7.3 A PROBABILISTIC-SPATIAL APPROACH TO THE QUALITY CONTROL OF CLIMATE OBSERVATIONS

Christopher Daly*, Wayne Gibson, Matthew Doggett, Joseph Smith, and George Taylor
Oregon State University

1. INTRODUCTION

Surface weather and climate observations are the backbone of a multitude of analyses, studies and assessments. Errors in the observations do occur; if left unidentified, they can have severe and adverse affects on analyses and the decisions that result from them. Rigorous quality control of climate data is probably the single most important thing an analyst can do to ensure a successful outcome. Traditionally, quality control (QC) systems have been designed to emphasize the identification of human errors in COOP-like (Cooperative Observer) observational records that occur while reading gauges and thermometers, recording the data on paper, and keypunching these data into electronic format. Traditional QC methods have employed a series of 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. Examples of such quality checks include range checks and internal and external consistency checks, some of which are operated manually. In general, this type of QC system can be termed categorical and deterministic; that is, they employ categorical validity checks, from which a determination of validity results.

An ever-increasing number of climate observations are being made electronically at automated weather stations. The development of the ASOS (Automated Surface Observing System), SNOTEL (Snowpack Telemetry), RAWS (Remote Automated Weather Station), and Agrimet networks, and the prospect of a COOP modernization program, all reflect the increased importance of electronic sensors and automated data delivery systems now and in the future. These new measurement technologies are presenting a number of challenges for traditional QC systems that include the following:

- Errors from electronic measurement systems are more often manifested as continuous drift, rather than categorical mistakes. Therefore, continuous estimates, rather than categorical tests, of observation validity are most meaningful.

- Increased usage of computer models (e.g., hydrologic models) that use climate observations as input has increased the need for quantitative estimates of observational uncertainty.

- The range of applications for climate data is increasingly rapidly, and each application has a difference tolerance for outlier data points. Probabilistic information from which a decision of validity can be made is more useful than a predetermined, "one size fits all" declaration of validity.

- Data generated by automated electronic systems are often more voluminous (e.g., shorter time step) and disseminated in a more timely manner than those from manual systems, favoring automated QC methods over those involving manual inspection.

A new generation of QC systems is being formulated in response to these changing needs and priorities. This paper describes the development of such a new-generation QC system for USDA-NRCS SNOTEL data that uses climate mapping technology and climate statistics to provide a continuous, quantitative confidence probability for each observation, estimate a replacement value, and provide a confidence interval for that replacement. An overview of system structure and operation is given, and the paper concludes with a series of questions that require further research.

2. BACKGROUND

In the mid and late 1990s, Oregon State University's Spatial Climate Analysis Service (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, data that are increasingly important for water supply assessment, climate analysis, power generation planning, and other uses in the West.

SNOTEL data are recorded electronically and transmitted via meteor burst technology 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 NRCS and at the Western Regional Climate Center has attempted to accomplish this, but has never been fully completed. Temperature, in particular, has posed problems for data quality assessment.

Over the past decade, SCAS has been developing tools for conducting "spatial quality control" as part of its ongoing climate mapping work in the United States and abroad (Daly et al., 2000). The tools are based on

* *Corresponding author address*: Christopher Daly, Spatial Climate Analysis Service, Oregon State University, 316 Strand Ag Hall, Corvallis, OR 97331; email: daly@coas.oregonstate.edu

PRISM (Parameter-elevation Regressions on Independent Slopes Model), developed at OSU (Daly et al., 1994, 2002, 2003). The spatial QC tools operate in a largely automated mode, with a graphical user interface to help aid QC decisions in difficult situations. These tools have proven to be effective in identifying invalid, incorrect, or missing data. SCAS spatial QC tools operate on the premise that spatial interpolation can be used to identify bad data values if there is sufficient skill in the interpolation process. 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, then the station value may be in error. PRISM provides a high degree of skill to the spatial interpolation process, making the identification of many erroneous data values possible.

In 2001 the USDA Natural Resource Conservation Service (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.

3. OVERVIEW OF THE PRISM PROBABILISTIC-SPATIAL QUALITY CONTROL (PSQC) SYSTEM

A process schematic of the new QC system, termed the PRISM Probabilistic-Spatial Quality Control (PSQC) System, is shown in Figure 1. The system consists of two main components: (1) climatological grid development, and (2) the QC iteration process. These are discussed below as they apply to the QC of SNOTEL daily temperature observations.

