Performance of Quality Assurance Procedures for an Applied Climate Information System

K. G. HUBBARD

High Plains Regional Climate Center, University of Nebraska–Lincoln, Lincoln, Nebraska

S. GODDARD

Computer Science and Engineering, University of Nebraska–Lincoln, Lincoln, Nebraska

W. D. SORENSEN

High Plains Regional Climate Center, University of Nebraska–Lincoln, Lincoln, Nebraska

N. WELLS AND T. T. OSUGI

Computer Science and Engineering, University of Nebraska–Lincoln, Lincoln, Nebraska

(Manuscript received 29 January 2004, in final form 29 April 2004)

ABSTRACT

Valid data are required to make climate assessments and to make climate-related decisions. The objective of this paper is threefold: to introduce an explicit treatment of Type I and Type II errors in evaluating the performance of quality assurance procedures, to illustrate a quality control approach that allows tailoring to regions and subregions, and to introduce a new spatial regression test. Threshold testing, step change, persistence, and spatial regression were included in a test of three decades of temperature and precipitation data at six weather stations representing different climate regimes. The magnitude of thresholds was addressed in terms of the climatic variability, and multiple thresholds were tested to determine the number of Type I errors generated. In a separate test, random errors were seeded into the data and the performance of the tests was such that most Type II errors were made in the range of ± 1°C for temperature, not too different from the sensor field accuracy. The study underscores the fact that precipitation is more difficult to quality control than temperature. The new spatial regression test presented in this document outperformed all the other tests, which together identified only a few errors beyond those identified by the spatial regression test.

1. Introduction

The quality assurance (QA) procedures discussed herein were developed and applied to the data systems of the National Oceanic and Atmospheric Administration's Regional Climate Centers. The National Climatic Data Center (NCDC) began semiautomated review of the data validation for the cooperative climatological stations in 1982 (Guttman and Quayle 1990). Although NCDC's validation process became somewhat automated, many data continue to be inspected manually (Guttman et al. 1988).

Generally, there are two categories of tests: those that use data from a single site (Meek and Hatfield 1994) and those that use data from multiple sites. The second type compares a station's data against neighboring stations' (Hubbard 2001; Reek et al. 1992). Statistical decisions play a large role in quality control efforts, but increasingly there are rules introduced that depend upon the physical system involved. Examples of these are the testing of hourly solar radiation against the clear sky envelope (Allen 1996; Geiger et al. 2002) and the use of soil heat diffusion theory to determine soil temperature validity (Hu et al. 2002). It is now realized that QA is best suited when made a seamless process between staff operating the quality control software at a centralized location where data is ingested and technicians in the field (Hubbard 2001; Shafer et al. 2000).

Quality assurance software consists of procedures or rules against which data are tested. Each procedure will either accept the datum as being true or reject the datum and label it as an outlier. This hypothesis (H_o) testing of each datum and the statistical decision to accept the datum or to note it as an outlier can have the outcomes

Corresponding author address: Kenneth G. Hubbard, High Plains Regional Climate Center, 244 Chase Hall, University of Nebraska–Lincoln, Lincoln, NE 68583-0728.
E-mail: khubbard1@unl.edu

© 2005 American Meteorological Society

TABLE 1. The error classification in testing of a quality assurance hypothesis.

Statistical decision	True situation	
	H_o true	H_o false
Accept H_o	No error	Type II error
Reject H_o	Type I error	No error

shown in Table 1. If the datum is valid and is accepted as such or the datum is invalid and rejected, the QA procedure is working appropriately. When the datum is valid and is rejected by QA, a Type I error is committed. If the datum is not valid but is accepted by QA, a Type II error is committed.

