2.11 A CLIMATOLOGY OF EXTREME HOURLY TEMPERATURE VARIABILITY ACROSS THE UNITED STATES: APPLICATION TO QUALITY CONTROL

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

1. INTRODUCTION

Long-term hourly temperature data are digitally available from NOAA via the National Climatic Data Center, generally covering the period 1948 to present. Currently, the digital hourly-resolution climatic record is in the process of being extended backward from 1948 into the late 1920s, through digitization of the original Surface Airways Observations (SAO) paper forms. An important component of that work is extending and applying quality control (QC) for all variables being processed. Due to the widespread interest among the scientific community in long-term trends and variability of temperature and moisture variables, this study focuses on dry-bulb, dew point, and wet-bulb temperatures and on methods developed to improve their QC.

Conventional temporal consistency checks (Meek and Hatfield 1994) in the QC of hourly temperature data incorporate only the magnitude of the change from one hour to the next. A threshold magnitude for hour-to-hour temperature jumps is employed, and the QC algorithm flags any hourly temperature for which the first differential exceeds this threshold in magnitude. In an exploratory analysis using currently available data, nearly all hourly temperatures so flagged were actually one-hour spikes or dips; very few were associated with step changes (i.e., strong frontal passages). Most of the spikes or dips probably resulted from simple errors of observation or recording, such as the transposition or omission of digits. Such errors are of great interest in the historical SAO conversion project, which involves a substantial amount of manual digitization. Therefore, an approach is needed that addresses this interest and explicitly defines an hourly temperature flagging threshold in terms of a one-hour spike or dip.

In this study, two methods are developed and tested that identify extreme hourly temperature variability in terms of hourly spikes or dips, using 1949–1958 SAO from 28 stations throughout the United States. A threshold climatology is presented and applied in the QC of a test subset covering 1959–1963 SAO. Performance of each method, along with two traditional methods, is compared, based upon this QC and experiments on randomly selected temperature records utilizing a deliberate modification scheme that mimics typical errors. Because the focus is on the QC of historical data, units must also be historical.

* *Corresponding author address:* Daniel Y. Graybeal, Cornell University, Northeast Regional Climate Center, Ithaca, NY 14853; e-mail: dyg2@cornell.edu.

2. METHODS

The focus of QC in this study is confined to temporal consistency checks. Meek and Hatfield (1994) describe a three-pronged approach to the QC of meteorological data measured at sub-daily resolution. First, a limits consistency (LC) check is performed; this is a common QC procedure used with a wide variety of weather elements and temporal resolutions (Eischeid et al. 1995; Reek et al. 1992). Examples include checks against daily or climatological extremes and physical bounds (i.e., "plausibility checks," after Gandin (1988)). Second, the datum in question is passed through an internal consistency (IC) check, using other information from the data record at that same hour or adjoining hours. A simple example is a check to see whether the dew point exceeds the dry-bulb temperature (Reek et al. 1992).

The objective of the third type, a temporal consistency (TC) check, is to ensure that variability of the temperature in question is neither excessively high nor low, with respect to other observations within the temporal vicinity. Implementations tend to refer to excessively high variability in terms of spikes, or dips, and to excessively low variability as flat-liners, or runs (Reek et al. 1992; Meek and Hatfield 1994).

TC checks of hourly temperature observations have historically been restricted to the hour-to-hour (first difference) rates of change, flagging rates whose magnitude exceeds a predefined, static threshold (Meek and Hatfield 1994). Such an hour-to-hour threshold approach is problematic for the following reasons:

- How the threshold was determined is unclear. Was climatological information used, and was the threshold experimentally tested?
- The threshold does not vary with season.
- The threshold does not vary over space.
- The approach does not precisely address the problem of a one-hour anomaly due to the types of error of interest here; step changes (fronts) are flagged as well as spikes.

Therefore, in this study, each of these shortcomings were addressed.

A spike or a dip is defined here as an hourly observation of temperature about which the successive first differences of the hourly time series are of opposing (and nonzero) sign. Two models are developed that measure the magnitude of the spike or dip in different ways. The first measures by the minimum absolute value of the two consecutive first differences about the hourly observation in question (model MDH2). The second measures by the residual from a five-hour median smooth centered on the hour in question (model MSR5). The first model is intolerant of missing hourlies, whereas the second allows no more than one missing

hourly in the five-hour span of the moving window. Twenty-eight first-order stations from across the United States were selected for analysis and testing (Figure 1). At most of these stations, less than 0.01% of the 1949–1963 hourly temperatures were missing, and all exhibited below 0.10% missing.