3.1. Climatological Grid Development

The PRISM PSQC system requires running PRISM to produce a high-quality spatial estimate of temperature at each SNOTEL station location each day. Experience has shown that the highest interpolation skill for daily temperature is obtained by running PRISM using a 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 maximum predictive skill for a given day.

The highest-quality existing climatological temperature grids for maximum and minimum temperature were produced in 1998 by SCAS. These grids are the best currently available for the western US, and would be suitable as PSQC predictive grids, except for two main deficiencies: (1) They represent the 1961-1990 climatological averaging period, and thus do not best reflect the time frame of the SNOTEL periods of record; and (2) were developed at a spatial resolution of 2.5 arc-minutes (~4 km), which does not capture all of

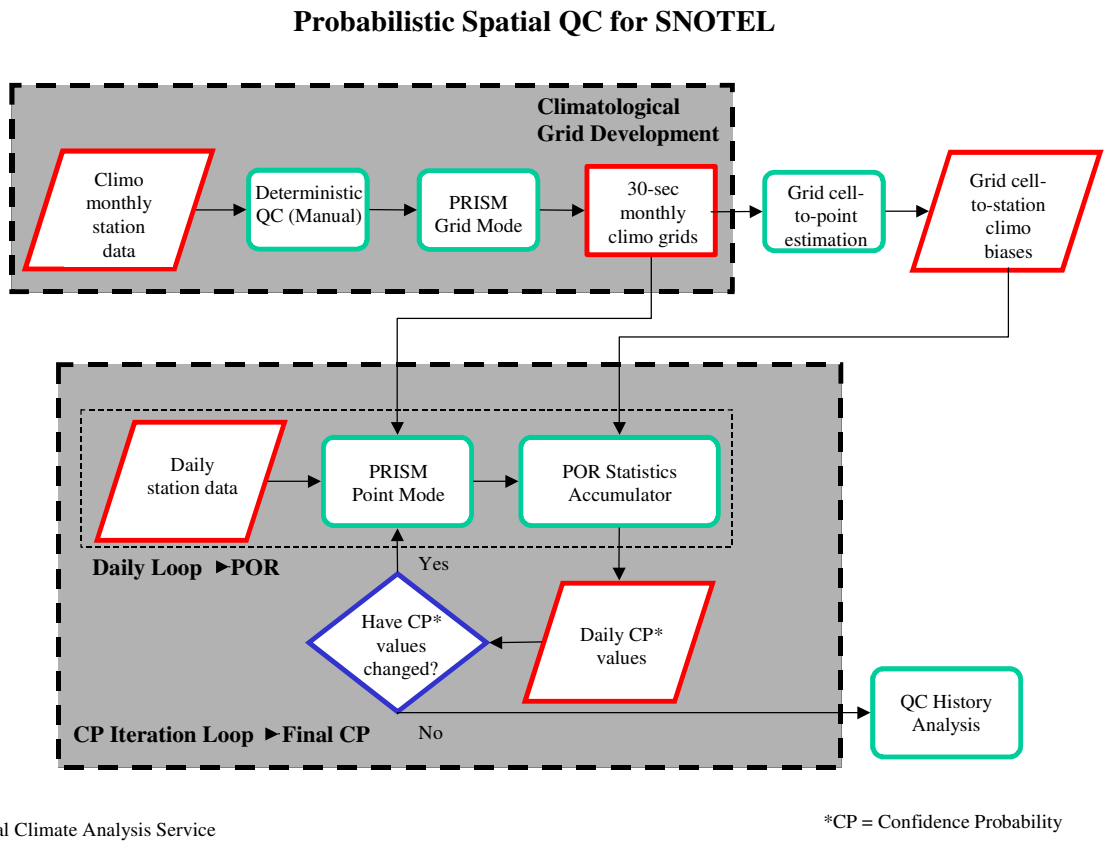

Figure 1. Process schematic of the PRISM Probabilistic Spatial Quality Control (PSQC) System.

the detailed temperature patterns known to exist in mountainous terrain. For example, the height of Mt. Hood on the 2.5-minute resolution digital terrain grid is only 2662 m, 765 m lower than the actual height of 3427 m. This height discrepancy translates into an approximately 5°C error in temperature, assuming an average lapse rate. Such large errors would be incorporated as noise into the SQC system.

Under USDA-NRCS funding, work is underway to produce new 1971-2000 monthly average minimum and maximum temperature grids at 30-sec (0.8-km) resolution for the western United States (see Doggett et al. (2004) for details). 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 to test the PRISM PSQC system. Figure 2 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. The tight regression fit suggests that the spatial patterns of the gridded July climatological mean maximum temperatures are very good at approximating those of 10 July 2000.

