

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

185,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



---

# **Toward a Better Quality Control of Weather Data**

---

Kenneth Hubbard, Jinsheng You and  
Martha Shulski

Additional information is available at the end of the chapter

<http://dx.doi.org/10.5772/51632>

---

## **1. Introduction**

Previous studies have documented various QC tools for use with weather data (26; 4; 6; 25; 9; 3; 10; 16; 18). As a result, there has been good progress in the automated QC of weather indices, especially the daily maximum/ minimum air temperature. The QC of precipitation is more difficult than for temperature; this is due to the fact that the spatial and temporal variability of a variable (2) is related to the confidence in identifying outliers. Another approach to maintaining quality of data is to conduct intercomparisons of redundant measurements taken at a site. For example, the designers of the United States Climate Reference Network (USCRN) made it possible to compare between redundant measurements by specifying a rain gauge with multiple vibrating wires in order to avoid a single point of failure in the measurement process. In this case the three vibrating wires can be compared to determine whether or not the outputs are comparable and any outlying values can result in a site visit. CRN also includes three temperature sensors at each site for the purpose of comparison.

Generally identifying outliers involves tests designed to work on data from a single site (9) or tests designed to compare a station's data against the data from neighboring stations (16). Statistical decisions play a large role in quality control efforts but, increasingly there are rules introduced which 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 quality assurance (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 responsible for maintenance of sensors in the field (16; 10).

Quality assurance software consists of procedures or rules against which data are tested. Each procedure will either accept the data as being true or reject the data and label it as an

outlier. This hypothesis ( $H_0$ ) testing of the data and the statistical decision to accept the data or to note it as an outlier can have the outcomes shown in Table 1:

Statistical Decision	True Situation	
	$H_0$ True	$H_0$ False
Accept $H_0$	No error	Type II error
Reject $H_0$	Type I error	No Error

**Table 1.** The classification of possible outcomes in testing of a quality assurance hypothesis.

Take the simple case of testing a variable against limits. If we take as our hypothesis that the data for a measured variable is valid only if it lies within  $\pm 3\sigma$  of the mean ( $X$ ), then assuming a normal distribution we expect to accept  $H_0$  99.73% of the time in the absence of errors. The values that lie beyond  $X \pm 3\sigma$  will be rejected and we will make a Type I error when we encounter valid values beyond these limits. In these cases, we are rejecting  $H_0$  when the value is actually valid and we therefore expect to make a Type I error 0.27% of the time assuming for this discussion that the data has no errant values. If we encounter a bad value inside the limits  $X \pm 3\sigma$  we will accept it when it is actually 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 quality assurance 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 and offer users an input into specifying the limits  $\pm 3\sigma$  beyond which the data will be marked as potential outliers.

In this chapter 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 comparison 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 sub-regions of a wide-spread network. Finally, we introduce a new spatial regression test (SRT) which uses a subset of the neighboring stations to provide the “best fit” to the target station. This spatial regression weighted procedure produces non-biased estimates with characteristics which make it possible to specify statistical confidence intervals for testing data at the target station.

## 2. A Dataset with seeded errors

A dataset consisting of original data and seeded errors (18) is used to evaluate the performance of the different QC approaches for temperature and precipitation. The QC procedures

can be tracked to determine the number of seeded errors that are identified. The ratio of errors identified by a QC procedure to the total number of errors seeded is a metric that can be compared across the range of error magnitudes introduced. The data used to create the seeded error dataset was from the U.S. Cooperative Observer Network as archived in the National Climatic Data Center (NCDC). We used the Applied Climate Information (ACIS) system to access stations with daily data available for all months from 1971~2000 (see 24). The data have been assessed using NCDC procedures and are referred to as “clean” data. Note, however, that “clean” does not necessarily infer that the data are true values but, means instead that the largest outliers have been removed.

About 2% of all observations were selected on a random basis to be seeded with an error. The magnitude of the error was also 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 range of  $\pm 3.5$ . This number was then multiplied by the standard deviation ( $\sigma_x$ ) of the variable in question to obtain the error magnitude  $E$  for the randomly selected observation  $x$ :

$$E_x = \sigma_x r \quad (1)$$

The variable is not used when the error would produce negative precipitation, ( $E_x + x < 0$ ). Thus the seeded error value is skewed distributed when  $r < 0$  but roughly uniformly distributed when  $r > 0$ . The selection of 3.5 for the range is arbitrary but does serve to produce a large range of errors ( $\pm 3.5\sigma_x$ ). This approach to producing a seeded data set is used below in some of the comparisons.

### 3. The spatial regression test (estimates) and Inverse Distance Weighted Estimates (IDW)

When checking data from a site, missing values are sometimes present. For modeling and other purposes where continuous data are required, an estimate is needed for the missing value. We will refer to the station which is missing the data as the target station. The IDW method has been used to make estimates ( $x'$ ) at the target stations from surrounding observations ( $x_i$ ).

$$x' = \frac{\sum_{i=1}^N (x_i / f(d_i))}{\sum_{i=1}^N 1 / f(d_i)} \quad (2)$$

Where  $d_i$  is the distance from the target station to each of the nearby stations,  $f(d_i)$  is a function relying on  $d_i$  (in our case we took  $f(d_i) = 1/d_i$ ). This approach assumes that the nearest stations will be most representative of the target site.

*Spatial Regression (SRT)* is a new method that provides an estimate for the target station and can be used to check that the observation (when not missing) falls inside the confidence in-

terval formed from  $N$  estimates based on  $N$  “best fits” between the target station and neighboring stations during a time period of length  $n$ . The surrounding stations are selected by specifying a radius around the station and finding those stations with the closest statistical agreement to the target station. Additional requirements for station selection are that the variable to be tested is one of the variables measured at the target site and the data for that variable spans 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 (e.g. more than 12 missing days where  $n=24$ ). First non-biased, preliminary estimates  $x_{it}$  are derived by use of the coefficients derived from linear regression, so for any time  $t$ , and for each surrounding station ( $y_{it}$ ) an estimate is formed.

$$x'_i = a_i + b_i y_i \quad (3)$$

The approach obtains an un-biased estimate ( $x'$ ) by utilizing the standard error of estimate ( $s$ ) for each of the linear regressions in the weighting process.

$$x' = \sum_{i=1}^N (x'_i / s_i^2) / \sum_{i=1}^N 1 / s_i^2 \quad (4)$$

$$N / s'^2 = \sum_{i=1}^N 1 / s_i^2 \quad (5)$$

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:

This approach provides more weight to the stations that are the best estimators of the target station. Because the stations used in (4) are a subset of the neighboring stations the estimate is not an areal average but a spatial regression weighted estimate

The 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.

$$x' - fs' \leq x \leq x' + fs' \quad (6)$$

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. An example of the estimates obtained from the SRT is given in Table 2.

ating yi based on x				20E 35S	Havelock	82E 20S	12W 55N	51E 13S									
A254739	A254749		days	x	y1	y2	y3	y4	x'1	x'2	x'3	x'4	x'1/s'1'^2	x'2/s'2'^2	x'3/s'3'^2		
83.696	85.586		6/1/2011	85.1	85.5	83.4	83.7	85.6	85.51	84.30	84.82	84.62	47.016	92.680	170.315		
85.604	87.584		6/2/2011	86.2	86.2	85.3	85.6	87.6	86.28	86.33	86.78	86.62	47.438	94.906	174.255		
89.942	92.282		6/3/2011	91.9	89.5	90.0	89.9	92.3	89.73	91.33	91.24	91.30	49.338	100.408	183.214		
85.478	85.1		6/4/2011	84.1	85.9	83.5	85.5	85.1	85.91	84.42	86.65	84.14	47.238	92.806	173.995		
94.46	97.286		6/5/2011	96.3	94.9	94.1	94.5	97.3	95.49	95.67	95.89	96.29	52.504	105.175	192.545		
97.574	100.994		6/6/2011	99.8	98.0	97.7	97.6	101.0	98.83	99.51	99.09	99.99	54.341	109.395	198.977		
95.918	98.726		6/7/2011	97.2	96.3	96.4	95.9	98.7	97.03	98.10	97.39	97.73	53.349	107.841	195.557		
83.066	86.288		6/8/2011	83.5	86.4	84.8	83.1	86.3	86.41	85.81	84.17	85.32	47.512	94.339	169.014		
69.674	72.878		6/9/2011	71.0	71.8	71.9	69.7	72.9	70.92	72.18	70.40	71.95	38.994	79.345	141.355		
66.2	67.766		6/10/2011	66.2	69.8	67.6	66.2	67.8	68.77	67.59	66.82	66.86	37.812	74.306	134.181		
75.758	76.694		6/11/2011	76.2	76.2	74.8	75.8	76.7	75.53	75.19	76.65	75.76	41.527	82.663	153.921		
77.324	78.98		6/12/2011	78.8	77.9	77.7	77.3	79.0	77.43	78.29	78.26	78.04	42.572	86.065	157.155		
69.314	70.97		6/13/2011	69.2	70.3	69.9	69.3	71.0	69.23	69.98	70.03	70.05	38.066	76.930	140.612		
76.028	78.728		6/14/2011	78.1	79.5	78.1	76.0	78.7	79.12	78.67	76.93	77.79	43.501	86.485	154.478		
84.632	86.396		6/15/2011	86.4	85.0	85.3	84.6	86.4	84.97	86.35	85.78	85.43	46.720	94.927	172.248		
85.118	86.27		6/16/2011	86.8	85.3	84.0	85.1	86.3	85.24	84.94	86.28	85.31	46.868	93.373	173.252		
90.266	92.732		6/17/2011	91.3	92.5	90.9	90.3	92.7	92.92	92.33	91.58	91.75	51.090	101.500	183.884		
80.312	82.904		6/18/2011	81.5	82.9	81.4	80.3	82.9	82.71	82.22	81.34	81.95	45.475	90.391	163.326		
85.118	87.458		6/19/2011	85.6	86.6	85.5	85.1	87.5	86.66	86.60	86.28	86.49	47.649	95.200	173.252		
86.81	88.448		6/20/2011	87.9	88.2	86.7	86.8	88.4	88.35	87.88	88.02	87.48	48.578	96.607	176.746		
71.258	72.788		6/21/2011	72.0	72.9	71.9	71.3	72.8	72.07	72.16	72.03	71.87	39.628	79.324	144.627		
74.948	76.586		6/22/2011	76.7	75.0	74.4	74.9	76.6	74.26	74.83	75.82	75.65	40.831	82.264	152.248		
76.604	78.62		6/23/2011	77.1	78.9	76.4	76.6	78.6	78.45	76.87	77.52	77.68	43.132	84.511	155.668		
78.17	80.168		6/24/2011	79.4	79.4	78.3	78.2	80.2	78.96	78.92	79.13	79.22	43.417	86.758	158.902		
80.564	82.544		6/25/2011	82.0	80.8	80.6	80.6	82.5	80.52	81.33	81.60	81.59	44.272	89.404	163.846		
81.302	82.814		6/26/2011	82.1	82.3	82.1	81.3	82.8	82.09	82.91	82.36	81.86	45.137	91.147	165.370		
78.044	80.06		6/27/2011	79.1	79.8	77.9	78.0	80.1	79.37	78.54	79.00	79.12	43.638	86.338	158.642		
79.61	81.716		6/28/2011	81.1	80.2	79.1	79.6	81.7	79.87	79.80	80.62	80.77	43.913	87.724	161.876		
89.78	91.76		6/29/2011	91.3	89.7	89.3	89.8	91.8	89.96	90.55	91.08	90.78	49.465	99.547	182.880		
98.78	101.48		6/30/2011	100.0	100.3	98.4	98.8	101.5	101.25	100.29	100.33	100.47	55.671	110.256	201.467		
				</													

