures were used to identify unusual values ("outliers"). Outliers were then examined by trained climatologists to assess their validity. A basic

TABLE 1. Size of overlap in digitized data between TD3206 and TD3200/TD3205, and the rate of discrepancies within the overlapping data. The rates are given for the Midwest and for the rest of the country. The Midwest includes the nine states of Illinois, Indiana, Iowa, Kentucky, Minnesota, Michigan, Missouri, Ohio, Wisconsin, and also New Mexico, which was included in TD3205.

	No. of overlapping values		Percent overlap		Discrepancies in overlap	
Element	Midwest	Rest of country	Midwest	Rest of country	Midwest	Rest of country
Max temp	5 556 000	791 000	54.3%	2.6%	8.1%	6.8%
Min temp	5 554 000	791 000	54.2%	2.6%	2.0%	2.2%
Precipitation	10 238 000	868 000	69.5%	2.0%	0.9%	0.8%

set of procedures was applied to data for all precipitation stations and for all temperature stations with at least 3 yr of data. A more detailed set of procedures was applied to long-term stations, defined earlier as those with less than 10% missing data for the period of 1895–2000. These stations will be heavily utilized to study climate trends, thus, warranting a greater allocation of quality control resources.

a. Basic procedures

The basic procedures identified the most extreme values in the dataset using either absolute thresholds or thresholds based on the station's own climatology. For precipitation, any value in the database that exceeded 254 mm (10 in.) was flagged as an outlier (accumulated values were not excluded from being flagged, but the support tools, specifically, a list of values on the days prior to the flagged day, that were available to the assessor provided the necessary information for recognition of the possibility of an accumulated value). This test was performed in order to identify (and flag as invalid) obvious erroneous values for all stations, not just those with long records. For $T_{\max}$ and $T_{\min}$, a daily value T_i was flagged as an outlier if its standardized anomaly from the monthly mean exceeded 5.0 in absolute value, that is,

$$\left|\frac{T_i - T_m}{\sigma_m}\right| > 5.0, \tag{1}$$

where m is the month, T_m is the monthly mean $T_{\max}$ or $T_{\min}$, and σ_m is the standard deviation of daily $T_{\max}$ and $T_{\min}$ for the month. As noted above, the temperature tests were applied only to stations with at least 3 yr of data. The threshold of 5 in Eq. (1) was determined empirically; as indicated in section 5a, the percentage of invalid values for anomalies less than 5 was quite low, and the decision was made to not expend limited validation resources on such values. The calculation of σ_m was performed using all values, including possible invalid ones. A more precise approach would be an iteration in which the process is repeated by recalculating σ_m after the initial set of invalid values is removed from the dataset, thereby adding additional outliers (because σ_m would be lower). Although this iteration was not performed, it is unlikely to have a major impact, again, because the percentage of invalid values for anomalies around 5 is quite low.

b. Procedures applied to long-term temperature stations

The second set of procedures identified outliers by performing spatial comparisons using nearby stations, along with double checks based on temporal continuity and extremes. Daily gridded fields (2/3° latitude × 1° longitude) of $T_{\max}$ and $T_{\min}$ for the period of 1895–1948 were produced using the objective analysis scheme of Barnes (1964), as modified by Achtemeier (1987, 1989). For each station, each daily temperature value T_i was compared with an estimate E_i from the corresponding gridded field using a bilinear interpolation from the four nearest grid points. A daily difference D_i was calculated as

$$D_i = (E_i - E_m) - (T_i - T_m), \tag{2}$$

where E_m is the monthly mean of the gridded estimates interpolated to the station location and T_m is the monthly mean of the station temperatures. Next, 12 cumulative distribution functions, one for each month, were generated from the set of D_i values. An example is shown in Fig. 4 for the month of December for Grand Marais, Michigan. Respectively, $D_{0.01}$ and $D_{0.99}$ are the difference limits for the fractional cumulative frequency values of 0.01 and 0.99, based on approximately 1050 observations. Because there often was a lack of symmetry between the positive and negative sides of the distribution, the mean of the two difference limits, defined as

$$D_{\text{mean}} = (D_{0.01} + D_{0.99}/2), \tag{3}$$

was used in the following metric.

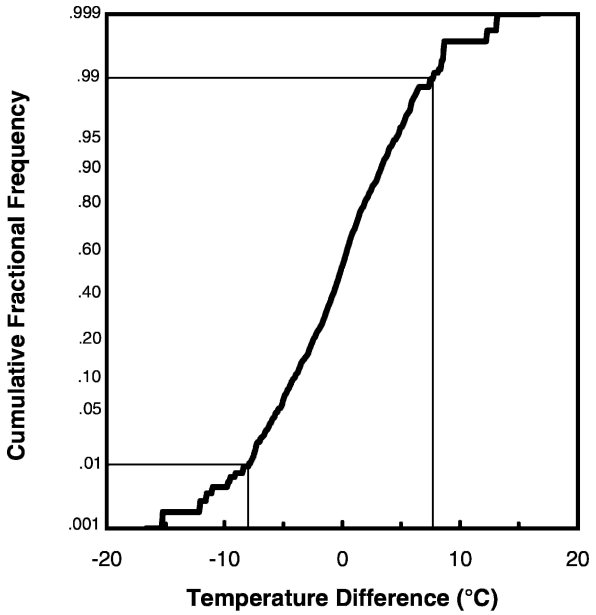

FIG. 4. Cumulative frequency (expressed as a fraction) of difference between station temperature anomalies and estimated temperature anomalies from the gridded data for Grand Marais, MI, in Dec. The light lines show the temperature difference values at cumulative frequencies of 0.01 and 0.99.

