Automated Weather Sensor Quality Control

Douglas Galarus, Rafal Angryk and John Sheppard

Computer Science Department and Western Transportation Institute, Montana State University
dgalarus@coe.montana.edu, angryk@cs.montana.edu, john.sheppard@cs.montana.edu

Abstract

In this paper, we investigate the application of data mining to existing techniques for quality control/anomaly detection on weather sensor observations. Specifically we adapt the popular Barnes Spatial interpolation method to use time-series distance rather than spatial distance to develop an online algorithm that uses readings from similar stations based on current and historical observations for interpolation and we demonstrate that this new algorithm exhibits less model error than the Barnes Spatial interpolation-based method. We focus on interpolation, which is a basis for this popular quality control method and other related methods, and examine a dataset of over 233 million temperature observations from California and surrounding areas. Our approach shows improved performance as indicated by mean squared error reduced by approximately one half for predicted values versus reported values.

Introduction

With the advancement of computing and communication capability, the near-real-time information available from weather station sensors has increased dramatically in recent years and will continue to increase. These stations range from those maintained meticulously by the National Weather Service at airports nationwide to personal weather stations operated and maintained by individuals at their homes. The quality and accuracy of readings from these sensors can vary dramatically, as no system is immune to failure or mis-calibration. It may be desirable for certain applications, including basic assessment of conditions, to use as much of the sensor information available as possible.

There are many challenges associated with the problem of quality control (anomaly detection) for weather sensors because the types and sources of error are many. An error source may be a faulty sensor, but the error may not become apparent for some time if the readings drift slowly away from expected values. A sensor may be buried in snow or ice, reporting the temperature of the snow or ice rather than the ambient air temperature. A controller may have faulty logic and produce incorrect output. A weather station may have incorrect metadata associated with it and report an incorrect location while the readings its sensors produce are otherwise valid.

There are a number of automated quality control techniques that have been applied to assess the quality of individual sensor readings. Many of the more advanced techniques are geo-spatial in nature, based on spatial and temporal consistency assumptions, as well as an implicit assumption that most sensors provide accurate readings. Interpolation methods using readings in a geographically-defined neighborhood can be used to compute expected readings, and expected readings can then be compared to actual readings. In the event of a large deviation between actual and expected, a reading will be flagged as failing quality control or suspect of failing. General shortcomings include the challenge of selection of suitable tolerance levels and other parameters or bounds to determine outliers and the consequential balance of false positives (accurate readings classified as inaccurate) and false negatives (inaccurate readings classified as accurate).

We hypothesize that given historical data from stations and sensors, existing techniques can be enhanced using only the historical sensor data and data mining techniques to give better results in terms of mean-squared error (MSE) for interpolation of values. We hypothesize specifically that grouping stations based on similarity of sensor time-series and weighting observations from these stations accordingly results in lesser error when interpolating to predict the value at given station. We focus our attention in this study on temperature sensor readings, although the methods investigated should generalize to other weather sensor reading types that exhibit some degree of spatial and temporal consistency.

In this paper we present background, our approach, results, conclusions and future work.

Background

The WeatherShare system (http://www.weathershare.org/) was developed by the Western Transportation Institute at Montana State University in partnership with the California Department of Transportation (Caltrans) to provide a single, all-encompassing source for road weather information throughout California. Caltrans operates approximately 170 Road Weather Information Systems (RWIS) along state highways, thus their coverage is limited. With each deployment costing in the neighborhood of $100,000, it is unrealistic to expect pervasive coverage of the roadway from RWIS alone. WeatherShare aggregates weather data from other third-party aggregation sources such as NOAA's Meteorological Assimilation Data Ingest System (MADIS) (http://madis.noaa.gov/), along with Caltrans RWIS to present a unified view of weather data from approximately 2,000 stations within California. A primary benefit of the system is far greater spatial coverage of the state, particularly roadways, when compared to the Caltrans RWIS network alone. A secondary benefit of the system is the ability to compare RWIS readings with those of other nearby sensors to assess accuracy. Formal, automated quality control procedures have been implemented in the WeatherShare system to assess sensor accuracy for not only Caltrans RWIS, but also all other sensor readings stored in the system. (Richter et al. 2009)

The WeatherShare system has not fully achieved the secondary goal of increasing Caltrans ability to assess RWIS sensor accuracy in an efficient manner guided by automated procedures. This shortcoming stems from both limitations in implemented procedures, as well as unrealized potential to deliver this information through an easy to use and informative interface.

