7.2 NEW TECHNIQUES IN QUALITY ASSURANCE OF HOURLY METEOROLOGICAL DATA: RESOLVING MULTIPLE FLAGS THROUGH A DECISION TREE

Daniel Y. Graybeal*, Arthur T. DeGaetano, and Keith L. Eggleston
Northeast Regional Climate Center, Cornell University, Ithaca, New York

1. INTRODUCTION

Beginning in 2001, historical hourly surface airways observations (SAOs) have been digitized from their original paper forms, as part of the Climate Database Modernization Program (CDMP), National Oceanic and Atmospheric Administration, U.S. Department of Commerce. This particular data recovery mission has extended the period of record for hourly meteorological data back an additional two decades, from the previously available start date of 1948 into the late 1920s. Prior to making these data available, quality assurance (QA) procedures were developed, with the goals of improving the individual component checks themselves (Graybeal et al. 2002) as well as the decision-making process by which flags are ultimately mapped into a data base. This paper provides an overview of the resultant QA system, describes the decision tree for mapping flags, summarizes its performance, and discusses lessons learned.

2. COMPLEX QA AND THE DECISION TREE

Complex QA evolved from more traditional QA, not only by providing individual component checks in greater number and/or complexity, but also by using a decision-making algorithm that weighs all the evidence from flags thrown by the individual tests, also called component checks (Gandin 1988). Such a treatment can handle varying degrees of severity of nonconformance of observations with QA models, rather than simply flagging in the event any component check failed. In Fig. 1, the implementation of complex QA in the present effort is diagrammed. An hourly record (meaning one line or row in a data base flat file) is examined, containing temperature, humidity, wind, present weather, visibility, cloud cover, and pressure information. This hourly record is then submitted to a battery of component checks that evaluate the elements in that record. Checks are made for limits consistency (LC; e.g., climatological) as well as for internal consistency (IC; e.g., dry bulb is at least as great as dew point). A third type of component check is for temporal consistency (TC) and generally looks for blips (excess variability) or runs (excess invariability).

Flags may be thrown by all three types of component checks on a given element in that hourly record. This situation necessitates a process to decide which of the several elements that may be involved is the most likely responsible, the most suspect. For example, if an IC flag is thrown on dry bulb and dew point, as well as a TC flag on dry bulb, chances are that the dry bulb is truly suspect and the dew point can be dropped from further consideration for flagging.

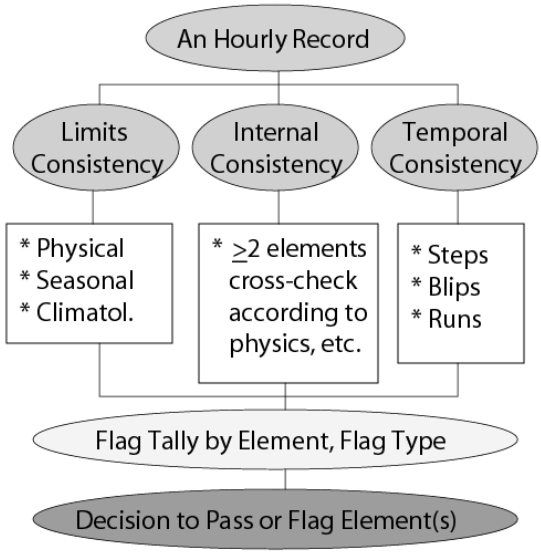

Fig. 1. Complex QA of CDMP SAOs diagrammed.

Critical to the operation of the decision tree developed here is a table of component flag counts by element and by component type (the "flag type table"). Consider the example of a dry bulb observation of –5.5°C; one lone misdigitization of that dry bulb as –55°C can raise as many as seven component flags (Table 1). Note that each component flag is indexed by its type, which is generally (but not always) defined by the number of elements involved in its check procedure. The three different types are needed to provide independent information. LC and TC checks generally consider one element at a time, whereas IC checks consider two or more. Table 2 illustrates the flag type table for the case of the fallacious –55°C dry bulb. In this case, both the LC and TC flag tallies provide independent information about the character of the dry bulb element. For dry bulb, not only is the highest flag count per element recorded, but also flags from all three types. Thus, the value for this element at this hour is considered highly suspect.

*Corresponding author address: Dr. Daniel Y. Graybeal, 1123 Bradfield Hall, Cornell University, Ithaca, NY 14853; e-mail: dyg2@cornell.edu.