**Table 2** An example of QC using Spatial Regression Test (SRT) method for daily maximum temperature estimation (unit: F). Stations are from the Automated Weather Data Network and locations are on an East-West by North South street naming convention. The original station (Lincoln 20E 35S) is labeled x while the four neighboring stations are y1,y2, y3, and y4. Equation 3 is used to derive the unbiased estimates  $x'_1$ ,  $x'_2$  etc. for n=30. The final estimate x(est) is determined from the unbiased estimates using equations 4 and 5.

Using the above methodology, the rate of error detection can be pre-selected. 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 which keep the potential Type I errors to about two percent. A representative set of stations and years can be pre-analyzed prior to QC to determine the  $f$  values appropriate to achieve this goal. The SRT method implicitly resolves the bias between variables at different stations induced by elevation difference or other attributes.

Tables 2 and 3 show the use of SRT (equations 3, 4 and 5 above). The data in the example are retrieved from the AWDN stations for the month of June 2011. Only one month was used in this



example. The stations are located in the city of Lincoln, NE, USA. The station being tested is Lincoln 20E 35S and is labeled  $x$  while the neighboring stations are labeled  $y_1, y_2, y_3$ , and  $y_4$ . The slope ( $a_i$ ), interception ( $b_i$ ), and standard errors of the linear regression between the  $x$  and  $y_i$  are computed. The non-biased estimation of  $x$  from data at neighboring stations ( $y_i$ ) are shown as  $x'_1, x'_2, x'_3$ , and  $x'_4$ . The values normalized  $s$  by the standard errors ( $x'_i/s_i$ ) are used in equation 4 to create the estimation  $x(\text{est})$ . The last column shows the bias between the true  $X$  value and the estimated value ( $x(\text{est})$ ) from the four stations. We see that the sum of bias of the 30 days has a value of 0.00, which is expected because the estimates using the SRT method are un-biased. The standard error of this regression estimation is 0.83 F. Here, for instance, where  $f$  was chosen as 3, any value that is smaller than  $-2.5 F$  or larger than  $2.5 F$  will be treated as an outlier. In this example no value of  $x-x(\text{est})$  was marked as an outlier.

days	Original data at Stations, Lincoln NE, USA					estimated x from y				Normalized by s'				X(est)	x-x(est)
	20E 35S	Havelock	82E 20S	12W 55N	51E 13S	$x'_1$	$x'_2$	$x'_3$	$x'_4$	$x'_1/s_1'^2$	$x'_2/s_2'^2$	$x'_3/s_3'^2$	$x'_4/s_4'^2$		
6/1/2011	85.1	85.5	83.4	83.7	85.6	85.64	84.39	84.84	84.66	54.055	98.238	164.200	171.671	84.8	-0.31
6/2/2011	86.2	86.2	85.3	85.6	87.6	86.39	86.39	86.80	86.64	54.533	100.577	167.980	175.693	86.6	0.46
6/3/2011	91.9	89.5	90.0	89.9	92.3	89.80	91.36	91.24	91.31	56.686	106.360	176.575	185.152	91.1	-0.81
6/4/2011	84.1	85.9	83.5	85.5	85.1	86.03	84.50	86.67	84.18	54.306	98.370	167.731	170.692	85.3	1.14
6/5/2011	96.3	94.9	94.1	94.5	97.3	95.49	95.67	95.86	96.28	60.274	111.370	185.527	195.227	95.9	-0.35
6/6/2011	99.8	98.0	97.7	97.6	101.0	98.79	99.48	99.05	99.96	62.356	115.806	191.697	202.692	99.4	-0.32
6/7/2011	97.2	96.3	96.4	95.9	98.7	97.00	98.07	97.36	97.71	61.231	114.173	188.416	198.126	97.6	0.35
6/8/2011	83.5	86.4	84.8	83.1	86.3	86.53	85.88	84.20	85.36	54.617	99.981	162.951	173.084	85.2	1.69
6/9/2011	71.0	71.8	71.9	69.7	72.9	71.23	72.35	70.49	72.04	44.964	84.223	136.417	146.085	71.5	0.47
6/10/2011		69.8	67.6	66.2	67.8	69.11	67.80	66.93	66.97	43.624	78.926	129.534	135.792	67.4	
6/11/2011	76.2	76.2	74.8	75.8	76.7	75.78	75.34	76.72	75.83	47.835	87.710	148.472	153.768	76.0	-0.15
6/12/2011	78.8	77.9	77.7	77.3	79.0	77.66	78.41	78.32	78.10	49.019	91.285	151.575	158.370	78.2	-0.65
6/13/2011	69.2	70.3	69.9	69.3	71.0	69.57	70.17	70.12	70.15	43.911	81.685	135.704	142.243	70.1	0.85
6/14/2011	78.1	79.5	78.1	76.0	78.7	79.32	78.79	76.99	77.85	50.071	91.727	149.007	157.863	77.9	-0.20
6/15/2011	86.4	85.0	85.3	84.6	86.4	85.10	86.41	85.80	85.46	53.720	100.599	166.054	173.301	85.7	-0.67
6/16/2011	86.8	85.3	84.0	85.1	86.3	85.37	85.01	86.30	85.34	53.887	98.966	167.017	173.048	85.6	-1.18
6/17/2011		92.5	90.9	90.3	92.7	92.95	92.35	91.57	91.75	58.672	107.508	177.217	186.058	91.9	
6/18/2011	81.5	82.9	81.4	80.3	82.9	82.87	82.32	81.38	82.00	52.308	95.832	157.495	166.271	81.9	0.47
6/19/2011	85.6	86.6	85.5	85.1	87.5	86.77	86.66	86.30	86.52	54.772	100.886	167.017	175.440	86.5	0.93
6/20/2011	87.9	88.2	86.7	86.8	88.4	88.44	87.93	88.03	87.50	55.825	102.365	170.370	177.433	87.9	0.01
6/21/2011	72.0	72.9	71.9	71.3	72.8	72.37	72.33	72.11	71.95	45.682	84.201	139.556	145.904	72.1	0.13
6/22/2011	76.7	75.0	74.4	74.9	76.6	74.53	74.98	75.89	75.72	47.045	87.291	146.867	153.550	75.5	-1.18
6/23/2011	77.1	78.9	76.4	76.6	78.6	78.66	77.01	77.58	77.74	49.653	89.652	150.148	157.645	77.6	0.50
6/24/2011	79.4	79.4	78.3	78.2	80.2	79.17	79.04	79.19	79.28	49.976	92.014	153.251	160.762	79.2	-0.24
6/25/2011	82.0	80.8	80.6	80.6	82.5	80.71	81.43	81.64	81.64	50.945	94.795	157.994	165.546	81.5	-0.46
6/26/2011	82.1	82.3	82.1	81.3	82.8	82.26	83.00	82.39	81.91	51.925	96.627	159.456	166.089	82.3	0.24
6/27/2011	79.1	79.8	77.9	78.0	80.1	79.57	78.66	79.06	79.17	50.227	91.572	153.001	160.545	79.1	0.00
6/28/2011	81.1	80.2	79.1	79.6	81.7	80.06	79.91	80.66	80.82	50.538	93.029	156.104	163.879	80.5	-0.61
6/29/2011	91.3	89.7	89.3	89.8	91.8	90.03	90.58	91.07	90.79	56.830	105.455	176.254	184.101	90.8	-0.57
6/30/2011	100.0	100.3	98.4	98.8	101.5	101.17	100.25	100.29	100.44	63.863	116.711	194.087	203.671	100.4	0.45
	Slope	1.053	1.053	1.024	0.993										0.00

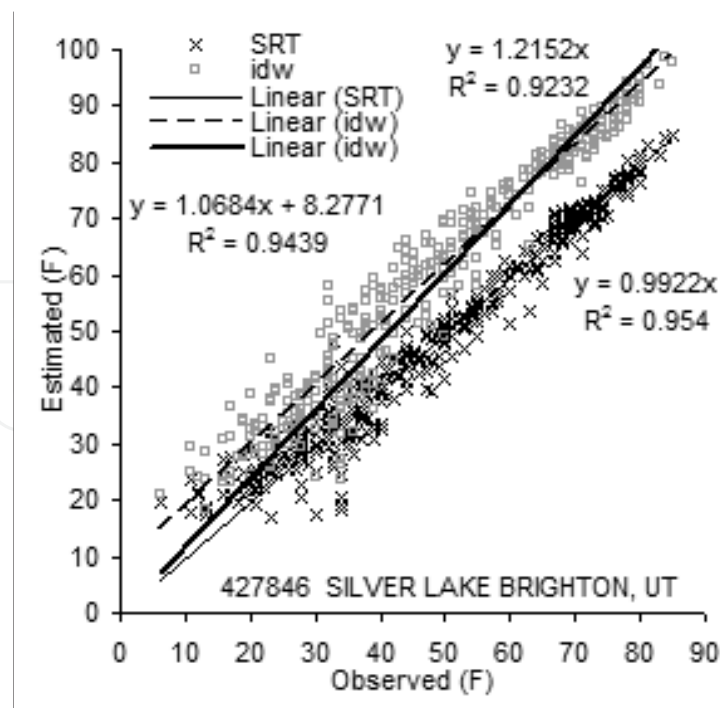
**Table 3** An example of estimating missing data Spatial Regression Test (SRT) method for daily maximum temperature estimation (unit: F). In this example, two days were assumed missing: 6/10 and 6/17 and were estimated using equations 3, 4, and 5 (see highlighted values in the  $x(\text{est})$  column). Stations are from the Automated Weather Data Network and locations are on an East-West by North South naming convention. The original station (Lincoln 20E 35S) is labeled  $x$  while the four neighboring stations are  $y_1, y_2, y_3$ , and  $y_4$ . Equation 3 is used to derive the unbiased estimates  $x'_1, x'_2$  etc. for  $n=28$ . The final estimate  $x(\text{est})$  is determined from the unbiased estimates using equations 4 and 5.

