INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **27**: 1257–1263 (2007)
Published online 22 May 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1484

Short Communication

Relationship of flagging frequency to confidence intervals in the statistical regression approach for automated quality control of $T_{\max}$ and $T_{\min}$

Jinsheng You and Kenneth G. Hubbard*
High Plains Regional Climate Center, University of Nebraska, Lincoln, Nebraska 68583-0997, USA

Abstract:

With the widespread use of electronic interfaces in data collection, many networks have increased, or will increase, the sampling rate and add more sensors. The associated increase in data volume will naturally lead to an increased reliance on automatic quality assurance (QA) procedures. The number of data entries flagged for further manual validation can be affected by the choice of confidence intervals in statistically based QA procedures, which in turn affects the number of bad entries classified as good measurements. At any given station, a number of confidence intervals for the Spatial Regression Test (SRT) were specified and tested in this study, using historical data for both the daily minimum ($T_{\min}$) and maximum ($T_{\max}$), to determine how the frequency of flagging is related to the choice of confidence interval. An assessment of the general relationship of the number of data flagged to the specified confidence interval over a set of widely dispersed stations in the High Plains was undertaken to determine whether a single confidence factor would suffice, at all stations, to identify a moderate number of flags. This study suggests that using a confidence factor 'f' larger than 2.5 to specify the confidence interval will flag a reasonable number of measurements (<1%) for further manual validation and a single confidence factor can be applied for a state. This paper initially compares two formulations of the SRT method. This comparison is followed by an analysis of the percentage of observations flagged as a function of confidence interval. Copyright © 2007 Royal Meteorological Society

KEY WORDS inverse distance weighting; spatial regression test; RMSE; quality control; weather data

Received 14 August 2006; Revised 27 November 2006; Accepted 27 November 2006

INTRODUCTION

Two types of error may occur in the quality control (QC) of weather data. A Type I error is the flagging of good data and a Type II error is failure to flag bad data. The frequency of occurrence of these two types of error is a good indicator for evaluating the performance of QC methods. The purpose of research on QC methods is to produce optimal techniques to identify bad data while minimizing the frequency of Type I and Type II errors.

One of the objectives of the development of automatic quality assurance (QA) procedures for climate data is to reduce the manual workload of human validators. QA procedures have been applied by the National Climatic Data Center (NCDC) (Guttman and Quayle, 1990) in a mix of manual and automatic checks to assess the validity of weather data from the cooperative climatological stations. The statistical literature is replete with general guidance about identifying outliers in data (e.g. Barnett and Lewis, 1994), but literature concerning the application of techniques specific to quality assessment of climatological data is scant. General testing approaches such as using threshold and step change criteria have been designed for the single station review of data to detect potential outliers (Wade, 1987; Reek *et al.*, 1992; Meek and Hatfield, 1994; Eischeid *et al.*, 1995; Shafer *et al.*, 2000). Recently QC has been expanding from procedures based on the in-station checking to include procedures for interstation checking (Wade, 1987; Gandin, 1988; Eischeid *et al.*, 1995; Hubbard *et al.*, 2005). The latter conducts the tests by comparing observations to the reference estimates obtained from the spatial techniques such as inverse weighting or statistical regression between stations.

The spatial regression test (SRT, Hubbard *et al.*, 2005; Hubbard and You, 2005) assigns weights according to the root mean square error (RMSE) associated with the linear regression between the station of interest and the neighboring stations. The SRT method proved to be robust owing to its implicit accounting of the systematic differences associated with temperature lapse rate with elevation in complex terrain (Hubbard *et al.*, 2005; Hubbard and You, 2005). These estimates by the SRT method are

* Correspondence to: Kenneth G. Hubbard, High Plains Regional Climate Center, University of Nebraska, Lincoln, Nebraska 68583-0997, USA. E-mail: khubbard@unlnotes.unl.edu

critical to the processes of identifying suspect temperature data, otherwise referred to as potential outliers, and the procedure is used operationally to ensure quality data in the Applied Climate Information System (ACIS; Hubbard *et al.*, 2004). The term potential outlier should not be confused with the extreme values of record for it is instead a value that is improbable based on the recent weather at the site and in the surrounding area. The estimates are also used to form a continuous dataset by filling in missing values. Research has demonstrated that the SRT is superior to other tests in identifying seeded errors (Hubbard *et al.*, 2005).