Table 1. Component flags raised by a single gross misdigitization of dry-bulb temperature.

Flag Type	Flag Message
LC	Dry bulb out of bounds
IC	Wet bulb exceeds dry bulb
IC	Dew point exceeds dry bulb
IC	Dew point depression inconsistent with diurnal temperature range
IC	Mismatch among dry and wet bulbs, dew point, and station pressure
IC	Mismatch among station and sea-level pressures, dry bulb, and elevation
TC	Blip in dry bulb

Table 2. Example of the flag type table used in mapping component flags to individual elements.

Element	LC Flags	IC Flags	TC Flags	Flag Total
Dry bulb	1	5	1	7
Wet bulb	0	2	0	2
Dew point	0	3	0	3
Station pressure	0	2	0	2
Sea-level pressure	0	1	0	1
Elevation	0	1	0	1

The decision tree evaluates the initial flag type table by a three-tiered process. The first tier has just been illustrated; the procedure looks for elements on which component flags have been thrown from more than one flag type. This represents the most suspicious case. If any are found, those elements are automatically considered erroneous. Processing continues with the element having the highest overall flag count. In the example above, the procedure begins with dry bulb. Next, all component flags (Table 1) associated with that element are removed, and the flag type table is recalculated from any remaining component flags. At that point, the procedure searches again at the top tier. In the example given, no other component flags remain after those associated with dry bulb have been removed from consideration, and the procedure terminates, flagging only the dry bulb with the code for "Erroneous."

Suppose two component flags are thrown on a given hourly record, indicating that wet bulb exceeds dry bulb and that dew point exceeds dry bulb (both IC checks). This condition illustrates the second tier in the decision tree, executed when the flag type table indicates at least one element is associated with more than one flag from any type. In this example, dry bulb is associated with two component flags and the other two elements with one flag each. The procedure continues as outlined above, removing all component flags associated with the element having the highest overall flag count (here, dry bulb), and repeating the test on the recalculated flag type table. In this example, no component flags are left, and the procedure terminates, flagging only dry bulb, also with the code for "Erroneous," as more than one line of evidence suggests it is suspect.

Occasionally there may be a tie of two or more IC flags per element, for two or more elements. In that case, the condition of "Erroneous" is no longer assumed. Suppose that, instead of the flag on dew point, a flag showing a mismatch among dry bulb, wet bulb, dew point, and station pressure was thrown. Then, both dry bulb and wet bulb would have two IC component flags each, but with no other types of flags. Not enough evidence is presented to the decision tree to choose one element over the other as more suspicious, so the flag code for "Suspect" is given to both elements involved. The same action is taken if only one flag per element is found; all such elements are flagged "Suspect," if no other information is available to suggest otherwise. Thus, the third tier of the decision tree is illustrated.

3. CLEANING

Before any meteorologically or climatologically based QA can be implemented, the data base must be in chronological order and its records keyed by unique and complete sets of identifiers. Guttman (2002) provides a nice discussion of cleaning problems, many of which were encountered here as well. For example, the station identifier and the year must match what the station-yearly file name indicates, and they must key the records for the correct station and year. None of the identifier elements was left missing, such as would result from the digitizer's code for illegibility. Impossible time stamps, such as 31 February or 7300 hours, were also treated as missing. Multiple or duplicate time stamps were discarded in separate files for later inspection, following a first-found, first-kept rule; subsequent manual analysis allowed some blocks to be reinserted or indexed to a different station.

One interesting aspect of these historical data was that the observation schedule, in minutes past the hour, evolved over the 20-year period of focus. Instead of making an abrupt transition from the 20s minutes to the 50s minutes, as occurred in June 1957 (Steurer and Bodosky 2000), the schedule varied fluidly over time, from the 40s and 50s minutes schedule in use during the early 1930s, to the 20s minutes during the late 1940s. The time stamp was mapped to the nearest hour based on a moving window directional mean minute (Mardia 1972) and using the results of extensive frequency analysis to determine the cutoff (34 minutes) for rounding forward or backward. The entire pre-QA cleaning process resulted in only approximately 0.25% of more than 14 million hourly records being eliminated. This percentage is similar to Guttman's (2002).

