J3.9 OPPORTUNITIES FOR IMPROVEMENTS IN THE QUALITY CONTROL OF CLIMATE OBSERVATIONS

Christopher Daly*, K. Redmond, W. Gibson, M. Doggett, J. Smith, and G. Taylor, P. Pasteris and G. Johnson
15th AMS Conf. on Applied Climatology, Amer. Meteorological Soc., Savannah, GA, June 20-23, 2005

1. INTRODUCTION

The methods by which climate data are being collected and used are expanding rapidly. Increasingly, data are being collected by automated electronic systems, with a wider variety of platforms, at higher temporal resolution, at more locations, and in more difficult and remote environments. Automated climate observations at various time steps are being disseminated over near-real time communication networks, for use by a variety of software applications, to fill a wider range of needs, in an increasingly automated and digital world. The development of the ASOS (Automated Surface Observing System), SNOTEL (Snowpack Telemetry), RAWS (Remote Automated Weather Station), Agrimet, innumerable mesonets, and the prospect of COOP modernization, all reflect an increased reliance on electronic sensors, data from remote environments, and automated, real-time data delivery systems.

This shift in observational strategy is largely the result of technological advances in areas such as electronics, communications, and computing, that have increased both the supply of, as well as demand for, climate data. Some of these same technological advances now provide opportunities for qualitatively different approaches to quality control, methods that are sophisticated, largely automated, data-rich, updatable, and capable of furnishing quantitative error and confidence information. Such methods were infeasible to implement and operate just a few years ago. This paper explores some of the characteristics of such methods, and discusses why they are both beneficial and practical. The characteristics discussed here are summarizes as follows:

- Data-rich QC frameworks (cross-network comparisons, variety of data sources)
- Continuous, quantitative estimates of observation validity and uncertainty in estimation
- Transparent dissemination of results that allow end-user interaction, education, and decision-making
- Automated QC systems that can improve through experience and feedback, and are applied retrospectively at regular intervals.

The ultimate goal is a QC approach that is self-consistent and physically plausible, in accord with known principles of how the atmosphere works, and that can be updated to reflect changes in our knowledge base.

The first generation of a spatial QC system recently developed for USDA Natural Resources Conservation Service (NRCS) SNOTEL temperature data is presented as an example of a QC system that is in the early stages of incorporating these characteristics. While this one example does not adequately represent the current state of the science pertaining to all the characteristics listed, it does provide a useful perspective on the needs and requirements of a QC system operating in today's environment The focus of the examples in this paper is primarily on daily totals and extremes (max/min temperature and precipitation), but the ideas discussed should be applicable to other time scales.

1.1. Data-rich QC Frameworks

Automated data collection systems, by definition, do not require a constant human presence, and many operate without AC power. Thus, the number and variety of potential locations is quite large, and these systems are often used to monitor difficult and extreme (e.g., mountains, wilderness, marine) environments far from permanent human habitation. Intelligently assessing the quality of the data collected in such environments is difficult, and requires a sophisticated, data-rich QC system that can account for factors controlling the spatial and temporal patterns of climate in the area. Simple, data-sparse QC systems sometimes rate observations taken in remote areas as poor, not because the data are actually poor, but because data values may be at the edges of the "typical" range, and the data may not seem to be directly consistent with other observations in the regions.

Data that could be used to improve a QC system are numerous and varied. Four types are discussed here: (1) *in situ* observations; (2) satellite and radar data; (3) diagnostic grids interpolated from observations; and (4) values and QC results from other observed elements. There are many national surface station networks available in the US, including SNOTEL, COOP, ASOS, and RAWS. Also valuable to the QC effort, especially at high elevations, are upper-air observations taken by the National Weather Service. Data from various mesonets and other sources could improve the quality of estimates through increased initialization and verification density, especially in previously data-sparse areas. However, there is a balance that must be struck between inclusion of additional data sources, and the quality of the added data. Additional poor-quality data can increase, rather than decrease, uncertainties in the QC system.

* *Corresponding author address*: Christopher Daly, Spatial Climate Analysis Service, Oregon State University, 326 Strand Ag Hall, Corvallis, OR 97331; email: daly@coas.oregonstate.edu

Non-station data sources include satellite radiances, snow cover, and infrared skin temperature; and radar precipitation estimates. All have serious limitations, and are useful only under certain conditions. Precipitation estimation from satellite and radar is very difficult in complex terrain, where snow dominates on an annual basis, winter clouds are shallow, bright bands are present, and sub-cloud air is very dry, so may be of limited usefulness for quality controlling networks such as SNOTEL. Infrared skin temperature and snow cover estimation suffer from cloud interference nationwide, resulting in intermittent spatial and temporal coverage.

However, if used carefully and restricted to the appropriate regions and seasons, these kinds of data can provide useful independent evaluation of observations. During the winter season, there can be severe undercatch of precipitation from surface networks using unshielded tipping bucket gauges (such as the current ASOS system), in some cases rendering these observations useless. The National Centers for Environmental Prediction (NCEP) are currently creating Stage IV precipitation estimates using a combination of surface and radar based data. The radar-based precipitation estimates could be used to detect suspicious surface based observations showing little or no precipitation. However, radar data also suffer from snow underestimation due to typically poor radar returns from snowfall or inability to see shallow winter clouds. Radar can be more helpful in summer, with tall convection not as susceptible to terrain blocking, to establish precipitation presence/absence.

The National Climate Data Center (NCDC) is taking a multi-sensor approach to the real-time QC of precipitation data in their PrecipVal system (Urzen et al. 2004). Data layers used (when available) are station data, radar, and satellite data, and rapid update cycle (RUC) model output. Observation confidence is generally based on the number of independent data layers agreeing with the observation. There is much work to be done to quantitatively determine, given the location and weather situation, uncertainties in each of these independent data layers. Here again, adding poor-quality data can increase, rather than decrease, uncertainties in the QC system.

As will be discussed in the presentation of the SNOTEL QC system in Section 2, higher interpolation skill is often obtained by employing in the interpolation process a predictive grid that represents long-term climatological patterns for that day or month. The only grids commonly available for this purpose represent thirty-year means. This may be problematic on days when the spatial patterns deviate significantly from the mean. A more robust solution would be to use spatial climatologies that are targeted to the conditions at hand. A useful first step for temperature would be to construct two sets of monthly climatologies: one for clear, dry, stable conditions, and one for moist, well-mixed storm conditions. These would account for many of the day-to-day differences in temperature patterns, and thus improve the QC estimates. Precipitation patterns are more varied and not as easily defined. Oregon State University's Spatial Climate Analysis Service (SCAS) is currently working with the National Weather Service Western Region's River Forecast Centers to define and create gridded precipitation climatologies that are targeted to specific storm conditions of interest to flood forecasting. However, these climatologies are not easily generalized over large areas.

