Complex Quality Assurance of Historical Hourly Surface Airways Meteorological Data

DANIEL Y. GRAYBEAL, ARTHUR T. DEGAETANO, AND KEITH L. EGGLESTON

Northeast Regional Climate Center, Cornell University, Ithaca, New York

(Manuscript received 14 October 2003, in final form 10 February 2004)

ABSTRACT

A new complex quality assurance (QA) procedure is developed for historical hourly surface airways meteorological data, recently digitized in a U.S. government–sponsored effort that extends the digital period of record back as early as the late 1920s. The procedure builds upon an existing, three-pronged framework and incorporates several checks previously employed for hourly data. New or enhanced checks were also developed, partly with the goal of accommodating the peculiarities of historical observation practices or reporting standards, but also to improve our ability to flag suspicious data. The latter goal is sought by use of more elaborate physics or empirical relationships than are employed in existing checks, or by modifying or replacing them to sharpen their intended focus.

Each of the three prongs checks for consistency, whether for internal or temporal consistency or against limits. Six internal consistency and three temporal consistency checks are developed or adapted from recent research and implemented as components in complex QA. These involve temperature, humidity, pressure, wind, present weather, and cloud cover from synoptic reports, as well as station elevation from metadata. In order to determine, from the many component-check flags activated, which element(s) are most likely responsible for the flag(s), a decision tree was developed. Resultant flag levels ''suspect'' and ''erroneous'' reflect the degree of confidence.

Flag rates averaged 9.5% overall but leveled off at about 2% after 1937 (0.3% for ''erroneous'' flags), comparing favorably with previously published flag rates. Systematic errors raise many ''suspect'' flags, suggesting that hybrid QA–inhomogeneity tests are needed.

1. Introduction

In the last 15–20 years, complex quality assurance (QA) procedures for meteorological data have been evolving in both scope and application. Complex QA differs from traditional QA by providing not only a greater number or complexity of consistency checks but also a decision tree to resolve the component flags into flags on particular data that are most likely suspect, given the weight of the evidence (Gandin 1988; Eskridge et al. 1995). Among the first complex QA systems to be developed was for radiosonde data (Gandin 1988), for which complex QA continues to be most highly developed (Lanzante 1996; Gaffen et al. 2000; Collins 2001; Lanzante et al. 2003). In recent years, however, complex QA has been extended to daily (Reek et al. 1992; Kunkel et al. 1998) and hourly (Meek and Hatfield 1994; DeGaetano 1997; Shafer et al. 2000) surface meteorological data. For hourly meteorological data, Meek and Hatfield (1994) developed a three-pronged framework, building on O'Brien and Keefer (1985), for implementing basic component checks and a simple decision-making process weighting flags by hierarchical flag type. Gandin (1988) and DeGaetano (1997) illustrate the potential for developing sophisticated component checks. Shafer et al. (2000) demonstrate a more advanced decision-making procedure and its application to real-time data from a mesoscale network, as well as show how an automated QA can function as one of several steps in a larger QA protocol, such as involving manual instrument inspections.

Beginning in 2001, the Climate Database Modernization Program (CDMP) has been digitizing historical meteorological observations from original paper forms (Fig. 1). As a result, for surface airways observations (SAOs), the period of record has been extended backward two decades, from the previous limit of 1948 into the late 1920s. Thus, QA procedures for hourly data were reviewed, with the goal of updating existing techniques and making them amenable to processing historical data. Quality assurance procedures for the hourly SAOs that had been used at the National Climatic Data Center (NCDC) may be characterized as either simple and traditional, as for the TD-3280 dataset, or more complex, as for the more recently developed TD-3505 (Integrated Surface Hourly) dataset (NCDC 2001). Some component checks needed refining, for example, a blip check instead of a step check (Graybeal et al. 2002), while others needed to be added, for example, a

Corresponding author address: Dr. Daniel Y. Graybeal, Northeast Regional Climate Center, Department of Earth and Atmospheric Sciences, 1123 Bradfield Hall, Cornell University, Ithaca, NY 14853.
E-mail: dyg2@cornell.edu

© 2004 American Meteorological Society

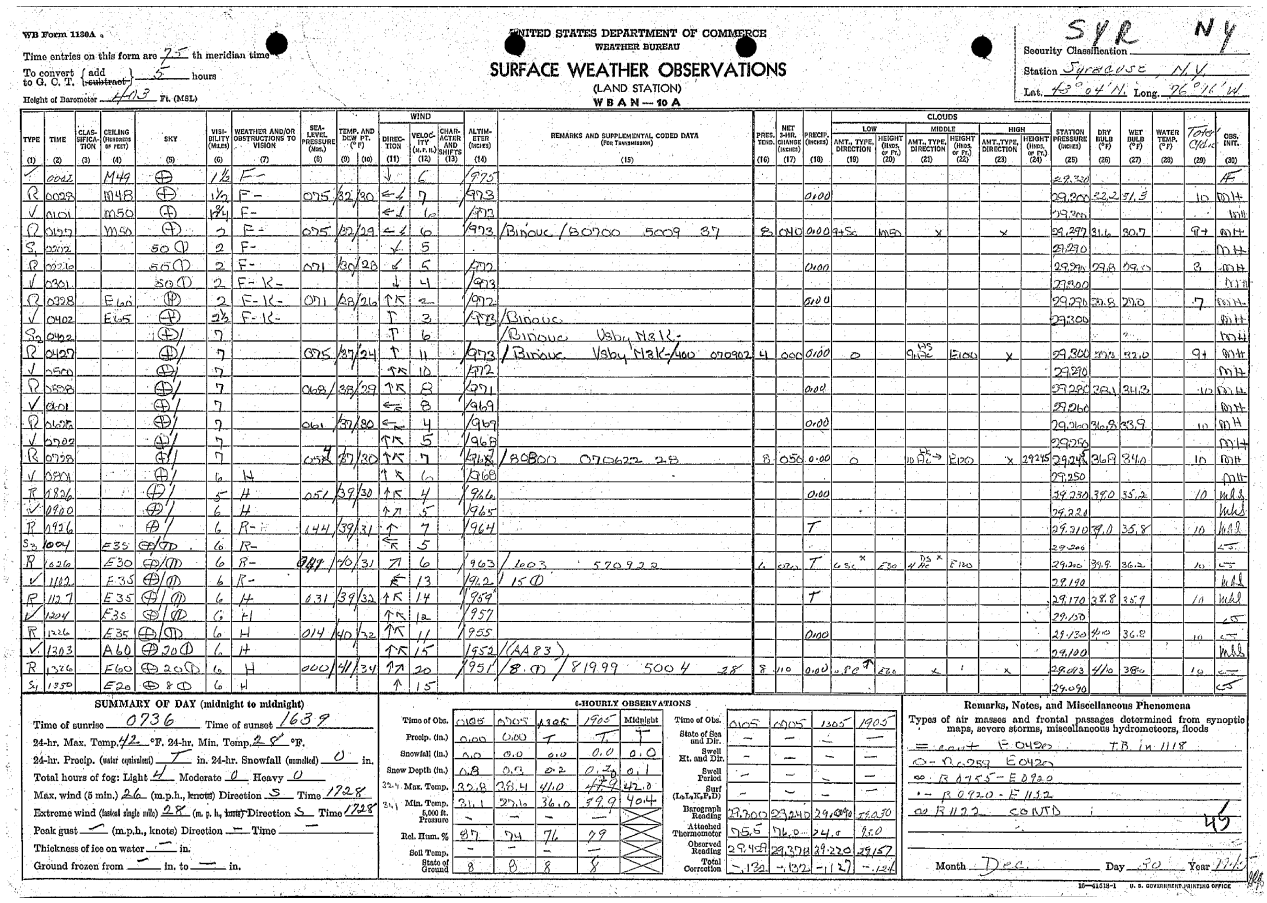

FIG. 1. Scanned image of an original paper weather observer form, containing SAOs from Syracuse, NY, on 30 Dec 1945.

physical cross check among temperature, humidity, and station pressure using redundant humidity information. The potential for varying some check thresholds spatially or seasonally, as opposed to treating them statically, was recognized. Component checks for temperature and humidity were rudimentary, and those for wind needed to incorporate DeGaetano's (1997) more advanced checks, for which no suitable updates existed. Furthermore, as practices for measuring SAO data elements have changed over the course of the century, some checks that may have been useful on those data may be obsolete in the Automated Surface Observing System (ASOS) era. Therefore, it was decided to develop a complex QA for historical SAO data that would improve on existing techniques by developing new or improved component checks and by expanding the decision tree to incorporate these new checks.