If one value or several values at the station  $x$  is missing, the  $x(\text{est})$  will provide an estimate for the missing data entry (see Table 3). The example in Table 3 shows that the val-

ue of  $x$  is missing in June 10 and June 17, 2011, through the SRT method we can obtain the estimates as 67.4 F and 91.9 F for the two days independent of the true values of 66.2 F and 91.3 F with a bias of 1.2 F and 0.6 F, respectively. Here we note that the estimated values of the two days are slightly different than those estimated in Table 2 because there are 2 less values to include in the regression.

#### 4. Providing estimates: robustness of SRT method and weakness of IDW method

The SRT method was tested against the Inverse Distance Weighted (IDW) method to determine the representativeness of estimates obtained (29). The SRT method outperformed the IDW method in complex terrain and complex microclimates. To illustrate this we have taken the data from a national cooperative observer site at Silver Lake Brighton, UT. The elevation at Silver Lake Brighton is 8740 ft. The nearest neighboring station is located at Soldier Summit at an elevation of 7486 ft. This data is for the year 2002. Daily estimates for maximum and minimum temperature were obtained for each day by temporarily removing the observation from that day and applying both the IDW (eq. 1) and the SRT (eq.2) methods against 15 neighboring stations. The estimations for the SRT method were derived by applying the method (deriving the un-biased estimates) every 24 data.

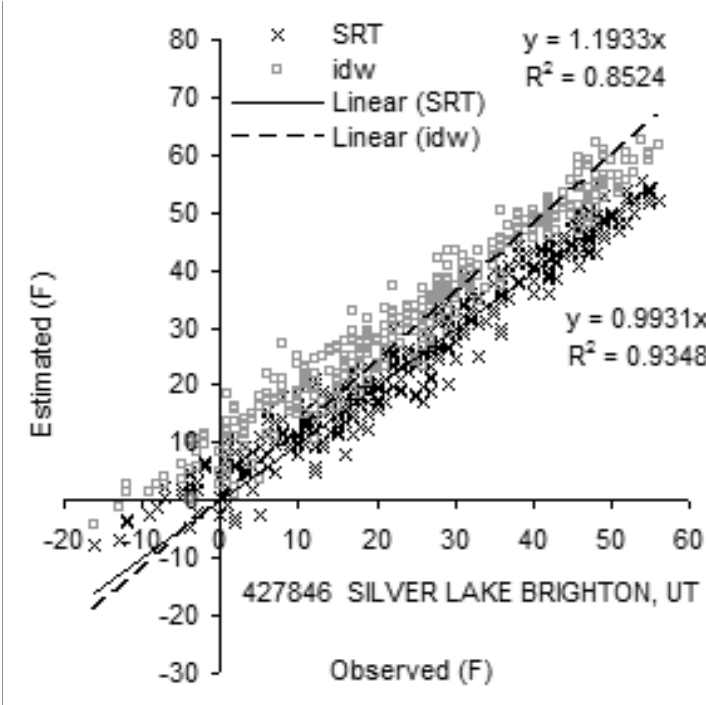


**Figure 1.** The results of estimating maximum temperature at Silver Lake Brighton, UT for both the IDW and the SRT methods.



Fig. 1 shows the result for maximum temperature at Silver Lake Brighton, Utah. The IDW approach results in a large bias. The best fit line for IDW indicates the estimates are systematically high by over 8 F (8.27); the slope is also greater than one (1.0684). When the best fit line for IDW estimates was forced through zero, the slope was 1.2152. On the other hand the estimates from the SRT indicate almost no bias as evidenced by the best-fit slope (0.9922).

For the minimum temperature estimates a similar result was found (Fig. 2). The slope of the best-fit line for the SRT indicates an unbiased (0.9931) while the slope for the IDW estimates indicates a large bias on the order of 20% (slope = 1.1933). The reader should note the SRT unbiased estimators are derived every 24 days (see ) and that applying the SRT only once for the entire period will degrade the results shown (7).



**Figure 2.** The results of estimating minimum temperature at Silver Lake Brighton, UT for both the IDW and the SRT methods.

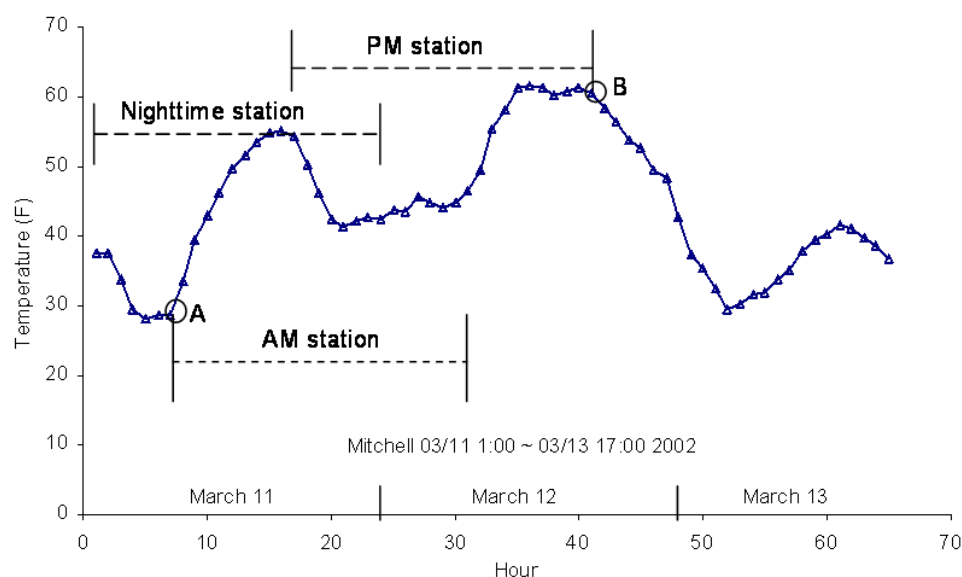
**5. Techniques used to improve the quality control procedures during the extreme events.**

Quality of data during the extreme events such as strong cold fronts and hurricanes may decrease resulting in a higher number of "true" outliers than that during the normal climate conditions. (28) carefully analyzed the sample examples of these extreme weather conditions to quantitatively demonstrate the causes of the outliers and then developed tools to reset the Type II error flags. The following discussion will elaborate on this technique.

### 5.1. Relationship between interval of measurement and QA failures

Analyses were conducted to prepare artificial max and min temperature records (not the measurements, but the values identified as the max and min from the hourly time series) for different times-of-observation from available hourly time series of measurements. The observation time for coop weather stations varies from site-to-site. Here we define the AM station, PM station, and nighttime station according to the time of observation (i.e. morning, afternoon-evening, and midnight respectively). The cooperative network has a higher number of PM stations but AM measurements are also common; the Automated Weather Data Network uses a midnight to midnight observation period.

The daily precipitation accumulates the precipitation for the past 24 hours ending at the time of observation. The precipitation during the time interval may not match the precipitation from nearby neighboring stations due to event slicing, i.e. precipitation may occur both before and after a station's time of observation. Thus, a single storm can be sliced into two observation periods.



**Figure 3.** Example time intervals for observations at Mitchell, NE (after 28).

The measurements of the maximum and the minimum temperature are the result of making discrete intervals on a continuous variable. The maximum or minimum temperature takes the maximum value or the minimum value of temperature during the specific time interval. Thus the maximum temperature or the minimum temperature is not necessarily the maximum or minimum value of a diurnal cycle. Examples of the differences were obtained from three time intervals (see Fig 3) after 28)). The hourly measurements of air temperature were retrieved from 1:00 March 11 to 17:00 March 13, 2002 at Mitchell, NE. The times of observation are marked. Point A shows the minimum air temperature obtained for March 11 for AM stations, and B is the maximum temperature obtained for March 13 at the PM stations. The minimum temperature may carry over to the following interval for AM stations and the

maximum temperature may carry over to the following interval for PM stations. We have therefore marked these as problematic in Table 4 to note that the thermodynamic state of the atmosphere will be represented differently for AM and PM stations. Through analysis of the time series of AM, PM and midnight calculated from the high quality hourly data we find that measurements obtained at the PM station have a higher risk of QA failure when compared to neighboring AM stations. The difference in temperature at different observation times may reach 20 °F for temperature and several inches for precipitation. Therefore the QA failures may not be due to sensor problems but, to comparing data from stations where the sensors are employed differently. To avoid this problem AM stations can be compared to AM stations, PM stations to PM stations, etc. Note this problem will be solved if modernization of network provides hourly or sub-hourly data at most station sites.

	AM station	PM station	Nighttime station (AWDN)
Time intervals (e.g.)	~7:00	~ 17:00	~midnight
Maximum temperature		Problematic	
Minimum temperature	Problematic		
Precipitation	Good	Good	Good

**Table 4.** Time interval and possible performance of three intervals of measurements.

5.2. 1993 floods

Quality control procedures were applied to the data for the 1993 Midwest floods over the Missouri River Basin and part of the upper Mississippi River Basin, where heavy rainfall and floods occurred (28). The spatial regression test performs well and flags 5~7 % of the data for most of the area at  $f=3$ . The spatial patterns of the fraction of the flagged records do not coincide with the spatial pattern of return period. For example, the southeast part of Nebraska does not show a high fraction of flagged records although most stations have return periods of more than 1000 years. While, upper Wisconsin has a higher fraction of flagged records although the precipitation for this case has a lower return period in that area.