For QC systems having access to multiple climate elements, there is the opportunity to create relationships between the values and QC results of these other elements. Obvious inconsistencies between variables can be easily checked (e.g., ensure that maximum temperature is greater than minimum temperature). Beyond these simple checks, it is possible to develop joint QC results that feed on the values of several elements, and thus strengthen the overall results. Values of other elements, such as precipitation, are also useful in determining the targeted temperature climatologies discussed in the previous paragraph.

1.2. Continuous, quantitative estimates of observation validity and estimation uncertainty

Traditional QC methods for surface observations have employed a series of categorical quality checks that an observation must pass if it is to be considered valid. The outcomes of these checks are typically of a "yes" or "no" nature, and the observation is flagged with notations based on these outcomes. For example, the NOAA Forecast Systems Lab is operating a near real-time Meteorological Assimilation Data Ingest System (MADIS) for providing quality controlled surface observations. Flags are available for each observation based on different levels of quality, such as: 1) validity check of data range; 2) internal consistency between various elements; 3) spatial or temporal consistency; and 4) subjective override of automated checks. For some elements, such as precipitation, only temporal flags are present.

There is much benefit in estimates of observational validity and estimation uncertainty that are quantitative and continuous, rather than categorical. Needs for quantitative and continuous estimates stem from two main sources: (1) errors from electronic measurement systems can suffer from drift (i.e., calibration issues) as well categorical mistakes (e.g., using an incorrect algorithm to convert voltage to temperature); and (2) computer models (e.g., hydrologic models) that rely on climate observations as input benefit from quantitative estimates of uncertainty. Providing such quantitative uncertainties requires that the statistical performance of a QC system be known, and incorporated into the results (see Section 1.4).

1.3. Transparent dissemination of results

The range of applications for climate data is rapidly increasing, and each application has a different tolerance for low-confidence data points. One user's unusual observation may be just what another user is looking for. Unfortunately, there is often very little information available to users about how QC systems arrived at their conclusions, (e.g., assumptions made,

uncertainties in the QC results, thresholds for data inclusion, etc.), making it difficult to make educated choices about how the data should be used and interpreted for a particular application.

We suggest a new approach, where full quantitative QC results are presented to the user in a transparent, interactive format. Full disclosure allows users to make alternate decisions about the QC results and an interactive system allows users to generate their own data sets, based on suggested default thresholds for data inclusion, or their own thresholds. This approach implies that there is no single data set that everyone must use. However, there will always be a need for an "official" QC'ed data set for the majority of users who do not choose to deviate from the suggested guidelines. This kind of product can and should be provided, so long as the process and assumptions used in creating the data set are documented in an easily understandable way, and accessible to any interested user.

1.4. Automated QC systems that are regularly improved and re-applied retrospectively

Data generated by automated electronic systems are often more voluminous (e.g., shorter time step, or higher temporal resolution) and disseminated in a much more timely manner than those from manual systems. Historical climate data archives are growing rapidly, requiring historical QC systems to handle large amounts of data. Whether operating on near-real time data, historical data, or both, largely automated QC approaches are required.

In order for QC systems to be optimally effective, climate data QC must be designed as an ongoing process of application and improvement of the system, not a one-time decree of data validity. Traditionally, a QC'ed data set has been thought of by users as one that has been "cleaned up," and that no "bad" values remain. This implies that there is a central authority that knows good data from bad, and that QC is just a matter of subjecting the data to this authority and getting a thumbs up or thumbs down. This kind of misleading and oversimplified perception places an unreasonable burden on the QC system to judge a data point usable or unusable for everyone and for all time. In treating data QC as process, the expectation will be that QC systems can and will be improved, and will be re-applied retrospectively on a periodic basis to reflect these improvements.

The first step in improving the system is to evaluate its current performance. QC systems should not be evaluated based on how many data values they find to be bad. As discussed earlier, good data are often flagged as bad by data-poor QC systems, because they are unable to account for the factors that produce spatial or temporal differences in climate data.

One effective way to determine the skill of a QC system is to go back to basics, and assess its ability to estimate for stations whose observations are known with reasonable confidence. One could argue that most stations maintained by national networks such as NWS COOP fall into this category, because the observational errors associated with these stations are probably much smaller than the estimation errors of a QC system. Unlike climate interpolation, where the actual field is unknown and evaluation must be done indirectly, we know what was observed at each station, which allows direct comparisons between observations and estimates. If the system can't estimate the observations well, then it will have little skill in determining data validity.

Because there are a seemingly infinite number and variety of individual situations a QC system must handle well, overall performance statistics only tell part of the story. Classifying performance statistics into weather types as discussed in Section 1.2 may give specific information on when and where the QC system needs improvement.

Equally important input into the evaluation process comes from manual spot check evaluations, in which the decisions of the QC system are compared with those of human decision makers. This requires much time and effort over a long period, usually coming in the form of feedback from users and developers as the QC system is in operation. A key part of the feedback from human evaluation, especially when the QC system appears to be making erroneous assessments, is to identify what information the person accessed, and how it was processed, to make this determination. Finding ways to include this information in the QC system's repertoire of data sets and algorithms is a major part of the improvement process.

In the end, the fundamental dilemma with nearly all quality control is a tension between the relative merits and costs of accidentally rejecting good data, or accidentally accepting bad data. A tradeoff is usually involved.

2. EXAMPLE: QUALITY CONTROL OF SNOTEL DATA

2.1. Background

In the mid and late 1990s, the SCAS developed new precipitation maps for the United States (USDA-NRCS, 1998; Daly and Johnson, 1999). SNOTEL was the primary high-elevation network used for the mapping and proved to be essential for map development. In addition to precipitation data, the more than 700 SNOTEL stations report temperature and snow water equivalent. SNOTEL data are recorded electronically and transmitted to data collection centers. The stations are in remote areas with limited winter access, and thus must operate unattended for long periods of time in difficult weather conditions. The data have never undergone complete spatial quality assurance and quality control corrections. Work within the USDA-NRCS and the Western Regional Climate Center had attempted to accomplish this, but was never fully completed.