Take the simple case of testing a variable against limits. Suppose that the hypothesis is that a datum for a measured variable is valid only if it lies within ± 3 standard deviations (σ) of the mean (μ), then, assuming a normal distribution, the expectation is that H_o will be accepted 99.73% of the time with no error. The values that lie beyond $\mu \pm 3\sigma$ will be rejected with a resulting Type I error if valid values are encountered beyond these limits. In these cases (H_o is rejected when the value is actually valid) the expectation is that a Type I error will be made 0.27% of the time, assuming for this discussion that the data have no errant values. If a ''true'' value is replaced with an ''errant'' value, then the hypothesis will properly be rejected, only if the ''errant'' value falls outside the range $\mu \pm 3\sigma$. It would otherwise be accepted, when it actually is false (the value is not valid), and this would lead to a Type II error. In this simple example, reducing the limits against which the data values are tested will produce more Type I errors and fewer Type II errors, while increasing the limits leads to fewer Type I errors and more Type II errors. For QA software, study is necessary to achieve a balance wherein one reduces the Type II errors (mark more ''errant'' data as having failed the test) while not increasing Type I errors to the point where valid extremes are brought into question. Because Type I errors cannot be avoided, it is prudent for data managers to always keep the original measured values regardless of the quality testing results.

In this manuscript we point to three major contributions. The first is the explicit treatment of Type I and Type II errors in the evaluation of the performance of quality control procedures to provide a basis for intercomparison of procedures. The second is to illustrate how the selection of parameters in the quality control process can be tailored to individual needs in regions or subregions of a widespread network. Finally, we introduce a new spatial regression test that uses a subset of the neighboring stations that provide the ''best fit'' to the target station. The spatial regression weighted estimate has characteristics that make it possible to build statistical confidence intervals for testing data at the target station.

2. Data and methodology

The tests performed in this study were conducted for six stations. These six stations are part of the cooperative weather observer network (TD3200 dataset at NCDC). Table 2 shows the location, the average annual maximum and minimum temperatures, the annual total precipitation, and the elevation of each station. The stations were chosen to represent different climate regimes. Crete, Nebraska, and Dickinson, North Dakota, are two sites in the High Plains where the latter is cooler and drier and at a higher elevation. Fort Myers and Key West, Florida, are both warm sites located in the vicinity of the Gulf of Mexico, although the latter is completely surrounded by water and the former is on the west side of the Florida peninsula. Tucson, Arizona, has a warm and dry climate, while Yellowstone Lake, Wyoming, has a cooler climate. Both Tucson and Yellowstone Lake are located in more complex terrain (deeper ridges and valleys) than the other sites (flatter terrain). The elevation range is from near zero at Key West to nearly 2400 m at Yellowstone Lake.

This study uses four procedures. Three tests are tuned to the prevailing climate: seasonal thresholds, seasonal rate of change, and seasonal persistence. The thresholds and limits for these tests are related to station climatology at the monthly level (period 1971–2000) as compared to previous efforts, which mainly used one set of limits for a variable, regardless of time of year (Shafer et al. 2000; Hubbard 2001). The fourth test is a spatial comparison, using linear regression to estimate confidence intervals for the station in question. Only valid

TABLE 2. The location and climate of weather stations included in this study.

Location	Lat (°N)	Lon (°W)	Avg annual max temperature (°C)	Avg annual min temperature (°C)	Avg annual total precipitation (mm)	Station elevation (m)
Crete, NE	40.62	96.95	17.3	4.9	738	437
Dickinson, ND	46.80	102.80	12.5	−0.4	415	750
Fort Myers, FL	26.59	81.86	29.2	18.4	1376	5
Key West, FL	24.55	81.75	28.3	22.9	989	1
Tucson, AZ	32.23	110.95	28.1	12.7	309	777
Yellowstone Lake, WY	44.56	110.40	7.6	−8.2	518	2399

(nonmissing) data are exposed to the tests described below.