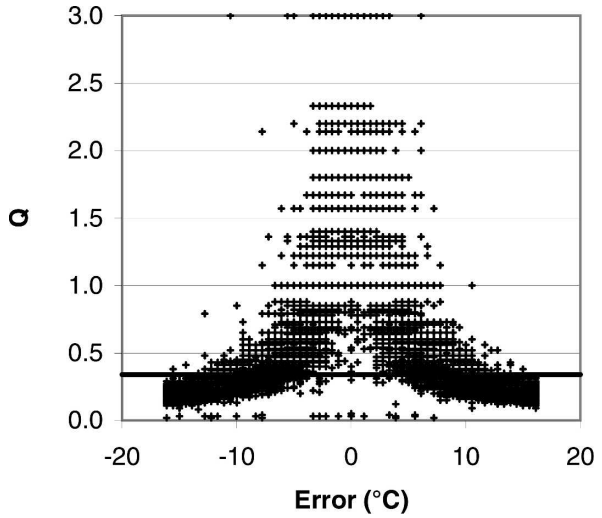

FIG. 5. The Q value vs error distribution for the daily maximum temperature values for the station at Urbana, IL. The random errors were uniformly distributed over the range from −16.7°C (−30°F) to +16.7°C (+30°F), and were applied to each daily maximum temperature within the period of 1896–1948. For each error value, Q was calculated using the original climatology of the station; Q ranges from 0 to infinity with lower values being more extreme. For this project, all outliers with $Q \leq 0.34$ (horizontal line) were manually validated.

For all values D_i, a "quality index" Q_i was calculated to rank the values in order of likely validity. This index was defined as

$$Q_i = (D_x - D_{\text{mean}})/(D_i - D_{\text{mean}}), \tag{4}$$

where D_x is $D_{0.01}$ if the denominator is negative, or $D_{0.99}$ if the denominator is positive, such that Q_i is always greater than 0. Values with lower Q are more extreme and are more likely to be invalid. Values with higher Q are likely to be valid, including a Q of infinity, when $D_i = D_{\text{mean}}$.

To illustrate the magnitude of the potential errors in the outliers as defined by the Q value, a random number generator was used to simulate "observer errors" in the daily values for the station at Urbana, Illinois. The random errors were uniformly distributed over the range from −16.7°C (−30°F) to +16.7°C (+30°F), and these values were added to each daily T_{max} and T_{min} in the entire period of record. For each error value, Q was calculated using the original climatology of the station. The Q value versus error distribution is shown in Fig. 5 for all T_{max} within the period of 1896–1948. For large errors of magnitude 11.1°C (20°F) or more, Q is low—less than 0.5. For small errors with magnitudes of less than 2.8°C (5°F), Q is high—greater than 0.5. The Q value versus error distribution is similar for T_{min} (not shown).

For this project, all outliers with Q less than or equal to 0.34 were manually evaluated; this cutoff was empirically determined by the accumulated number of values that could be manually assessed with the available resources. As indicated in Fig. 5 (horizontal line) for Urbana, this Q value cutoff will include almost all (99%) of the errors with magnitudes of greater than 11.1°C (20°F), as well as a significant portion (66%) of those with magnitudes between 5.6°C (10°F) and 11.1°C (20°F). For stations with a lower correlation between the daily station data and the daily grid estimates, such as those in the mountainous west, the range in Q of the errors with magnitudes of greater than 11.1°C (20°F) is greater, so the Q value cutoff of 0.34 used here results in a greater percentage of those large errors being excluded from the manual verification process.

c. *Procedures applied to long-term precipitation stations*

A similar methodology, using gridded estimates, was tested for daily precipitation. However, an initial test indicated that there were many valid precipitation values for which the calculated Q values were very low, thus, requiring much unnecessary manual assessment. This was due to the high spatial variability of precipitation during convective situations. An alternate method was developed that proved to be superior at selectively identifying invalid values. For each station, a set of nearest-neighbor stations was identified based on geographical distance. All nonzero daily values were ranked from lowest to highest. Outlier values were defined as those exceeding the 95th percentile threshold and were subjected to further tests to identify those values that were most likely to be invalid.

For each outlying value P_i, two indicators of Q were calculated. The first indicator incorporated the actual daily precipitation amounts as follows:

$$Q_{\text{amt}}(i, n) = P_n/P_i, \tag{5}$$

where Q_{amt} (i, n) is the Q indicator, using precipitation amounts for day i (P_i) and nearest-neighbor station n, and P_n is the precipitation amount for station n. The second indicator was calculated from a daily percentile rank as follows:

$$Q_{\text{per}}(i, n) = (100 - R_i)/(100 - R_n), \tag{6}$$

where Q_{per} (i, n) is the Q indicator using precipitation percentiles, and R_n and R_i are the monthly percentile ranks for the nearest neighbor and the station being evaluated, respectively. The monthly percentiles were

obtained by ranking all nonzero precipitation values for the month.

The final value of Q, Q_i, for P_i is the maximum individual Q value among the set of Q_{amt} and Q_{per} values. The key aspect of the procedure is that a high Q value will be calculated if any single nearest-neighbor station has a precipitation value that is seasonably high. Values with very low Q only occur when no nearby station has a high precipitation value. Our tests indicated that this procedure was effective at selecting invalid values and maximizing use of personnel resources for manual assessment.

d. Manual assessment

The application of the manual assessment was developed by having a group of experienced climatologists that was familiar with observational data independently examine a small subset of outliers (50–100), and then discuss the differences in any individual assessments until the group was in agreement as to the application of the quality flags to be applied to the outliers. This was an iterative process that included the development of procedures for calculating Q. The general consensus of the group was that the manual assessment would give the observations the benefit of the doubt, that is, an outlier was assumed to be valid if there was at least one piece of confirming evidence. In addition, written guidelines were developed to assist the assessors; these are given in the appendix. Each outlier was assessed and assigned one of four flags described as follows.