2. Data preparation

Quality assurance is distinguished here from quality control (QC), which is taken to include data replacement or elimination. In QA, data are instead merely flagged, retaining data that may yet contain useful information, and leaving the ultimate QC decision to the user. A principal partner in this project is a general climatic data center that must satisfy a very broad spectrum of users. It employs a mix of QA and QC in its products and services, and the database format allows extensive support for either, such as by preserving the original data in the event of an edited replacement or deletion.

Prior to making the data available for QA, an extensive preliminary "cleaning" was necessary. Each SAO record is keyed by a U.S. Weather Bureau, Air Force, and Navy (WBAN) station identifier and the time–date stamp, given to the nearest hour and sometimes to the minute. Ensuring that each database key is a unique, valid, and ordered combination is the main purpose of the cleaning. Principal problems are summarized in Table 1; for a detailed exposition, see Guttman (2002). In processing records with duplicate time–date stamps, an automated cleaning procedure followed a first-found–first-kept rule, placing the others in a discard pile for manual inspection and correction. Of more than 43 000 discarded records, more than 90% have been inspected manually, then reindexed and reinserted where needed.

For development of models and algorithms employed to perform checks for QA components, hourly data from

TABLE 1. Common or principal problems encountered in the digital database produced from historical SAO paper forms.

Principal database problems requiring cleaning prior to QA
1. Records with missing or invalid key elements
2. Records out of chronological order
3. Records improperly formatted
4. Records incorrectly indexed
5. Records with duplicate keys
Common observer or keyer mistakes that may be flagged in QA
1. Keyer misreads handwritten digits (e.g., 6 is mistaken for 0, 1 for 7, etc.)
2. Observer misrecords element (e.g., dewpoint depression instead of dewpoint)
3. Observer or keyer accidentally multiplies by 10 (or 0.1)
4. Observer or keyer accidentally adds or subtracts 100 or integer multiples of 10
5. Observer or keyer transposed digits (e.g., 53 entered as 35)
6. Observer or keyer omits or adds a negative sign
7. Observer or keyer records data in incorrect column on form or in database

1949–63 and 25 stations approximately evenly distributed around the contiguous United States were used, unless noted otherwise. Data from that period were available prior to availability of the CDMP datasets and contain hourly observations with relatively consistent practices. The 25 stations were chosen by CDMP as high in priority for processing, prior to the start of the present work. Independent data were always used to test the models and algorithms; until the CDMP data became available in large volume, the development and test periods were 1949–58 and 1959–63, respectively. Afterward, the fuller 1949–63 period was used for development and the pre-1949 data for testing. Table 2 compares the spatial density of the CDMP with those of existing digital hourly station networks. So, as the CDMP data gradually became available, the number of stations and records available for sampling in testing grew large, relative to the pool drawn from in development. Original data were recorded in traditional units (e.g., °F for temperature), except sea level pressure (sometimes in mb); QA checks were developed in the same units but are presented here in SI.

3. Component checks

Three components are represented in the complex QA system (Fig. 2). These are checks for consistency with limits (LC), checks for internal consistency (IC), and checks for temporal consistency (TC). These three prongs are common in many recent QA procedures (Gandin 1988; Meek and Hatfield 1994; DeGaetano 1997). A fourth type of check, for spatial consistency, is also commonly employed, as in the dense Oklahoma Mesonetwork (Shafer et al. 2000). However, first-order station density in the early twentieth century is so low as to bring into question the feasibility of a spatial check. Spatial checks for hourly data are implemented by Miller and Benjamin (1992) but on a then-current network of average interstation distance of only about 80 km; by comparison, the CDMP network averages nearly 260 km between stations. Quality assurance of radiosonde data, also from course spatial station density, has also long employed spatial checks (Gandin 1988), but free-atmosphere data are generally smoother than surface data. Whether spatial checks for sparse-resolution SAOs need augmentation from the denser U.S. Cooperative Observer Network, also recently becoming available (Kunkel et al. 1998), is a subject for future research, not developed in this paper.

a. Limits consistency checks

Physical limits and station-based climatological limits are employed for elements as given in Table 3. It should

TABLE 2. Comparison of counts of hourly reporting stations in the CDMP SAO network with those of later (1949–2002) U.S. hourly reporting networks, excluding Alaska and Hawaii. Years in parentheses indicate the year when the minimum/maximum occurred. (Source: National Climatic Data Center.)

Network	CDMP[a]	WMO–NWS/FAA[b]	NWS/FAA[b]	All hourly[c]
Period covered	1928–48	1949–2002	1949–2002	1949–2002
Total available	199	270	729	2815
Minimum yearly	(1928) 6	(1949) 43	(1949) 161	(1949) 1184
Maximum yearly	(1946) 135	(1988) 241	(1992) 525	(2001) 1486
Average yearly	104	178	408	1269
Median yearly	116	204	444	1238
Average coverage (km^2 $station^{-1}$)	67 200	39 200	17 100	5500
Average interstation distance (km)	259	198	131	74

[a] Includes two stations from Alaska (Eagle and Sitka) and one from Hawaii (Honolulu).
[b] Stations counted with a National Weather Service (NWS) or Federal Aviation Administration (FAA) identifier must also have a Cooperative Observer Network identifier.
[c] Stations have at least a WBAN identifier.

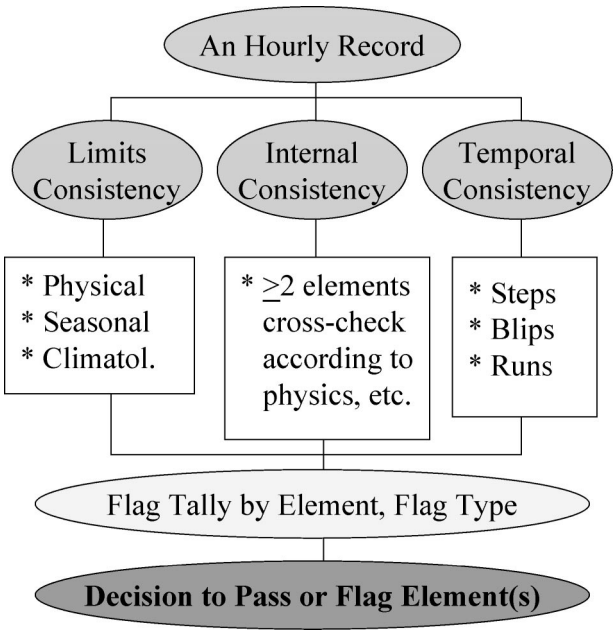

FIG. 2. Diagram of the complex QA operational and decision-making process.

be noted that the TD-3280 data format of the NCDC requires storage of wind speed and direction as one element. Thus, one check that was originally classified as for IC, comparison of wind speed and direction, has been reclassified as an LC; it ensures that both speed and direction are either 0 or non-0 (0 meaning calm for direction).

For each station, climatological limits were determined by ranking all hourly values in the 1948–present period of record and then subjecting the 10 lowest and highest (approximately the 0.002 and 99.998 percentiles in a 50-yr record, respectively) to manual inspection by a meteorologically trained technician, in order to screen for outliers. For suspicious cases, the full synoptic report was examined to support acceptance or rejection. For stations with limited periods of record for hourly data (e.g., fewer than 20 yr), published climatological extremes were consulted instead, for elements for which these were available. In most cases, the values so determined were reasonable. However, in the few cases when all the top or bottom 10 are outliers, limits were set according to more general theoretical or observational extremes (e.g., statewide). If the top 10 are outliers, chances are the next top 10 are as well, so this course of action was chosen arbitrarily and to manage technician resources.

Physical limits were set to accommodate instrument or reporting constraints. For example, the ceiling- and cloud-layer-height bounds (Table 3) are due to the limitations of the ceiliometer then in use (U.S. Weather Bureau 1935). In the case of invalid codes or alphanumeric characters found, the element in question was set to ''missing,'' rather than triggering a flag for violating limits.