The SRT method has an input parameter to determine the confidence level of the test, namely, the 'f factor', which is similar to the cutoff in other QA methods. In a separate study, the investigators used the SRT to identify the potential outliers during unique weather events (You and Hubbard, 2006). In the case of hurricanes, cold front passage, floods, and droughts, the number of QA failures were largely due to the different times of observation coupled with the ambiguity associated with position relative to tight gradients of temperature or precipitation (Eischeid *et al.*, 1995; Belcher and DeGaetano, 2005; Wu *et al.*, 2005; You and Hubbard, 2006). The original SRT method was found to flag a considerable number of $T_{\max}$ and $T_{\min}$ measurements when a cold front passes or precipitation measurements when a hurricane event occurs. You and Hubbard (2006) introduced modifications to the original SRT method to create an enhanced SRT relying on the measurements at neighboring stations to recognize these excessive flagging situations and, in special circumstances, to reset the 'bad data' flags back to 'good data' flags. The ratio of flags reset to the number of flags identified by the original SRT method for the unique events is referred to as the 'reset fraction.'

The original version of the SRT method (original SRT) takes the square of intermediate estimates used in the weighting process, which does not preserve the sign for the final estimate when one or more intermediate estimates are negative. The additional effort to determine the sign introduces further demands on computer resources when otherwise the algorithm is quite efficient. This study presents a modified or new SRT method, called SRT2, that eliminates the need to determine the sign of the estimates. To determine the effectiveness of the modified SRT2 method, the differences between the accuracy of the new method and the original SRT method are evaluated for all stations in the contiguous states of the USA for both the $T_{\max}$ and $T_{\min}$ for the year 2002.

Other factors, besides f, can affect the precision of the estimates. Hubbard and You (2005) explored the sensitivity of the SRT method to such factors as the radius of inclusion, the regression time-window, the regression time-offset and the number of stations used to make the estimates. The performance of the SRT method stabilized when ten or more stations were applied in the estimates and therefore we recommend 10 stations be used in this QC procedure. Questions still remain, such as whether a single value of the confidence factor (f) can be employed to flag the same fraction of data on a region-wide basis. This study explores how the general relationship between a specified confidence interval and the fraction of total records flagged depends on the value of f and whether or not the effect of the f value on flagging frequency varies geographically.

MATERIALS AND METHODS

Data

The data from stations within the NOAA Cooperative Observer Weather Data Network (National Weather Service, 2000), a regional automated weather data network (Hubbard, 2001), and other networks such as the Hourly Surface Airways Network, and the Historical Climatology Network were retrieved through the Applied Climate Information System (ACIS, Hubbard *et al.*, 2004), a distributed data management system. This study includes estimation of the maximum air temperature ($T_{\max}$) and minimum air temperature ($T_{\min}$) for the time period 1971–2000.

The sensitivity analysis of the SRT 'f' factor to the fraction of data flagged and the reset fraction by the modified SRT method were carried out over the states Colorado, Wyoming, North Dakota, South Dakota, Nebraska, Kansas, and Iowa, for the 30-year period 1971–2000. The inclusion of a long time period in the analysis reduces the potential for the relationship to be dominated by a particular pattern from a specific year.

The comparisons between the original SRT method and the new SRT method (SRT2) were carried out for all continental stations of the USA in the year 2002 for both $T_{\max}$ and $T_{\min}$. The results obtained by both methods are compared by examining widely used statistical indices such as R^2, RMSE, and Nash and Sutcliffe coefficient of efficiency (NSCE, Nash and Sutcliffe, 1970).

Currently NCDC and other regional climate centers in the United States of America archive the $T_{\max}$ and $T_{\min}$ in degrees Fahrenheit. To be consistent with the widespread use of this data in Fahrenheit and consistent with the database, we use degrees Fahrenheit in this paper.

The station density is limited by the availability of stations in the existing network(s). Many stations have a distance of 20 km or less to the nearest stations. In less populated areas like mountains, deserts, or wetlands, some stations have a distance of separation of 50 km or greater.

Methods

Three SRT methods were employed in various stages of this study: the original SRT (Hubbard *et al.*, 2005), the enhanced SRT (You and Hubbard, 2006), and a new method described below (SRT2). You and Hubbard (2006) introduced a technique to reset 'false' flags, thereby enhancing the ability of the SRT method to reset the flags of the 'good' data during cold front passage. The technique (enhanced SRT) introduces a new parameter 'f'' which essentially affects the percentage of reset flags out of the total flags. In this study, a new version of the SRT2 directly weights the estimates obtained from surrounding stations using the inverse RMSE(s) between the measured time series of the station and the reference stations:

$$x' = \sum_{i=1}^{N} x_i s_i^{-2} \Big/ \sum_{i=1}^{N} {s_i}^{-2}. \tag{1}$$

Where x' is the estimate for station of interest, N is the number of neighboring stations used in the estimation, i is the index of reference station, and x_i is the regression estimate. When temperatures in an area hover around zero, this approach (SRT2) inherently preserves the correct sign on the final estimate in contrast to the original SRT method (Hubbard *et al.*, 2005) wherein the sign required a separate assessment. In the SRT2 method, the determining factors are the root mean square error between the station of interest and the neighboring stations and the sign of the observations at neighboring stations.