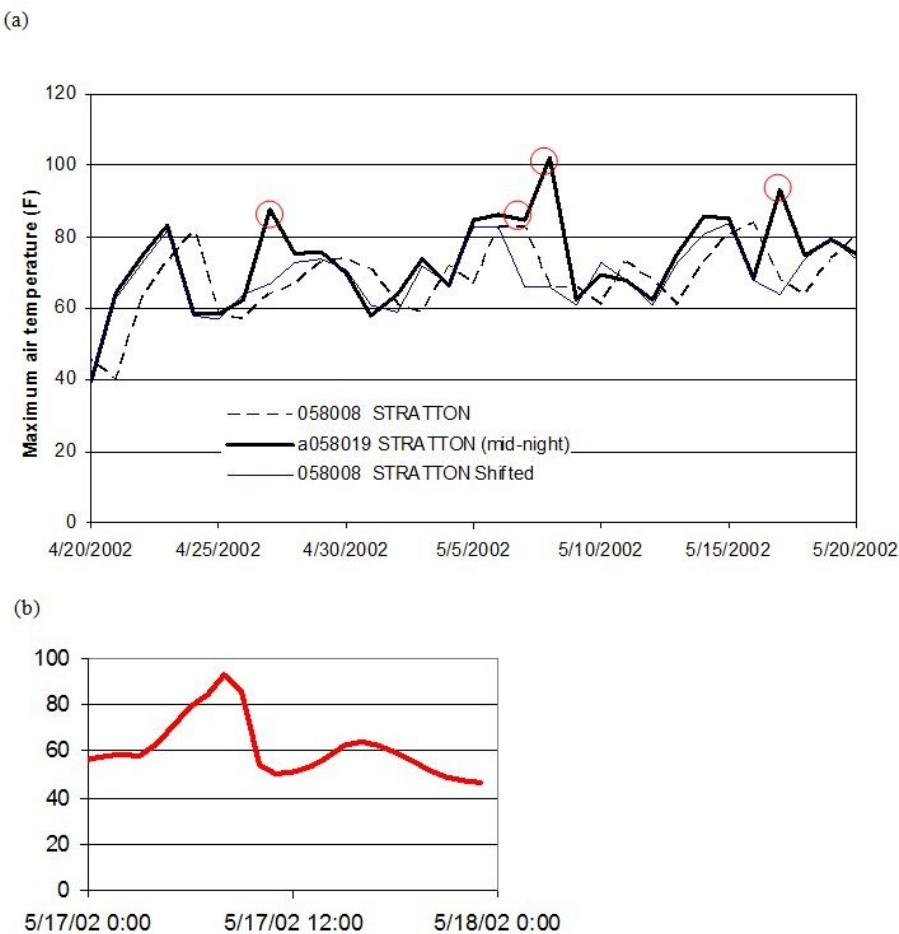
The analysis shows a significantly higher fraction of flagged records using AWDN stations in North Dakota than in other states. This demonstrates that the differences in daily precipitation obtained from stations with different times of observation contributed to the high fraction of QA failures. A high risk of failure would occur in such cases when the measurements of the current station and the reference station are obtained from PM stations and AM stations respectively. The situation worsens if the measurements at weather stations were obtained from different time intervals and the distribution of stations with different time-of-observation is unfavorable. This would be the case for an isolated AM or PM station.

Among the 13 flags at Grand Forks, 9 flags may be due to the different times of observation or perhaps the size and spacing of clouds (28). Four other flags occurred during localized

precipitation events, in which only a single station received significant precipitation. Higher precipitation entries occurring in isolation are more likely to be identified as potential outliers. These problems were expected to be avoided by examining the precipitation over larger intervals, e.g. summing consecutive days into event totals.

5.3. 2002 drought events

No significant relationship is found between the topography and the fraction of flagged records. Some clusters of stations with high flag frequency are located along the mountains; however, other mountainous stations do not show this pattern. Moreover, some locations with similar topography have different patterns. For the State of Colorado, a high fraction of flags occurs along the foothills of the Rocky Mountains where the mountains meet the high plains. A high fraction was also found along interstate highways 25 and 70 in east Colorado. These situations may come about because the weather stations were managed by different organizations or different sensors were employed at these stations. These differences lead to possible higher fraction of flagged records in some areas.



**Figure 4.** Time series of Stratton and a neighboring station during 2002 droughts. a) The daily time series of Tmax for Stratton and Stratton AWDN station (a058019). b) Hourly time series at Stratton AWDN station. (after 28).

Instrumental failures and abnormal events also lead to QA failures. Fig. 4 shows the time series of the Stratton Station in Colorado as part of the automated weather network. This station has nighttime (midnight) readings while all of the neighboring sites are AM or PM stations. Stratton thus has the most flagged records in the state (6); the highlighted records in Fig. 4 were flagged. We checked the hourly data time series to investigate the QA failure in the daily maximum temperature time series for the time period from April 20 to May 20, 2002. No value was found to support a Tmax of 88 for May 6 in the hourly time series, thus 88 °F appears to be an outlier. On May 7 a high of 85 °F is recorded for the PM station observation interval, in which the value of the afternoon of May 6 is recorded as the high on May 7. The 102 °F observation of May 8 at 6:00 AM appears to be an observation error caused by a spike in the instrument reading. The observation of 93 °F at 8:00 AM May 17 is supported by the hourly observation time series (see Fig. 4 (b)) and is apparently associated with a down burst from a decaying thunderstorm.

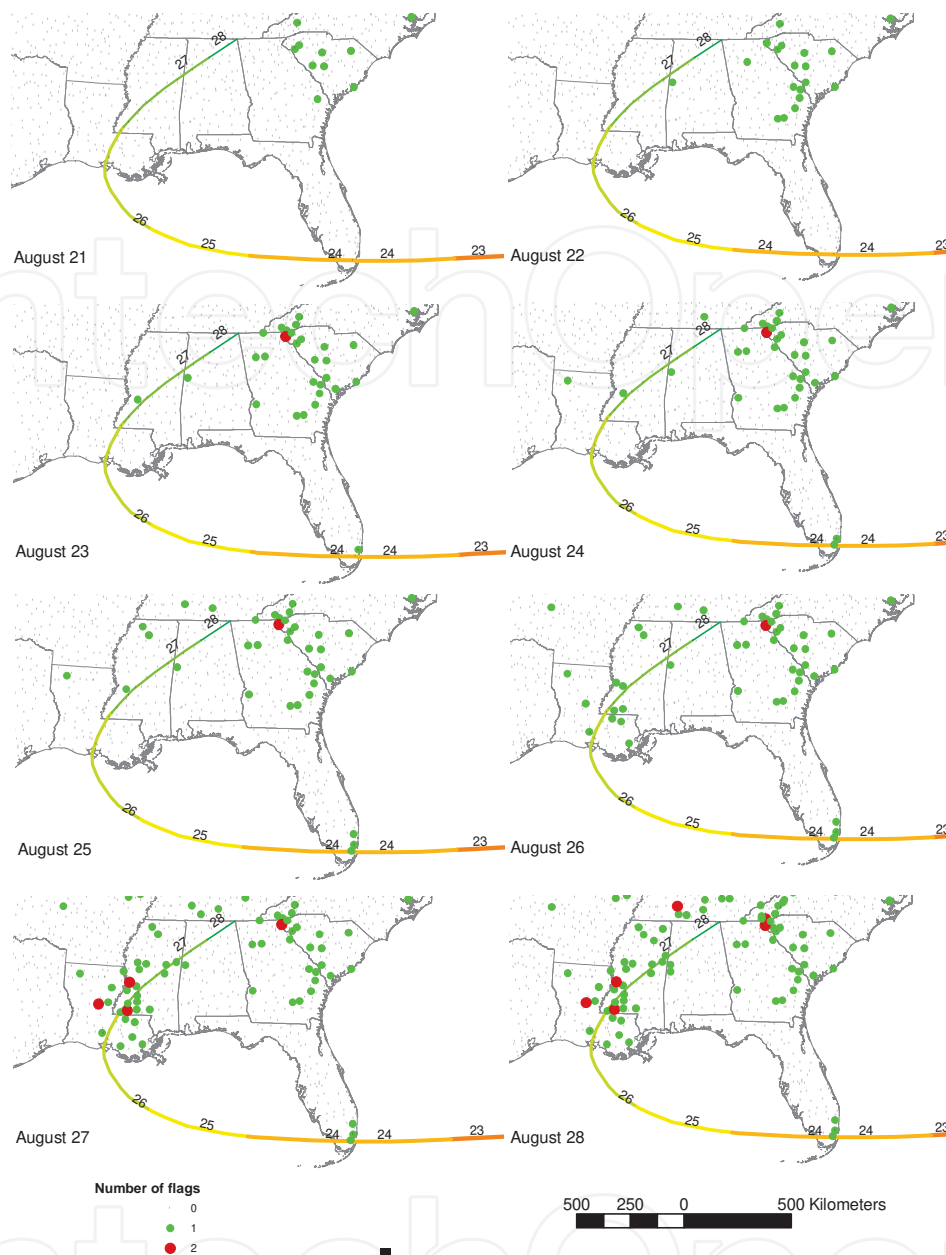
#### 5.4. 1992 Andrew Hurricane

In Fig. 5 the evolution of the spatial pattern of flagged records from August 25 to August 28, 1992 during Hurricane Andrew and the corresponding daily weather maps shows a heavy pattern of flagging. The flags in the spatial pattern figures are cumulative for the days indicated. The test shows that the spatial regression test explicitly marks the track of the tropical storm. Starting from the second land-fall of Hurricane Andrew at mid-south Louisiana, the weather stations along the route have flagged records. The wind field formed by Hurricane Andrew helps to define the influence zone of the hurricane on flags. Many stations without flags have daily precipitation of more than 2 inches as the hurricane passes, which confirms that the spatial regression test is performing reasonably well in the presence of high precipitation events.