TABLE 3. Weather elements subject to checks on physical and climatological limits, including specific general physical and station-example climatological limits.

Physical limits	Climatological limits
Wind speed	Dry-bulb temperature
0–112 m s^{-1} (250 mph)	Boise, ID: −31.6°–43.9°C
0 if direction code is calm	
Wind direction	Wet-bulb temperature
Valid alphanumeric characters or codes	St. Louis, MO: −28.3°–29.4°C
Code for calm if speed is zero	
Cloud-layer height	Dewpoint temperature
0–2972 m (9750 ft)	Albany, NY: −38.3°–26.1°C
Ceiling height	Station pressure
0–2972 m (9750 ft)	Denver, CO: 805.0–848.3 mb
Horizontal visibility	Sea level pressure
0–160 km (100 mi)	Bismarck, ND: 974.3–1056.9 mb
	Altimeter
	Macon, GA: 997.7–1037.0 mb

Since climatological limits were determined using data from the second half of the twentieth century, extremes in an earlier period may exceed by a few degrees the limits so determined. Therefore, an arbitrary tolerance of 2.8°C (5°F) was extended about the envelope provided by the predetermined limits to weed out such extremes and focus on large departures from range expectation, such as miskeying by a factor of 10 or 100. The tolerance about pressures was likewise set to 3.4 mb (0.10 in. of Hg.).

b. *Internal consistency checks*

1) TEMPERATURE AND HUMIDITY

The physical relationship $T_{dry} \geq T_{wet} \geq T_{dew}$ (in which T is temperature and the subscripts refer to dry bulb, wet bulb, and dewpoint, respectively) is evaluated by three IC checks that operate in bivariate comparison. These ensure that $T_{dry} \geq T_{wet}$, $T_{wet} \geq T_{dew}$, and $T_{dry} \geq T_{dew}$. The relationship was broken down in this manner so that if two or more checks are violated additional information as to the degree of suspicion warranted for each of the elements involved may be given. For example, suppose, a valid T_{dry} of 25°C is mistranscribed as −25°C; this value would most likely fall below both T_{wet} and T_{dew} and trigger two flags as violating $T_{dry} \geq T_{wet}$ and $T_{dry} \geq T_{dew}$. From those two flags, the weight of the evidence suggests that T_{dry}, rather than T_{wet} or T_{dew}, is the most suspect of the three elements involved.

2) TEMPERATURE, HUMIDITY, AND STATION PRESSURE

If redundant humidity information is available, along with station pressure, a more elaborate physical model

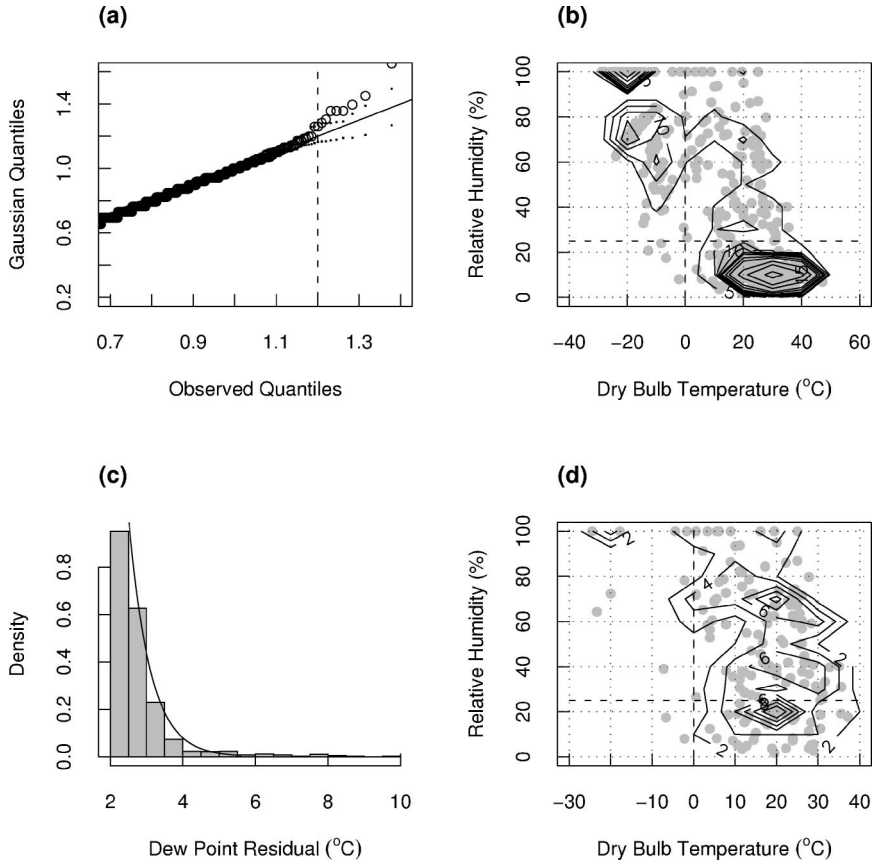

FIG. 3. (a) Upper-tail close-up of a Gaussian Q–Q plot of 0.33-power-transformed residuals (distance of observed dewpoint from wet–dry tolerance envelope about the estimate; open circles), with 95% confidence intervals (dots) and outlier threshold (dashed line); (b) contoured scatterplot of residuals above 2.2°C, by dry bulb and RH, with 0°C and 25% thresholds (dashed lines); (c) trimmed histogram of points in (b) ≤0°C or ≤25% RH, with an exponential fit; (d) as in (b) but for residuals outside the tiered thresholds.

may be employed to crosscheck these four elements, estimating one given the other three. For simplicity, dewpoint was selected as the target in all cases. Derivation of the check, including details of the particular approximations used, is beyond the scope of this paper. However, the check is simply an application of a thermodynamics exercise (Rogers and Yau 1989).

A flag is activated if the dewpoint estimate and the observation disagree beyond a tolerance envelope determined by both theoretical and empirical considerations. This tolerance is necessary because of the errors in observation, recording, and physical estimation. A first tolerance window is constructed from dry and wet estimates, recognizing principally the effects of errors due to observing and rounding. The dry estimate is obtained by adding 0.3°C (0.5°F, one-half the minimum resolution in the original units available for development and testing) to the dry bulb and subtracting the same from the wet bulb, increasing the depression; the wet estimate is obtained by reversing the directions of the respective adjustments, decreasing the depression. A second tolerance extension is by 4.4°C (8°F) if the dry bulb is 0°C or lower, or the relative humidity is 25% or lower; otherwise, an extension of 2.2°C (4°F) is used. These tolerance extensions were determined by inspecting histograms of residuals from a random sample of real data, as described below.

Figure 3a shows how the threshold of 2.2°C about the wet–dry tolerance envelope is determined, as the point at which the residuals from a random sample ($n = 1000$) begin to depart significantly from normality, following transformation of the residuals by raising them to the 0.33 power (Barnett and Lewis 1994). In Fig. 3a, the outlier threshold is shown as approximately 1.2 (1.7°C, following an inverse transform). The 99.9% confidence interval from the normal distribution fit to the subset of the residuals below 1.7°C is upper bounded at approximately 2.2°C. Residuals above that 2.2°C threshold are clustered into two main groups, as shown in Fig. 3b, one below 0°C (41.3% of the sample) and the other below 25% RH (47.5%). The two groups are meaningful physically, as the estimation of dewpoint from observed dry- and wet-bulb readings is subject to increasingly larger errors under cold or dry conditions