The period 1949–1963 was chosen for two reasons. First, the protocol for temperature and humidity observation was mostly constant throughout the period; dew point was calculated from dry-bulb and wet-bulb temperatures read directly (Robinson 2000; U.S. Dept. Commerce 1962). Second, only three-hourly observations are currently available from the mid-1960s into the early 1980s (Robinson 2000). For model development, the ten-year period 1949–1958 was selected, and for testing, the remaining five-year period 1959–1963 was reserved.

For each three-month season (winter is December–February, etc.), and for each station, the distribution of all hourly spikes and dips was examined (the number being generally about one-tenth the total number of hourlies), and flagging thresholds were determined. That step involves the substantial problem of potential outlier identification (Barnett and Lewis 1984; Grant and Leavenworth 1988). Most historical methods for this purpose assume only one outlier is present in a set, and their iterative application in finding more than one outlier is highly problematic (Barnett and Lewis 1984; Davies and Gather 1993). More recent methods are multivariate (e.g., Hadi 1992); we include multivariate data within the LC-IC-TC framework, specifically as an IC check (Meek and Hatfield 1994; DeGaetano 1997).

In developing QC for hourly wind data, DeGaetano (1997) found values flagged generally lay beyond the 99.9 to 99.95 percentiles in his data sets. Extrapolating to these data sets, which include only temperature spikes and dips, a 99.95 percentile threshold flags approximately one or two spikes or dips every decade, in each season. Visual inspection of spike/dip histograms generally supported this flagging rate, except outlying groups sometimes occurred. With the understanding that IC checks would likely catch outliers that were meteorologically valid, the number of outliers to allow in processing the developmental subset of the climatological record was increased to four. To minimize the sensitivity of the threshold choice to outliers, the percentile corresponding to a four-outlier pass was used as the threshold for flagging. Thus, a climatology of extreme hourly temperature variability was developed.

In testing these models, an experimental procedure was designed that mimics errors likely in digitization (Reek et al. 1992). Seven types of deliberate errors were introduced on randomly selected samples of 100 hourly observations from each season and each station, and the percentage of these flagged was noted in each test. The seven types are:

1. Digit transposition (e.g., 53°F becomes 35°F);
2. Sign omission or commission;
3. Add/subtract 100°F;
4. Add/subtract integer multiple of 10°F;

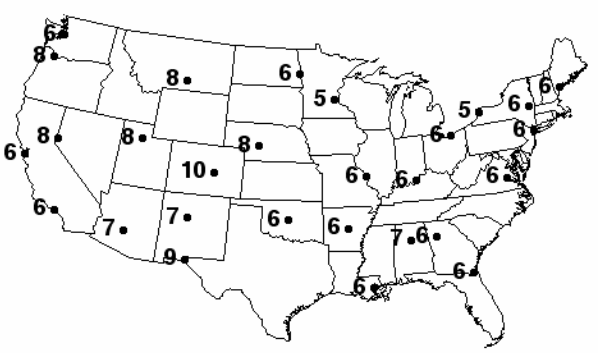

Figure 1. Flagging thresholds (°F) determined for dry-bulb temperature using model MDH2, averaged across all four seasons.

5. Scale shift (e.g., 62°F becomes 6.2°F);
6. Units shift (e.g., 68°F becomes 20°C, as °F);
7. Both subtract and scale (e.g., 36°F as 3°F).

In addition, a test was performed ("Test 0") with no deliberate modifications within the test data set, to evaluate the original data flagged by these methods. Original records whose temperature, dew point, or wet-bulb was flagged were compared with their temporal neighbors, and the character of the hourly observation as a whole was noted (e.g., a spike in temperature was flagged, but a thunderstorm and wind gust co-occurred).

Finally, two traditional methods were compared with these two new methods. These are MH94, which incorporates an hour-to-hour threshold of 11°F (Meek and Hatfield 1994), and DT18, an 18°F hour-to-hour threshold previously employed by NCDC.

Table 1. Flagging thresholds (°F) for model MDH2, averaged over all stations, by variable and by season. Variables are TMPD=dry-bulb, DPTP=dew point, and TMPW=wet-bulb temperatures.

	Winter	**Spring**	**Summer**	**Autumn**	***(Mean)***
TMPD	6.4	6.6	7.6	6.3	*6.7*
DPTP	9.8	10.2	9.0	9.3	*9.6*
TMPW	5.3	4.9	5.3	4.8	*5.0*
(Mean)	*7.1*	*7.2*	*7.3*	*6.8*	***7.1***

3. RESULTS AND DISCUSSION

Spatial patterns of threshold values determined for temperature, dew point, and wet-bulb were similar (Figure 1). Generally, flagging thresholds of both models MDH2 and MSR5 were minimal over the eastern United States, maximal over the High Plains, Rocky Mountains, Great Basin, and Sierra Nevada, and minimal again along the Pacific coast.