RESULTS

Estimates obtained using the original SRT method and the new SRT2 method at all stations for the year 2002 were evaluated for their precision across continental USA for both $T_{\max}$ and $T_{\min}$. The year 2002 was the latest year with complete data when we started the research and it is assumed that the weather over the USA for this period presents an adequate test environment for the comparisons shown. The R^2, NSCE, and RMSE were calculated separately between the actual observations and the estimates obtained for both the original SRT and SRT2 methods. The difference of R^2, NSCE, and RMSE between the original SRT and the new SRT2 method was calculated for all stations and the distribution of these differences are plotted for all continental stations in the USA. More details on these indices are given in Hubbard and You (2005). In this comparison, the difference between the two methods is negligible when the difference of R^2 and NSCE is smaller than 0.01 and the difference of RMSE is less than 0.1 F. Nearly all stations have a negligible difference between the indices as calculated using the original and new SRT2 method (Figure 1). Thus the methods are interchangeable in estimating $T_{\max}$ and $T_{\min}$. A sensitivity analysis of the original SRT to f was conducted and because SRT and SRT2 were found to be interchangeable, a separate analysis for SRT2 is not presented here. The results of the study could be tested using more complex statistical methods, e.g. T-test and Variance test; however, these tests are not necessary since the differences of RMSE are smaller than 0.1 F for nearly all stations (i.e. less than typical measurement errors of perhaps 0.5 F).