In 2002 the NRCS asked the SCAS to develop a formal QC system for their SNOTEL data products, based upon SCAS spatial QC tools. The system was to

be used to QC historical daily data over the SNOTEL period of record (beginning in about 1980), and subsequently installed and operated at NRCS to QC daily data in near real-time. The project was to address temperature first, then move to precipitation and snow water equivalent. The resulting, first-generation SNOTEL QC system for temperature, termed the SNOTEL Probabilistic-Spatial Quality Control (PSQC) System, is described below.

2.2. Overview of the SNOTEL PSQC System

The PSQC system for SNOTEL is spatially-oriented, uses a knowledge-based system to make predictions, and ingests a variety of spatial data sets (Daly et al. 2004, Gibson et al. 2004). It operates on the premise that spatial consistency, if assessed accurately, is a useful indicator of data validity. A climate estimate is made at a station location when the station's data value is withheld from the interpolation. If there is a large discrepancy between the station value and the estimate at the station's location, and the ability of the system to judge data quality is accounted for, the probability of the observation being correct may be low. The goal of the QC process is to, through a series of iterations, gradually and systematically "weed out" spatially inconsistent observations from consistent ones. This process is necessarily an iterative one, because the validity of an observation is assessed through surrounding observations, which themselves may be in error.

The predictive tools are based on PRISM (Parameter-elevation Regressions on Independent Slopes Model), a knowledge-based climate mapping system developed at Oregon State University (Daly et al., 1994, 2002, 2003). PRISM provides a relatively high degree of skill to the spatial interpolation process, especially in complex regions.

Experience has shown that higher interpolation skill for daily temperature is obtained by running PRISM using a high-quality predictive, or "background," grid that represents the long-term climatological temperature for that day or month, rather than a digital elevation grid. Such background grids have the expected spatial patterns of climatological temperature built in to provide increased explanatory power. This is sometimes referred to as climatologically aided interpolation (CAI).

Under USDA-NRCS funding, work is underway at SCAS to produce new 1971-2000 monthly average minimum and maximum temperature grids at 30-sec (0.8-km) resolution for the United States (Doggett et al., (2004). The 0.8-km grid cell size captures a good deal of the topographic variability in mountainous regions. Initial drafts of these grids are being used as the predictive grids for the PRISM PSQC system. Figure 1 is an example of a PRISM regression function, showing a local regression between observed maximum temperatures for 20 July 2000 and their 1971-2000 climatological mean values for the month of July. Over long periods, the slope of this linear regression function should average out to approximately 1.0, but may vary appreciably during individual days. As discussed in Section 1.2, further interpolation improvements could be realized by employing gridded climatologies targeted to a specific type of weather pattern.

An extensive array of spatial information interacts with the PRISM knowledge base to weight stations entering the regression function. These include grids of elevation, topographic facets (at six different scales), coastal proximity, inversion height, topographic index, and effective terrain height (see Daly et al. 2002 for details).

2.3. The PSQC Process

The QC process consists of two nested loops: a daily loop inside an iterative loop (Figure 2). In the daily loop, PRISM is run for each station location for each day within the period of record, and summary statistics accumulated. Once all days have been run, confidence probabilities (*CP*) for each daily station observation are estimated (discussed below). In the outer *CP* iteration loop, these *CP* values are used to weight the daily observations in a second series of PRISM daily runs. Observations (*O*) that have lower *CP* values are given lower weight, and thus have less influence, in the second set of PRISM estimates, and are also given lower weight in the calculation of the second set of summary statistics. *CP* values are again calculated and passed back to the daily PRISM runs. This iterative process continues until the change in *CP* values between the present and previous iterations falls below a threshold "equilibrium" level, at which time the process stops and summary QC information is produced. The number of iterations required to reach equilibrium typically ranges from one to five.

Variables calculated during the QC process are listed in Table 1. They fall into three main categories: (1) PRISM variables, (2) summary statistics, and (3) probability statistics. During the daily loop, PRISM is run in point mode to obtain an estimate, referred to here as a "prediction," *P*, for each station location for each day. First, a prediction is made for the target station in its absence, using all available observations from surrounding stations for the PRISM regression function (Figure 1),. The process is then repeated several times while deleting nearby observations, first singly, then in pairs, with replacement. The cycles of deletion are performed to preclude highly erroneous observations from contaminating the predictions. It is assumed that the chances of more than two erroneous observations occurring in the immediate vicinity of each other on a given day are small. The residual, *R, (R = P - O)* and the PRISM regression standard deviation, *S*, are calculated and summed to obtain a score for each station deletion scenario. The scenario that produces the lowest score is accepted, the associated values of *P, R*, and *S* recorded for that day, and the deleted observations replaced in the data set. *T* and *V*, the temporal variability statistics, are also calculated for each day. These are discussed as improvements to the PSQC system in Sections 2.5.1 and 2.5.4.

Daily values for the PRISM variables are accumulated in a database, and summary statistics for

these variables are calculated for each day of each year (Table 1; Figure 2). A 31-day moving window, centered on the target day, within a five-year moving window, centered on the target year (N=155), is used to calculate localized "long-term" means and standard deviations of *O* ($\overline{O}$, s_o), *P* ($\overline{P}$, s_p), *R* ($\overline{R}$, s_r), *S* ($\overline{S}$, s_s), and *V* ($\overline{V}$, s_v). For example, summary statistics for July 15, 1995 are accumulated from all non-missing days within the period July 1-30, 1993-1997. The 30-day and 5-year windows were thought to represent a good compromise between including enough days to produce a stable mean and standard deviation, but not so many as to dilute seasonal and inter-annual trends in spatial climate patterns and nearby station availability. Multiple years were required to allow the identification of periods of bad SNOTEL observations that sometimes persisted unnoticed for many months or longer in these remote, automated systems.

Once the summary statistics are calculated for each day of the year, each daily observation, prediction, residual, and standard deviation is compared to its "long-term" mean and standard deviation with a t-test, and a p-value is calculated. The p-value estimates the (two-tailed) proportion of observations that can be expected to fall at least as far away from the mean as the daily value (Figure 3). The daily p-values for observation, prediction, residual, and standard deviation are multiplied by 100 to express them as percentages, and are denoted *OP, PP, RP,* and *SP,* respectively. In addition, an overall confidence probability for the observation, *CP*, is calculated from these probability statistics (discussed below).