#### 5.5. Cold front in 1990

Flags for the cold front event during October, 1990 were examined. The maximum air temperature dropped by as much as 40 °F during the passage of the cold front. Spatial patterns of flags on October 6 coincide with the area traversed by the cold front and many stations were flagged in such states as North Dakota, South Dakota, Iowa, and Nebraska. On October 7, the cold front moved to southeast regions beyond Nebraska and Iowa. Of course nearby stations on opposite sides of the cold front may experience different temperatures thus leading to flags. This may be further complicated when different times of observation are involved. The cold front continues moving and the area of high frequency of flags also moves with the front correspondingly.

A similar phenomenon can be found in the test of the precipitation and the minimum temperature. A spatial regression test of any of these three variables can roughly mark the movements of the cold front events. The identified movements of the cold fronts and associated flagging of “good records” may lead to more manual work to examine the records. Simple pattern recognition tools have been developed to identify the spatial patterns of these flags and reset these flags automatically (see Fig. 6).



**Figure 5.** Daily weather maps and spatial pattern of flagged records for 1992 Andrew Hurricane events. (after 28).

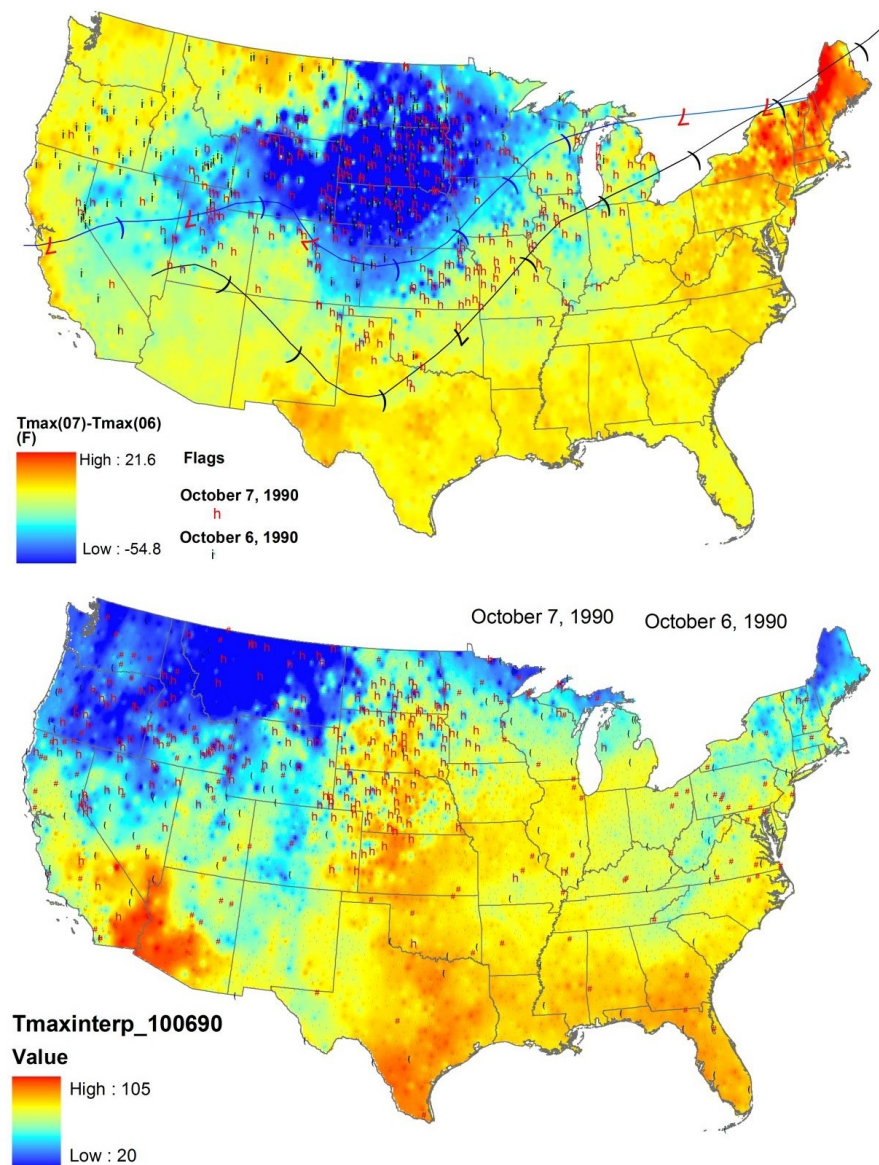
The spatial patterns of flagged records are significant for both the spatial regression test of the cold front events and the tropical storm events. However, most of these flagged records are type I errors, thus we tested a simple pattern recognition tool to assist in reducing these flags. Differences still exist between the distribution patterns of the flagged records for the cold front event and the tropical storm events due to the characteristics of cold front events and tropical storm events. These differences are:

- Cold fronts have wide influence zones where the passages of the cold fronts are wider and the large areas immediately behind the cold front may have a significant flagged frac-



tion of weather stations. The influence zones of the tropical storms are smaller where only the stations along the storm route and the neighboring stations have flags.

- Cold fronts exert influences on both the air temperature and precipitation. The temperature differences between the regions immediately ahead of the cold fronts and regions behind can reach 10~20 °C. The precipitation events caused by the cold fronts may be significant, depending on the moisture in the atmosphere during the passage. The tropical storms generally produce a significant amount of precipitation. A few inches of rainfall in 24 hours is very common along the track because the tropical storms generally carry a large amount of moisture.



**Figure 6.** Spatial patterns of flagged records for cold front events and related fronts. The temperature map is the interpolated maximum temperature difference between October 6 and October 7, 1990. The color front is on October 7, and the black one is on October 6. The flags are the QA failures on that day.

## 5.6. Resetting the flags for cold front events and hurricanes

Some measurements during the cold front and the hurricane were valid but flagged as outliers due to the effect of QC tests during times of large temperature changes caused by the cold front passages and the heavy precipitation occurring in hurricanes. A simple spatial scheme was developed to recognize regions where flags have been set due to Type I errors. The stations along the cold front may experience the mixed population where some stations have been affected by the cold fronts and others have not. A complex pattern recognition method can be applied to identify the influence zone of the cold fronts through the temperature changes (e.g. using some methods described in Jain et al, 2000). In our work, we use the simple rule to reset the flag given that significant temperature changes occur when the cold front passes. The mean and the standard deviation of the temperature change can be calculated as:

$$\overline{\Delta T} = \frac{1}{n} \sum_{i=1}^n \Delta T_i \quad (7)$$

$$\sigma_{\Delta T}^2 = \frac{1}{n} \sum_{i=0}^n (\Delta T_i * \Delta T_i) - \overline{\Delta T} * \overline{\Delta T} \quad (8)$$

where  $\overline{\Delta T}$  is the mean temperature change of the reference stations,  $\Delta T_i$  is the temperature change at the  $i^{\text{th}}$  station for the current day,  $n$  is the number of neighboring stations, and  $\sigma_{\Delta T}$  is the standard deviation of the temperature change for the current day. A second round test is applied to records that were flagged in the first round:

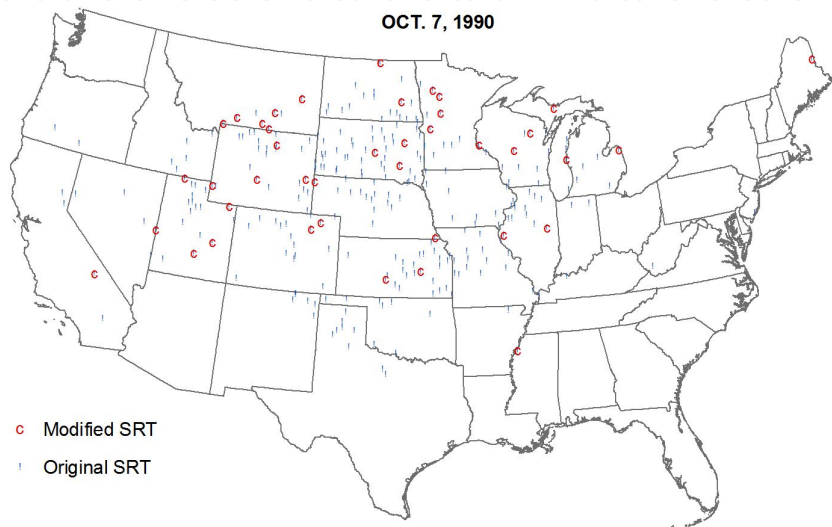
$$\overline{\Delta T} - f' \sigma_{\Delta T} \leq \Delta T \leq \overline{\Delta T} + f' \sigma_{\Delta T} \quad (9)$$

where  $\Delta T$  is the difference between maximum/minimum air temperature for the current day and the last day. The cutoff value  $f'$  takes a value of 3.0. The test results with this refinement for  $T_{\max}$  are shown in Fig. 7 for Oct. 7, 1990. The results obtained using the refinements described in this section were labeled “modified SRT” and the results using the original SRT were labeled “original SRT” in Fig. 7 and 8. Of the 291 flags originally noted only 41 flags remain after the reset phase. The daily temperature drops more than 20 °F at most stations where the flags were reset and the largest drop is 55 °F.

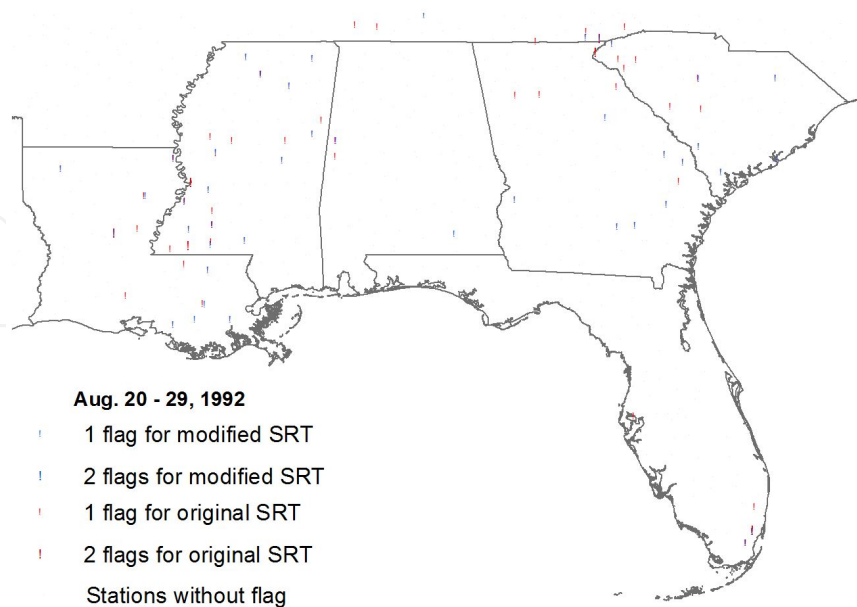
For the heavy precipitation events, we compare the amount of precipitation at neighboring stations to see whether heavy precipitation occurred. We use a similar approach as for temperature to check the number of neighboring stations that have significant precipitation,

$$\zeta = \text{count}(p_i \geq p_{\text{threshold}}) \quad (10)$$

where the  $p_i$  is the daily precipitation amount at a neighboring station, and  $p_{threshold}$  is a threshold beyond which we recognize that a significant precipitation event has occurred at the neighboring station, e.g. 1 in. When  $\zeta \geq 2$  and  $p \geq p_{high}$ , we reset the flag. Here  $p$  is the precipitation amount of the current station, and  $p_{high}$  is the upper threshold beyond which the threshold will flag the measurement. Fig.8 shows maps of flags after the reset process. Of the 78 flags originally noted only 41 flags remain after the reset phase. Most of the remaining flags are due to the precipitation being higher than the upper threshold.