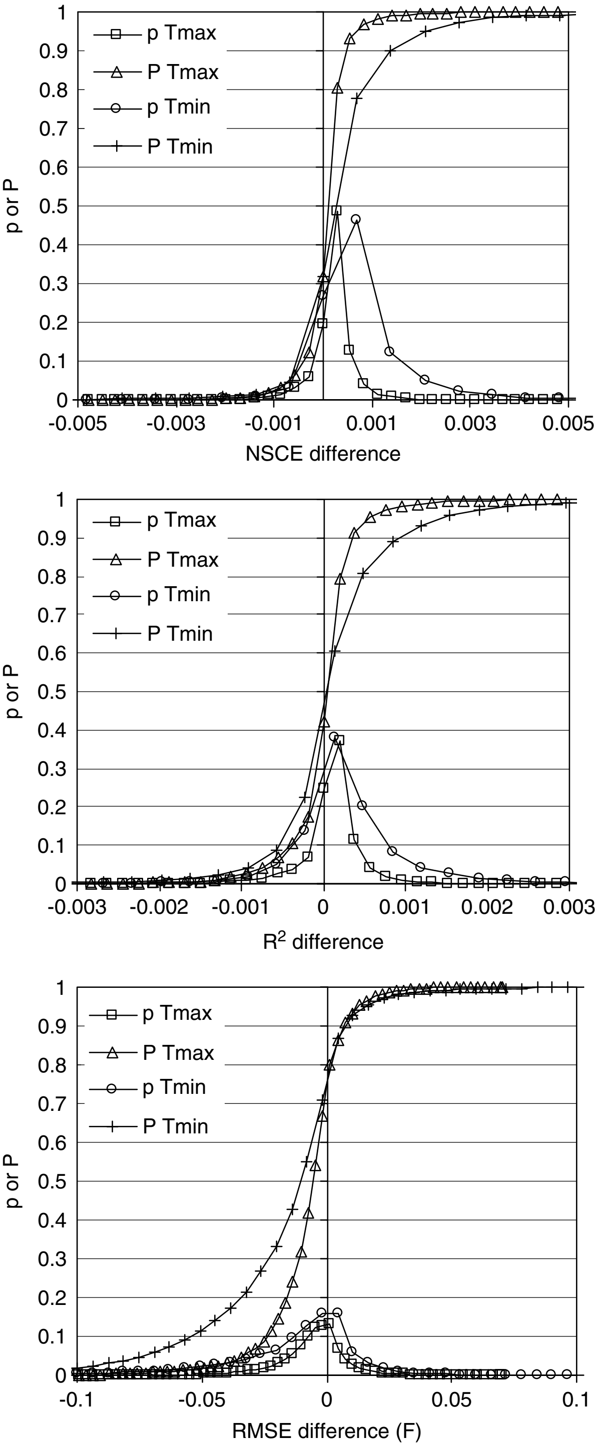

Figure 1. The differences on the x-axis are between the NSCE (top), R^2 (middle), and RMSE (bottom) obtained from the SRT and SRT2 methods (e.g. R^2(SRT)- R^2(SRT2)). Both the frequency distributions (p) of the differences and the proportion of stations (P) having differences smaller than the given value are shown.

Climatologists and data validators are concerned with the effort required to validate the weather data. Because the number of 'bad' values in the $T_{\max}$ and $T_{\min}$ measurements is unknown *a priori*, a totally manual operation requires that all data be assessed. However, with automated QC it may be possible to single out a subset of data that requires manual checking and thus reduce the workload while achieving the same result. It would be desirable to reduce the values that must be manually checked to say 1% of the incoming data stream and this value was taken as a guide for selection of the f value in

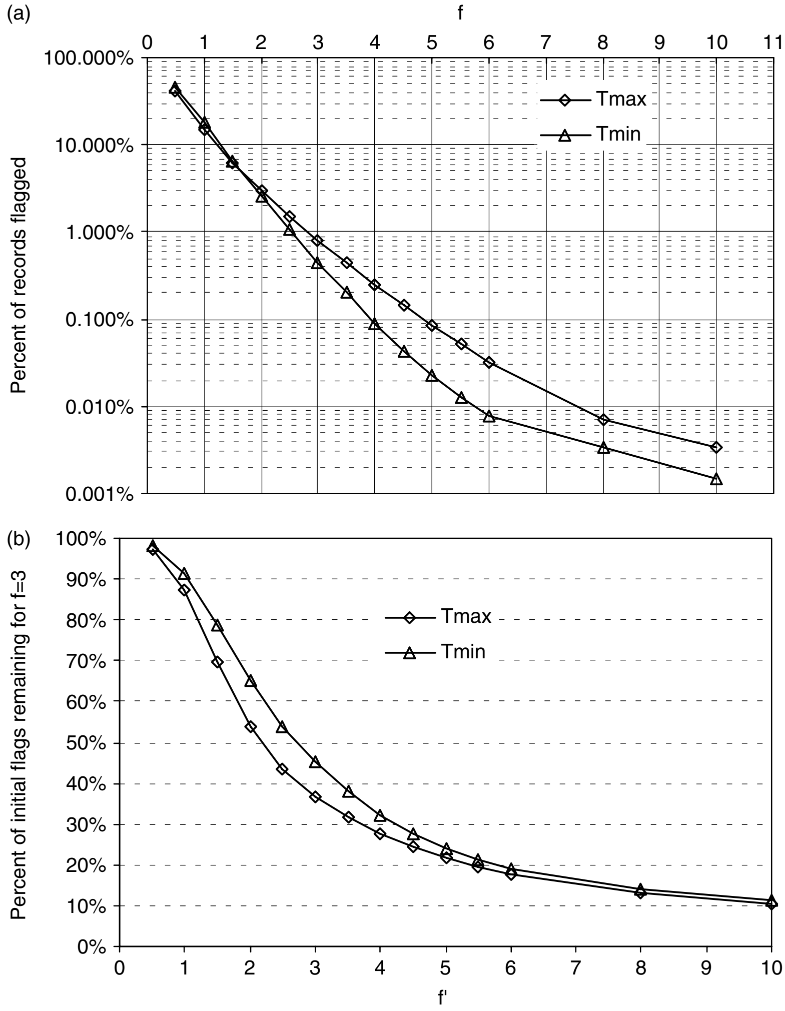

Figure 2. Percent of (a) original data flagged (SRT) as a function of f and (b) original flags remaining after employing the enhanced SRT as a function of f'.