Of particular importance is *RP*, the residual probability, because it has the most relevance to the consistency, and hence validity, of the observation. *RP* is a measure of the relative success of the model prediction in approximating the observation. A low residual probability indicates that PRISM is having an unusually difficult time predicting for a station on a particular day. *RP* implicitly accounts for the overall ability of PRISM to predict for a daily station observation; if the residual for that time of year is highly variable, with many large values, the standard deviation of the distribution of R will be large, and *RP* will be accordingly larger for a given deviation of *R* from $\overline{R}$ (see Figure 3). The overall confidence probability, *CP*, is currently set to the value of *RP*.

The QC system also uses a similar probabilistic approach to assessing whether potential flatliners are caused by data errors (see Section 2.5.1).

2.4. Dissemination of Results

Integral to the SNOTEL PSQC system is a Web interface that provides developers and users alike with the capability to plot, list, and generally explore the observations and QC results. Station locations and metadata can be found through an Internet mapserver, and plots and tables are constructed and viewed with PHP-based applications. The interface has both basic and advanced views. The advanced view gives users access to dozens of variables produced by the QC system, with time series, scatterplot, and histogram plotting capabilities. The basic view limits and simplifies plotting and display choices to those expected to be used by more casual users. An example plot from the Web interface is shown in Figure 4.

The download section of the interface is devoted to the access and delivery of QC'ed data to users. Users can choose to download data from one or more stations within a given period of record. Blending of observed and predicted values into the downloaded data set is possible by specifying upper and lower thresholds for acceptable *CP* values; an upper *CP* threshold defines the *CP* value above which observations are accepted as-is, and a lower *CP* threshold defines the *CP* value below which predictions should be used instead of the observations. Between the upper and lower thresholds, the observations are blended with the predictions with a linear weighting scheme that weights the observations more highly as the *CP* value increases from the lower to the upper threshold. The upper and lower default settings of these *CP* thresholds are currently set to 30 and 10, respectively based on internal evaluation of the QC system, but may be changed to alternate values by the user.

Data are downloaded in a final results table for each station that has a header indicating the station ID, table creation date, period covered, etc. In the data portion of the table, each daily record includes the following:

- STNID: Station ID
- DT: Date
- O: Observation value
- PREDICTED: PRISM predicted value
- CP: Observation confidence (*CP*). 100= Highest confidence, 0=lowest confidence (or missing).
- FINAL: Final QC'ed value (with blending if specified)
- CPMAX: *CP* value below which the FINAL value is a blend between the observation and the predicted value. At CPMAX, the observation receives full weight.
- CPMIN: *CP* value at which the predicted value receives full weight.
- RSD: Standard deviation of the residual distribution
- MAE: Mean absolute prediction error
- SIGMA: Standard deviation used to determine *CP*
- COUNT: Number of obs used in determining *CP*
- FLAG: Pre-processing flags

This simple record provides the user with vital information about the observation, as well as the QC system used to assess its validity. If needed, additional information about an observation can be obtained from the Web interface.

2.5. Improvements to the System

The SNOTEL QC system uses the PRISM climate mapping system, a knowledge-based system that has been continually improved and updated since its

conception in 1991. Major challenges faced in improving PRISM or any other spatial model include: (1) identifying what data are not available to PRISM by asking the question: "what additional information do I have to be able to say that PRISM has made an error?"; (2) finding viable ways to make this information available to PRISM in a reliable and usable format; and (3) developing algorithms that transform this information into decisions and calculations that help create better spatial predictions in a variety of situations.

Although it is relatively new, the SNOTEL PSQC system itself has already undergone a number of changes that make the system better simulate how a human being would QC an observation. In the end, we are the best judge of performance, and a system that mimics our thought process is most likely to succeed.

Some examples of improvement through feedback follow. There will undoubtedly be many more changes made to the QC system as experience is gained through continued use and development.

2.5.1. Flatliners

In the US, it is difficult to find a realistic example of the same exact maximum or minimum temperature persisting for ten days or more; these are nearly always caused by erroneous data. Flatliners persisting for no more than 5 days occur fairly often. However, flatliners that persist for less than 10 days but more than 5 days are more difficult to assess. In the PSQC system, the assessment of these "potential flatliners" was not handled well by solely evaluating the difference between the prediction and observation, because sometimes the value of the flatliner was a reasonable one. The human expert usually assesses a potential flatliner by comparing the station's temperature variability to that of the surrounding stations; if all are relatively constant, the potential flatliner is considered real. If not, the data are considered suspicious. This information was provided to the QC system in the form of the variability ratio, $V = log_{10}(T_o / T_s)$, where T_o is the 5-day running standard deviation of the station's daily values and T_s is the average 5-day running standard deviation of the surrounding stations (see Table 1). Statistics of this ratio over the summary period (±15 days and ±2 years) were accumulated, and statistical distributions developed, as was done for the residuals. Then, a p-value for the variability ratio for each potential flatliner period was determined to form a *VP* (variability probability) value. For potential flatliners with 5-9 persistent observations, the final *CP* value was then taken as the minimum of *RP* and *VP.*

2.5.2 Accounting for bias in $\overline{R}$

As discussed above, in non-flatline situations, *RP* is used to approximate *CP*. In the calculation of *RP*, $\overline{R}$ and s_r are the operative mean and standard deviation. $\overline{R}$ may show a tendency for bias over the "long-term" period. If the mean is biased 1 or 2 degrees from zero, a daily *R* of zero (perfect prediction) would be 1 or 2 degrees from the mean, and receive a relatively low *RP* value, which seems counterintuitive; perhaps a nearby station which was causing the long-term prediction bias is missing that day. Therefore, the difference between *R* and $\overline{R}$ is now calculated as the minimum of the difference between *R* and $\overline{R}$ and *R* and zero.

2.5.3. A more liberal substitute for s_r

The *RP* value for a daily observation is largely dependent on s_r, which characterizes the variability in the distribution of *R*. If s_r is very small, low *RP* values can result for relatively small differences between *R* and $\overline{R}$. s_r tends to be small for a number of reasons, including the fact that a "best" prediction, which tries to match the observation, is used in the calculation of s_r (see Section 2.3 for details). A more robust calculation of distribution variability was implemented, which calculates a new standard deviation as the maximum of s_r, *S*, $\overline{S}$ and 2°C. *S* and $\overline{S}$ represent the daily and average standard deviation of the PRISM regression function, and can be thought of as the "prediction precision." A 2°C minimum represents the practical notion that distributions with standard deviations less than about 2°C are "splitting hairs," and would be considered too narrow when evaluated by an expert QC operator.