MADIS is an online database of real time and archived weather data including sensor readings from nearly 40,000 stations in North America, including Hawaii and Central America. For the meteorological surface dataset, MADIS implements three levels of automated quality control. The "Level 1" quality control checks are also referred to as "validity checks" or "range checks". They check that a sensor reading is within a range of predetermined values indicating the "tolerance limits" of that sensor reading type. The range for air temperature is given as [-60° F, 130° F]. There are three "Level 2" quality control checks: "internal consistency," "temporal consistency," and "statistical spatial consistency."

The Level 3 quality control check is referred to as the "spatial consistency" or "buddy" check, and is a variant of the Optimal Interpolation (OI), technique (Belousev et al 1968). For a given station and observation, an interpolated value is determined for that station using neighboring stations and excluding the station being analyzed. If the difference between the actual value and the interpolated value is "small," then the station is considered to be in agreement with its neighbors and it passes the spatial consistency check. However, if the difference is not small, then the interpolation and analysis is repeated with one of the neighboring observations removed. If the removal of the neighbor results in a small difference between the interpolated and observed value, then the observation is flagged as "good" and the neighboring observation is flagged as "bad".

The *Clarus* initiative (http://www.clarusinitiative.org/) was established in 2004 by USDOT Federal Highway Administration Road Weather Management Program and the Intelligent Transportation Systems Joint Program Office to "reduce the impact of adverse weather conditions on surface transportation users." Specifically, *Clarus* was built to collect atmospheric and pavement observations from state-owned road weather information systems in near real time.

The *Clarus* System (http://www.clarus-system.com/) provides ESS data from participating states and Canadian provinces. At present, there are 38 participating states and 4 participating provinces. Data is available for California and it is provided by Caltrans to the *Clarus* system in the same manner that it is provided to WeatherShare. Current data is provided via an online graphical user interface, and archived data is also available for download. *Clarus* implements nine quality control algorithms including a Barnes Spatial Test (Pisano et al. 2007). The Barnes Spatial Test is based on the Barnes Spatial Interpolation Scheme (Barnes 1964), which uses a Gaussian filter to interpolate values over a two dimensional area using known readings within that area.

The Barnes Interpolation Scheme is used as a basis for the Barnes Spatial Quality Control test, as applied by the Oklahoma Mesonetwork (Shafer et al. 2000), which uses one pass of the Barnes Interpolation Scheme to estimate values for each observation. The Barnes Interpolation Scheme does not account for elevation. Since it was developed in Oklahoma, its application to areas with little variation in terrain may be reasonable. However, its use may be limited in areas with mountain terrain.

There are other notable approaches to modeling weather sensor data for the purposes of quality control and anomaly detection. The Utah Mesonet (Mesowest) uses linear regression (Split et al) to incorporate elevation into an interpolation model and subsequent quality control checks for temperature, dewpoint and pressure. PRISM (Precipitation-elevation Regressions on Independent Slopes Model), developed at Oregon State University, accounts for elevation and general topographic impact on weather variation, creating a grid of estimated precipitation using station readings that fall within topographically-similar facets (Daly et al. 1994).

Our Approach

For our experiment, we investigated air temperature observations only. Data were used from MADIS covering a rectangular region that includes all of California and portions of Oregon, Nevada, Idaho and Arizona. We restricted our attention to stations for which their locations, including elevation, were consistent throughout the entire period for which data was available. Note that locations of many individual stations changed over time. This is a consequence either of stations being mobile in nature, including ship-based maritime stations, as well as stations for which locations were not and may still not be reported accurately. We used temperature data from July 2001 through December 2010.

Table 1 shows quality control descriptors associated with each sensor reading in the MADIS data set. The dataset provides additional detail indicating which tests were applied and which resulted in failure for the reading.

Table 1: MADIS Quality Control Descriptors

B	subjective bad
C	coarse pass, passed level 1
G	subjective good
Q	questioned, passed level 1, failed level 2 or level 3
S	screened, passed level 1 and level 2
V	verified, passed level 1, level 2 and level 3
X	Rejected/erroneous, failed level 1
Z	preliminary, no quality control check

Although not an emphasis of this study, we pre-processed data using a [-60° F, 130° F] range check, in conformance to that used for MADIS. It is recognized that such preliminary checks and filters are key to the performance of the more advanced (Level 3) spatial checks. In effect, this pre-processing removed all observations having an "X" quality control descriptor. There were some observations in the data set having quality control descriptors other than "X" which also failed this range test and these were removed also.