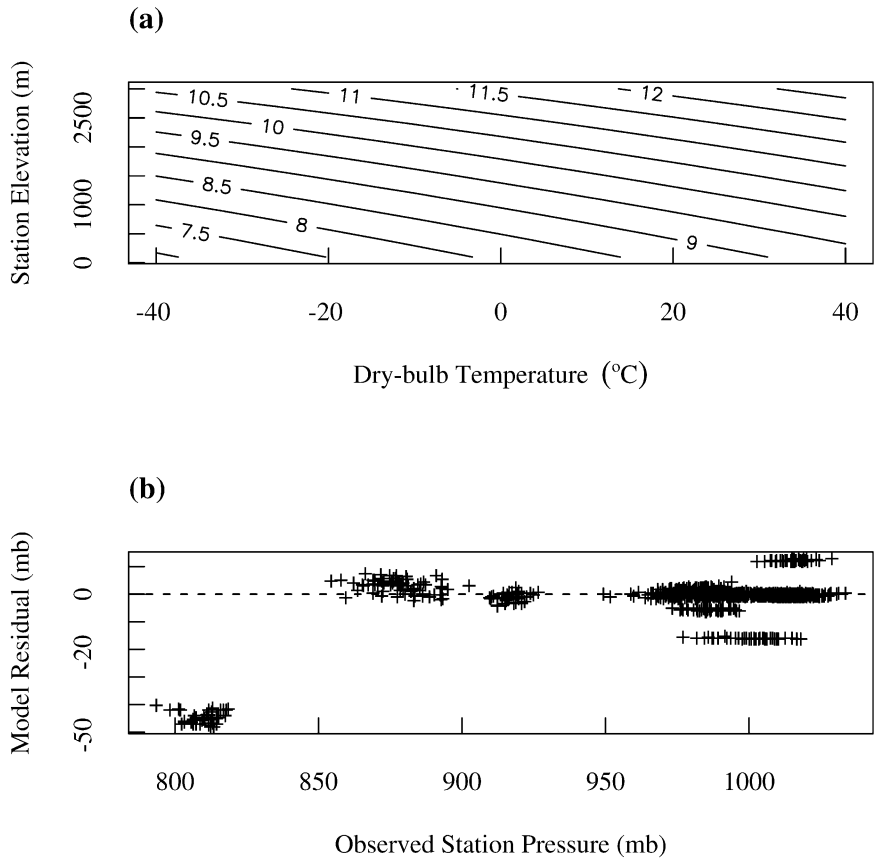

FIG. 4. (a) The maximum perturbation in station elevation necessary to cause the estimated station pressure to vary by up to 1.0 mb at a given combination of station elevation and dry-bulb temperature (and a sea level pressure of 1000 mb assumed). (b) Residuals between observed and estimated station pressure found in a random sample (n = 1000) from hourly observations, with perfect prediction shown as a dashed line.

(WMO 1954). An additional allowance in the QA tolerance under those conditions is, thus, sensible. The 4.4°C extended tier was chosen so as to lower the expected flag rate by one order of magnitude, as shown in Fig. 3c, a histogram of residuals when dry bulb is 0°C or less or RH is 25% or less.

Figure 3d shows a contoured scatterplot of residuals falling outside the two-tiered threshold (2.2°C if dry bulb exceeds 0°C and RH exceeds 25%, 4.4°C otherwise). In Fig. 3c, 91.3% of the points, deemed flaggable data, lie above 0°C, while the main peak in terms of humidity is still below 25% RH, with secondary peaks near 70% and 100% RH. This pattern suggests a dewpoint or wet-bulb depression substitution error, or a random error, such as miscoding by a factor of 10, that may produce a similar effect. The top 10 residuals were inspected manually, of which 8 were found erroneous, according to common digitizer mistakes (Graybeal et al. 2002). In some other, more subtle cases, dewpoint depression substitution was found as the cause for flagging.

3) PRESSURE, TEMPERATURE, AND STATION ELEVATION

Sometimes both station and sea level pressure are simultaneously available, read, and computed from the station barometer. In that case, a relationship of station and sea level pressures with dry-bulb temperature and station elevation may be used, derived from the hydrostatic equation and equations of state (Rogers and Yau 1989). As above, to account for errors in measurement, recording, and physical estimation, a tolerance about the estimated station pressure was used in comparing the estimate with the observation; departures greater than this tolerance were flagged. In Fig. 4a, the contoured field indicates the maximum perturbation in station elevation needed to cause the estimate of station pressure to vary by up to 1.0 mb from the estimate obtained by using the station elevation and dry-bulb temperature given in the figure axes. This station elevation perturbation is sensitive to elevation and temperature, as shown. For example, at 500-m elevation and 0°C, the elevation need change by only about 8.5 m to cause the

estimate of station pressure to vary by 1.0 mb from that obtained using the original station elevation. Inspection of an outlier-trimmed histogram of residuals from a random sample of real data (not shown) suggested an initial tolerance envelope of 6.8 mb (0.2 in.) about the estimate, or an elevation perturbation threshold of 57.8 m (190 ft) at 0°C and 500-m elevation.

Occasionally, a station move or a discrepancy in location of different instruments (e.g., barometer on a 100-m rooftop but thermometer at sea level) may cause a long run of this IC flag to occur over many consecutive hourly records. To avoid this systematic flag placing, the tolerance was simply increased arbitrarily to 30.5 mb (0.9 in.). Multiplying the approximate field average of 10 m in Fig. 4a by this threshold gives an average elevation tolerance of approximately 300 m (1000 ft), sufficiently stringent to catch a ± 33.9 mb (1.0 in.) miscoding, fairly common among flagged pressure data; for example, Fig. 4b illustrates a random sample. Treating geographic metadata as time- and element-dependent in QA can afford more discriminatory power in this check. Alternatively, a supplementary check may be designed, more toward the scope of inhomogeneity analysis.

4) Station pressure, altimeter, and elevation

The hydrostatic equation and ideal gas law may be combined to relate station pressure, altimeter, and station elevation (Rogers and Yau 1989). As with the other pressure IC check, a tolerance window was constructed about the station pressure estimate, to allow for errors in measurement, recording, and estimation. Also, as above, practicality necessitated setting a wider tolerance about the estimate than inspection of a residuals histogram may suggest. The same tolerance envelope of 30.5 mb (0.9 in.) was employed.

5) Checks involving present weather

Six IC checks were developed or adapted relating weather to other elements. First, if a freezing precipitation report is found in the present weather field, the temperature must not exceed 4.4°C. Second, if snow or sleet is reported, the temperature should not exceed 7.8°C. Third, if a liquid precipitation report is found, the temperature should be at least -2.8°C. Thresholds set for these checks were established by analyzing histograms of temperature during hours when reports of the weather phenomena of interest were found. The checks, if not the thresholds, were adapted from the list of present weather checks employed for production of the TD-3505 (Integrated Surface Hourly) database (NCDC 2001). Fourth, if blowing or drifting snow is reported, wind speeds of less than 4.1 m s^{-1} are flagged. This check is incorporated directly from the TD-3505 QA routine. Fifth, if visibility is less than 8.0 km and no obstruction to horizontal vision is reported, a flag is activated. This IC check results from the rules for reporting (U.S. Weather Bureau 1935), with an allowance for variations in local practice of about 3.0 km inward. Sixth, if the visibility exceeds 16.0 km, the dewpoint depression is less than 5.5°C, and fog, haze, mist, smoke, or dust is indicated, a flag is activated. The visibility and dewpoint depression thresholds were determined by inspecting histograms during hours reporting some obstruction to visibility.

6) Cloud cover

Internal consistency checks were developed for cloud cover that conformed to the rules that had become established over the early SAOs' period of record (U.S. Weather Bureau 1935), and flags were activated on records that violated these rules. Depending on the format, either two or three cloud layers were allowed in the report, and layer heights may or may not be available. Layer amounts are given in terms of clear (less than 0.1 coverage, hereafter CLR), scattered (0.1–0.5 coverage, hereafter SCT), broken (0.6–0.9 coverage, hereafter BKN), or overcast (greater than 0.9 coverage, hereafter OVC), and they may be further detailed by thin or dark modifiers (except CLR). The observers were directed (U.S. Weather Bureau 1935) to use not more than two layers in their reports, listing the higher one first; if a third, scattered layer is found, they were to indicate the condition in the remarks section. Thus, the only allowable possibilities for two-layer reports were SCT–SCT, BKN–SCT, SCT–BKN, BKN–BKN, OVC–SCT, or OVC–BKN. Three-layer reports were checked to see if their lowest layer was SCT. Other IC checks involve layer heights and amounts and the ceiling height. Ceiling is defined as present if at least one layer amount below 2972 m is at least BKN, and the ceiling height is that of the lowest such layer (U.S. Weather Bureau 1935). Thus, if a ceiling report is expected from the layer amount reports but not found (i.e., height is missing, not coded as unlimited, in both the ceiling and expected layer fields), a flag is activated, as for when the opposite is true.