This is an excellent example of the QC system learning and improving, based on subjective interpretation of the results. In an ideal situation, the 2°C threshold, or something similar, should be calculated by the system. It may be possible to do so by calculating the distribution of differences between observations from stations with known data quality in close proximity, but there are many factors to consider, including station siting and configuration, sensor type, and difference in the timing of meteorological events. In the end, even the most "objective" QC system must be subjectively parameterized to produce what is perceived as optimal performance.

2.5.4. Inconsistent observation times

A difficulty in performing spatial QC on a daily time step is dealing with differences in observation time. While this is not typically an issue for automated observing systems (00-24 being standard), spatial QC of these systems requires that all surrounding stations be QC'ed as well. National Weather Service COOP stations have observation times which vary from station to station, and when processing inconsistencies are considered, may produce time shifts of +/- 2 days. The problem is most serious when there are large daily temperature variations. The QC system, considering each day in isolation, was assigning relatively low *CP*

values to observations in situations where temperature changes were out of phase due to time shifting, despite the fact that the observations were otherwise correct. This was remedied by using T_s, the average 5-day running standard deviation of the surrounding stations (discussed in Section 2.5.1), as a measure of the day-to-day variability in temperature. If the variability was high, it would be necessary to widen the distribution of *R* to accommodate the possibility that time shifting was occurring. This became yet another term in the calculation of the standard deviation used to determine the p-value, and hence, *CP*. Now, the new standard deviation was calculated as the maximum of s_r, *S*, $\overline{S}$, 2°C, and T_s.

3. SUMMARY AND QUESTIONS TO CONSIDER

The methods by which climate data are being collected and used are expanding rapidly. Increasingly, data are being collected by automated electronic systems, with a wider variety of platforms, at higher temporal resolution, at more locations, and in more difficult and remote environments. This shift in observational strategy is largely the result of technological advances that have increased both the supply of, as well as demand for, climate data. Some of these same technological advances now provide opportunities for qualitatively different approaches to quality control, methods that are sophisticated, largely automated, data-rich, updatable, and capable of furnishing quantitative error and confidence information. This paper explores some of the characteristics of such methods, and discusses why they are both beneficial and practical. These are summarized below:

- Data-rich QC frameworks. Automated data collection systems are often used to monitor difficult, often mountainous environments far from permanent human habitation. Intelligently assessing the quality of the data collected in such environments is difficult, and requires a sophisticated, data-rich QC system that can account for factors controlling the spatial and temporal patterns of climate in the area. Data that could be used to improve a QC system are numerous and varied. Four types were discussed: (1) *in situ* observations; (2) satellite and radar data; (3) diagnostic climatological grids interpolated from observations; and (4) values and QC results from other observed variables.

- Continuous, quantitative estimates of observation validity and estimation uncertainty. There is much benefit in estimates of observational validity and estimation uncertainty that are quantitative and continuous, rather than categorical. For example, errors from electronic measurement systems can suffer from continuous drift (i.e., calibration issues) as well categorical mistakes (e.g., using an incorrect algorithm to convert voltage to temperature). In addition, computer models that rely on climate observations as input benefit from quantitative estimates of uncertainty. Providing such quantitative uncertainties requires that the statistical performance of a QC system be known, and incorporated into the results.

- Transparent dissemination of results that allow end-user interaction, education, and decision-making. The range of applications for climate data is rapidly increasing, and each application has a different tolerance for low-confidence data points. In addition, users need to be informed about how the QC system arrived at its conclusion, so they can make better choices about how the data should be used. A new approach is required, where full quantitative information is presented to the user in a transparent, interactive format. Full disclosure allows users to make intelligent decisions about the QC results and an interactive system allows users to generate their own data sets, based on suggested default thresholds for data inclusion, or their own thresholds.

- Automated QC systems that improve through experience and feedback. Data generated by automated electronic systems are often more voluminous (e.g., shorter time step) and disseminated in a much more timely manner than those from manual systems. Whether operating on near-real time data, historical data, or both, automated QC systems are needed. This does not mean that humans should be excluded from the QC process. Experts are needed to parameterize the system (often done subjectively), provide feedback, and make updates and improvements to the automated system. In order for QC systems to be optimally effective, climate data QC must be designed as an ongoing process of improvement, not a one-time assessment. In this way, the expectation will be that QC systems can and will be improved, and will be re-applied retrospectively on a periodic basis to reflect these improvements.

The first generation of a spatial QC system recently developed for USDA-NRCS SNOTEL temperature data, called the SNOTEL Probabilistic-Spatial Quality Control (PSQC) System, was presented as an example of a system that is in the early stages of incorporating these characteristics. The SNOTEL PSQC System is spatially-oriented, uses a knowledge-based system (PRISM) to make predictions, and ingests a variety of spatial data sets. It operates on the premise that spatial consistency, if assessed accurately, is a useful indicator of data validity. Continuous, quantitative estimates of observation validity and prediction uncertainty are made by evaluating the statistical performance of the system. Results are provided via a Web interface that allows users to make intelligent decisions about the QC results and generate their own data sets, if desired. The

system is automated, and is being upgraded through feedback from users and developers.

The development of this QC system has raised a number of issues. Examples of some of the issues and questions we are currently considering include:

- By using a probabilistic approach, the SNOTEL PSQC System accounts for PRISM's ability to predict in a station's absence. But unusual situations occur in which the observation appears to be valid, but also spatially inconsistent. How can information about these unusual situations be incorporated into the QC system's knowledge base? More broadly, is there a limit to how far one can take the assumption that spatial inconsistency equates with validity?

- Spatial QC depends on "long-term" information on the ability of PRISM to predict in a station's absence. This ability can be affected by the presence or absence of nearby station observations. How do we account for intermittencies in station reporting, especially if we are to operate the system in near real-time, where observations are often missing?

- Non-spatial validity tests have also been incorporated into the SNOTEL PSQC system. For example, the probability of a station "flat-lining" (having the same observation repeated) for a specified period of days is now calculated and subjected to the same p-value calculation. Are there other non-spatial checks that could be made and assessed probabilistically? Perhaps the tendency for errors in electronic measuring systems to occur in persistent temporal blocks could be utilized.