The *''upper and lower'' threshold* test checks whether a given variable (e.g., daily maximum temperature) falls in a specific range for the month in question. This test has been in use for some time. Where relatively new stations are involved the threshold test is often employed by considering the climate extremes for the area (Shafer et al. 2000). When the limits are determined based on the statistics of the distribution it has been called the sigma test (Guttman et al. 1988). The threshold test for variable x is

$$\overline{x} - f\sigma_x \le x \le \overline{x} + f\sigma_x, \tag{1}$$

where $\overline{x}$ is the daily mean (e.g., mean of daily maximum, 30×31 days for January) and σ_x is the standard deviation of the daily values (e.g., daily maximum values) for the month in question. The variable x may represent maximum temperature, minimum temperature, or rainfall. An analysis was performed on the data (1971–2000) to determine the relationship between the ''percent of data passing'' the test and various values of f. This procedure allows an informed choice regarding how many data points will be flagged in the natural datastream. If the datastream contained no errors, the values not passing would be Type I errors. In operational use, the data so flagged as potential Type I errors will be considered suspect and subjected to further manual checking, so a realistic determination of f is critical to project staff requirements. Graphs were developed to display the potential Type I errors versus f for the threshold test.

The *step change (SC)* test checks to see whether or not the change in consecutive values of the variable fall within the climatologically expected lower and upper limits on daily rate of change for the month in question. In this case the step is defined as the difference between values on day i and $i - 1$, for example, $x_i = d_i - d_{i-1}$. Utilizing this definition of x and calculating the associated mean and the variance allows Eq. (1) to again be used, and an analysis of the data (1971–2000) determines the relationship between f and the potential Type I errors for the SC test.

The *persistence* test checks the variability of the measurements. When a sensor fails it will often report a constant value; thus the standard deviation (σ) will become smaller, and if the sensor is out for an entire reporting period, σ will be zero. In other cases the instrument may work intermittently and produce reasonable values interspersed with zero values, thereby greatly increasing the variability for the period. Thus, when the variability is too high or too low the data should be flagged for further checking. The first step is to calculate the standard deviation from daily values for each month (j) and year (k) of the 30-yr record, σ_{jk}. Then the mean standard deviation is calculated for each month $\overline{\sigma_j}$ by averaging σ_{jk} over the years. Likewise, the standard deviation of these monthly values (σ_{jk}) is calculated over all years $\sigma_{j\sigma}$ is calculated. The persistence test compares the standard deviation for the time period being tested to the limits expected as follows:

$$\overline{\sigma_j} - f\sigma_{j\sigma} \le \sigma \le \overline{\sigma_j} + f\sigma_{j\sigma}. \tag{2}$$

The period under consideration passes the persistence test if the above relation holds for the specified value of f.

The *spatial weighted regression* test checks that the variable falls inside the confidence interval formed from estimates based on N ''best fit'' neighboring stations during a time period of length n, which is taken as 24 for this study, and N was set to 5. The surrounding stations were selected by specifying a radius around the station and finding those stations with the closest statistical agreement to the target station. This was taken as 50 km for all stations with the exception of Key West (150 km) and Yellowstone Lake (100 km). These latter stations were in lower-station-density areas, thus prompting larger radii. Additional requirements for station selection were that the variable to be tested is one of the variables measured at the candidate site and the data for that variable span the data period to be tested. A station that otherwise qualifies could also be eliminated from consideration if more than half of the data is missing for the time span (more than 12 missing days).

First preliminary estimates x_{lt} are derived by use of the coefficients derived from linear regression, so for any time t and for each surrounding station (y_{lt}) an estimate is formed:

$$x_l = a_l + b_l y_l.$$

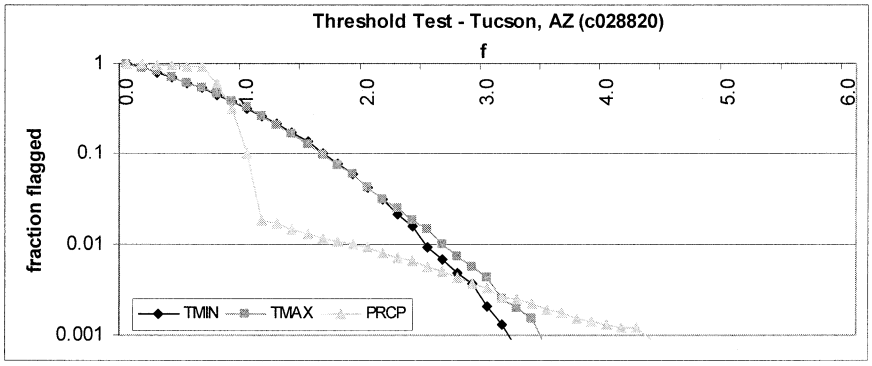
The new approach here is to obtain a weighted estimate (x') by utilizing the standard error of estimate (s) for each of the linear regressions (also known as root-mean-square error) in the weighting process. The surrounding stations are ranked according to the magnitude of the standard error of estimate, and the N stations with the lowest s values are used in the weighting process; in this case N is taken as 5:

$$x'^2 = \frac{\sum_{i=1}^{N} \frac{x_i^2}{s_i^2}}{\sum_{i=1}^{N} \frac{1}{s_i^2}}. \tag{3}$$

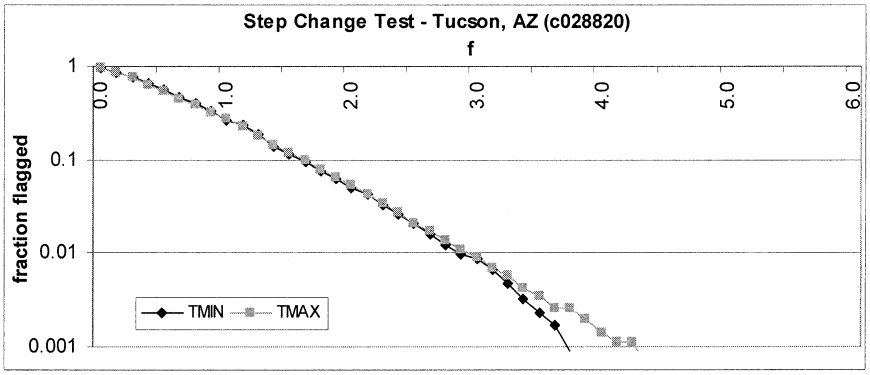
This new approach provides more weight to the stations that have the best fit with the target station. Because the stations used in (3) are a subset of the neighboring stations we maintain that the estimate is not an areal average but a spatial regression weighted estimate. Care must be taken to preserve the correct sign on x'. The weighted standard error of estimate (s') is calculated from

$$\frac{1}{s'^2} = \frac{\sum_{i=1}^{N} \frac{1}{s_i^2}}{N}. \tag{4}$$

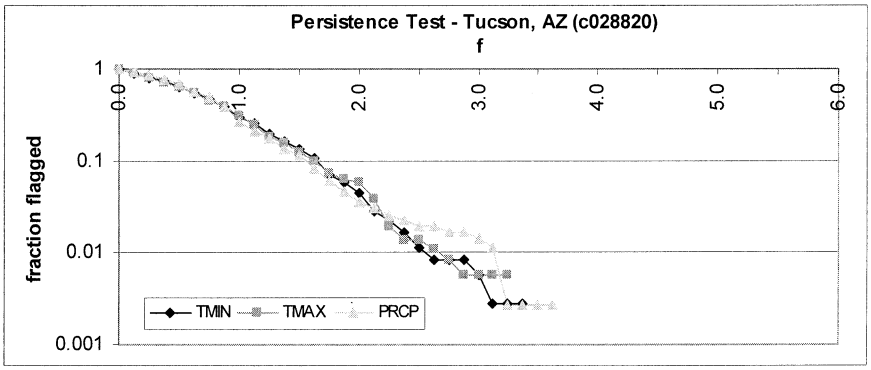
(a)


Threshold Test - Tucson, AZ (c028820)
f
fraction flagged
1
0.1
0.01
0.001
0.0
1.0
2.0
3.0
4.0
5.0
6.0
TMIN
TMAX
PRCP

(b)


Step Change Test - Tucson, AZ (c028820)
f
fraction flagged
1
0.1
0.01
0.001
0.0
1.0
2.0
3.0
4.0
5.0
6.0
TMIN
TMAX

(c)


Persistence Test - Tucson, AZ (c028820)
f
fraction flagged
1
0.1
0.01
0.001
0.0
1.0
2.0
3.0
4.0
5.0
6.0
TMIN
TMAX
PRCP