Even at 0.8-km resolution, there is not a perfect match between the station point climatological values and the climatological values of the grid cells containing the stations. This misalignment of point and grid cell values can produce noise in the regression functions. These discrepancies, termed "grid biases," are accounted for by calculating the point-grid cell differences for the 1971-2000 climatology at each station, and adjusting the background climatological values by these differences when calculating each PRISM PSQC regression function.

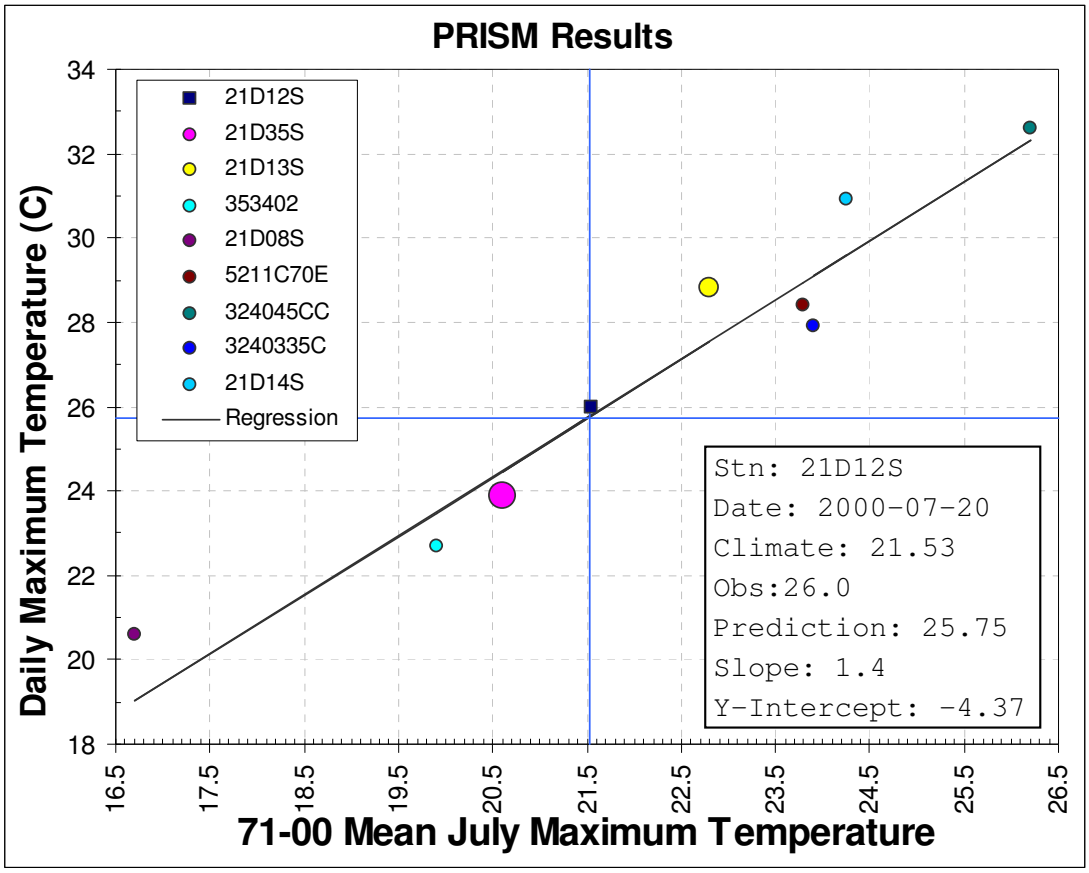

Figure 2. Scatterplot and PRISM regression line for 20 July 2000 maximum temperature and 1971-2000 mean July maximum temperature for the location of SNOTEL station 21D12S in the Cascades Mountains south of Mt. Hood, Oregon. Size of circles represents a station'

4. THE QC PROCESS

The goal of the QC process is to, through a series of iterations, gradually and systematically "weed out" spatially inconsistent observations from consistent ones. The QC process consists of two nested loops (Figure 1). In the daily loop, PRISM is run for each station location for each day within the period of record, and summary statistics are accumulated. Once all days have been run, confidence probabilities (*CP*) for each daily station observation are estimated (discussed below). These *CP* values are used to weight the daily observations in a second series of PRISM daily runs. Observations that have lower CP values are given lower weight, and thus have less influence, in the second set of PRISM predictions, 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.