- Continuous and probabilistic QC systems are beneficial for assessing the quality of data from electronic observing systems. Are they also useful and beneficial for manual observing systems?

4. REFERENCES

Daly, C., Gibson, W.P., M. Doggett, J. Smith, and G. Taylor. 2004. A Probabilistic-Spatial Approach To The Quality Control Of Climate Observations. Proc., *14th AMS Conf. on Applied Climatology, 84th AMS Annual Meeting Combined Preprints,* Amer. Meteorological Soc., Seattle, WA, January 13-16, 2004, Paper 7.3. http://ams.confex.com/ams/pdfpapers/71411.pdf

Daly, C., E.H. Helmer, and M. Quinones. 2003. Mapping the climate of Puerto Rico, Vieques, and Culebra. *International Journal of Climatology*, 23: 1359-1381.

Daly, C., W. P. Gibson, G.H. Taylor, G. L. Johnson, P. Pasteris. 2002. A knowledge-based approach to the statistical mapping of climate. *Climate Research*, **22**: 99-113.

Daly, C. and G.L. Johnson. 1999. PRISM spatial climate layers: their development and use. *Short Course on Topics in Applied Climatology*, 79th Annual Meeting of the American Meteorological Society, 10-15 January, Dallas, TX. 49 pp. http://www.ocs.orst.edu/prism/prisguid.pdf.

Daly, C., R.P. Neilson, and D.L. Phillips. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology,* 33: 140-158.

Doggett, M., C. Daly, J. Smith, W. Gibson, G. Taylor, G. Johnson, and P. Pasteris. High-resolution 1971-2000 mean monthly temperature maps for the western United States. Proc., *14th AMS Conf. on Applied Climatology, 84th AMS Annual Meeting Combined Preprints,* Amer. Meteorological Soc., Seattle, WA, January 13-16, 2004, Paper 4.3. http://ams.confex.com/ams/pdfpapers/71450.pdf

Gibson, W.P., C. Daly, M. Doggett, J. Smith, and G. Taylor. 2004. Application of a probabilistic spatial quality control system to daily temperature observations in Oregon. Proc., *14th AMS Conf. on Applied Climatology, 84th AMS Annual Meeting Combined Preprints,* Amer. Meteorological Soc., Seattle, WA, January 13-16, 2004, Paper 4.4. http://ams.confex.com/ams/pdfpapers/71434.pdf

Urzen, M., S. Anzari, and S. Del Greco. 2004. Automated spatial precipitation estimator (PrecipVal). *Proc., 14th AMS Conf. on Applied Climatology,* Amer. Meteorological Soc., Seattle, WA, January 13-16. http://ams.confex.com/ams/pdfpapers/70681.pdf

USDA-NRCS. 1998. *PRISM Climate Mapping Project--Precipitation. Mean monthly and annual precipitation digital files for the continental U.S.* USDA-NRCS National Cartography and Geospatial Center, Ft. Worth TX. December, CD-ROM.

Table 1. Variables calculated by the SNOTEL PSQC system.

Abbreviation	Description	Notes
PRISM Variables		
O	Observation	Observed station value on a given day
P	Prediction	"Best" PRISM spatial prediction for a station, on a given day, that most closely matches the observation after systematic deletion of surrounding stations, individually and in pairs
R	Residual (P-O)	Difference between the prediction and the observation for a station on a given day
S	Regression standard deviation	Standard deviation of the PRISM regression function for a station on a given day
T_o	Temporal variability of station	5-day running standard deviation of O; represents the local day-to-day variability of O
T_s	Temporal variability of surrounding stations	Average 5-day running standard deviation of O for the 10 surrounding stations, weighted by PRISM calculated weights; represents the local day-to-day variability of observations from surrounding stations
V	Variability ratio	Ratio of the station's T_o to that of the surrounding stations' T_s, calculated as $log_{10}(T_o / T_s)$
Summary Statistics		
$\overline{O}, s_o$	"Long-term" mean and standard deviation of observation	Mean and standard deviation of the observation for a given day of the year, calculated as the mean of observations for a station centered on the current day, ±15 days and ± 2 years
$\overline{P}, s_p$	"Long-term" mean and standard deviation of prediction	Same as above, except for prediction
$\overline{R}, s_r$	"Long-term" mean and standard deviation of residual	Same as above, except for residual
$\overline{S}, s_s$	"Long-term" mean and standard deviation of regression standard deviation	Same as above, except for regression standard deviation
$\overline{V}, s_v$	"Long-term" mean and standard deviation of variability ratio	Same as above, except for variability ratio
Probability Statistics		
OP	Observation probability	P-value*100 from a t-test comparing *O* to the distribution of O, parameterized by $\overline{O}, s_o$. Represents the percent of observations within $O-\overline{O}$ of the mean. Measure of how unusual the observation is compared to others at this station at the same time of year
PP	Prediction probability	Same as above, except for prediction
RP	Residual probability	Same as above, except for residual
SP	Standard deviation probability	Same as above, except for regression standard deviation
VP	Variability Probability	Same as above, except for variability ratio
CP	Overall confidence probability	Overall confidence probability for the station observation on a given day. Currently, *CP=RP*