None of the techniques documented so far in this paper are perfect, and there may be room for improvement in each. In this project, we selected one of these methods, the Barnes Spatial Interpolation Scheme, for comparison and prospective enhancement using data mining techniques. In general, California offers an ideal setting for the evaluation of quality control procedures because of its geographic and meteorological diversity. California includes coastal areas, mountains, deserts, rain forests, and both the highest and lowest points in the contiguous 48 states. While the Barnes Spatial Interpolation Scheme is widely applied, it is also susceptible to the challenges of varying terrain.

For our work, we represent a temperature observation o as a 3-tuple $o = (s, t, v) = (o_s, o_t, o_v)$, consisting of the station, time and the value (° F) of the observation. Since the time and frequency of observations vary from station to station, we adopt a convention for interpolation at time t of using the most recent reading $\rho(s, t, c)$ from any given station s at time t and within a time cutoff c:

$$\rho(s, t, c) = \arg\max_{o \in O} \{o_t : o_s = s, t - c < o_t \leq t\}.$$

For the experiments presented here we use $c = 60$ minutes. We compute the distance between two stations $\delta(s_1, s_2)$, as the great circle distance between the stations.

The Barnes Spatial Interpolation Scheme uses a Gaussian filter to interpolate values over a two dimensional area using known readings within that area. Using our notation and conventions from above, then the value interpolated to correspond to an observation o is:

$$g(o) = \frac{\sum_{s' \in S'} \eta\big(\delta(o_s, s')\big) \cdot \rho(s', o_t, c)}{\sum_{s' \in S'} \eta\big(\delta(o_s, s')\big)}$$

where $\eta(d) = e^{\frac{-d^2}{4k}}$ is the filter and

$$S' = \{s' \in S : s' \neq o_s, |\rho(s', o_t, c)| > 0, \delta(o_s, s') < r\}$$

is the set of stations for which there are observations within the time cutoff c and within a distance cutoff r. We use $r = 70$ miles as the distance cutoff. The parameter k determines the shape (wide versus narrow) of the filter. Guidance is given in selecting k by using the equation $\frac{r^2}{4k} = -\ln \epsilon \equiv E$, where r is the radius of influence and $1 - \epsilon$ is the desired influence. Selection of E=4 results in approximately 98 percent influence ($e^{-4} \approx .0183$). We use the corresponding value $k = 70^2/(4 \cdot 4)$.

Note that our objective here for quality control assessment is not the interpolation over the entire area of interest but rather interpolation at the location of each observation to compare the interpolated value against the actual value of the observation. If the interpolated value deviates greatly from the actual value, then the actual value may be flagged as suspect. Further note that Barnes applied multiple iterations of interpolation to improve the fit of the model to the underlying data. We follow the convention of others in using a single iteration of interpolation for quality control assessment.

Background information appears to be a key to dealing with diversity of terrain and climate in some methods. The PRISM system, while using a very simple model at grid point level, makes extensive use of background information to improve its accuracy, including the incorporation of human intervention and tuning. As a

whole, it is a rather complex model. We believe that with the ever increasing number and distribution of environmental sensors, there is an opportunity to develop comparable models based solely on historical data for the purpose of quality control. We further believe that such an approach is advantageous because it does not require a domain expert for development, interpretation or tuning.

Given basic assumptions of spatial and temporal consistency, and given some reasonable assumption about the trusted operation of a high proportion of sensors in a probabilistic sense or perhaps a lesser number of trusted stations and sensors distributed throughout a region, the generation of such models should be dependent only on the amount of historical data available and the spatial distribution and density of the sensors. We believe that with nearly 2,000 stations in California, there is sufficient density for application in a large portion of the state. Further, we recognize that the number of available stations will increase over time, particularly if data from unofficial, personal weather stations is used.

We believe such an approach can be advantageous over the other methods that make extensive use of background information and also over methods that make assumptions related to uniform spatial consistency. For these methods, the naïve assumptions of uniform spatial proximity might be replaced with time-series distance to indicate (dis)similarity of stations based on historical data. Further, stations can be grouped based on this same time-series distance or (dis)similarity measure to form a radius-based or nearest neighbor-based grouping, per station, as is used by the Barnes Spatial and Optimal Interpolation methods, respectively.