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 a "best" 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. The process is then repeated several times while deleting nearby observations, first singly, then in pairs. The prediction that most closely matches the observation, *O*, is accepted. The cycles of deletion are performed to preclude bad observations from contaminating the predictions. It is assumed that the chances of more than two bad observations occurring in the immediate vicinity of each other on a given day are small. The residual, *R, (R = P - O)* and the regression standard deviation, *S*, are also calculated by PRISM. *S* is of notable importance, in that it quantifies the standard error for the daily temperature prediction at the target station, and represents the uncertainty of a value that could be used as a replacement for the observation.

Daily values for *O, P, R,* and *S* are accumulated in a database, and summary statistics for these variables are calculated for each day of the year. A 30-day moving window, centered on the target day, within a five-year moving window, centered on the target year (N=150), are 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), and *S* ($\overline{S}$, s_s). 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.

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

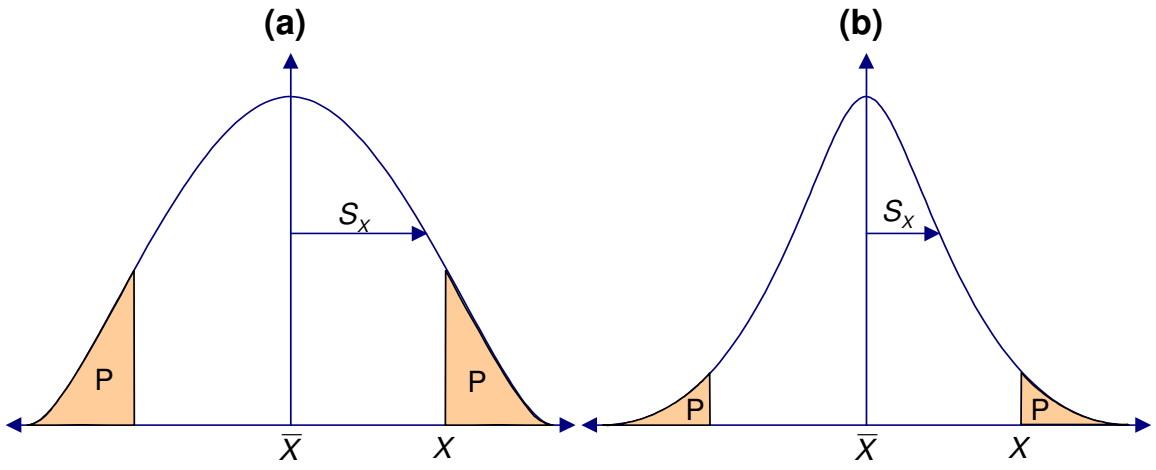

Figure 3. Two-tailed p-values (orange shaded areas) for a generic daily value (*X*) its mean ($\overline{X}$), and standard deviation (s_x) for: (a) a distribution with a large s_x , and (b) a distribution with a small s_x . Note that the p-value is much higher for a given deviation from $\overline{X}$ when s_x is large.

observation, *CP*, is calculated from these probability statistics (discussed later).

OP is a measure of the unusualness of an observation compared to its normal distribution for that time of year. Invalid observations are often associated with low *OP* values, but not always. A valid observation that accurately tracks a cold wave would have a low *OP* value because it is an unusual reading, but an invalid observation that records a near average temperature during that same cold wave would have a high *OP* value, because of its lack of unusualness.

PP is a measure of the unusualness of a prediction compared to its normal distribution for that time of year. As in the case of *OP*, *PP* is often associated with poor predictions, but not always. A good prediction that accurately tracks a cold wave would have a low *PP* value because its an unusual prediction, but a poor prediction of a near average temperature during that same cold wave would have a high *PP* value, because of its lack of unusualness.

SP is a measure of the unusualness of the standard deviation of the PRISM regression function compared to its normal distribution for that time of year. A low *SP* value indicates that the regression function is unusually noisy, and may indicate that an invalid observation is in the data set. However, it could also mean that the relationship between the daily temperatures and their 1971-2000 climatologies is unusually noisy.

RP has particular importance to the QC process because is 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, s_r will be large, and *RP* will be accordingly larger for a given deviation of *R* from $\overline{R}$ (see Figure 3a). As such, a small *RP* indicates that the observation is unusually inconsistent with its neighbors (which are used as predictors), and this lowers confidence. Because *RP* represents the single most useful estimate of confidence in the observation, *CP*, the overall confidence probability, is currently set to the value of *RP*.