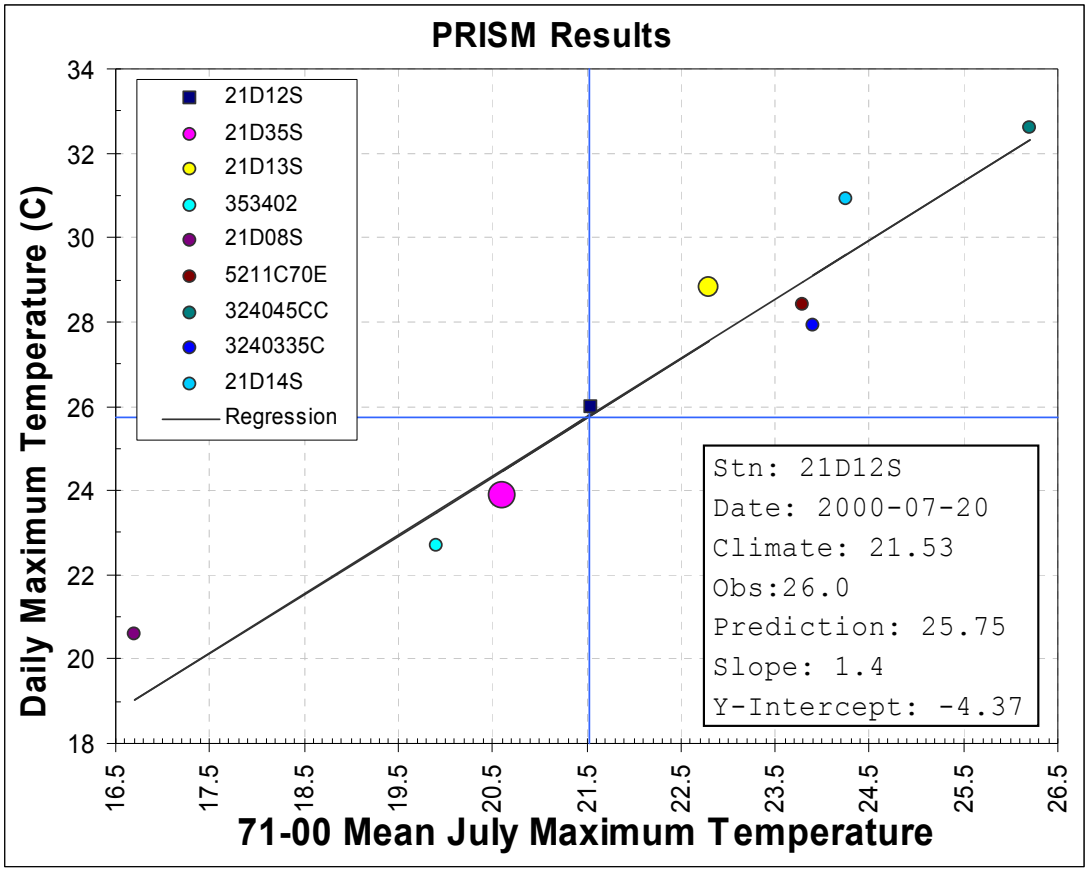


Figure 1. Scatterplot and PRISM regression line for 20 July 2000 maximum temperature and 1971-2000 mean July maximum temperature for the location of the Mt. Hood Test Site SNOTEL station (21D12S) in the Cascades Mountains south of Mt. Hood, Oregon. Cross network --- Size of circles represents a station's relative weight in the regression function. Note that data from three difference networks are employed in the regression function: SNOTEL (6-character alpha-numeric ID), COOP (6-digit ID), and RAWS (8-character alpha-numeric ID).

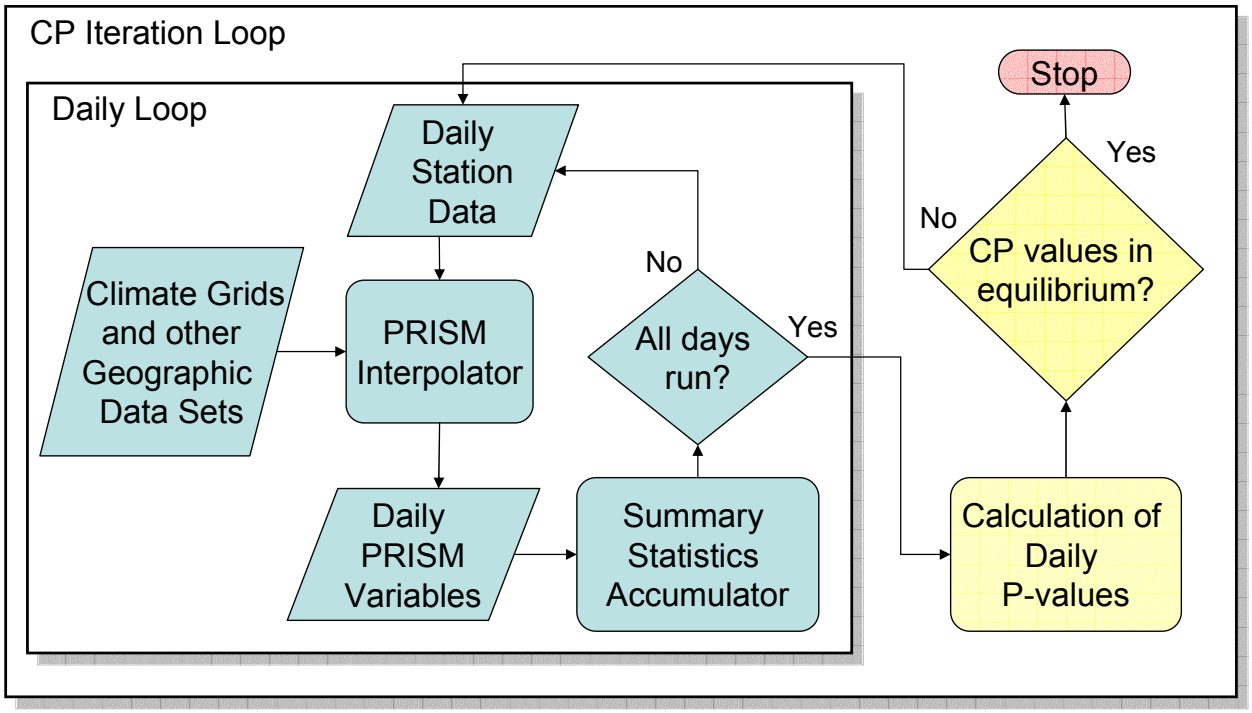
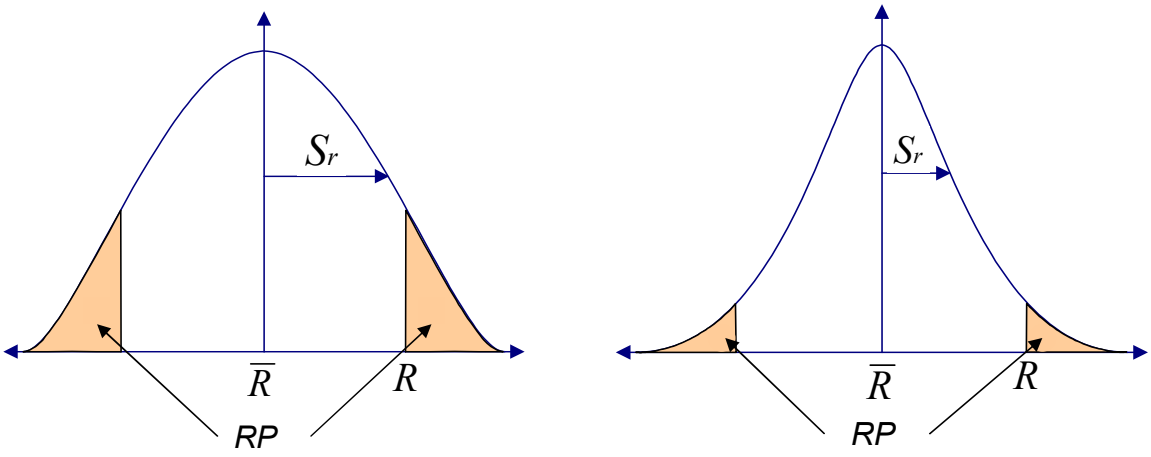