Valid: there is some confirming evidence. (Usually, this evidence consisted of similar values at one or more nearby stations; or, the observed spatial pattern of values was recognized as a commonly occurring one for the region and time of year.)

Plausible: there may be no nearby stations with similar values, but the assessor recognizes that such a pattern has occurred in the past at the location and time of year.

Questionable: the assessor judges that the observed pattern is not a regularly occurring one and the value is unlikely to be valid, but cannot discount the physical possibility of the observed pattern.

Invalid: the assessor judges that the observed value is outside a physically possible range or that the observed spatial pattern is not likely to be physically possible.

To aid in these assessments, several tools were available to the assessor. A Web site was developed to provide for the simultaneous display of these tools. The Web site provided the advantage of allowing access by geographically distributed assessors. Also, the assessors' flags and comments were automatically recorded by the Web site, so that postprocessing did not require any further digitization of the information. The tools provided on the Web site were as follows.

1) A table displaying 15 days of data centered on the outlier day for that station. (All data available in that time period, including precipitation, snowfall, and snow depth were printed in the table.)
2) A time series graph, containing T_{max} and T_{min} extremes, as well as accumulated precipitation and accumulated liquid snow equivalent values for the period from 60 days before to 60 days after the day of the value being assessed.
3) Maps displaying the value being assessed and the values of up to 50 nearby stations for the day of the outlier and a day on either side of the outlier to encompass different times of observations and possible shifted observations.
4) A map displaying the neighboring and outlier station elevation and their estimated time of observation only for those nearest neighbors that were plotted in the above map. (The outlier station was highlighted in bold.)
5) Maps displaying the difference between the plotted values and the climatological mean for the month in which the day falls, for the outlier day and 1 day on either side.
6) Maps displaying the normalized daily anomaly of the value being assessed and the neighboring stations for the outlier day and 1 day on either side.
7) Table of precipitation and plots of temperatures for the outlier station's nine nearest neighbors. (The temperature plots included 15-day time series plots centered on the day in question.)
8) Web links to state topographic maps and historical daily weather maps, which opened into a new browser window.

5. Summary of validations

a. Invalid rates

The basic temperature test was applied to over 82 million daily T_{max} and T_{min}. A total of 4380 values were identified that exceeded the limit on the standardized anomaly [see Eq. (1)]. The results of the manual assessment (Fig. 6) show a clear and expected relationship to the magnitude of the standardized anomaly. For standardized anomalies of greater than 7, more than 80% of the 153 values were judged to be invalid. This percentage drops to about 20% for the 642 values in the 5.0–5.5 category. As the magnitude of the standardized anomaly category decreases, the number of values in

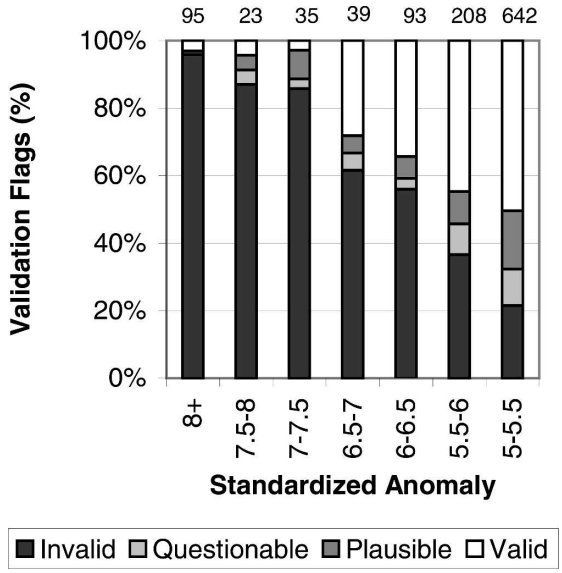

FIG. 6. The percentage of manual outlier assessments in each category (valid, plausible, questionable, and invalid) as a function of the standardized deviation for temperature outliers. The total number of outliers in each standard deviation bin is shown at the top of the bar.

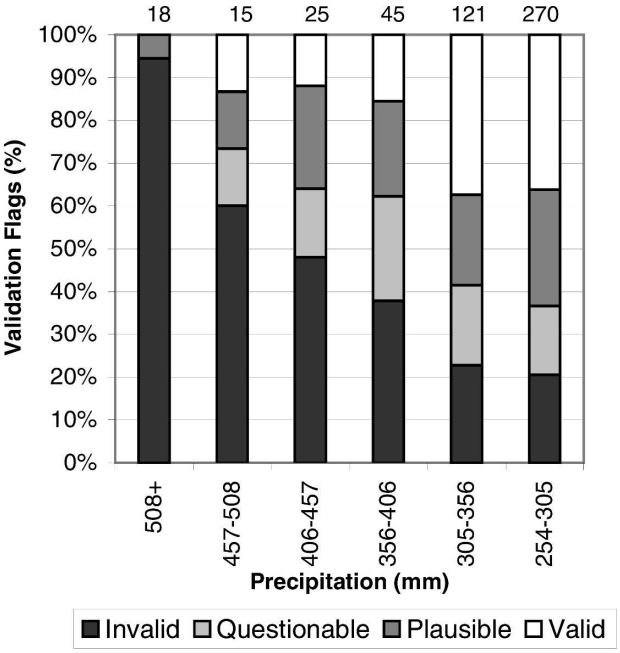

FIG. 7. The percentage of manual outlier assessments in each category (valid, plausible, questionable, and invalid) as a function of amount for precipitation outliers. The total number of outliers in each precipitation bin is shown at the top of the bar.

the fixed-width category increases, from 95 in the 8+ category to 642 in the 5.0–5.5 category. For this test, an invalid rate of 20% or less for the 5.0–5.5 category represents a point of diminishing returns beyond which the cost in time increases significantly, in part, because of the much larger number of candidates to process, and, in part, because the values are much more plausible, and, thus, less likely to obviously be wrong, and thereby require more attention from the assessor.