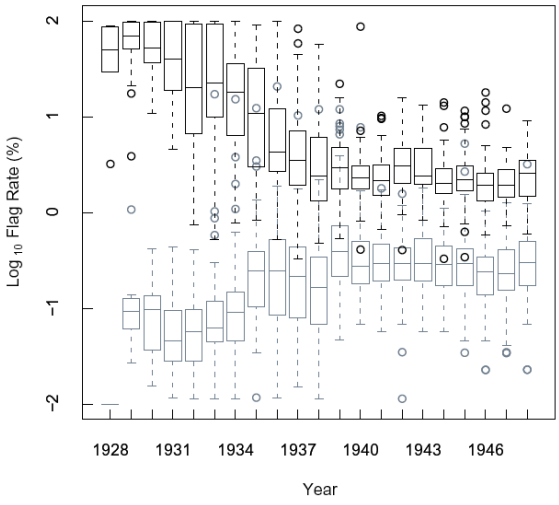

Fig. 2. Boxplots of (log, base 10, of percent) flag rate per year over all stations. Black boxes indicate all flags ("Suspect" and "Erroneous"), while grey boxes indicate "Erroneous" flags.

4. PERFORMANCE, WITH DISCUSSION

Of the hourly records passing cleaning, about 9.5% have a flag on at least one of its elements. Although this flag rate is comparable to those reported from QA of other data sets (Eskridge et al. 1995), it is still fairly high. However, only about 0.4% of all records have an "Erroneous" flag on at least one of its elements, illustrating the value of the tiered flag structure to users. Again, this more stringent flag rate is comparable to those obtained in QA of other data sets (Kunkel et al. 1998; Guttman 2002). Fig. 2 illustrates the asymptotic changes in flag rate (expressed as percent of hourly records having at least one element flagged) over time, for all stations examined. Boxplots are fairly standard, featuring the interquartile range (IQR) by the box and the median by the middle line. Beginning about 1937 or 1938, the flag rates level off at about $10^{0.5}$% (3%) overall ("Suspect" and "Erroneous" flags), or $10^{-0.5}$% (0.3%) for "Erroneous" flags. Prior to that time, overall flag rates were significantly higher.

The large flag rates, both overall and prior to 1937, suggest the presence of systematic errors, in addition to the random errors that most component checks were designed to catch. Both types of errors should be caught in a complete QA (Eskridge et al. 1995); however, the scale of focus required for catching systematic errors is longer than for random errors. This is because systematic errors usually persist for some time.

Early in processing these data, blocks of time were found during which the dew point depression (DPD) had been substituted for dew point by the observer on the original form, or *vice versa*. One of the more elaborate component checks developed for this complex QA examines daily maximum DPD in relation to the diurnal temperature range (DTR). A nearly 1:1 linear relationship was found, when both variables are transformed logarithmically (after adding to each one tenth the precision in recording the temperatures, as precaution against attempting to transform zeros). Details of this procedure are given elsewhere (manuscript submitted to *J. Appl. Meteor.*). Although it performed suitably in detecting random errors, it was designed with systematic dew point-DPD reversals in mind. Toward the latter end, it flagged about 40% of individual hourly records that were part of long, contiguous blocks of systematic reporting error.

This flagging practice illustrates two general problems with these types of components that focus on hour-to-hour conditions. First, less than half the systematic errors are flagged, due perhaps to the differences in scale between the focus of the test and the manifestation of the systematic error. Second, even flagging nearly half the systematic errors presents the user with the problem of having to handle so many hundreds or thousands of flagged data. In other words, applying hour-to-hour checks to systematic errors amounts to imprecision in catching them and inefficiency in dealing with them. These problems are not limited to the DPD-DTR check; IC checks for pressure that involve station elevation behave similarly. Pressure estimation is sensitive to changes in elevation, such as may accompany station moves. If the move is not recorded in the metadata and causes the pressure estimate to fall outside the tolerance of the check, a long run of pressure flags is generated.

5. CONCLUSIONS AND FUTURE RESEARCH

The issues raised in the previous section are not easily resolved by existing techniques. Random errors have been the focus of traditional QA checks that operate on the time scale of one to tens of the temporal units of resolution, e.g., hour-to-hour or a one-day moving window. On the other hand, inhomogeneity analysis (IA) techniques exist that address systematic errors persisting over the scale of years to decades. The kinds of systematic errors encountered in this hourly data set require solutions developed on an intermediate time frame. Fortunately, some work is already being done for daily and monthly data sets (Menne and Duchon 2002) that recognizes this same need for an intermediate-term time frame for IA. Most IA practitioners agree a reference time series is needed to optimize its performance on a candidate series (Peterson et al. 1998), and Menne and Duchon use such a series composited from a spatial window. While a spatially interpolated reference series is plausible for the dense networks they consider, it is not feasible for the sparse, early-twentieth century airways network such as is under QA here. For the present work,

reference series must come from the same station, but another weather element.