Thresholds varied by season as well (Table 1), although the variation was more noticeable at some stations than at others. Overall, thresholds for temperature exhibited the strongest interseasonal variability, with those for dew point second. Dew point thresholds averaged higher than those for temperature,

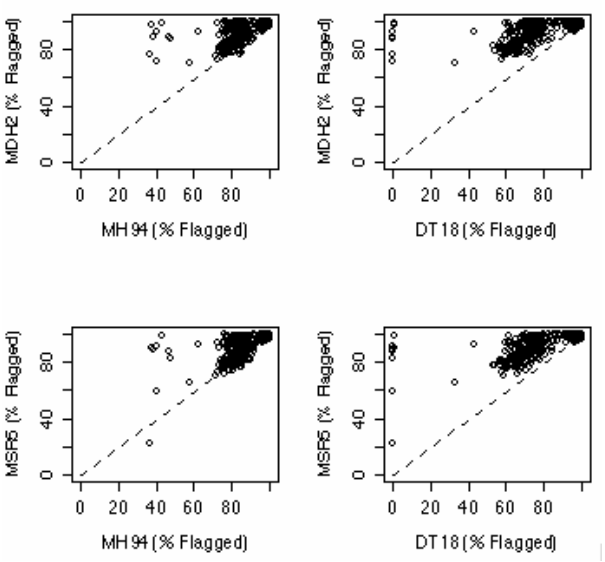

Figure 2. Pairwise model performance comparison across class of models, for dry-bulb temperature. There are 784 data points, each representing a test type (of 7), a season (of 4), and a station (of 28).

a result not unexpected given its historical means of computation and extreme sensitivity of the formula at low relative humidity. Time series of wet-bulb are quite well-behaved; both its threshold magnitude and variability were low. The flagging threshold of 7.1°F in model MDH2 (Table 1), averaged across stations, seasons, and variables, was much lower than thresholds previously utilized, and it is theoretically low enough to catch errors of digit transposition. Thresholds in model MSR5 showed similar patterns but averaged about 1.1°F larger in magnitude. Inclusion of the interseasonal variability in thresholds dramatically reduced the scatter in flagging rates in the error-peppering experiments, compared to the models that employ an invariant threshold.

Intermodel comparisons demonstrate the overall superiority of models MDH2 and MSR5, that explicitly define a spike or a dip, use climatological information, and allow spatial and seasonal variation in flagging threshold determination, over models MH94 and DT18, that search for single hour-to-hour jumps and incorporate constant, predefined thresholds. For dry-bulb temperature, pairwise comparisons of flagging rates on a test-by-test case basis are plotted in Figures 2 and 3. Both MDH2 and MSR5 outperformed each of MH94 and DT18 (Figure 2). This result held for the other two variables, except in the case of dew point and model MH94, in which MH94 performance was comparable to that of MDH2 and slightly better than that of MSR5. In comparing within model class, MDH2 usually outperformed MSR5 (Figure 3), except in the case of wet-bulb, in which MSR5 exhibited a slight advantage. Trimmed (less the bottom 2.5% and top 2.5%) mean bias error (MDH2–MSR5) was +0.081%, +0.167%, and –0.240% flagged for dry-bulb, dew point, and wet-bulb temperatures, respectively.

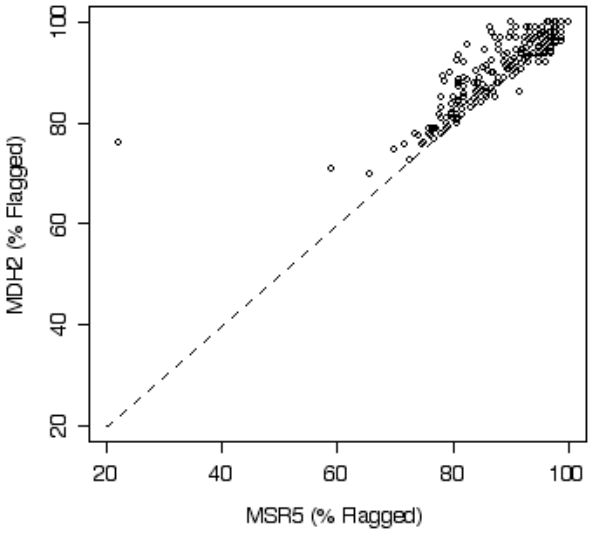

Figure 3. Pairwise model performance comparison within class of models, for dry-bulb temperature and the two new models. See Figure 2 caption for details.

Evaluating model performance on flagging errors deliberately thrown in experimentation is only one aspect of total model performance considered. Model performance in flagging unaltered data from the test subset of the climatological record also deserves attention. In terms of the number of flags thrown on unaltered, original data, model MH94 was by far the most liberal, whereas model DT18 was the most conservative. Of five years' worth of original hourly dry-bulb temperatures at 28 stations, models MH94, MDH2, MSR5, and DT18 flagged 0.262% (3224), 0.033% (401), 0.028% (348), and 0.011% (135), respectively. Model MH94 was also an order of magnitude more active in flagging dew point, but in that case, model DT18 increased its activity, and the most conservative model was instead MDH2. In flagging wet-bulb, all models exhibited a similar 0.025–0.032%, except DT18, which was an order of magnitude more conservative.