The basic precipitation test—flagging values greater than or equal to 254 mm (10 in.)—was applied to over 29 million nonzero precipitation values. A total of 498 precipitation values exceeded 254 mm and were manually assessed. The results of the manual assessment (Fig. 7) indicate that the percentage of invalid values decreased with decreasing amount, from about 95% for the 18 values greater than 508 mm (20 in.) to about 20% for the 270 values in the 254–305-mm (10–12 in.) category. As was the case for the basic temperature test, the manual validation of outliers flagged from the basic precipitation test was not applied beyond the fixed-width category, giving an invalid rate of about 20%.

The spatial tests were applied to over 28 million daily T_{max} and T_{min} values for 884 long-term temperature stations. A total of 7390 values with Q less than or equal to 0.34 were manually assessed (0.03% of the values tested). The results of the manual assessment (Fig. 8) indicate that the percentage of invalid values decreased with increasing Q, from 100% for $Q < 0.10$ to about 70% for the 0.30–0.34 category. The next category, 0.35–0.40, includes 13 248 values. The validation of this category would have required a significant portion of the resources that were available for the project. There

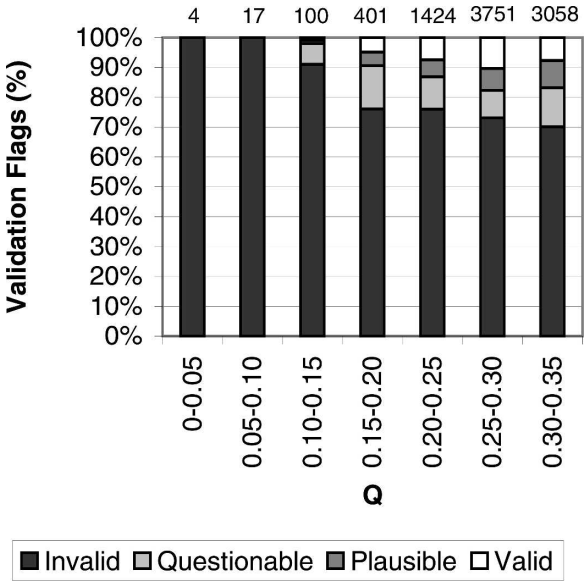

FIG. 8. The percentage of manual outlier assessments in each category (valid, plausible, questionable, and invalid) as a function of Q for temperature outliers. The total number of outliers in each Q bin is shown at the top of the bar.

are a very large number of values (973 036) in the Q range of 0.35–1.0. Thus, there remain many invalid values in the dataset that could not be examined because of resource limitations. However, the values in this category will be less extreme, as indicated by their higher Q value, and, thus, will have a lesser impact on the extreme event focus of this study.

Two additional algorithms to identify T_{max} and T_{min} outliers that are more likely to be invalid were tested on the outliers in the Q range of 0.35–1.0. The first was an extremes test, with the monthly cutoff limits of 1% and 99% generated from each station's climatology. The second was a temporal (spike) test, with the monthly cutoff limits of 5% (on each end of the distribution) also generated from each station's climatology. These two tests are described in Kunkel et al. (1998). There were 7083 (of the 973 036) values flagged by the extremes test, and 472 flagged by the spike test. Resources were not available to validate all of these values, so a portion of them—those with lowest Q—were validated. Of the values flagged by the extremes test 1209 were validated, and 72 of the values were flagged by the spike test. Of these validated values, well over 80% were assessed as invalid, which is a higher rate than that suggested by the assessment of the outliers identified by Q alone. This suggests that, if more, but limited, resources were available for continued assessment, a combination of tests for identifying outliers would be helpful.

For the nearest-neighbor tests applied to the 1044 long-term precipitation stations, a total of 8459 values with Q less than or equal to 0.50 were manually assessed. The results of the manual assessment (Fig. 9) indicate that the percentage of invalid values decreased with increasing Q, from roughly 40% to less than 10% at a Q value of 0.50. Precipitation outliers were more difficult to assess due to the greater spatial and temporal variability of precipitation and, as a result, many of the values were flagged as being plausible or questionable. The results of the manual assessment (Fig. 9) include the interesting feature that a greater portion of the lower Q outliers was assessed as being valid than for higher Q outliers. Valid outliers with lower Q included those along coastlines, which could be influenced by tropical systems, and those in mountainous regions, where orographic lift was the primary forcing mechanism for precipitation. The higher Q outliers included values associated with summertime convection, and were not necessarily limited to coastal stations. For these events, the typical spatial distribution of the convection made it very difficult to declare an outlier as being valid; rather, these outliers were much more likely to be assessed as being plausible.