(d)

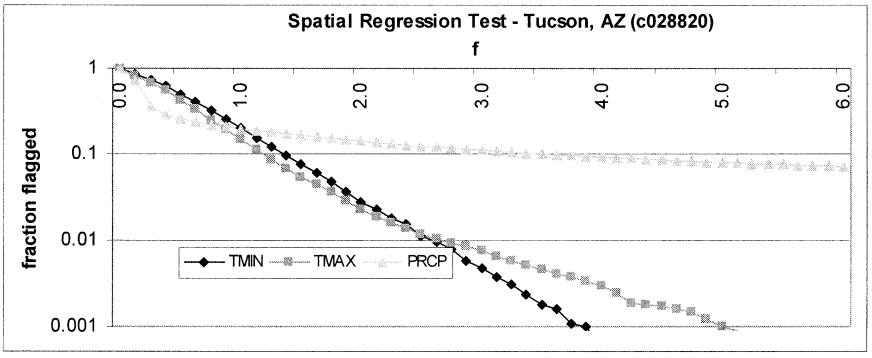

Spatial Regression Test - Tucson, AZ (c028820)
f
fraction flagged
1
0.1
0.01
0.001
0.0
1.0
2.0
3.0
4.0
5.0
6.0
TMIN
TMAX
PRCP

TABLE 3. The fraction (Type I) of maximum temperature data flagged (%) at $f = 3$ for each test and each site.

	Flagged (%) for max temperature, $f = 3$			
Station	Threshold	Step change	Persistence	Spatial regression
Tucson, AZ	0.4	0.9	0.6	0.8
Fort Meyers, FL	1.1	1.1	0.6	1.8
Key West, FL	1.0	1.4	0.6	1.9
Crete, NE	0.3	1.1	0.3	1.8
Dickinson, ND	0.2	0.7	0	2.1
Yellowstone Lake, WY	0.4	0.9	0.3	3.3

TABLE 4. The fraction (Type I) of minimum temperature data flagged (%) at $f = 3$ for each test and each site.

	Flagged (%) for min temperature, $f = 3$			
Station	Threshold	Step change	Persistence	Spatial regression
Tucson, AZ	0.2	0.9	0.6	0.5
Fort Meyers, FL	0.6	0.8	0.6	1.7
Key West, FL	0.5	0.9	0	1.4
Crete, NE	0.3	0.5	0.6	1.7
Dickinson, ND	0.3	0.6	0	0.9
Yellowstone Lake, WY	0.5	0.9	0	3.3

This approach differs from inverse distance weighting in that the standard error of estimate has a statistical distribution; therefore, confidence intervals can be calculated on the basis of s' and the station value (x) can be tested to determine whether or not it falls within the confidence intervals:

$$x' - fs' \leq x \leq x' + fs'. \quad (5)$$

If the above relationship holds, then the datum passes the spatial test. This relationship indicates that with successively larger values of f, the number of potential Type I errors decreases. Unlike distance weighting techniques, this approach does not assume that the best station to compare against is the closest station but instead looks to the relationships between the actual station data to settle which stations should be used to make the estimates and what weighting these stations should receive.

Using the above methodology, the rate of error detection can be preselected. The reader should note that the results are presented in terms of the fraction of data flagged against the range of f values (defined above) rather than selecting one f value on an arbitrary basis. This type of analysis makes it possible to select the specific f values for stations in differing climate regimes that would keep the Type I error rate uniform across the country. For example, for sake of illustration, suppose the goal is to select f values that keep the potential Type I errors to about 2%. A representative set of stations and years can be preanalyzed prior to QC to determine the f values appropriate to achieve this goal.

To document the performance of the various procedures in a controlled situation, a set of ''seeded'' errors were introduced to the datasets and the performance of the various procedures in regard to catching these errors were recorded. By a random process, 2% of the dates were selected to receive a ''seeded'' error.