A random sample revealed interesting patterns of character in the whole hourly observation or its temporally adjacent observations (Table 2). In most

Table 2. Of a random sample of 100 flagged original dry-bulb temperatures, 25 per model, the number belonging to different weather conditions and flagged by the different models is shown. Condition "Legitimate spike" is for cases involving a report of a thunderstorm, moderate to heavy precipitation, showers of any kind, a 45° wind shift, or at least a 5 mph spike in wind speed.

Condition	**MDH2**	**MSR5**	**MH94**	**DT18**	***(Total)***
Legitimate spike	20	15	8	17	*60*
Frontal passage	0	0	15	6	*21*
Suspicious	2	5	0	2	*9*
Other	3	5	2	0	*10*
(Total)	*25*	*25*	*25*	*25*	***100***

cases (70% for MDH2 and MSR5 combined, and 50% for MH94 and DT18 combined), values flagged were also coincident with a report of a thunderstorm, a burst or shift in wind, heavy precipitation, etc. This is useful information to include in a subsequent IC check, as in DeGaetano (1997). Of the 50 flags in the sample thrown by either MH94 or DT18, nearly as many (41% combined) were of frontal passages as of legitimate spikes or dips. By contrast, the sample contained no frontal passages flagged by MDH2 or MSR5. This discrepancy is almost certainly due to the construction of the flagging threshold as a spike or dip, as opposed to a single hour-to-hour jump.

Conditions classified as "other" in Table 2 include the following situations. In 16% of cases, MSR5 flagged a dip that was a legitimate morning low under clear skies with calm or light and variable winds. Flags of temperatures whose values were deemed suspicious occurred when the rest of the record indicated no special case for such a spike or dip. One exception occurred in the case of a 4°F threshold, which is approximately the minimum threshold found in this study. Whether a minimum threshold should be set remains to be determined. Three cases involved wind speed variability, not as a spike or dip, but as a jump from calm to a speed of at least 6 mph, or as a persistent, strong wind (three consecutive hours over 20 mph). Persistent, strong winds may contribute to a temperature spike or dip through eddy turbulence in the downward transfer of energy, moisture, and momentum toward the surface.

4. SUMMARY AND CONCLUSIONS

Historical hourly meteorological data, temperature in particular, are widely used in climatic research. Current efforts in extending this record backward in time through digitization of paper forms affords a fresh opportunity to improve the quality control of these hourly data. The vast majority of hourly temperatures flagged using a traditional focus on hour-to-hour jumps were one-hour spikes or dips. In this study, therefore, a climatology of extreme hourly temperature variability was developed, using two new models incorporating an explicit definition of spikes and dips as the target. Historical data from across the United States were used to develop and test these models and compare their flagging performance with two traditional models.

Results indicate the new models, employing thresholds defined in terms of spikes or dips, determined using climatological information, and that vary with season and over space, were more efficient than traditional models in trapping deliberate errors believed to mimic digitizing errors. In a sample of cases where original data were flagged, the large majority co-occurred with reports of a thunderstorm, heavy rain, or a gust or shift in the wind. Inclusion of an internal consistency check for such information in the temporal vicinity of temperatures flagged by the MDH2 method is recommended in the temporal consistency aspect of quality control of these data.

5. ACKNOWLEDGMENTS

This research is a part of the ongoing Climate Database Modernization Program of NOAA-NESDIS-NCDC. Python and R software were used for analysis.

6. LITERATURE CITED

Bartlett, V. and T. Lewis, 1984: *Outliers in Statistical Data*. 2d ed. Wiley & Sons, 463 pp.

Davies, L. and U. Gather, 1993: The identification of multiple outliers. *J. Amer. Stat. Assoc.*, **88**, 782--794.

DeGaetano, A. T., 1997: A quality-control routine for hourly wind observations. *J. Atmos. Oceanic Technol.*, **14**, 308--317.

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

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

Grant, E. L. and R. S. Leavenworth, 1988: *Statistical Quality Control*. 6th ed. McGraw Hill, 714 pp.

Hadi, A. S., 1992: Identifying multiple outliers in multivariate data. *J. Royal Stat. Soc.*, **B54**, 761--771.

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

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

Robinson, P. J., 2000: Temporal trends in United States dew point temperatures. *Int. J. Climatol.*, **20**, 985--1002.

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.

United States Department of Commerce, Weather Bureau, 1962: *Instructions for Climatological Observers. Circular B*. 11th ed. U.S.G.P.O.