Time series distance can be computed in many different ways. In our experiment, we implemented an online approach, which continually updates the time series distance between stations as new observations are reported by using the sum-of-squares difference between each observation and the most recent observations from other stations. We define the time series distance τ between stations s_1 and s_2 at time t as:

$$\tau(s_1, s_2, t) = \sqrt{\frac{\sum_{o \in O_{s_1,t}} \left(o_v - \rho_v(s_2, o_t, c)\right)^2}{\left|O_{s_1,t}\right|}}$$

where $O_{s_1,t} = \{o \in O : o_s = s_1, o_t \leq t\}$ is the set of all observations from station s_1 at or prior to time t.

Using this measure, we developed a new method called Time-Series-Distance-Filter Interpolation (TSDFI) as a variation of the Barnes Spatial method by replacing the station to station distance measure δ with τ. Then the interpolated value $h(o)$ corresponding to an observation o is:

$$h(o) = \frac{\sum_{s' \in S'} \eta\left(\tau(o_s, s')\right) \cdot \rho(s', o_t, c)}{\sum_{s' \in S'} \eta\left(\tau(o_s, s')\right)}$$

where $\eta(d) = e^{\frac{-d^2}{4k}}$ is the filter and

$$S'' = \{s'' \in S : s'' \neq o_s, |\rho(s'', o_t, c)| > 0, \tau(o_s, s'') < q\}$$

is the set of stations for which there are observations within the time cutoff c and within a time series distance cutoff q. We use $r = 5^\circ$ F corresponding to a station-to-station root-mean-squared error of 5. $k = 5^2/(4 \cdot 4)$ is computed accordingly.

Both the Barnes Spatial method and our new TSDFI method use Gaussian Filters to interpolate values, creating a model of reported data. Both were implemented in an online fashion, with the dataset processed chronologically, modeling each sensor value in the dataset. The error of reported versus predicted was recorded for each sensor value and subsequently aggregated as mean-squared-error for comparison.

Results

Over 233 million temperature observations from 2001-2010 and the corresponding predicted values using Barnes Spatial (Barnes) and the Time Series Distance Filter Interpolation (TSDFI) methods were analyzed. The TSDFI method had an overall mean-squared error of less than half that of Barnes over the entire data set. The mean-squared error for TSDFI is consistently less than that for Barnes Spatial over time by nearly a factor of two. Figure 1 shows peaks and troughs in the MSE for both methods, with peaks occurring in proximity to June of each year and troughs in proximity to December. Further investigation is necessary to determine if this is a consequence of normal seasonal variability in the underlying data and whether there is a need to account for such variability in the models. For instance, should there be separate summer and winter models?

The mean-squared errors for the Barnes and TSDFI methods yield promising results when grouped by observations according to the MADIS quality control descriptors. Recall that all readings flagged with quality control descriptor X, rejected/erroneous, were removed prior to application of the interpolation methods. For those readings that passed all three levels of MADIS quality control, labeled V, Table 2 shows that the mean-squared error for TSDFI is very small, and approximately one-third that of the Barnes Spatial method. For those flagged Q for

questionable, having failed level 2 or level 3 in MADIS, the MSE for the TSDFI method is very high and comparable to that of Barnes Spatial. This is not problematic since the high MSE is attributable to questionable and likely erroneous sensor readings rather than model error, and would be indicative of such readings. Results corresponding to other MADIS quality control descriptors show similar results and appear to indicate that the TSDFI method may discriminate valid versus erroneous readings relatively well.

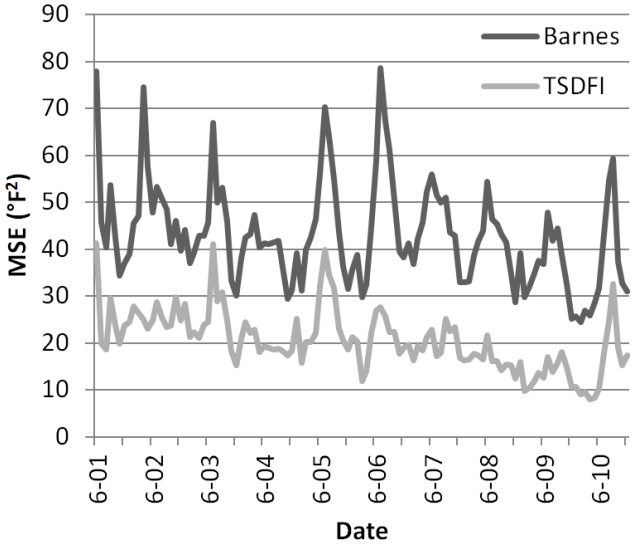

Figure 1: MSE (°F^2) Over Time for Barnes versus TSDFI. The MSE for TSDFI is approximately half that for Barnes.