Sometimes a 0-ceiling height is recorded, in the event of dense fog, heavy precipitation, obstructions to horizontal vision, or very low horizontal visibility. In such cases the ceiling is usually below the reporting resolution (30.5 m). Zero-ceiling reports may also be accompanied by blanks in all layer-amount and height fields, or by a CLR report in one layer. Internal consistency checks for these conditions are performed.

7) Dewpoint depression and diurnal temperature range

During the mid-1930s, a switch in the observation form directed that the dewpoint depression (DPD) be recorded instead of dewpoint. After a few years of this practice, a reversal was instituted and maintained. How-

ever, incorrect substitutions of one element for the other were noted, not uncommonly (and are suspected to contaminate post-1948 data as well). Thus, an IC check was developed to catch violations of consistency between the DPD observed at a given hour and the diurnal temperature range (DTR) observed in the preceding 24 h. It was hypothesized that the greater the DTR, the greater the diurnal maximum DPD; a linear relation in logarithmic space was developed and confirmed the hypothesis. If the DPD at the given hour exceeded by some tolerance the maximum DPD expected from the DTR, a flag is activated. Details of this procedure and its performance are given elsewhere (Graybeal et al. 2004).

c. *Temporal consistency checks*

A TC check involves two components: a check for excess variability (a ''blip'' check) and a check for excess invariability (a ''flat-liner'' or ''runs'' check). Both are implemented here.

1) TEMPERATURE AND HUMIDITY

A pilot study toward developing improved QA checks for these historical hourly data indicated that a large majority of flags activated by a traditional step check were single-hour blips. Thus, an important early drive in improving the QA of hourly data focused on catching these blips, while letting through valid but strong frontal passages, such as were often caught with the traditional step checks. Blip-flagging thresholds were developed for dry-bulb, wet-bulb, and dewpoint temperatures by examining the empirical distributions of blips and applying a target flag rate. Blip-flagging thresholds were allowed to vary by season and location and averaged overall about 3.5°C for dry bulb and about 6.5°C for dewpoint. If a blip threshold was exceeded, a flag was activated only if other aspects of meteorology, normally expected to be associated with large blips in these elements, were not also found in the record and within 1 h of the blip. These include a thunderstorm-related report (including hail), heavy precipitation, or a shower of any kind (DeGaetano 1997). Details of the development and testing of this procedure are given elsewhere (Graybeal et al. 2002).

Temperature and humidity elements were also subject to runs tests, employing a static threshold of 24-h duration, determined from analysis of histograms of run lengths in these elements at select stations.

2) PRESSURE

A blip check for station and sea level pressures and altimeter readings employed a static 3.0-mb flagging threshold, adapted from the jump test used in QA of the NCDC TD-3505 database (NCDC 2001). Based on analysis of histograms, a static runs check was instituted that flagged consecutive runs of constant pressure longer than 12 h in duration.

3) WIND SPEED AND DIRECTION

Temporal consistency checking of wind elements implements the methods outlined by DeGaetano (1997). This procedure utilizes a 24-h moving window approximately centered on the hour being checked (12 windowed hours before and 11 following). Checks for excess variability and invariability are performed on both speed and direction. The excess variability check employs the test statistic

$$D_{\max 2} = \max_i |w_i - w_{i+2}|, \qquad -11 \geq i \geq 10,$$
$$i \in \{-2, 0\},$$

in which w is wind speed or direction and i is the hour of observation, relative to the hour of interest (0). The test statistic is evaluated with respect to 2-h wind differentials, to avoid trapping frontal passages. The value of $D_{\max 2}(i)$ is to be at least as great as the absolute difference of w_i and w_0, for wind speed, and the smallest angular difference, for wind direction. Furthermore, the absolute (smallest) difference of wind speed (direction) must not exceed 15 kt (80°). Besides a check for the absolute difference being an integer multiple of 10 in the prevalent units, a few other, relatively minor checks for excess variability are added, as described in detail by DeGaetano (1997). If these conditions are not met, the hour in question becomes a candidate for flagging. The flag is activated if the reports within 1 h do not also indicate a thunderstorm-related report (including hail), heavy precipitation, or a shower of any kind.

4) PRESENT WEATHER

Although the need for a TC check on present weather may not be intuitive, a runs-type check was developed in order to complete the support structure for the flag-mapping decision tree. An analysis, described below, suggested a consecutive-runs threshold of 12 h. A flag is not activated, however, unless the 24-h median absolute deviation (MAD) of dewpoint also exceeds 9°C. Although this scheme could be considered an IC check for involving two elements, the dewpoint component of this check does not employ an outlier-sensitive method of detection and was developed to support flagging a run of constant present weather code.

The dewpoint MAD cothreshold was developed by finding its maximum in each integer run length in hours. Starting at 2 h and incrementing forward until 10 h less than the maximum run length, successive regressions of the maximum dewpoint MAD on the run length were performed. This process stopped when the slope of the regression line was statistically indistinguishable from 0 at the 95% confidence level. A threshold was established at the run length corresponding to the detection

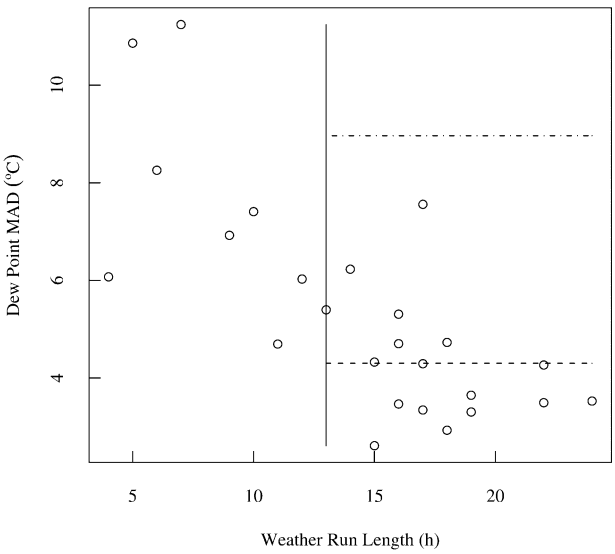

FIG. 5. Station-specific thresholds of present-weather run length and 24-h MAD of dewpoint. The vertical solid line is drawn at 13 h, beginning at which point there is no significant ($\alpha \leq 0.05$) relationship between the run length thresholds and the dewpoint MAD thresholds (dotted line). The upper 99% prediction interval about the 0-slope regression is indicated by the dash–dot line at approximately 9°C.

of an insignificant slope. A related threshold for dewpoint MAD was set as the upper 99% prediction interval about the (0 slope) regression line above the threshold run length. To determine a pair of thresholds for general use, the process was repeated using the threshold pairs from all stations. The two thresholds finally selected are indicated in Fig. 5.

4. Decision tree

Following the cleaning process described above, the database is ready for QA. The first step in QA is for each component check to activate flags as needed during a pass through the chronologically sorted database. For example, suppose a dry-bulb temperature of −5.5°C is mistranscribed to −55°C in the digital database. This one gross error may raise seven flags, as given in Table 4. These flags are stored, along with the station and time-stamp identifier linking it to the original observation, in a component-flag database. The next problem is substantial, which is to map these seven flags to one flag onto the one truly suspect element, dry-bulb temperature. This mapping is required by the format descriptors of both the standard target datasets, the TD-3280 (SAOs) and the TD-3505. Those databases cannot incorporate (nor would their users wish to handle) seven or more flags associated with each element, such as found in the component-flag database produced in the first step of QA. The mapping from component-check flags to individual elements deserving of flags is accomplished through a decision tree. The guiding principle of complex QA underlies this decision, namely, that all available evidence must be considered before flagging a datum under consideration (Gandin 1988; Eskridge et al. 1995).

TABLE 4. Component-check flags raised by a single gross mistranscription of dry-bulb temperature, where LC, IC, and TC represent limits, internal, and temporal consistency, respectively.