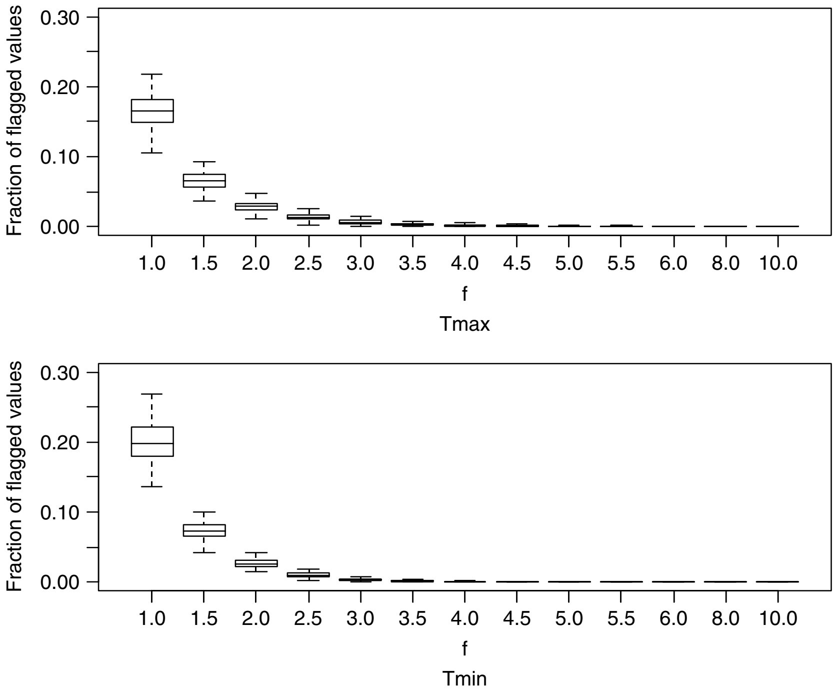

Figure 3. Box plot of the fraction of data flagged by original SRT at individual stations for different f values for $T_{\max}$ and $T_{\min}$. The median fraction is indicated by the black center line, and the upper and lower edges of the box represent the interquartile range (IQR). The extreme values (within 1.5 times the IQR from the upper or lower quartile) are the whiskers extending from the IQR.

DOI: 10.1002/joc

the SRT method. For the High Plains stations, the number of data flagged for the SRT method as a function of f was determined using the data period from 1971 to 2000. The average fraction of data flagged over all stations for each value of the confidence factor for both $T_{\max}$ and $T_{\min}$ were determined (see Figure 2(a)). Figure 3 presents the box plot of the proportion of flagged data out of the total measurements based on all stations. The fraction of data flagged decreases considerably when f is greater than 2.5. The mean of the fraction of data flagged is less than 1% when f is greater than 3.0. A few stations were found to exceed the whiskers in the box plot (not plotted). These stations require further examination to determine why they do not fit within the whiskers. Some possibilities are: (1) those below the bottom whisker are the stations of highest quality or are associated with exceptionally high spatial correlation structure, (2) those above the top whisker may have underlying problems with station equipment or observing practices, or (3) some physical factor is at work. Until a determination can be made it may be necessary to use a station specific f value for these stations.

After the reset procedure (You and Hubbard, 2006) was employed with given f' values for both $T_{\max}$ and $T_{\min}$, there was a reduction in flagged data values as shown in Figure 2(b). In this analysis of the effect of f' on the resetting of flags, the initial flagging was accomplished with $f = 3.0$ so that no more than 1% of the data would be initially identified as potential outliers. Several stations were found to fall beyond the whiskers (not plotted) in the f' box plots (Figure 4). However, the fractions of data flagged were small compared to 1%. Two possible reasons may cause the stations to fall outside the whiskers: (1) the flags are not reset by the simple rules at some locations; (2) the low station density around some stations leads to a breakdown in the correlation structure and thus not enough information pertaining to temperature changes from the surrounding stations. For example, some outliers may be associated with identified flags located in relative isolation from the influence zone of the cold fronts shown in Figure 10 of You and Hubbard (2006).

Usually the manual checking rate for the automatic QA procedures is not expected to exceed some percentage, e.g. 1% of the measurements. Here, the confidence factor, f, was calculated for each station for 1% of the data flagged for both $T_{\max}$ and $T_{\min}$. The box plot of the f value was created for each state (Figure 5). The mean values of the f value for 1% of the data flagged are slightly lower than 3.0 for $T_{\max}$ and are slightly higher than 2.5 for $T_{\min}$. We found the daytime air temperature has higher variation than the nighttime temperature. The differences of f values for $T_{\max}$ and $T_{\min}$ may be caused because long wave radiation is not the only factor driving temperature during the daytime. Solar radiation and albedo are also factors during daytime. In addition, mixing and advection are more prevalent than at nighttime. During nighttime, long wave radiation balance may lead to a slightly higher correlation between stations than during daytime. In some instances, to flag 1% of the data at some stations requires f values beyond the whiskers in some states like Colorado and Kansas. The variance of f for each state is small with a value around 0.1, which is also the case for Colorado if the largest outlier in both $T_{\max}$ and $T_{\min}$ are excluded. Based on this