The magnitude of the error was determined in a random manner. A random number, r, was selected using a random number generator operating on a uniform distribution with a mean of zero and a range of ± 3.5. This number was then multiplied by the standard deviation of the variable in question to obtain the error magnitude:

$$E_{ix} = \sigma_x r_i.$$

Thus, the expected distribution of the error magnitudes has a mean of zero and a range equal to 3.5 times that of the observed standard deviation of the variable (σ_x). The selection of 3.5 ensures that the tests include cases that are close to the extremes of the period 1971–2000.

←

FIG. 1. Results from (a) the threshold test showing the fraction of all data flagged (1971–2000) for maximum temperature (TMAX), minimum temperature (TMIN), and precipitation (PRCP) for the values of f shown; (b) the step change test showing the fraction of all data flagged (1971–2000) for maximum temperature (TMAX) and minimum temperature (TMIN) for the values of f shown; (c) the persistence test showing the fraction of all data flagged (1971–2000) for maximum temperature (TMAX), minimum temperature (TMIN), and precipitation (PRCP) for the values of f shown; (d) the spatial regression test showing the fraction of all data flagged (1971–2000) for maximum temperature (TMAX), minimum temperature (TMIN), and precipitation (PRCP) for the values of f shown.

TABLE 5. The fraction (Type I) of precipitation data flagged (%) at $f = 3$ for each test and each site.

Station	Flagged % for precipitation, $f = 3$			
	Threshold	Step change	Persistence	Spatial regression
Tucson, AZ	0.3	–	1.4	11.0
Fort Meyers, FL	0.7	–	0.8	10.9
Key West, FL	0.6	–	2.0	11.4
Crete, NE	0.5	–	1.1	10.9
Dickinson, ND	0.5	–	1.7	8.8
Yellowstone Lake, WY	0.9	–	0.8	9.8

The results of running the procedures on the ''modified'' dataset were cataloged for those days on which errors were introduced. The fraction of errors caught by each procedure was compared across the range of error magnitudes introduced.

3. Results

a. Type I errors

It is important to examine the number of potential Type I errors that would occur when using the specified procedures with various f factors. The general shape of the relationship between f and the fraction of data flagged is shown in Figs. 1a–d. Although we show the fraction on a log scale, the results obtained here have a resemblance to the results of Eischeid et al. (1995), although their work was with respect to monthly data and the interquartile range. The result for the threshold analysis at Tucson indicates that approximately 2% of the data would be flagged for maximum and minimum temperature if f values of 2.4 and 2.3 are used, respectively. For precipitation, 2% of the data were flagged in the threshold test for an f value of 1.13. These results are shown in Fig. 1a. Similar figures for Tucson are shown for the step change test (Fig. 1b), the persistence test (Fig. 1c), and the spatial test (Fig. 1d). Other stations show similar relationships between ''fraction of data flagged'' and f. Zero is not shown on the vertical axes of Figs. 1a–d, but where the curves have an endpoint inside the box, there were no values flagged by the test beyond that point (e.g., for the persistence test there were no minimum temperature values flagged beyond f of about 3.5).

An across-site comparison is shown for an f value of 3 in all the procedures in Tables 3–5. For maximum temperature (Table 3), $f = 3$ would flag less than 2% of the data, with the exception of the spatial regression tests at Dickinson and Yellowstone Lake. For minimum temperature (Table 4), $f = 3$ would flag less than 2% of the data, except at Yellowstone. For precipitation (Table 5), the step change test is not implemented because of the discontinuous nature of precipitation. For precipitation the value of $f = 3$ resulted in less than 2% of the data being flagged for the threshold and persistence tests. In the case of the spatial test anywhere from 5% to 7% of the data were still flagged at $f = 6$.

These results show that it will be possible to select dynamic f values for each station and season that will result in a specific but quasi-fixed rate of Type I error generation (say 2% or 0.5%) across the nation.

On first glance these error detection rates may not look stellar, but it should be recognized (see below) that

TABLE 6. The percentage of seeded errors flagged by each test as a function of error magnitude ($r\sigma$) for maximum temperature at Crete, NE.