Flag type	Flag message
LC	Temperature out of bounds
IC	Wet bulb exceeds dry bulb
IC	Dewpoint exceeds dry bulb
IC	Dewpoint depression inconsistent with diurnal temperature range
IC	Mismatch among dry and wet bulbs, dewpoint, and station pressure
IC	Mismatch among station and sea level pressure, dry bulb, and station elevation
TC	Blip in dry bulb

TABLE 5. Weather elements associated with the seven component-check flags in Table 4.

Element code (TD-3280)	Element name
TMPD	Dry-bulb temperature
TMPW	Wet-bulb temperature
DPTP	Dewpoint temperature
PRES	Station pressure
SLVP	Sea level pressure
ELEV*	Station elevation

* ELEV is not a TD-3280 element code but is rather inserted for convenience and to facilitate future expansion to handle time-dependent metadata. Elevation occupies a mandatory field in the TD-3505 database.

In the language of Holsapple and Whinston (1996), the principal effort in developing this complex QA as a decision support system (DSS) has been in expanding the constituent knowledge system, by improving the ability of the component checks to activate flags suitable to the phenomena toward which they are designed to react. The subject of this section is the problem processing system, the central decision-making subsystem within a DSS. Holsapple and Whinston (1996) describe many different DSSs and expert systems (rule-based DSSs, of which this is an example); the one developed here arose organically and pragmatically, to handle the increased output from the expanded knowledge system's component checks.

A three-tiered flag code is adopted that is compatible with the TD-3505 format and can be accommodated by the TD-3280 format. A datum is declared ''good,'' ''suspect,'' or ''erroneous'' to reflect the degree of concern warranted by the evidence in component-check flags. The first step in the flag translation process is to construct a table of component-check flag counts, indexed by element and flag type. In the previous example, the miscoded −55°C resulted in seven component-check flags affecting, by association, the six elements in Table 5. The component-check flags fall into all three flag

TABLE 6. Example of an element-flag-type table used in mapping component-check flags to element-based flags for inclusion into a formal database, such as TD-3280. Refer to Table 5 for element code descriptions.

Element code	LC flags	IC flags	TC flags	Flag total
TMPD	1	5	1	7
TMPW	0	2	0	2
DPTP	0	3	0	3
PRES	0	2	0	2
SLVP	0	1	0	1
ELEV	0	1	0	1

types. Table 6 shows the component-check flag counts broken down by element and flag type, with a flag total by element. Of all elements, dry bulb is involved in the most flags (all seven).

Layers in the decision-tree hierarchy represent different, and descending, levels of concern about data quality. The top layer, first encountered in the algorithm, represents the strongest concern, the bottom layer last and least. The hierarchy follows the recommendation of Meek and Hatfield (1994). A worst-case scenario, such as the factor-of-10 miscoding described above, would likely raise flags of all three categories. On the other hand, a mild departure from expectation, such as a daily maximum temperature only a few degrees higher than the climatological limit set from a shorter period of record, might raise a flag in only one category. A datum in the former case is most probably erroneous, or in the latter case, merely suspect. For sake of efficiency, the algorithm begins by looking for any elements that violate component checks from more than one class, the highest level of concern. One class, the IC checks, identifies more than one element as potentially suspect, while the other two classes provide independent information that focuses identification on one element at a time. Thus, a combination of flag classes activated on a given element is an accumulation of different kinds of evidence, suggesting that it is highly suspect. If any elements are found in the component-check-type table that satisfy this condition, the procedure then ranks the elements by total component-flag count. Finding a tie, it proceeds by element code in alphabetical order. Beginning with the element with the highest count, the procedure removes all component flags associated with that element and automatically flags it erroneous. This first round is repeated until all such elements have been identified; it is bypassed if none qualify.

In the −55°C miscoding example given above, TMPD is associated with all three component-flag types and caught in the first step of the decision process. It is the only such element in this case, with seven component flags activated involving it. Upon removal of all associated component flags, there are none left to test, so the procedure stops. The other five elements tallied as potentially suspect in this case are dropped from further consideration, leaving TMPD to be flagged erroneous in that record in the final database. To illustrate the value in employing three classes of component flags, suppose that station pressure (PRES) is also miscoded by a factor of 10, such that LC and TC flags on PRES (one in each class) are activated in addition to the two IC checks listed above. The first stage of processing catches TMPD and PRES but proceeds with TMPD first, as its count of seven flags exceeds the count of four associated with PRES. After all TMPD-related flags are removed, two flags still remain, from different classes, that involve PRES. Thus, PRES is also identified as erroneous in the final database. No component flags remain after these two elements are processed, and the procedure terminates.

Sometimes more than one flag will be activated on an element from a given component class. Consider an error in element DPTP, of sufficient magnitude as to raise two component flags: DPTP > TMPW and DPTP > TMPD. Three elements are involved: TMPD, TMPW, and DPTP. However, one element (DPTP) is involved in two flags, both from the same class (IC). This situation falls through the first stage and is caught by the second, which looks for elements with more than one flag from any one component class. As more evidence points to the fallibility of DPTP than to that of either TMPD or TMPW, the element DPTP is more highly suspect than the others. Reaching this stage and finding no other evidence, the procedure affixes the flag erroneous onto the element DPTP on this record in the final database and drops the other two elements from consideration for flagging.

Another example will illustrate the final step in the procedure, should there be a need for further testing after completion of the first two rounds. Suppose that, in addition to the seven component flags activated in the −55°C TMPD example, an IC component flag is activated involving the present weather element (PWTH) and horizontal visibility element (HZVS). Following removal of all TMPD-related component flags, this IC component flag remains. As there is no more evidence to consider, in the form of component flags, and it cannot be determined which of the two is the more suspect, the procedure stops here, keeping both remaining elements PWTH and HZVS flagged suspect.

Occasionally, a tie in counts under the same component class will be found for different elements; this situation is resolved in the second stage of processing. Consider a mild error in the element TMPW, not severe enough to raise LC or TC flags but enough to raise two:

- (IC) dewpoint that exceeds wet bulb; and
- (IC) mismatch among dry bulb, wet bulb, dewpoint, and station pressure.

The elements TMPD, TMPW, DPTP, and PRES are involved, but TMPW and DPTP are each linked to two flags from the same component class. The second stage in the process identifies both elements and finds a tie in their flag counts. Removing all component flags as-

sociated with one leaves no other flags. There is not enough evidence from the flags activated to judge which of the two elements is the more suspicious. Thus, both TMPW and DPTP are flagged suspect in the final database, and TMPD and PRES are dropped from consideration.

5. Operational performance, with discussion

During cleaning, approximately 0.25% of all records received were removed from further consideration, because of problems such as missing or duplicate time stamp. Of more than 1.16×10^7 hourly SAO records processed, approximately 1.11×10^6 (or 9.51%) contained a flag on at least one element. The number of records flagged erroneous is between one and two orders of magnitude lower, at approximately 46 300, or 0.397%. These flag rates are obtained after full processing through the complex QA framework and decision tree.

A flag rate of nearly 10% is quite high, especially for checks developed mainly to catch random errors. Further inspection reveals patterns in the occurrence of high flag rates. Counts of records processed and of records with flags activated were obtained by station years of data, and the flag-rate percentages (count of records with flags divided by count of all records, times 100) were examined by year, to illustrate trends in time. In Fig. 6a, boxplots by year of the $\log_{10}$ percentage rates for all stations and all flags reveals a marked discontinuity about the year 1938. Beginning that year, the median flag rate over all stations per year was fairly consistent, at approximately $10^{0.4}$% [11-yr median 2.29%, mean 3.17%, intraquartile range (IQR) 1.61%–3.39%]. On the other hand, prior to 1938, flag rates were much higher, at generally $10^{1.5}$% (31.6%) and higher for the 1928–31 period and falling approximately linearly through 1937. Flag rates were approximately an order of magnitude lower for flags of erroneous (Fig. 6b; 11-yr median 0.285%, mean 0.434%, IQR 0.171%–0.459%).

In a separate paper (Graybeal et al. 2004), the entire pre-1949 record at Columbus, Ohio, was examined manually, finding long runs of systematic errors, particularly dewpoint having been substituted for its depression, or vice versa. These errors occurred in contiguous blocks during 1935, 1936, and parts of 1937. In Fig. 7a, the effect of these systematic errors on the ($\log_{10}$) time in hours between flags activated is evident at Columbus, at which only 1934 in the pre-1938 period exhibits a median flag-time difference similar to those from 1938–48. By comparison, the mean (median) interflag-time difference for flags of erroneous at Columbus is much higher (Fig. 7b) at 685 h (343 h), or an (acceptable) average flag rate of 0.15%.