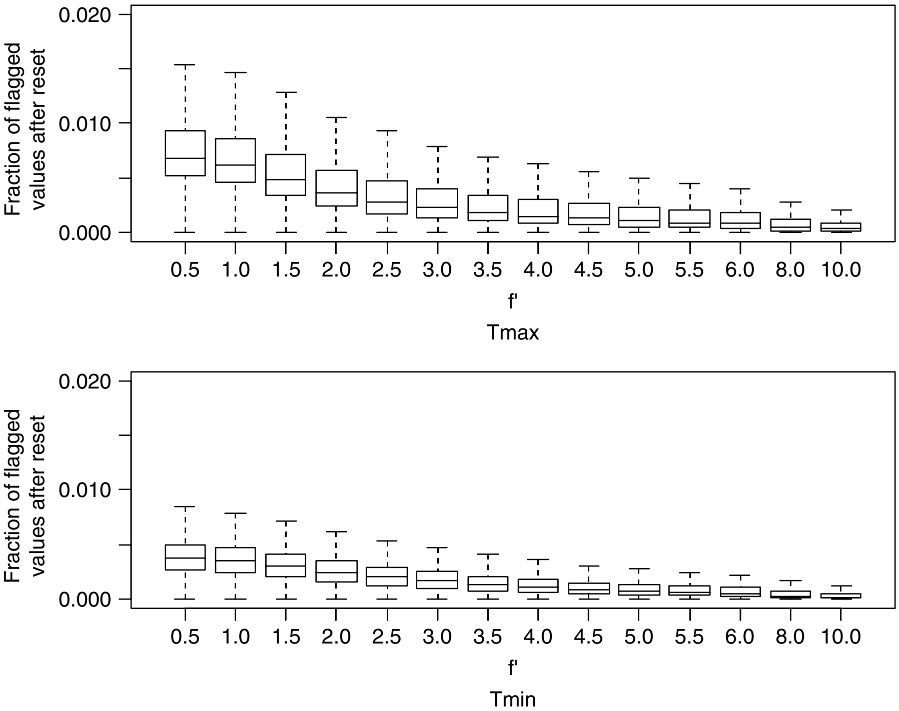

Figure 4. Box plot of the fraction of remaining flagged values at individual stations after employing the enhanced SRT and different f' values during the period 1971 to 2000 when f equal to 3 for both $T_{\max}$ and $T_{\min}$.

DOI: 10.1002/joc

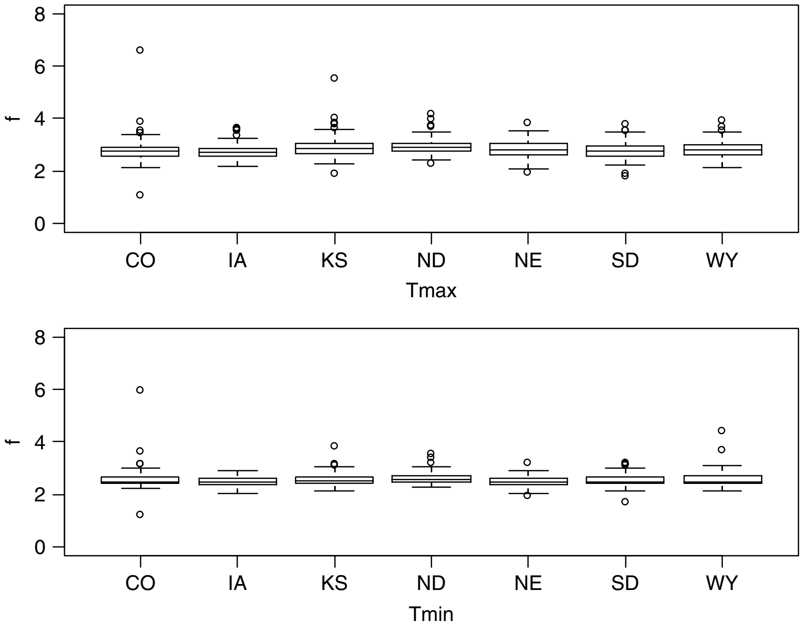

Figure 5. Box plot for all stations in each state of the f values that resulted in an original SRT flagging of 1% of the data for both $T_{\max}$ and $T_{\min}$. Circle symbols represent stations that do not fit within the whiskers and may require further study.