Table 2: MSE for Barnes versus TSDFI by MADIS QC Indicator. The MSE for TSDFI is less than that for Barnes in every case and less than one fourth that for Barnes for verified (V) readings.

		Barnes	TSDFI
	Count	MSE	MSE
B	300	55.82	60.54
C	439,630	109.88	51.57
G	206,873	23.01	15.52
Q	13,709,191	246.13	158.28
S	21,337,164	38.93	16.32
V	197,415,497	25.86	8.01
Z	28,862	49.12	22.42
	Overall	40.17	17.70

By comparing TSDFI error to Barnes Spatial error for individual stations, we can speculate on reasons for large differences, including those that may be attributable to quality control problems for stations. Most stations have modest MSE values (less than 100) for both Barnes and TSD. There are some that have large MSE for both. See Figure 2 and Figure 3.

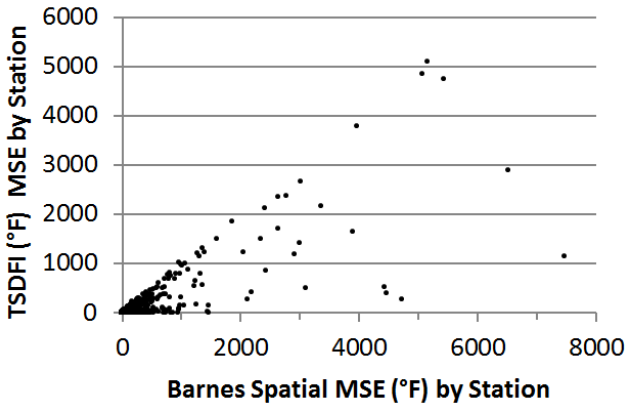

Figure 2: Plot of Barnes Spatial and TSDFI Mean-Squared Errors by Station. The MSE for TSDFI is less than that for Barnes with very few exceptions. Note some large errors for both methods.

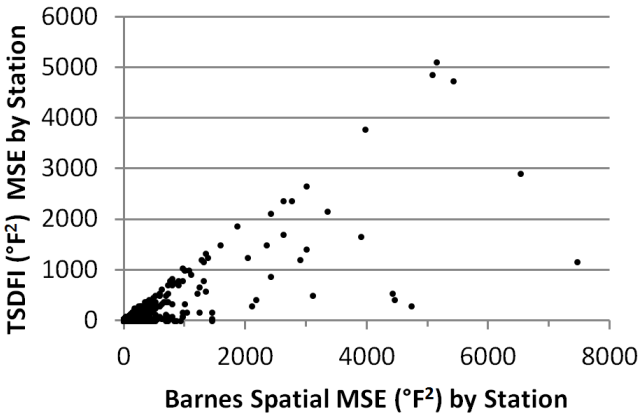

Figure 3: Collapsed Plot of Barnes Spatial and TSDFI Mean-Squared Errors by Station. The MSE for TSDFI is less than that for Barnes with very few exceptions. Note that most of the data for TSDFI has a MSE less than 20.

Consider station F2988, for which the reported location is 37.07°N, 119.03°W. MADIS reports an elevation near zero for this station, which is odd since the latitude/longitude corresponds to a point in the Sierra Nevada mountain range. The MSE for Barnes Spatial for this station is 89.321. The MSE for TSDFI for this station is 1.513. This indicates that other stations in proximity to this station do not match it well while there are other stations which are better matches. This seems to indicate that the reported station location is incorrect. However, there are only 22 readings for this station, and this discrepancy may simply be a consequence of lack of data. For station SCNC1, the MSE for Barnes Spatial is 5,155.291 and the MSE for TSDFI is 5,097.405. This station is located on San Clemente Island, which is over 60 miles off the California coast. While there are several other stations located on San Clemente Island, there are very few additional stations in proximity. Either this station is

problematic in general or there are not enough stations in proximity for comparison using either method. In fact, the "closest" station in terms of time series distance is station F1426, which is located at Camp Pendleton, north of San Diego and over 60 miles away.