Most data from prior to 1932–36 were reported using a different form than later data, through 1948, and were much more likely to be reported on a less regular (e.g., 10 A.M. and 2 P.M.) or frequent (6 hourly) schedule than the later, hourly data. As a result, TC checks were not activated as frequently in the earlier as in the later years, allowing more IC checks to pass the decision tree and enter the database. More significantly, as reporting standards were still somewhat fluid, systematic errors, such as swapping dewpoint depression for dewpoint, were not uncommon during the transition years of the mid-1930s. As many as 2.5% of the flagged records activated during the period 1928–37 may be attributable to an unconventional (by standards that crystalized toward the end of that period) treatment of variable cloud amounts. The forms (e.g., Fig. 1) indicate observer identity,

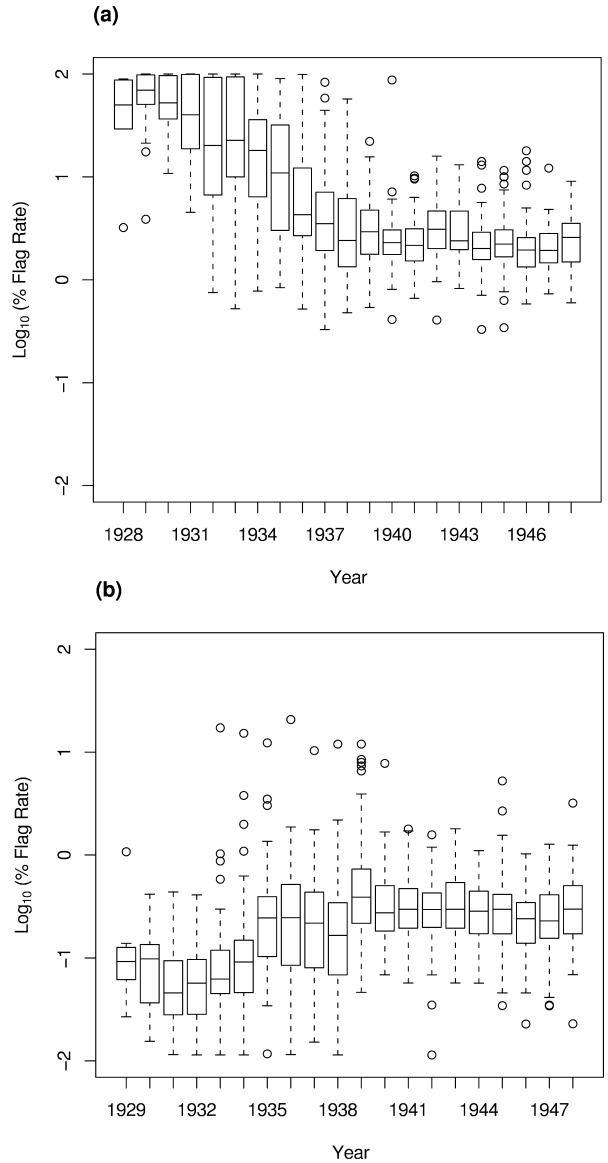

FIG. 6. Boxplots of station-year $\log_{10}$ percent flag rates, by year, for all stations and (a) all flags and (b) erroneous flags. Whiskers extend to the most extreme data not more than 1.5 IQR (given by the box) from the median (given by the horizontal line midbox).

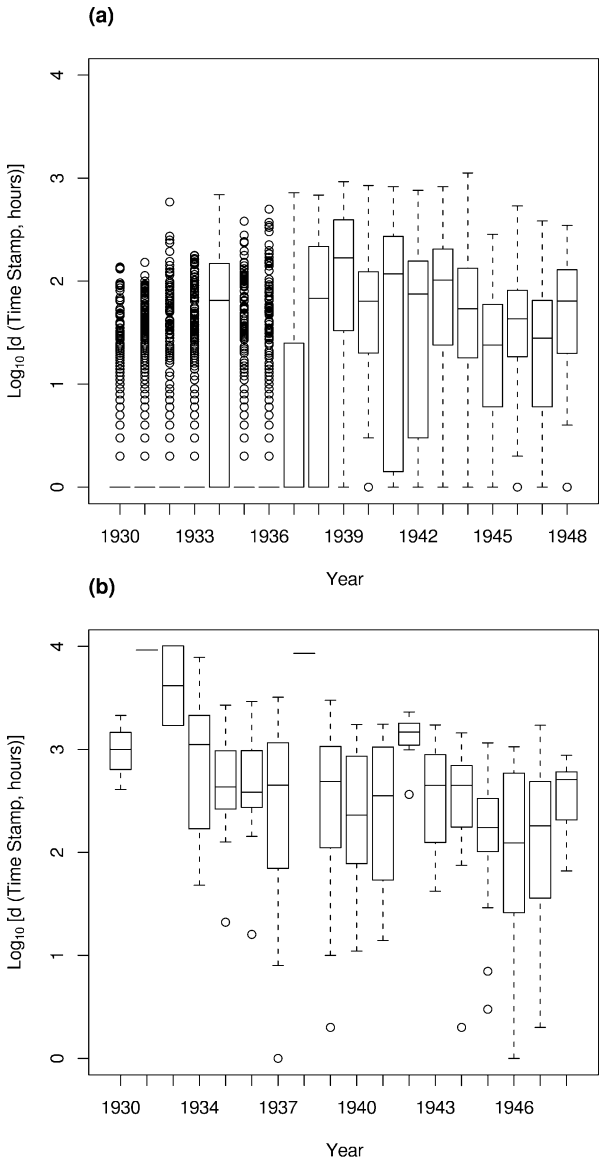

FIG. 7. Boxplots of station-year $\log_{10}$ difference in time (h) between flags activated, by year, for Columbus, OH, and (a) all flags and (b) erroneous flags. Boxplots are as in Fig. 6.

which, unfortunately, is not readily available to aid QA processing; such information may be helpful in detecting other systematic errors.

From narrowing the focus from all flags to flags on a specific element, for example dry-bulb temperature, it is found that approximately 2.32% of all records are flagged, and 0.028% are flagged ''erroneous,'' over all station years of data. For Columbus's 1930–48 record, these percentages are approximately 2.5% and 0.01%, respectively. The flag rates found here are comparable to the 0.2%–1.1% total keying error rates found in the double-keyed comparison of the TD-3200 and Midwest Climate Center Digitization Project datasets (Kunkel et al. 1998), although those datasets are of daily resolution. Kunkel et al. (1998) break down errors into small and large, roughly analogous to our suspect and erroneous categories, and find the keying error rate for the former to be about 4–5 times that of the latter. Eskridge et al. (1995), in evaluating their complex QA for aerological data, note flag rates of between 0.09% and 5.7% for period-of-record evaluations using different datasets, with the percentage as high as 16%–18% in some individual years or short sequences of years. Miller and Benjamin (1992), in their hourly data assimilation project, report an overall flag rate of about 0.4%.

Although the system presented in this paper is highly automated, it is nevertheless a semiautomated complex QA. Manual inspection is an important component of an overall QA as well (Shafer et al. 2000; Fiebrich and Crawford 2001; Sopoco et al. 2004). At present, manual effort is most extensively spent on ensuring that each database record is indexed to the correct station, date, and time, that is, cleaning and sorting prior to meteorological complex QA. The principal concern in balancing manual and automated elements of overall QA is efficient use of the far more expensive human resources. A good example of manual inspection and semiautomated intervention is in handling the frequent occurrence, in data from the mid-1930s, of long runs of the DPD–DTR flag, caused by switching dewpoint for DPD on the observation forms. As discussed above, this problem is implicated as a factor in the high early flag rates, and a semiautomated solution is planned. Tables by station-month of counts of this particular flag will be constructed, and station-months with unusually high counts will be tagged for inspection (Montgomery 1985). Manual inspection of the data will be used to determine whether a run has, in fact, been detected and, if so, the precise start and end dates and times. Using that information, a simple computer program will be applied that swaps whatever is in the dewpoint column with its depression from dry-bulb temperature, supplying the resultant value as an edited replacement. Pure manual inspection of meteorological records and their flags is limited by resources to small random samples for spot-checking. There exists in target databases limited capacity to record whether a flag was applied automatically or manually (NCDC 2002).