Relative magnitude of error	Flagged by threshold	Flagged by step change	Flagged by persistence	Flagged by spatial regression	Total flagged (%)
$r <= -3.0$	41	94	0	100	100
$-3 < r <= -2.5$	41	71	6	100	100
$-2.5 < r <= -2.0$	36	50	0	100	100
$-2.0 < r <= -1.5$	8	0	0	92	92
$-1.5 < r <= -1.0$	0	7	0	93	93
$-1.0 < r <= -0.5$	5	5	0	53	53
$-0.5 < r <= 0.0$	0	0	0	0	0
$0.0 < r <= 0.5$	0	0	0	0	0
$0.5 < r <= 1.0$	0	6	0	53	53
$1.0 < r <= 1.5$	6	6	0	83	83
$1.5 < r <= 2.0$	11	22	0	100	100
$2.0 < r <= 2.5$	7	36	0	100	100
$2.5 < r <= 3.0$	31	81	6	100	100
$r > 3.0$	76	94	6	100	100
Total	19	35	1	75	75

the worst errors are being caught, and it is only those errors that are down in the range comparable to sensor accuracy that are slipping through undetected.

b. Type II errors

The results of the seeding analysis are presented in terms of the percentage of errors that were correctly identified as a function of the size of the error. The percent of errors not identified (Type II) is actually 100 minus the percent of correctly identified errors. An example for maximum temperature at Crete is given in Table 6. This result is typical of the other sites. None of the tests were able to identify the smallest errors $-0.5 < r < 0.5$; however, each of the tests became more successful in identifying errors as the magnitude of the error increased. This is not a realistic assessment of the persistence test because seeding of one errant value during a period of 30 days is not likely to move the variability outside the acceptable limits. The spatial regression test identified the most errors (75%), followed by the step change (SC) at 35%, the threshold at 19%, and persistence with less than 2%. All tests combined together identified 75% of the errors introduced. In the case of the other sites, as in this case, most of the errors identified by the other three tests were a subset of those identified by the spatial test.

The analysis shown in Table 6 was repeated for maximum and minimum temperatures and precipitation and for other locations with similar results. The combined performances of the four tests are indicated in Figs. 2a–c. For maximum temperature, the best performance (70%–80%) was noted at the plains sites (Crete and Dickinson) and the desert site (Tucson). The performance at the island (Key West) and shore (Fort Meyers) sites as well as the site with low station density (Yellowstone Lake) was somewhat less (50%–60%).

For minimum temperature (Fig. 2b) the combined performance was about 60% except for the island site. For precipitation (Fig. 2c) the site with complex terrain and the island location gave poorer performance (30%–40%), while the other sites all came in from 40%–50%.

In each case, the spatial regression technique was responsible for identifying the majority of the errors found in the combined analysis.