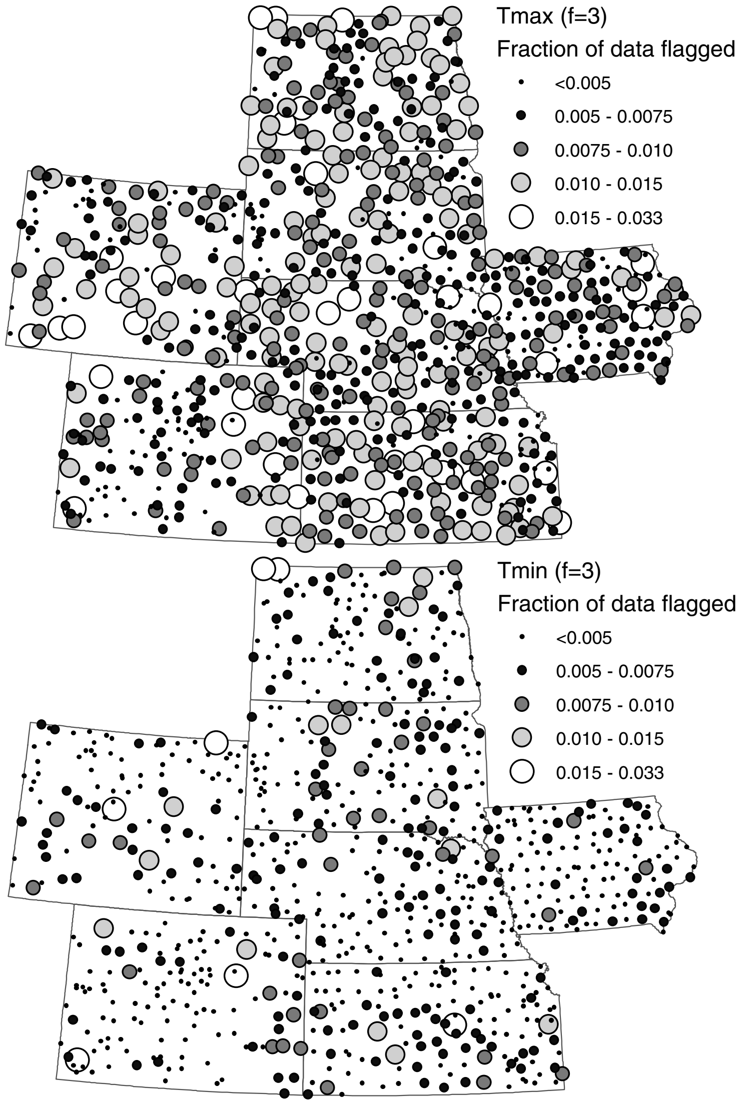

Figure 6. Fraction of data flagged by the SRT for f equal to 3.0 during 1971 ~ 2000 for both $T_{\max}$ and $T_{\min}$.

DOI: 10.1002/joc

analysis, for the ACIS database an f value of 3.0 will be used for both $T_{\max}$ and $T_{\min}$ at all High Plains states. This will result in roughly 1% of the records being flagged for $T_{\max}$ and around half percent for $T_{\min}$.

The QA tests of $T_{\max}$ and $T_{\min}$ with original SRT were also conducted using an f value of 3.0 (see Figure 6). Most stations have flagged fractions between 0.5 to 1.5% of the total records for $T_{\max}$ and flagged fractions less than 0.5% for $T_{\min}$. The flagged fraction for $T_{\min}$ is generally lower than that of $T_{\max}$. In the plot, we cannot identify strong patterns of the f value by regions. The areas with sparsely distributed stations have a relatively higher fraction of data flagged than the areas with densely distributed stations, although occasionally a relatively higher fraction may exist for the densely distributed stations. In general, we suggest that a single value of f for a state is acceptable for the QA of $T_{\max}$ and $T_{\min}$.

DISCUSSION AND CONCLUSIONS

A new version of the SRT method was developed to eliminate the need for special procedures to determine the sign of the estimates, which was a shortcoming of the original SRT method. The results demonstrate that the two approaches are interchangeable for the 48 conterminous states. The differences obtained by the two methods are negligible and thus utilization of the modified method, which is simple, is recommended.