Using the relationship previously found in log space, daily maximum DPD is subject to an exploratory IA, using coincident DTR as a reference series. IA is performed on the "DPD factor," simply the ratio of DPD to DTR in log space, each variable incremented by 0.01°F prior to transformation. Fig. 3 (a) shows the time series of monthly median, daily DPD factor over the CDMP period of record at Columbus, Ohio (ID #14821); Fig. 3 (b) shows the weekly median for just the year 1935. In both plots the expected ratio near 1 is given by a dashed line; in the latter, the medians of two different segments, the difference highly significant by a modified Wilcoxon-type rank sum test (p-level < 0.001%) (Lanzante 1996), are given by solid lines. The first segment, extending through week 35, is much above expected. Inspection of the 1935 records indicates the DPD switch was initiated on 20 January and maintained until 30 September; the latter coincides within one month of the changepoint identified by Lanzante's (1996) procedure. Thus, Menne and Duchon's "middle way" may be promising, if a suitable within-station reference series can be found from another weather element. Changepoints found can be used to guide manual intervention, such as simply flipping the DPD switch on a block of data. While such a technique is not available for the current data release, it may be developed for, and applied to, a later upgrade.

6. LITERATURE CITED

Eskridge, R. E., O. A. Alduchov, I. V. Chemykh, 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.

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. Appl. Climatol.*, Portland, OR, Amer. Meteor. Soc., 55–58.

Guttman, N. B., 2002: Digitization of historical daily cooperative network data. Preprints, *13th Conf. Appl. Clima-tol.*, Portland, OR, Amer. Meteor. Soc., 43–46.

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., 1996: Nonparametric techniques for analysis of climate data. *Int. J. Climatol.*, **16**, 1197-1226.

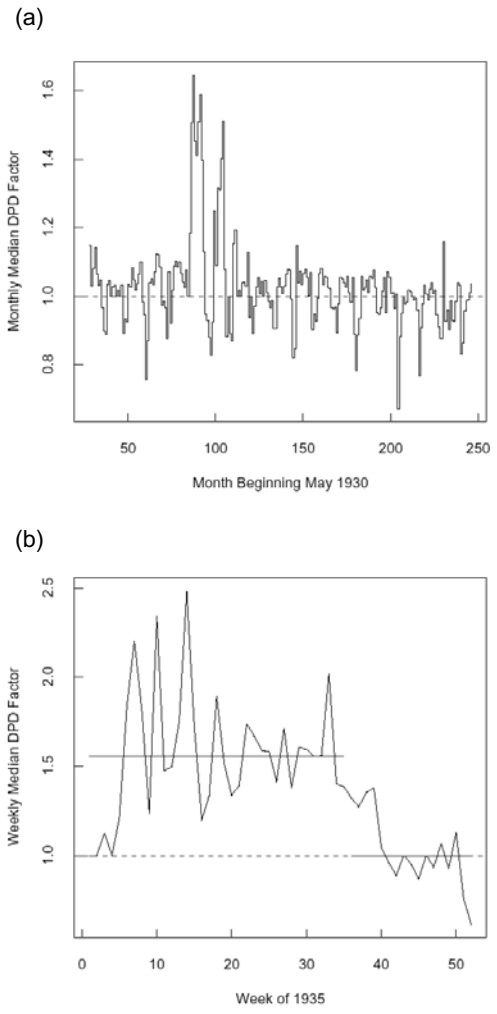

Fig. 3. (a) Monthly and (b) weekly median, daily DPD factor at Columbus, Ohio, for 1930–1948 and 1935, respectively.

Mardia, K. V., 1972: *Statistics of Directional Data*. Academic Press, 357 pp.

Menne, M. J. and C. E. Duchon, 2002: Quality assurance of monthly temperature observations at the National Climatic Data Center. Preprints, *13th Conf. Appl. Climatol.*, Portland, OR, Amer. Meteor. Soc., 18–23.

Peterson, and Coauthors, 1998: Homogeneity adjustments of in situ atmospheric climate data: A review. *Int. J. Climatol.*, **18**, 1493–1517.

Steurer, P. and M. Bodosky, 2000: *Surface Airways Hourly TD-3280 And Airways Solar Radiation TD-3281*. National Climatic Data Center, 53 pp.