Fiebrich and Crawford (2001) describe effects of some unique meteorological phenomena, such as mesohighs and heat bursts, on automated QA performance. Because they indicate that high winds or a wind burst often accompany these phenomena, we have included in our check for blips an IC subcomponent that turns off a flag initially activated if, within 1 h, there are reports of thunderstorm-related phenomena, hail, showers, heavy precipitation, squalls, blowing materials, wind burst, or variable wind direction. The blip check is known to misflag a few strong frontal passages and calm, clear morning lows, but the observed Type I error rate is only about 5% (details in Graybeal et al. 2004).

One way to improve the performance of the automated complex QA is by incorporating fuzzy logic into the determination of flagging thresholds and tolerances (e.g., Talbot et al. 1999). Development would likely begin with improving individual component checks, and initially for high-priority elements such as dry-bulb temperature. Component-check output would increase in complexity from binary to tiered, and the decision tree would need further development to accommodate the added layer of complexity, perhaps utilizing a neural network. It is unclear how much reduction in flag rates is expected to result from that approach, or how it would compare toward that end to a hybrid-inhomogeneity approach for detecting systematic errors.

6. Summary and conclusions

A complex QA algorithm for historical hourly meteorological data is described in this paper, from its individual components to its decision tree and operational results. Cleaning procedures prior to making the data available for QA are discussed, as well as the performance of the system on 20 yr of real data from approximately 150 stations. The complex QA is three pronged, incorporating component checks for limits consistency, internal consistency, and temporal consistency. No spatial checks are incorporated at this time, because of the sparse density of early stations in the hourly SAO network. A decision tree is developed that considers all the flags activated on a given hourly record and maps the elements most likely to be truly suspect or erroneous. The procedure uses the count of flags and of flag types to determine the level of flag severity. More than two flags or flag types generally indicate an erroneous datum, whereas flags of only one count or flag type are generally only powerful enough to indicate that the datum is merely suspect.

The overall flag rate (measured as hourly records with some type of flag present on any elements within) was quite high, at nearly 10%. However, a more detailed analysis shows that this figure is strongly influenced by time-varying observation practices as well as systematic errors. For example, long runs of dewpoint or sea level pressure substitution errors can substantially inflate the flag rate and are perhaps better caught using a hybrid procedure that focuses midway between traditional checks for random errors and inhomogeneity analysis [both are discussed by Eskridge et al. (1995)]. Such a procedure needs to be developed or applied to historical meteorological data. Aside from systematic errors, the flag rate is estimated as an order of magnitude less than when including them and is from one to two orders of magnitude less for ''erroneous'' than for ''suspect'' flags. Although the focus in development was on historical hourly datasets, the authors are working on portability of this complex QA to more recent hourly datasets, as well as to earlier historical datasets containing twice-daily observations, with only minor modification.

Acknowledgments. The authors thank all partners involved with the hourly SAOs and CDMP for their many discussions, suggestions, and directions, as well as two anonymous reviewers. This work was supported by NOAA Grant NA17RJ1222.

REFERENCES

Barnett, V., and T. Lewis, 1994: *Outliers in Statistical Data.* 3d ed. Wiley, 584 pp.

Collins, W. G., 2001: The operational complex quality control of radiosonde heights and temperatures at the National Centers for Environmental Prediction. Part I: Description of the method. *J. Appl. Meteor.,* **40,** 137–151.

DeGaetano, A. T., 1997: A quality control procedure for hourly wind data. *J. Atmos. Oceanic Technol.,* **14,** 308–317.

Eskridge, R. E., O. A. Alduchov, I. V. Chernykh, Z. Panmao, A. C. Polansky, and S. R. Doty, 1995: A comprehensive aerological reference data set (CARDS): Rough and systematic errors. *Bull. Amer. Meteor. Soc.,* **76,** 1759–1775.

Fiebrich, C. A., and K. C. Crawford, 2001: The impact of unique meteorological phenomena detected by the Oklahoma Mesonet and ARS Micronet on automated quality control. *Bull. Amer. Meteor. Soc.,* **82,** 2173–2187.

Gaffen, D. J., M. A. Sargent, R. E. Habermann, and J. R. Lanzante, 2000: Sensitivity of tropospheric and stratospheric temperature trends to radiosonde data quality. *J. Climate,* **13,** 1776–1796.

Gandin, L. S., 1988: Complex quality control of meteorological data. *Mon. Wea. Rev.,* **116,** 1137–1156.

Graybeal, D. Y., K. L. Eggleston, and A. T. DeGaetano, 2002: A climatology of extreme hourly temperature variability across the United States: Application to quality control. Preprints, *13th Conf. on Applied Climatology,* Portland, OR, Amer. Meteor. Soc., 55–58.

——, A. T. DeGaetano, and K. L. Eggleston, 2004: Improved quality assurance for historical hourly temperature and humidity: Development and application to environmental analysis. *J. Appl. Meteor.,* in press.

Guttman, N. B., 2002: Digitization of historical daily cooperative network data. Preprints, *13th Conf. on Applied Climatology,* Portland, OR, Amer. Meteor. Soc., 43–46.

Holsapple, C. W., and A. B. Whinston, 1996: *Decision Support Systems: A Knowledge-Based Approach.* Course Technology, 875 pp.

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.

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

——, S. A. Klein, and D. J. Seidel, 2003: Temporal homogenization of monthly radiosonde temperature data. Part I: Methodology. *J. Climate,* **16,** 224–240.

Meek, D. W., and J. L. Hatfield, 1994: Data quality checking for single station meteorological databases. *Agric. For. Meteor.,* **69,** 85–109.

Miller, P. A., and S. G. Benjamin, 1992: A system for the hourly assimilation of surface observations in mountainous and flat terrain. *Mon. Wea. Rev.,* **120,** 2342–2359.

Montgomery, D. C., 1985: *Introduction to Statistical Quality Control.* Wiley, 520 pp.

NCDC, 2001: ISH QC algorithms. NCDC Tech. Note, 12 pp. [Available from National Climatic Data Center, Federal Building, 151 Patton Avenue, Asheville, NC 28801-5001.]

——, 2002: Federal Climate Complex data documentation for Integrated Surface Hourly data. NCDC Tech. Note, 52 pp. [Available from National Climatic Data Center, Federal Building, 151 Patton Avenue, Asheville, NC 28801-5001.]

O'Brien, K. J., and T. N. Keefer, 1985: Real-time data verification. *Proc. ASCE Special Conf.,* Buffalo, NY, American Society of Civil Engineers, 764–770.

Reek, T., S. R. Doty, and T. W. Owen, 1992: A deterministic approach to the validation of historical daily temperature and precipitation data. *Bull. Amer. Meteor. Soc.,* **73,** 753–762.

Rogers, R. R., and M. K. Yau, 1989: *A Short Course in Cloud Physics.* 3d ed. Pergamon, 290 pp.

Shafer, M. A., C. A. Fiebrich, D. S. Arndt, S. E. Frederickson, and T. W. Hughes, 2000: Quality assurance procedures in the Oklahoma Mesonetwork. *J. Atmos. Oceanic Technol.,* **17,** 474–494.

Sopoco, T. H., L. D. Boudreau, and M. Armstrong, 2004: A new automated and interactive system for solar radiation data processing and quality control. Preprints, *20th Int. Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, 9.9.

Talbot, L. M., B. G. Talbot, R. E. Peterson, H. D. Trolley, and H. D. Mecham, 1999: Application of fuzzy grade-of-membership clustering to analysis of remote sensing data. *J. Climate,* **12,** 200–219.

U.S. Weather Bureau, 1935: *Instructions for Airway Meteorological Service, Circular N, Aerological Division.* 3d ed. U.S. Government Printing Office, 149 pp.

WMO, 1954: *Guide to International Meteorological Instrument and Observing Practice: Chapters 1 to 13.* WMO, 141 pp.