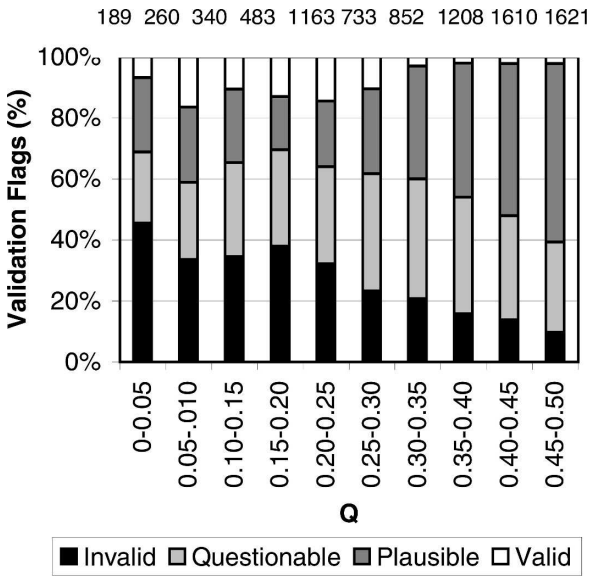

FIG. 9. The percentage of manual outlier assessments in each category (valid, plausible, questionable, and invalid) as a function of Q for precipitation outliers. The total number of outliers in each Q bin is shown at the top of the bar.

b. Spatial and temporal distribution of outliers generated by spatial tests

The distribution of manually assessed outliers among nine geographic regions of the United States is relatively uniform, both in the percent of outliers flagged from the available long-term stations for each region and in the proportion of each validation code assigned to the outliers (Table 2). The percentage of T_{max} and T_{min} values flagged as being outliers is higher in the eastern two-thirds of the country and is lower in the western third. The percent of precipitation values flagged as being outliers and tested is higher in the south and lower in the north. Within each of the three element types, the proportion of each validation code assigned to the outliers is relatively consistent among the regions. A somewhat larger number of minimum than maximum temperatures were assessed as being invalid, which implies either that the Q test was more effective at identifying invalid minimum temperatures, or that the maximum temperatures were more difficult to assess due to factors such as the observation time, or both.

Over the period for which outliers were generated, 1896–1948, the number of outliers per year decreased significantly for all three element types (Figs. 10a–c). One relevant point is that the outliers were not assessed in chronological order, but in inverse order of their Q values, so that the assessor was continually skipping around in time; thus, the trend in Fig. 10 is not caused

TABLE 2. Distribution of outliers from Q tests and assessments by climate region.

	Region	No. of outliers	Percent of tested values	Valid	Plausible	Questionable	Invalid
			Maximum temperature				
1	Northeast	363	0.031%	8%	7%	10%	75%
2	East–north–central	1360	0.042%	14%	7%	12%	68%
3	Central	957	0.031%	16%	10%	10%	64%
4	Southeast	358	0.023%	12%	11%	9%	68%
5	West–north–central	547	0.030%	7%	9%	13%	71%
6	South	724	0.037%	11%	12%	12%	65%
7	Southwest	201	0.026%	9%	9%	13%	69%
8	Northwest	194	0.026%	2%	7%	14%	77%
9	West	110	0.022%	6%	14%	9%	71%
			Minimum temperature				
1	Northeast	334	0.028%	4%	7%	10%	78%
2	East–north–central	1033	0.032%	7%	3%	9%	81%
3	Central	795	0.026%	6%	5%	10%	79%
4	Southeast	657	0.043%	9%	8%	12%	70%
5	West–north–central	463	0.026%	4%	6%	12%	79%
6	South	595	0.030%	4%	8%	16%	72%
7	Southwest	123	0.016%	6%	5%	7%	82%
8	Northwest	140	0.019%	4%	6%	8%	82%
9	West	84	0.017%	5%	7%	20%	68%
			Precipitation				
1	Northeast	635	0.16%	4%	35%	36%	25%
2	East–north–central	1309	0.16%	10%	28%	27%	34%
3	Central	1640	0.18%	9%	42%	27%	23%
4	Southeast	1560	0.35%	7%	40%	36%	17%
5	West–north–central	1037	0.28%	6%	36%	35%	23%
6	South	1713	0.39%	9%	42%	38%	10%
7	Southwest	531	0.45%	2%	49%	30%	18%
8	Northwest	312	0.14%	5%	40%	30%	25%
9	West	246	0.33%	9%	28%	33%	30%

by a change in the experience of the assessor with time. The relative proportion of outliers assessed as being valid/invalid, that is, the effectiveness of the Q test and validation process, did not change over this period. The density of stations that are available more than doubled over this period, which may affect the accuracy of the grid used in the objective application of the Q test. For the manual assessment, the higher density of stations provides more pieces of data for the validator to identify an outlier as being valid, which, if all else were equal, should change the effectiveness of the Q test and validation process as a whole. That the effectiveness did not change suggests that there were both more outliers and more invalid values in the early portion of the record.

The distribution of the manually assessed outliers over the year is relatively constant for T_{max} (Fig. 11a), while for T_{min} more outliers were flagged in the summer than in the winter. For both T_{max} and T_{min}, a small number of days includes a large number of outliers (spikes in Figs. 11a and 11b). When manually assessed, these outliers are generally found to be other than invalid, and are probably related to unique situations found with frontal passages. For precipitation (Fig. 11c), the distribution of the manually assessed outliers shows a large peak in the summer in the total number of outliers, as well as the number of outliers manually assessed as being plausible and questionable. The number of outliers assessed as valid or invalid was highest in the winter. The greater total number of outliers for precipitation in the summer is related to the greater spatial variability of precipitation from convective storms, which also contributes to greater difficulty in the manual assessment of outliers as either being obviously valid or invalid.

c. *Keying errors in validated values*

A set of 108 outliers were verified with the microfiche of the original documents to check the rate of keying errors within the outliers. At the same time, an additional 324 values on adjoining days were also