Figure 2. SNOTEL PSQC process flow. Within the inner daily loop, PRISM is run for each station-day in the absence of the observation (*O*), producing a prediction (*P*), residual (*R*=*P*-*O*), standard deviation of the PRISM regression function (*S*), and temporal variability statistics (*T* and *V*). Summary statistics, including mean and standard deviation, of *P, O, R*, *S,* and *V* for a 31-day, 5-year moving window around each day are accumulated. (The calculation of summary statistics is weighted by the confidence probabilities (*CP*) of the station observations assigned in the previous iteration.) Once all desired days are run, the process moves to the outer iterative loop. Here, each station-day's *P, O, R*, *S,* and *V* are compared to their statistical distributions for that day, and p-values for each calculated using a t-test, giving *PP, OP, RP, SP*, and *VP*, respectively. *CP* is set to the value of *RP*, except in cases of potential flatliners, where it is set to the minimum of *RP* and *VP* (see Section 2.5.1 for details on flatliners). If the current iteration's *CP* values are similar to those from the previous iteration, the process stops. If not, the *CP* value of each station observation is updated in the station listing, and PRISM weights the station observation accordingly in the next daily interpolation loop. In this way, observations with lower confidence (*CP*) have less influence on the subsequent PRISM predictions and summary statistics. The outer *CP* iteration loop is typically run 1-5 times before all station-days reach equilibrium.



(a) Relatively poor predictive performance (b) Relatively good predictive performance

Figure 3. Two-tailed p-values (shaded areas) for a daily residual (*R*) its mean ($\overline{R}$), and standard deviation (S_r) for: (a) a distribution with a large S_r , representing a wide distribution of differences between the PRISM prediction and the observation, indicating relatively poor predictive performance; and (b) a distribution with a small S_r , representing a narrow distribution of differences between the PRISM prediction and the observation, indicating relatively good predictive performance. Note that when the predictive performance is poor, the resulting *RP* (two-tailed p-value for *R*) is greater for the same daily deviation from $\overline{R}$ than when the predictive performance is good.

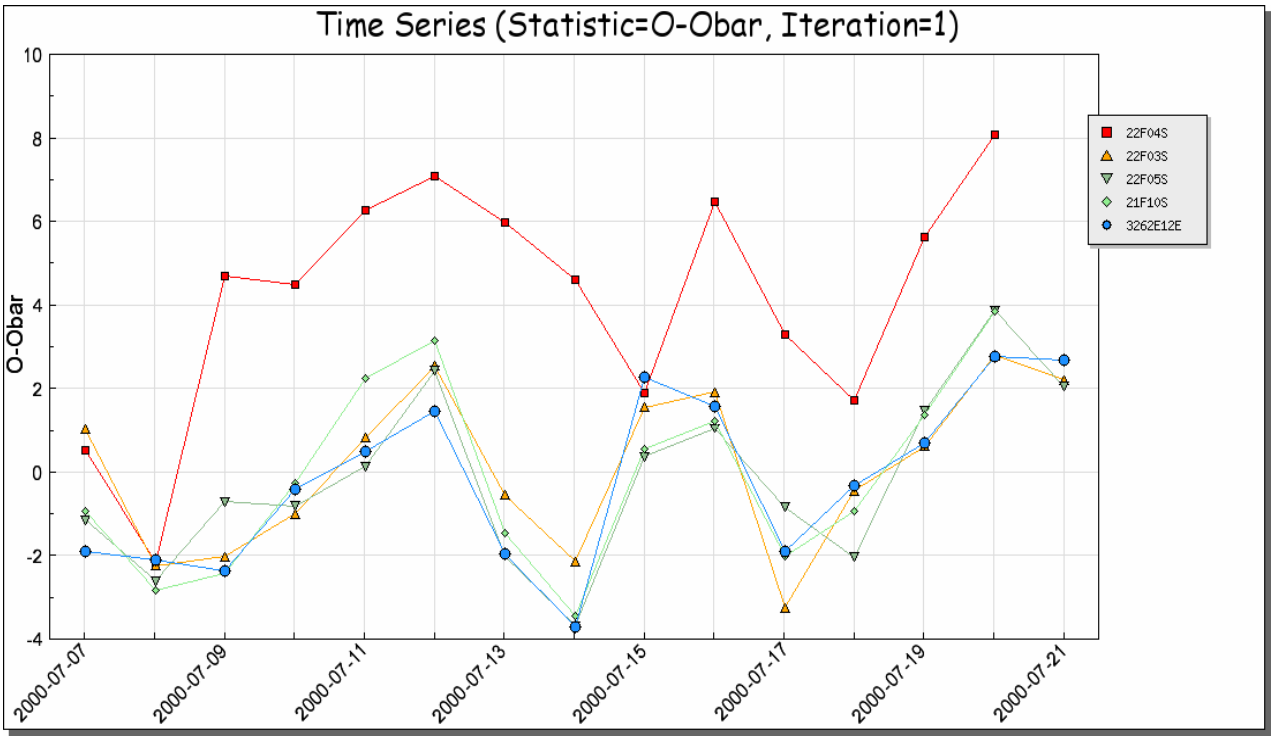


Figure 4. Example plot from the SNOTEL PSQC Web interface, showing erroneous maximum temperature observations at the Salt Creek Falls, Oregon SNOTEL site (22F04) over a two-week period in July 2000 (red line). Maximum temperatures are plotted as deviations from $\overline{O}$, the ±15 day, ±2 year average. Errors at this station persisted all summer before being remedied in the fall.