**Figure 7.** All points shown were flagged by the original SRT method while the red points were those that are flagged by the modified SRT method for maximum daily Temperature. Blue symbols are those that are reset by the modified SRT method.



**Figure 8.** This is the reset of flags for the Andrew 1992 hurricane. The flags are the cumulative flags starting from Aug. 20 to Aug. 29, 1992. The flags by the modified SRT method overlay the flags by the original SRT method.

Flags for the Andrew 1992 hurricane. The flags are the cumulative flags starting from Aug. 20 to Aug. 29, 1992. The flags by the modified SRT method overlay the flags by the original SRT method.

## 6. Multiple interval methods based on measurements from reference stations for precipitation.

One QC approach involved developing threshold quantification methods to identify a subset of data consisting of potential outliers in the precipitation observations with the aim of reducing the manual checking workload. This QC method for precipitation was developed based on the empirical statistical distributions underlying the observations.

The search for precipitation quality control (QC) methods has proven difficult. The high spatial and temporal variability associated with precipitation data causes high uncertainty and edge creep when regression-based approaches are applied. Precipitation frequency distributions are generally skewed rather than normally distributed. The commonly assumed normal distribution in QC methods is not a good representation of the actual distribution of precipitation and is inefficient in identifying the outliers.

The SRT method is able to identify many of the errant data values but the rate of finding errant values to that of making type I errors is conservatively 1:6. This is not acceptable because it would take excessive manpower to check all the flagged values that are generated in a nationwide network. For example, the number of precipitation observations from the cooperative network in a typical day is 4000. Using an error rate of 2% and considering the type I error rate indicates that several hundred values may be flagged, requiring substantial personnel resources for assessment.

(29) found the use of a single gamma distribution fit to all precipitation data was ineffective. A second test, the multiple intervals gamma distribution (MIGD) method, was introduced. It assumed that meteorological conditions that produce a certain range in average precipitation at surrounding stations will produce a predictable range at the target station. The MIGD method sorts data into bins according to the average of precipitation at neighboring stations; then, for the events in a specific bin, an associated gamma distribution is derived by fit to the same events at the target station. The new gamma distributions can then be used to establish the threshold for QC according to the user-selected probability of exceedance. We also employed the  $Q$  test for precipitation (20) using a metric based on comparisons with neighboring stations. The performance of the three approaches was evaluated by assessing the fraction of “known” errors that can be identified in a seeded error dataset(18). The single gamma distribution and  $Q$ -test approach were found to be relatively efficient at identifying extreme precipitation values as potential outliers. However, the MIGD method outperforms the other two QC methods. This method identifies more seeded errors and results in fewer Type I errors than the other methods.

### 6.1. Estimation of parameters for distribution of precipitation and thresholds from the Gama distribution

The Gamma distribution was employed to represent the distribution of precipitation. While other functions may provide a better overall fit to precipitation data our goal is to establish a reasonable threshold on values beyond which further checking will be required to determine if the value is an outlier or simply an extreme precipitation event. The precipitation events are fit to a Gamma distribution,  $G(\gamma, \beta)$ . The shape and scale parameters  $\gamma, \beta$  can be estimated from the precipitation events following (21) and (13),

$$\gamma = \frac{\bar{X}^2}{s^2} \quad (11)$$

$$\beta = \frac{s^2}{\bar{X}} \quad (12)$$

where  $\bar{X}$  and  $s$  are the sample mean and the sample standard deviation, respectively.

The data for each station in the Gamma distribution test include all precipitation events on a daily basis for a year. The parameters for left-censored (0 values excluded) Gamma distributions, on a monthly basis, are also calculated, based on the precipitation events for individual months in the historical record. To ascertain the representativeness of the Gamma distribution, the precipitation value for the corresponding percentiles ( $P$ ): 99, 99.9, 99.99, and 99.999% were computed from the Gamma distribution and compared with the precipitation values for given percentiles based on ranking (original data).

The criterion for a threshold test approach can be written as,

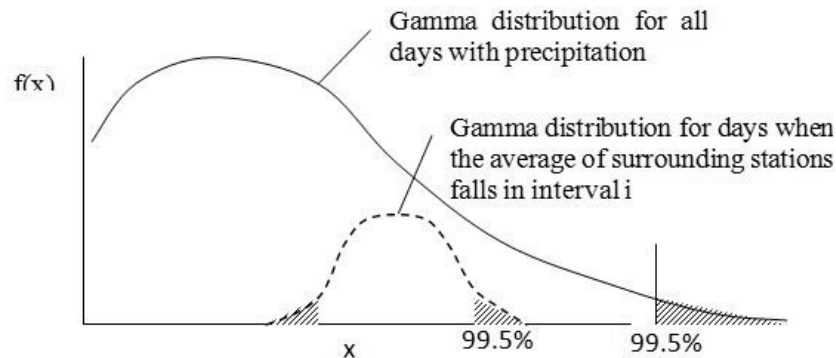
$$x(j, t) < I(p) \quad (13)$$

where  $x(j, t)$  is the observed daily precipitation on day  $t$  at station  $j$  and  $I(p)$  is the threshold daily precipitation for a given probability,  $p$  ( $=P/100$ ), calculated using the Gamma distribution. A value not meeting this criterion is noted as a potential outlier (the shaded area to the right of the  $p=0.995$  value for the distribution for all precipitation events in Fig. 9). The test function uses the one-sided test for precipitation, a non-negative variable.

### 6.2. Multiple interval range limit gamma distribution test for precipitation (MIGD)

Analysis has shown that precipitation data at a station can be fit to a Gamma distribution, which can then be applied to a threshold test approach. With this method only the most extreme precipitation events will be flagged as potential outliers so errant data at other points in the distribution are not identified.





**Figure 9.** Schematic of gamma distribution for all daily precipitation events and for the  $i^{\text{th}}$  interval of the MIGD approach.

The MIGD was developed to address these non-extreme points along the distribution. It assumes that meteorological conditions that produce a certain range in average precipitation at surrounding stations will produce a predictable range at the target station. Our concept is to develop a family of Gamma distributions for the station of interest and to selectively apply the distributions based on specific criteria. The average precipitation for each day is calculated for neighboring stations during a time period (e.g. 30 years). These values are ranked and placed into  $n$  bins with an equal number of values in each. The range for  $n$  intervals can be obtained from the cumulative probabilities of neighboring average time series,  $\{0, 1/n, 2/n, \dots, n-1/n, 1\}$ . For the  $i^{\text{th}}$  interval all corresponding precipitation values at the station of interest (target station) are gathered and parameters for the gamma distribution estimated. This process is repeated for each of the  $n$  intervals resulting in a family of Gamma curves ( $G_i$ ). The operational QC involves the application of the threshold test where the gamma distribution for a given day is selected from the family of curves based on the average precipitation for the neighboring stations. Each interval can be defined as  $(\xi(p(i/n)), \xi(p((i+1)/n))]$ , where  $p(i/n)$  is the cumulative probability associated with  $i/n$ ,  $i=0$  to  $n-1$ , and  $\xi(p(i/n))$  is the neighboring stations' average for a given cumulative probability.

Now for each precipitation event,  $x$ , at the station of interest, the neighboring stations' average is calculated. If the average precipitation falls in the interval  $(\xi(p(i/n)), \xi(p((i+1)/n))]$ , then  $G_i$  is used to form a test:

$$G_i(1-p) < x(j, t) < G_i(p) \quad (14)$$

where  $p$  is a probability in the range  $(0.5, 1)$ , and the  $G_i(p)$  is the precipitation value for the given probability  $p$  in the gamma distribution associated with the  $i^{\text{th}}$  interval. This equation forms a two sided test. Any value that does not satisfy this test will be treated as an outlier for further manual checking. The intervals and the estimation of this method were implemented using R statistical software (19).

The results indicate that the Gamma distribution is well suited for deriving appropriate thresholds for a particular precipitation event. The calculated extreme values provide a good basis



for identifying extreme outliers in the precipitation observations. The inclusion of all precipitation events reduces the data requirements for the quantification of extreme events which generally requires a long time series of observations (e.g. using Gumbel distribution.) Using the approach based on the Gamma distribution, a suitable representation of the distribution of precipitation can be obtained with only a few years of observation, as is the case with newly established automatic weather stations, e.g. Climate Reference Network. Further study is required for probability selection in the Gamma distribution approach.

sname	range	lower	upper	q999	q995	q99	q975	q95	q9	q1	q05	q025	q01	q005
028820	0~0.0510	0	0.0510	0.8441	0.6332	0.5431	0.4251	0.3369	0.2502	0.0055	0.0021	0.0008	0.0002	9.39E-05
028820	0.0510~0.1111	0.0510	0.1111	1.1533	0.8704	0.7493	0.5903	0.4711	0.3532	0.0097	0.0040	0.0017	0.0005	0.0002
028820	0.1111~0.1899	0.1111	0.1899	1.2989	0.9896	0.8567	0.6816	0.5495	0.4181	0.0156	0.0071	0.0033	0.0012	0.0006
028820	0.1899~0.3326	0.1899	0.3326	1.8631	1.4162	1.2243	0.9717	0.7815	0.5925	0.0206	0.0092	0.0042	0.0015	0.0007
028820	0.3326~2.1216	0.3326	2.1216	2.8514	2.1861	1.8996	1.5210	1.2346	0.9484	0.0429	0.0208	0.0102	0.0040	0.0020

**Table 5.** Multiple gamma distributions ( $n=5$ ) for the Multiple Interval Gamma Distribution (MIGD) method at Tucson, AZ. Lower and upper represent the upper and lower limits of each bin for surrounding station averages. The precipitation threshold for the target station can be selected from q999, q995, q99, q975, q95, q9, q1, q05, q025, q01, q005, and q001 as these are associated with gamma distribution for the station of interest.