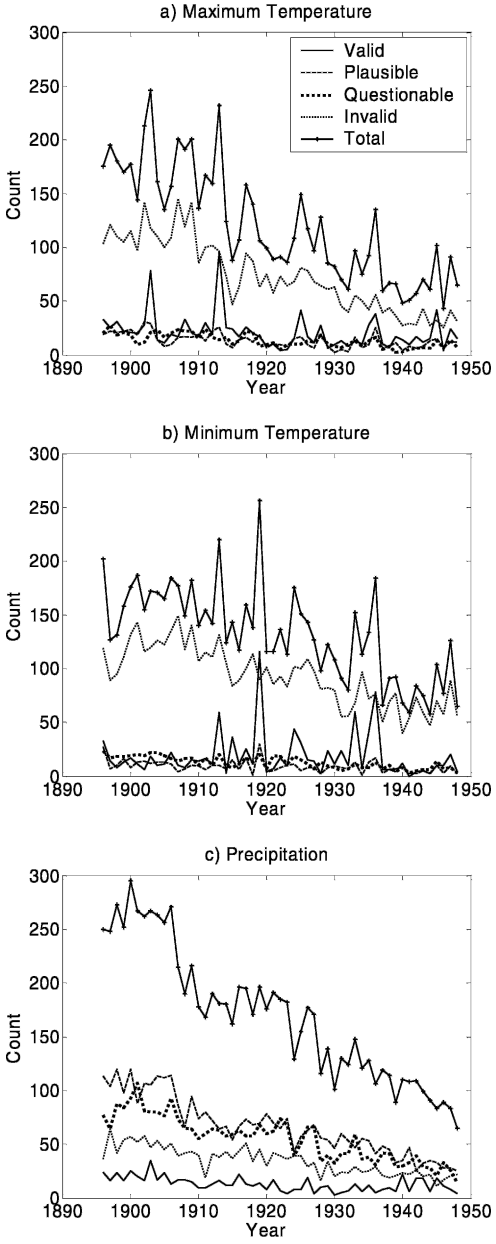

FIG. 10. The number of manually assessed outliers by year for each element type, including the number of each of the four validation codes assigned to the outliers by the assessors.

verified. Of the 31 maximum temperature outliers verified, two (6%) were found to have been keying errors; of the 77 minimum temperature outliers verified, five (6%) were found to have been keying errors. Of the 139 other maximum temperature values verified, one (less than 1%) was found to have been a keying error; and of the 185 other minimum temperatures verified, two (1%) were found to have been keying errors. Given the small sample verified here, the 95% confidence limits on the keying error rate in the outliers is approximately 1%–11%, and the rate in the adjoining values is 0%–2%. The magnitudes of the keying errors in flagged values were relatively high, generally 5.6°C (10°F) or 11.1°C (20°F), while the magnitudes of the errors in the adjoining values were smaller, 5.6°C (10°F) or less. That the keying error rate is a relatively small proportion of the outliers (<11% at the 95% level of confidence) suggests that the primary explanation for invalid outlier values is from other sources, for example, observer error.

d. Consistency of validations among assessors

The consistency of the assignment of the validation flags to the outliers during the validation process was checked by a blind test for both temperature and precipitation. Two assessors were given the same set of 100 randomly selected outliers within a much larger batch, so that they were not aware of which outliers were being used for the check. The distribution of validation flags that was assigned by the two assessors is shown in Table 3. For temperature, 73 of the 100 outliers were given exactly the same flag, and an additional 23 were different by one category. Of the outliers with the greatest category difference in flags assigned, a number were associated with stations in situations where there was some question on the timing of a frontal passage intertwined with uncertainties in the observing time. For precipitation, 50 of the 100 outliers were given exactly the same flag, and an additional 42 were different by one category. This consistency check suggests that, on average, the validation flags on the temperature outliers may be different by at least one category 27% of the time, and different by two or more categories 4% of the time; the validation flags on the precipitation outliers may be different by at least one category 50% of the time, and different by two or more categories 8% of the time. These differences among assessors are a result of both the subjective nature of the validation process and the experience of the assessors built up over the course of this project. In an attempt to facilitate the validation process, and to minimize these gross differences, a general list of guidelines was drawn up and distributed to each assessor. This list provided insightful clues as well as tips and hints from the experience of other assessors. Other informal comparisons of consistency among the assessors suggests that the greatest consistency may be produced among assessors located in the same physical office, such that they may continue to "train" each other informally as they discuss which flag to assign to outliers in curious or unusual climatic situations.