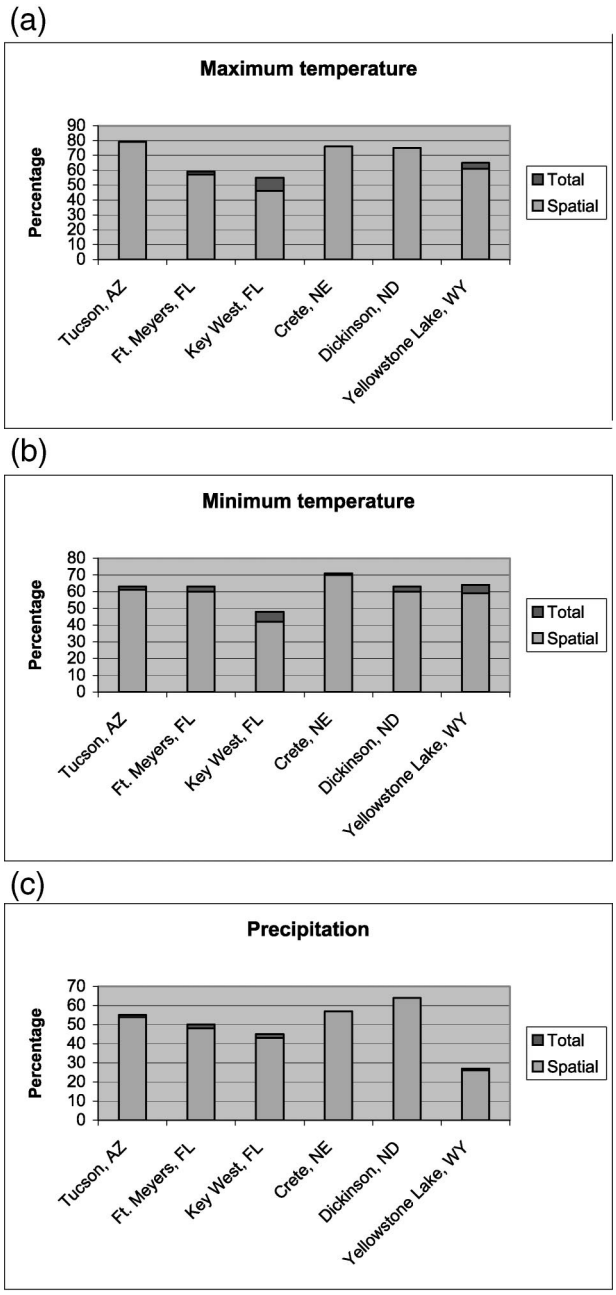


FIG. 2. The percentage of seeded errors discovered by the spatial tests and by all tests for (a) maximum temperature, (b) minimum temperature, and (c) precipitation.

4. Discussion and conclusions

It is essential to test the performance and capability of quality control procedures. In this study, the relative performance of quality control tests varied modestly with climate type and significantly with the variable tested. Seeded errors closer to zero were not detected by quality control tests; but, as the magnitude of error increased, so did the effectiveness of the quality control procedures. For large errors (comparable to f values >2.0), spatial regression was able to flag 100% of ''seeded'' errors. Continuous variables, such as temperature, produce fewer Type I errors under the test procedures used. The trade-off between Type I and Type II errors is very evident with the precipitation variable. Noncontinuous variables, such as precipitation, will produce more Type I errors, especially in the spatial test. The spatial test does, however, offer a means of reducing the Type II errors. Although daily precipitation is known to follow a gamma distribution, it was included in these tests to give a reference point. The authors intend to

focus on alternative QA procedures in follow-up studies, especially additional tests that recognize the non-normal distribution of precipitation. Future work should also include comparison of the techniques set forth here to nonparametric techniques suggested by Lanzante (1996). Pattern recognition seems to have a role to play as well, in that the Type I errors often appear in geographical groupings according to the location and passage of synoptic systems. An effective implementation of pattern recognition has potential to greatly reduce the number of Type I errors made in quality control and assurance.

REFERENCES

Allen, R. G., 1996: Assessing integrity of weather data for reference evapotranspiration estimation. *J. Irrig. Drainage Eng.*, **122,** 97–106.

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

Geiger, M., L. Diabate, L. Menard, and L. Wald, 2002: A web service for controlling the quality of measurements of global solar irradiation. *Solar Energy*, **73,** 475–480.

Guttman, N. B., and R. G. Quayle, 1990: A review of cooperative temperature data validation. *J. Atmos. Oceanic Technol.*, **7,** 334–339.

——, C. Karl, T. Reek, and V. Shuler, 1988: Measuring the performance of data validators. *Bull. Amer. Meteor. Soc.*, **69,** 1448–1452.

Hu, Q., S. Feng, and G. Schaefer, 2002: Quality control for USDA NRCS SM-ST network soil temperatures: A method and dataset. *J. Appl. Meteor.*, **41,** 607–619.

Hubbard, K. G., 2001: Multiple station quality control procedures. Automated weather stations for aplications in agriculture and water resources management. World Meteorological Organization Tech. Doc. AGM-3 WMO/TD No. 1074, 133–136.

Lazante, J. R., 1996: Lag relationships involving tropical sea surface temperatures. *J. Climate*, **9,** 2568–2578.

Meek, D. W., and J. L. Hatfield, 1994: Data quality checking for single station meteorological databases. *Agric. For. 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,** 753–765.

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