Sensitivity analysis was conducted for the f value of the SRT method and for the f' value associated with the technique to reset flags, e.g. in the event of a cold front passage where excessive ‘false flagging’ is initially prevalent. A value of 2.5 or larger for the f value is more suitable than using smaller threshold values, considering the relatively small number of potential outliers, which need further manual examination. Different f values can be applied to suit different utilizations of the climate data by setting corresponding confidence levels for the SRT method, e.g. 90% for one application and 95% for another. The estimates obtained using the SRT method or other methods, e.g. inverse distance weighted method, can replace the identified outliers or missing data to complete the dataset in specific model applications.

This study suggests that in the High Plains, a uniform confidence factor of, e.g. 3.0, for the SRT QA procedure can give a reasonable number of flagged data for further manual validation. Similar work can be implemented in other states and regions to determine any regional differences.

The SRT method uses linear regression and can be implemented using different time intervals over which the regressions are formed. This approach is relatively independent of future temperature trends because long-term memory of weather data is not incorporated in the estimates. One more concern may arise with the fraction of data flagged when more data sources are available. If additional networks are used in the QC process, factors like instrument type and sampling strategy may affect the correlation between stations and could lead to more flags. This may not be a serious problem because only the stations with the highest correlations are used in SRT or SRT2. It appears more likely to the authors that the effect of increased density of stations may lead to higher correlation between stations and in turn a lower fraction of data flagged. Of course, to realize the higher correlation the station data must be brought in individually and not as gridded data. For example, the performance of SRT2 may be greatly improved and thus flag fewer valid data when more stations are installed in mountainous regions. Therefore, more data entries are expected to be validated when additional networks are used for QC in the future.

References

Barnett V, Lewis T. 1994. *Outliers in Statistical Data*, 3rd edn. Wiley and Sons: New York; 584.

Belcher BN, DeGaetano AT. 2005. A method to infer time of observation at US Cooperative Observer Network stations using model analyses. *International Journal of Climatology* **25**(9): 1237–1251.

Eischeid JK, Baker CB, Karl T, Diaz HF. 1995. The quality control of long-term climatological data using objective data analysis. *Journal of Applied Meteorology* **34**: 2787–2795.

Gandin LS. 1988. Complex quality control of meteorological observations. *Monthly Weather Review* **116**: 1137–1156.

Guttman NV, Quayle RG. 1990. A review of cooperative temperature data validation. *Journal of Atmospheric and Oceanic Technology* **7**: 334–339.

Hubbard KG. 2001. The Nebraska and High Plains Regional experience with automated weather stations. *Automated Weather Station for Application in Agriculture and Water Resources Management*. High Plains Regional Climate Center: Lincoln, NE, AGM-3 WMO/TD No. 1074; 248.

Hubbard KG, You J. 2005. Sensitivity analysis of quality assurance using spatial regression approach – a case study of the maximum/minimum air temperature. *Journal of Atmospheric and Oceanic Technology* **22**(10): 1520–1530.

Hubbard KG, DeGaetano AT, Robbins KD. 2004. Announcing a modern applied climatic information system (ACIS). *Bulletin of the American Meteorological Society* **85**(6): 811–812.

Hubbard KG, Goddard S, Sorensen WD, Wells N, Osugi TT. 2005. Performance of quality assurance procedures for an applied climate information system. *Journal of Atmospheric and Oceanic Technology* **22**: 105–112.

Meek DW, Hatfield JL. 1994. Data quality checking for single station meteorological databases. *Agricultural and Forest Meteorology* **69**: 85–109.

National Weather Service. 2000. *Cooperative Observer Program (COOP)*. National Weather Service, Silver Spring, MD. (www.nws.noaa.gov/om/coop/Publications/coop.PDF).

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models. *Journal of Hydrology* **10**: 282–290.

Reek T, Doty SR, Owen TW. 1992. A deterministic approach to the validation of historical daily temperature and precipitation data from the Cooperative Network. *Bulletin of the American Meteorological Society* **73**: 753–762.

Shafer MA, Fiebrich CA, Arndt DS, Fredrickson SE, Hughes TW. 2000. Quality assurance procedures in the Oklahoma mesonetwork. *Journal of Atmospheric and Oceanic Technology* **17**: 474–494.

Wade CG. 1987. A quality control program for surface mesometeorological data. *Journal of Atmospheric and Oceanic Technology* **4**: 435–453.

Wu H, Hubbard KG, You J. 2005. Some concerns when using national weather service’s daily surface observations: a Nebraska case study. *Journal of Atmospheric and Oceanic Technology* **22**: 592–602.

You J, Hubbard KG. 2006. Quality control of weather data during extreme events. *Journal of Atmospheric and Oceanic Technology* **23**(2): 184–197.

DOI: 10.1002/joc