A simple gamma distribution can be fit to the daily precipitation values at a station. Upper thresholds can be set based on the cumulative probability of the precipitation distribution. This single gamma distribution (SGD) test will address the most extreme values of precipitation and flag them for further testing. However, to address non-extreme values of precipitation that are not out on the tail of the SGD another approach is needed. We have formulated the multiple interval gamma distribution test (MIGD) for this purpose. The main assumption is that the meteorological conditions that produce a certain range in average precipitation at surrounding stations will produce a predictable range of precipitation at the target station. It does not estimate the precipitation at the target station but estimates the range into which the precipitation should fit.

The average precipitation for each day is calculated for neighboring stations during a historical period, say 30 years. These values are then ranked and placed into  $n$  bins with an equal number of values in each. For all the values in a given bin, the daily precipitation at the target station are gathered and a gamma distribution formed. The process is repeated  $n$  times once for each bin resulting in a family of gamma distribution curves. A separate family of curves can be derived for each month or each season. In operation, the daily average of the precipitation at surrounding stations is calculated and used to point to the  $n$ 'th gamma distribution which in turn provides thresholds against which to test for that day. For instance, the upper threshold can be selected to correspond with the cumulative probability for the  $n$ 'th gamma distribution. The user is able to specify the threshold according the cumulative probability. For example we can be 99.5 % confident that values will not exceed the corresponding value on the cumulative probability curve. Values that exceed this are not necessarily wrong but flagged for further review. The MIGD will find more precipitation values that need to be reviewed than the single gamma distribution test.

Table 5 provides an example of the MIGD for  $n=5$  at Tucson, AZ, USA. We update this type of information on an annual basis. If the precipitation value falls outside the  $q$  value of a selected

confidence level, we mark the value as an outlier. For example, Suppose we select  $q_{999}$  for our confidence. The precipitation on August 2, 1987 was 1.3 inches while the average of neighboring stations had a value of 0.06 inches. The average falls between lower and upper in the 2<sup>nd</sup> row,  $n=2$ . ie. 0.05, 0.11. The rainfall value (1.3 inches) is larger than the  $q_{999}$  threshold (1.15 inches) thus we can say we are 99.9 % confident that the rainfall is an outlier and it should be flagged for further manual examination. Note that 1.3 inches is in no way an extreme precipitation value but, it's validity can be challenged on the basis of the MIGD test.

One other QC method for precipitation test is the Q-test (20). The Q-test approach serves as a tool to discriminate between extreme precipitation and outliers and it has proven to minimize the manual examination of precipitation by choice of parameters that identify the most likely outliers (20). The performance of both the Gamma distribution test and the Q-test is relatively weak with respect to identifying the seeded errors. The Q-Test is different from the Gamma distribution method because the Q-Test uses both the historical data and measurements from neighboring stations while the simple implementation of the Gamma distribution method only uses the data from the station of interest.

The MIGD method is a more complex implementation of the Gamma distribution that uses historical data and measurements from neighboring stations to partition a station's precipitation values into separate populations. The MIGD method shows promise and outperforms other QC methods for precipitation. This method identifies more seeded errors and creates fewer Type I errors than the other methods. MIGD will be used as an operational tool in identifying the outliers for precipitation in ACIS. However, the fraction of errors identified by the MIGD method varies for different probabilities and among the different stations. Network operators, data managers, and scientist who plan to use MIGD to identify potential precipitation outliers can perform a similar analysis (sort the data into bins and derive the gamma distribution coefficients for each interval) over their geographic region to choose an optimum probability level.

## **7. Quality control of the NCDC dataset to create a serially complete dataset.**

Development of continuous and high-quality climate datasets is essential to populate Web-distributed databases (17) and to serve as input to Decision Support Systems (e.g., 27).

Serially complete data are necessary as input to many risk assessments related to human endeavor including the frequency analysis associated with heavy rains, severe heat, severe cold, and drought. Continuous data are also needed to understand the climate impacts on crop yield, and ecosystem production. The National Drought Mitigation Center (NDMC) and the High Plains Regional Climate Center (HPRCC) at the University of Nebraska are developing a new drought atlas. The last drought atlas (1994) was produced with the data from 1119 stations ending in 1992. The forthcoming drought atlas will include additional stations and will update the analyses, maps, and figures through the period 1994 to the present time. A list was compiled from the Applied Climate Information System (ACIS) for

stations with a length of at least 40 years of observations for all three variables: precipitation (PRCP), maximum (Tmax), and minimum (Tmin) temperatures. Paper records were scrutinized to identify reported, but previously non-digitized data to reduce, to the extent possible, the number of missing data. A list of 2144 stations was compiled for the sites that met the criterion of at least 40 years data with less than two months continuous missing gaps for at least one of the three variables. The remaining missing data in the dataset were supplemented by the estimates obtained from the measurements made at nearby stations. The spatial regression test (SRT) and the inverse distance weighted (IDW) method were adopted in a dynamic data filling procedure to provide these estimates. The replacement of missing values follows a reproducible process that uses robust estimation procedures and results in a serially complete data set (SCD) for 2144 stations that provide a firm basis for climate analysis. Scientists who have used more qualitative or less sophisticated quantitative QC techniques may wish to use this data set so that direct comparisons to other studies that used this SCD can be made without worry about how differences in missing data procedures would influence the results. A drought atlas based on data from the SCD will provide decision makers more support in their risk management needs.

After identifying stations with a long-term (at least 40 years) continuous (no data gaps longer than two months) dataset of Tmax, Tmin, and/or PRCP for a total of 2144 stations, the missing values in the original dataset retrieved from ACIS were filled to the extent possible with the keyed data from paper record and the estimates using the SRT and IDW methods. Two implementations of SRT were applied in this study. The short-window (60 days) implementation provides the best estimates based on the most recent information available for constructing the regression. The second implementation of SRT fills the long gaps, e.g. gaps longer than one month using the data available on a yearly basis. The IDW method was adopted to fill any remaining missing data after the two implementations of SRT.

This is the first serially complete data set where a statement of confidence can be associated with many of the estimates, ie. SRT estimates. The RMSE is less than 1F in most cases and thus we are 95% confident that the value, if available, would lie between  $\pm 2F$  of the estimate. This data set is available <sup>1</sup> to interested parties and can be used in crop models, assessment of severe heat, cold, and dryness. Probabilities related to extreme rainfall for flooding and erosion potential can be derived along with indices to reflect impact on livestock production. The data set is offered as an option to distributing raw data to the users who need this level of spatial and temporal coverage but are not well positioned to spend time and resources to fill gaps with acceptable estimates.

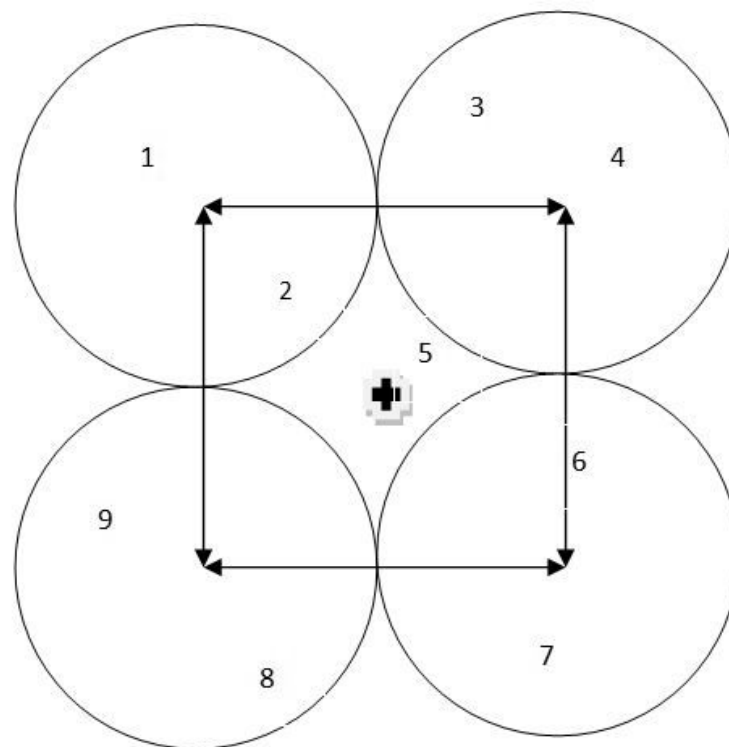
Analysis based on the long-term dataset will best reveal the regional and large scale climatic variability in the continental U.S., making this an ideal data set for the development of a new drought atlas and associated drought index calculations. Future data observations can be easily appended to this SCD with the dynamic data filling procedures described herein.

---

<sup>1</sup> Contact the High Plains Regional Climate Center at 402-472-6709

## 8. Issues relating QC to gridded datasets,

Gridded datasets are sometimes used in QC but, we caution against this for the following reasons. New datasets created from inverse distance weighted methods or kriging suffer from uncertainties. The values at a grid point are usually not "true" measurements but are interpolated values from the measurements at nearby stations in the weather network. Thus, the values at the grid points are susceptible to bias. When further interpolation is made to a given location within the grid, bias will again exist at the specific location between the gridded values. Fig. 10 provides an example of potential bias. Outside of a gridded data set the target location would give a large weight to the value at station 5. However, if the radius used for the gridded data is as in the Fig. 10, then the closest station to the target station (5) will not be included in the grid-based estimation.