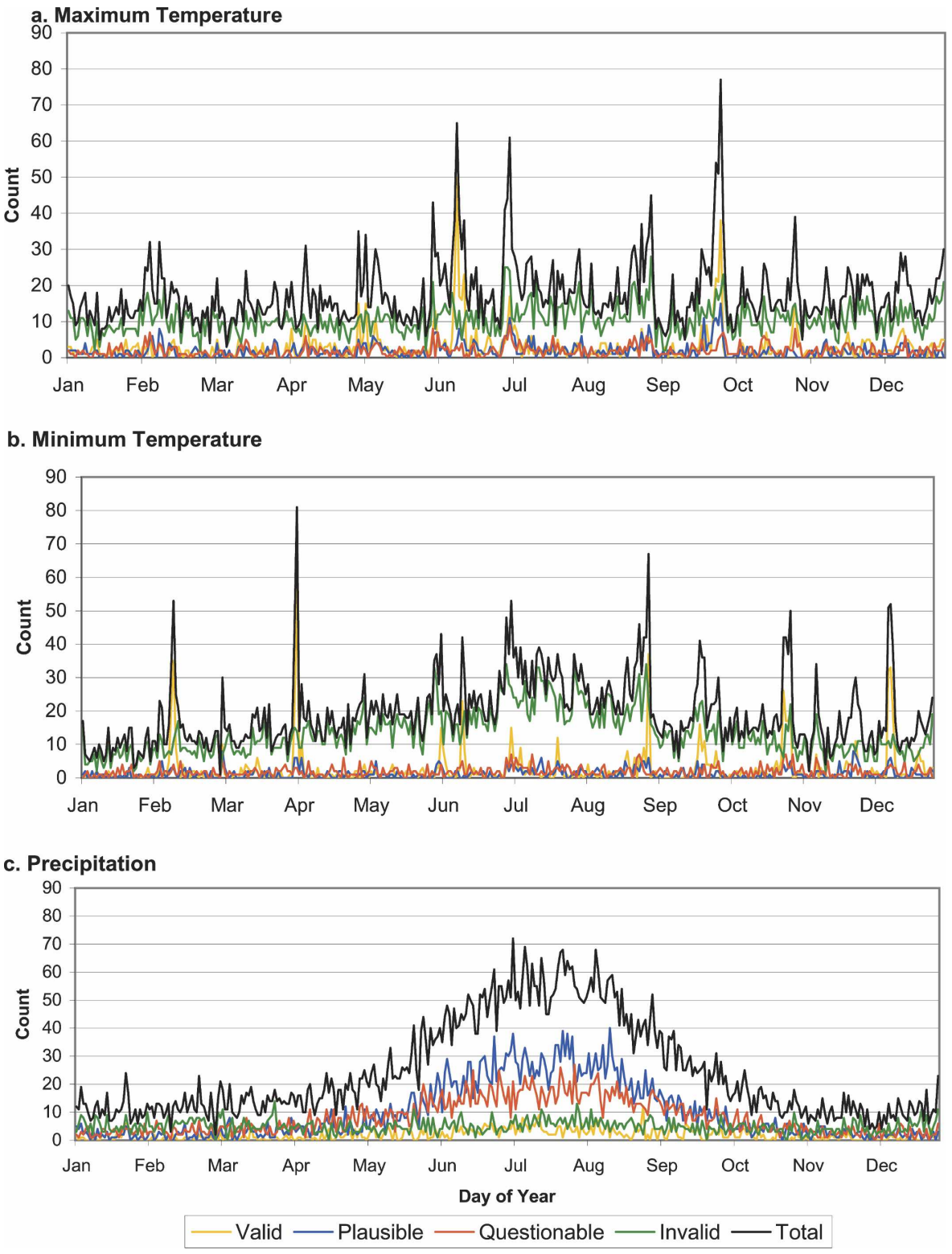

FIG. 11. The number of manually assessed outliers by day of the year for each element type, including the number of each of the four validation codes assigned to the outliers by the assessors.

TABLE 3. Comparison of validation codes assigned by two assessors to the same set of 100 temperature outliers and 100 precipitation outliers. The 100 sample outliers of each type were randomly selected out of a much larger set, and were not marked so that the assessors were not aware of which outliers were being compared. The validation codes are V: valid, P: plausible, Q: questionable, and I: invalid.

		Temperature comparison			
		Validator A			
		V	P	Q	I
Validator B	V	1		1	
	P	2	1	1	
	Q		3	3	8
	I	1	2	9	68
		Precipitation comparison			
Validator B	V				
	P	5	25	16	
	Q	4	9	17	3
	I	2	2	9	8

e. Validations and extreme values

The general effect of the QC process on extreme values was examined for daily values exceeding the threshold of a return period of 0.5 yr (or an average of two events per year). For temperature, there were 884 long-term stations to which the Q test was applied for the period of 1896–1948, resulting in 100 000 extreme temperature values each for T_{max} and T_{min} (extreme highs for the maximum temperature, extreme lows for the minimum temperature). Of these 200 000 extreme temperature values, only 323 (0.2%) were flagged as being outliers. By comparison, in the entire dataset of long-term stations, 0.03% of the values were flagged as being outliers. Of the 323 outliers that were also extreme values, over 70% were assessed as being invalid by the assessors (see Table 4). This rate of invalid values is somewhat higher than that for all of the outliers. Therefore, for the definition of extreme values used here, while the vast majority of the extreme values passed the automated QC, the extreme values were more likely to be flagged as being outliers, and also are somewhat more likely to be assessed as being invalid.

TABLE 4. Distribution (%) of the validation codes assigned to the assessed outliers for each of the three element types—precipitation and maximum and minimum temperature—for all outliers validated and for 2 days yr^{-1} extreme values that are also outliers.

	Precipitation		Max temperature		Min temperature	
Code	All	Extremes	All	Extremes	All	Extremes
V	8%	4%	17%	6%	14%	5%
P	39%	42%	11%	9%	7%	8%
Q	33%	34%	11%	12%	9%	10%
I	20%	20%	61%	73%	70%	77%
Count	8, 82	4091	6415	156	7065	129

For precipitation, there were 1044 long-term stations to which the nearest-neighbor test was applied for the period 1896–1948, resulting in just over 110 000 extreme precipitation values for the return rate of 2 days yr^{-1}. The great majority (96%) of these extreme values passed the objective nearest-neighbor test, with 4091 (4%) of these extreme values being flagged as outliers. Of these 4091 outliers that were also extreme values, a smaller proportion of them (4%) were assessed as being valid, as compared to all of the assessed precipitation outliers (8%) (see Table 4).