Conclusions and Future Work

The TSDFI method shows sufficient promise to merit subsequent research and development into a related method for anomaly detection. TSDFI reduces overall model error in comparison to Barnes Spatial by grouping stations based on similarity of sensor time series and weighting them accordingly rather than by using spatial distance. Intuitively this approach should not be prone to problems of over-fitting, which could mask sensor error. The results presented used parameters for Barnes Spatial that are cited in other efforts but that have not been optimized, and arbitrary parameter value choices for TSDFI. There is further room for improvement by optimizing parameters for these models, including the potential to vary parameters on a per-station basis. And, it would be worthwhile to investigate a hybrid method that combines both time-series distance and spatial distance. Elevation might also be accounted for directly using similar approaches.

It would also be desirable to investigate varying time periods for both time series distance calculation and prediction. In our investigation we used data from June 2001 through December 2010. It is important to determine how much data is necessary to develop a model that is applicable year-round and to subsequent years. We suspect that seasonal patterns will have an impact on performance.

We believe that these techniques should be applicable to other sensor reading types – wind and precipitation, for example. These reading types will certainly exhibit behavior different than temperature, although some degree of spatial and temporal consistency will still be assumed. Such readings may not be available from as many stations as for temperature and we note that not all stations offer the same suite of sensor types. Such challenges present further opportunity. It would be beneficial to investigate combining different sensor type readings to better characterize similarities between stations, and subsequently analyze individual readings for their validity.

Finally we note that ensemble methods may be applicable. Prior work indicates motivation for each of the models discussed relative to the original area in which it was applied – Barnes Spatial was developed in Oklahoma, which is relatively flat; Optimal Interpolation is said to be more robust than Barnes Spatial in that it forces comparison with stations in all directions; the regression techniques used by Mesowest and PRISM incorporate elevation into their models to account for the impact of terrain; and, PRISM incorporates other information to create “facets” containing related stations. It may be the case that an ensemble of several or all of these approaches would be more accurate. We believe, though, that a technique similar to the one we present in this study could be equally effective given sufficient historical data and spatial coverage.

Acknowledgments

We acknowledge the California Department of Transportation (Caltrans) for its sponsorship of the WeatherShare project and other related projects. In particular, we acknowledge Ian Turnbull, Mandy Chu and Sean Campbell from Caltrans. We further acknowledge staff at the Western Transportation Institute for their work on the WeatherShare project. The work presented in this paper has been conducted subsequent to and separate from that prior work.

References

Barnes, S. L. 1964. A Technique for Maximizing Details in Numerical Weather Map Analysis," *Journal of Applied Meteorology*, vol. 3, pp. 396-409.

Belousov, S.L., L.S. Gandin, and S.A. Mashkovich, 1968: *Computer Processing of Current Meteorological Data*. Ed. V. Bugaev. Meteorological Translation No. 18, 1972, Atmospheric Environment Service, Downsview, Ontario, Canada, 227 pp.

Daly,C.; Neilson, R. P. and Phillips, D. L. 1994. A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain. *Journal of Applied Meteorology*, vol. 33, pp. 140-158.

Pisano, P. A.; Pol, J. S.; Stem, A. D.; Boyce, B. C. B and Garrett, J. K.. 2007. Evolution of the U.S. Department of Transportation *Clarus* Initiative: Project Status and Future Plans. Web.. 20 Nov. 2011 <http://ops.fhwa.dot.gov/weather/resources/publications/fhwa/evolclarusams07.pdf>

Richter, D.; Wang, S. and Galarus, D. 2009. WeatherShare Phase 2 Final Report. Western Transportation Institute, Montana State University, Bozeman, MT. Web. 20 Nov. 2011 <http://www.westerntransportationinstitute.org/>

Shafer, M. A.; Fiebrich, C. A.; Arndt, D. S.; Fredrickson, S. E. and Hughes, T. W. 2000. Quality Assurance Procedures in the Oklahoma Mesonetwork, *Journal of Atmospheric and Oceanic Technology*, vol. 17, pp. 474-494.

Splitt, M. E. and Horel, J.. University of Utah Department of Meteorology. Use of Multivariate Linear Regression for Meteorological Data Analysis and Quality Assessment in Complex Terrain.Web. 20 Nov. 2011 <http://mesowest.utah.edu/html/help/regress.html>