**Figure 10.** An example of station distribution used in the grid method.

## 9. Quality control of high temporal resolution datasets

The Oklahoma Mesonet (<http://www.mesonet.org/>) measures and archives weather conditions at 5-minute intervals (Shafer et al., 2000). The quality control system used in the network starts from the raw data of the measurements for the high temporal resolution data. A set of QC tools was developed to routinely maintain data of the Mesonet. These tools depend on the status of hardware and measurement flag sets built in the climate data sys-

tem. The Climate Reference Network (CRN, Baker et al. 2004) is another example of the QC of high frequency data, which installs multiple sensors for each variable to guarantee the continuous operation of the weather station and thus the quality control can also rely on the multiple measurements of a single variable. This method is efficient to detect the instrumental failures or other disturbances; however the cost of such a network may be prohibitive for non-research or operational networks. The authors of this chapter also carried out QC on a high temporal resolution dataset in the Beaufort and Chukchi Sea regions. Surface meteorological data from more than 200 stations in a variety of observing networks and various stand-alone projects were obtained for the MMS Beaufort and Chukchi Seas Modeling Study (Phase II). Many stations have a relatively short period of record (i.e. less than 10 years). The traditional basic QC procedures were developed and tested for a daily data and found in need of improvement for the high temporal resolution data. In the modification, the time series of the maximum and the minimum were calculated from the high resolution data. The mean and standard deviation of the maximum and the minimum can then be calculated from the time series (e.g. max and min temperatures) as the  $(u_x, s_x)$  and  $(u_n, s_n)$ , respectively. The equation (6) using  $(u_x + f s_x)$  and  $(u_n - f s_n)$  forms limits defined by the upper limits of the maximum and lower limits of the minimum. The value falling outside the limits will be flagged as an outlier for further manual checking. Similarly, the diurnal change of a variable (e.g. temperature) was calculated from the high resolution (hourly or sub-hourly) data. The mean and standard deviation calculated from the diurnal changes will form the limits.

The traditional quality control methods were improved for examining the high temporal resolution data, to avoid intensive manual reviewing which is not timely or cost efficient. The identified problems in the dataset demonstrate that the improved methods did find considerable errors in the raw data including the time errors (e.g. month being great than 12). These new tools offer a dataset that, after manual checking of the flagged data, can be given a statement of confidence. The level of confidence can be selected by the user, prior to QC.

The applied in-station limit tests can successfully identify outliers in the dataset. However, spatial tests based information from the neighboring stations is more robust in many cases and identifies errors or outliers in the dataset when strong correlation exists. The good relationship between the measurements at station pairs demonstrates that there is a potential opportunity to successfully apply the spatial regression test (SRT, 18) to the stations which measure the same variables (i.e. air temperature or wind speed). The short term measurements at some stations may not be efficiently QC'ed with only the three methods described in this work. One example is the dew point measurements at the first-order station Iultin-in-Chukot. More than 90 percent of the dew point measurements were flagged, because the parameters for QC'ing the variable used the state wide parameters which cannot reflect the microclimate of each station.

## 10. Summary and Conclusions

Quality control (QC) methods can never provide total proof that a data point is good or bad. Type I errors (false positives) or Type II errors (false negatives) can occur and result in labeling



of good data as bad and bad data as good respectively. Decreasing the number of Type I and Type II errors is difficult because often a push to decrease Type I errors will result in an unintended increase in Type II errors and vice versa. We have derived a spatial technique to introduce thresholds associated with user selected probabilities (i.e. select 99.7% as the level of confidence that a data value is an outlier before labeling it as bad and/or replacing it with an estimate). We base this technique on statistical regression in the neighborhood of the data in question and call it the Spatial Regression Test (SRT). Observations taken in a network are often affected by the same factors. In weather applications individual stations in a network are generally exposed to air masses in much the same way as are neighboring stations. Thus, temperatures in the vicinity move up and down together and the correlation between data in the same neighborhood is very high. Similarly seasonal forcings on this neighborhood (e.g. the day to day and seasonal solar irradiance) are essentially the same. We have defined a neighborhood for a station as those nearby stations that are best correlated to it. We found that the SRT method is an improvement over conventional inverse distance weighting estimates (IDW). A huge benefit of the SRT method is its ability to remove systematic biases in the data estimation process. Additionally, the method allows a user selected threshold on the probability as contrasted to the IDW. Although the SRT estimates are similar to IDW estimates over smooth terrain, SRT estimates are notably superior over complex terrain (mountains) and in the vicinity of other climate forcing (e.g. ocean/land boundaries). Gridded data sets that result from IDW, Kriging or most other interpolation schemes do not provide unbiased estimates. Even when grid spacing is decreased to a point where the complexity of the land surface is well represented there remains two problems: what is the microclimate of the nearest observation points and what is the transfer function between points. This is a future challenge for increasing the quality of data sets and the estimation of data between observation sites.

## Author details

Kenneth Hubbard\*, Jinsheng You and Martha Shulski

\*Address all correspondence to: [khubbard1@unl.edu](mailto:khubbard1@unl.edu)

High Plains Regional Climate Center, University of Nebraska, Lincoln, NE, USA

## References

- [1] Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data* (3 ed.), J. Wiley and Sons, 604.
- [2] Camargo, M. B. P., & Hubbard, K. G. (1999). Spatial and temporal variability of daily weather variables in sub-humid and semi-arid areas of the U.S. *High Plains Ag. and Forest Meteor.*, 93, 141-148.



- [3] Eischeid, J. K., Baker, C. B., Karl, T., & Diaz, H. F. (1995). The quality control of long-term climatological data using objective data analysis. *J. Appl. Meteor.*, 34, 2787-2795.
- [4] Gandin, L. S. (1988). Complex quality control of meteorological observations. *Mon. Wea. Rev.*, 116, 1137-1156.
- [5] Guttman, N., Karl, C., Reek, T., & Shuler, V. (1988). Measuring the performance of data validators. *Bull. Amer. Meteor. Soc.*, 69(12), 1448-1452.
- [6] Guttman, N., & Quayler, R. G. (1990). A review of Cooperative temperature data validation. *J. Atmos. Ocean. Tech.*, 7(2), 334-339.
- [7] Hubbard, K. G., & You, J. (2005). Sensitivity analysis of quality assurance using spatial regression approach-a case study of maximum/minimum air temperature. *J. Atmos. And Oceanic Technology*, 22(10), 1520-1530.
- [8] Hubbard, K. G., Goddard, S., Sorensen, W. D., Wells, N., & Osugi, T. T. (2005). Performance of quality assurance procedures for an Applied Climate Information System. *J. Atmos. Ocean. Tech.*, 22(1), 105-112.
- [9] Meek, D. W., & Hatfield, J. L. (1994). Data quality checking for single station meteorological databases. *Agric. For. Meteor.*, 69, 85-109.
- [10] Shafer, M. A., Fiebrich, C. A., Arndt, D., Fredrickson, S. E., & Hughes, T. W. (2000). Quality assurance procedures in the Oklahoma Mesonet. *J. Atmos. Ocean. Tech.*, 17, 474-494.
- [11] Wade, C. G. (1987). A quality control program for surface mesometeorological data. *J. Atmos. Ocean. Tech.*, 4, 435-453.
- [12] Eischeid, J. K., Baker, C. B., Karl, T., & Diaz, H. F. (1995). The quality control of long-term climatological data using objective data analysis. *J. Appl. Meteor.*, 34, 2787-2795.
- [13] Evans, M. , Hastings, N., & Peacock, B. (2000). *Statistical Distributions* (3rd. Ed.), John Wiley and Sons, 221.
- [14] Gandin, L. S. (1988). Complex quality control of meteorological observations. *Mon. Wea. Rev.*, 116, 1137-1156.
- [15] Guttman, N. V., & Quayle, R. G. (1990). A review of cooperative temperature data validation. *J. Atmos. and Oceanic Tech.*, 7, 334-339.
- [16] Hubbard, K. G. (2001). Multiple station quality control procedures. In *Automated Weather Stations for Applications in Agriculture and Water Resources Management*. AGM-3 WMO./TD [1074], 248.
- [17] Hubbard, K. G., De Gaetano, A. T., & Robbins, K. D. (2004). Announcing a Modern Applied Climatic Information System (ACIS). *Bull. Amer. Meteorol. Soc.*, 85(6), 811-812.

- [18] Hubbard, K. G., Goddard, S., Sorensen, W. D., Wells, N., & Osugi, T. T. (2005). Performance of Quality Assurance Procedures for an Applied Climate Information System. *J. Atmos. and Oceanic Tech.*, 22, 105-112.
- [19] Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *J. of Computational and Graphical Statistics*, 5, 299-314.
- [20] Kunkel, K. E., Easterling, D. R., Hubbard, K., Redmond, K., Andsager, K., Kruk, M. C., & Spinar, M. L. (2005). Quality control of pre-1948 cooperative network observer data. *J. Atmos. Oceanic Technol.*, 22, 1691-1705.
- [21] Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous Univariate Distributions Volumes I and II* (2nd. Ed.), John Wiley and Sons, 761.
- [22] Martinez, J. E., Fiebrich, C. A., & Shafer, M. A. (2004). The Value of a Quality Assurance Meteorologist. Seattle, WA. *14th Conference on Applied Climatology, The 84th AMS Annual Meeting*.
- [23] Meek, D. W., & Hatfield, J. L. (1994). Data quality checking for single station meteorological databases. *Agric. and Forest Meteor.*, 69, 85-109.
- [24] National Weather Service. (1987). *Cooperative Program Management, Weather Service Operations Manual B-17 (revised)*, Silver Spring, MD, National Oceanic and Atmospheric Administration, 50.
- [25] Reek, T., Doty, S. R., & Owen, T. W. (1992). A Deterministic Approach to the Validation of Historical Daily Temperature and Precipitation Data from the Cooperative Network. *Bull. Amer. Meteorolog. Soc.*, 73(6), 753-765.
- [26] Wade, C. G. (1987). A quality control program for surface mesometeorological data. *J. Atmos. and Oceanic Tech.*, 4, 435-453.
- [27] Westphal, K. S., Vogel, R. M., Kirshen, P., & Chapra, S. C. (2003). Decision Support System for Adaptive Water Supply Management. *Journal of Water Resources Planning and Management*, 129(3), 165-177.
- [28] You, J., & Hubbard, K. G. (2006). Quality Control of Weather Data during Extreme Events. *J. Atmos. and Oceanic Tech.*, 23(2), 184-197.
- [29] You, J. K., Hubbard, G., & Goddard, S. (2007). Comparison of methods for spatially estimating station temperatures in a quality control system. *International J. of Climatology*, 28, 777-787, DOI: joc.1571.