6. General conclusions

The newly keyed pre-1948 data represent a major enhancement to the COOP dataset, which is widely used for analysis of climate variability and change. The quality control applied in this project increases its value by eliminating a sizeable number of errors in individual values. A total of 10 671 temperature and precipitation values (or 48% of the 22 462 outliers) were assessed to be invalid. Analysis of the effectiveness of the objective spatial test, developed to identify values with the greatest potential for error, suggests that many of these invalid temperature outliers were in error of magnitudes of 20°F or more.

Invariably, QC efforts are constrained by resources, and this project was no exception. A set of metrics was defined by trial and error and used to select the "worst" outliers for the resource-intensive manual assessment. In this case, worst meant values that were either climatologically highly unusual or not spatially consistent with neighboring stations. These metrics were used to rank values, and manual assessment proceeded according to the ranks.

This was a team effort involving experts from several institutions. Corporately, the team included a high level of expertise in all major climate regimes of the United States. The project was accomplished via spatially distributed participation, with all tools and data located on a single Web site. In addition, written guidelines (appendix) were developed to assist the assessors. These steps ensured that all participants followed the same ground rules and allowed everyone access, if needed, to the results of the assessments of all of the other participants. Early in the project, this promoted a rapid consensus building on the ground rules of the manual

assessment. As the project progressed, this also allowed for the ongoing tracking of progress and balancing of the workload among groups. This approach was a key to the project's timely completion and could serve as a model for other QC efforts.

The results for temperature outliers from the spatial tests indicate that further manual assessment of values with higher Q would likely result in a substantial number of additional invalid values. Further manual assessment of values with higher Q would also increase the effectiveness of the Q test at capturing outliers with the greatest potential magnitude errors.

For temperature, the effectiveness of the spatial tests using the Q test could be improved by more accurate representation of the annual cycle beyond the use of the climatological monthly mean. It could also be improved by the use of a more refined gridding scheme that provided higher correlations between the actual daily value and the grid estimate, especially in the mountainous regions of the West.

The manual assessment of the outliers was undertaken to avoid automatically removing valid values from the dataset. In general, the outliers assessed as being invalid are most likely so, although because the manual assessment was, by its nature, subjective, and prone to a certain level of human inconsistency or error, as shown by the comparison between assessors, a small fraction of them may be valid.

This dataset is available from NCDC. All values are assigned one of the flags described in section 4d. No value was actually removed from the dataset, including any that were flagged as being invalid. Thus, future users of this dataset are able to perform their own assessments, if desired. However, because of the conservative nature of the assessment performed here, it is recommended that for standard applications, users assign a value of "missing" to any values flagged as being invalid, and perhaps to those flagged as being questionable.

Acknowledgments. We thank Ned Guttman for his advice on the development of methods. This material is based upon work supported by the Office of Global Programs, National Oceanic and Atmospheric Administration (NOAA) under Award NA16GP1498. Additional support was provided by NOAA Cooperative Agreement NA17RJ1222. Partial support for David Easterling was provided by the Office of Biological and Environmental Research, U.S. Department of Energy (DOE). Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of NOAA or the Illinois State Water Survey.

APPENDIX

DSI-3206 Validation Guidelines

Following is the list of validation guidelines for dataset DSI-3206.

1) Is the station an A.M. or P.M. observing site?
2) Does the temperature pattern map reflect mostly A.M. or P.M. stations?
3) Are there any stations with the same observation time and, if so, do their data support the measurement in question?
   (a) Do the measurement values agree?
   (b) Are the deviations relative to the monthly mean comparable?
   (c) Are the standardized anomalies relative to the monthly mean comparable?
4) Are there significant elevation differences that may justify the outlier value?
5) Is it conceivable that unrecorded physical phenomena, such as a downburst, an inversion, or convection influenced the local environment? Is there evidence of precipitation during the period that might signal evaporational cooling? Have possible local mesoscale effects, including sea and lake breezes, downslope warming, and strong radiational cooling over a deep snowpack, been considered?
6) Is there any observable pattern in the time series plot (such as a large number of missing values or a step discontinuity) that may indicate a problem with the thermometer or a change in the instrumentation used at the station?
7) Enter a comment justifying the validation, for example, the observer may have reversed the digits and the value should likely be XY instead of YX.
8) Choose the validation flag that best describes the situation:
   V: The data meshes with surrounding values is attributable to reasonable physical phenomena, or is consistent with the station's climatology.
   P: There is less confidence that the datum is valid, but the value is still physically possible.
   Q: There is little confidence that the datum is valid, but it is not physically impossible, given the available information.
   I: The datum is physically impossible and is completely inconsistent with the surrounding stations.

Occasionally, assessors used additional information to put the observation in context, such as NCDC's Daily Weather Map Series, topographic maps of the region, various time series of station data, and station location and siting information (e.g., if the station lo-

cated along a river bank, in a forest, or near a highly urbanized setting).

REFERENCES

Achtemeier, G. L., 1987: On the concept of varying influence radii for a successive corrections objective analysis. *Mon. Wea. Rev.,* **115,** 1760–1772.

——, 1989: Modification of a successive corrections objective analysis for improved derivative calculations. *Mon. Wea. Rev.,* **117,** 78–86.

Barnes, S. L., 1964: A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteor.,* **3,** 396–409.

CDMP, 2001: Annual report. 8 pp. [Available online at http://www.ncdc.noaa.gov/oa/climate/cdmp/files/annualreport2001.pdf.]

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.

National Climatic Data Center, 2003: Data documentation for Data Set 3206 (DSI-3206) COOP Summary of the Day—CDMP—Pre 1948. 18 pp. [Available online at http://www4.ncdc.noaa.gov/ol/documentlibrary/datasets.html.]