Initial tests of the QC system using *RP* as the estimate of *CP* indicated that the system was too strict in its estimate of station confidence. In many cases, the *CP* values were lower than we thought they should be. As a result, several measures were implemented to "give the observation every chance of success." They are as follows:

- **Precision**: The precision of the daily station observations varies from 0.1°C for SNOTEL, Agrimet, and RAWS, to 1°F for COOP, to 1°F – 1°C for ASOS, depending on the station. Therefore, for probability calculations, 1°C (the most imprecise possibility) is now subtracted from differences between the daily value and its "long-term" mean before a t-test is conducted.

- **Accounting for bias in $\overline{R}$:** As discussed above, *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.

- **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. 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 1°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 1°C minimum represents the practical notion that distributions with standard deviations less than the precision of the observations make little sense.

5. SUMMARY AND QUESTIONS TO CONSIDER

An ever-increasing number of climate observations are being made electronically at automated weather stations. These new measurement technologies are presenting a number of challenges for categorical and deterministic QC systems designed for manual observing systems, such as COOP. Errors from electronic measurement systems are more often manifested as continuous drift, rather than categorical mistakes. Therefore, continuous estimates, rather than categorical tests, of observation validity are most meaningful. Increased use of computer models that require climate observations as input has heightened the need for quantitative estimates of observational uncertainty. The range of applications for climate data is increasingly rapidly, and each application has a difference tolerance for outlier data points. Probabilistic information from which a decision of validity can be made is more useful than a predetermined, "one size fits

all" declaration of validity. Data generated by automated electronic systems are often more voluminous (e.g., shorter time step) and disseminated in a more timely manner than those from manual systems, favoring automated QC methods over those involving manual inspection.

A new generation of QC systems is being formulated in response to these changing needs and priorities. This paper described the development of such a new-generation QC system for USDA-NRCS SNOTEL data called the PRISM probabilistic-spatial quality control system (PRISM PSQC System). It uses climate mapping technology and climate statistics to provide a continuous, quantitative confidence probability for each observation, estimate a replacement value, and provide a confidence interval for that replacement. System development followed the following principles:

- Spatial consistency with nearby observations is a useful and powerful proxy for observation validity
- While observations of unknown validity must be used to determine the validity of other observations, the weight of evidence should isolate the truly inconsistent observations
- Estimates of data validity should be continuous and quantitative
- A prediction, and a accompanying error estimate, should be made for each observation, allowing the user to choose which to use, based on the application
- The system must be fully automated for eventual use in near real-time QC

Initial tests of the system in western Oregon have produced informative results. Detailed examples of the performance of the PRISM PSQC System are available from Gibson et al. (2004). As work progresses, many questions will arise concerning the strengths and weaknesses of the system, and how it compares to traditional QC systems. Examples of some of the issues and questions we are currently considering include:

- A probabilistic QC system can, if desired, transfer the decision to "toss or keep" an observation to the user. However, it must be accompanied by guidelines on how the user should make these decisions. As a first effort to develop such guidelines, we plan to compare the PRISM PSQC results to the "toss or keep" decisions made by the Western Regional Climate Center's sophisticated, but deterministic, spatial QC system for SNOTEL temperature.

- If continuous and probabilistic QC systems are needed to address the needs of electronic observing systems, are they also useful for manual observing systems? If not, can they be modified or hybridized to be more useful?

- By using a probabilistic approach, the PRISM PSQC System accounts for PRISM's ability to predict in a station's absence. But highly unusual situations occur in which the observation appears to be valid, but also spatially inconsistent (see Gibson et al., 2004 for an example). How far can the assumption be taken 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?

- Spatial QC mixes observations from different networks with sometimes very different reporting protocols. The most obvious of these is the COOP network's variable definition of an observation day. Differences between a 4 PM COOP observation time and a nearby midnight SNOTEL measurement time will produce occasional but persistent spatial inconsistencies, as will two COOP stations with different observation times. How can these effects best be minimized for daily QC purposes?

- Non-spatial validity tests can also be incorporated into the probabilistic QC system. For example, the probability of a station "flat-lining" (having the same observation repeated) for a specified period of days can be calculated and subjected to the same p-value calculation. Another example is the probability of a station exceeding a given change in temperature from one day to the next. Work to incorporate such non-spatial tests is underway.

- *OP*, *PP*, *RP*, and *SP* are calculated under the assumption that *O*, *P*, *R*, and *S* are normally distributed. This is not always the case. Is there a need to consider using non-parametric statistical tests? What are the ramifications of doing this?

Table 1. Variables calculated by the PRISM 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
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
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-Obar of the mean. Measure of how unusual the observation is compared to others at this station at the same time of year
PP	Percent of predictions within P-Pbar of the mean	Same as above, except for prediction
RP	Percent of residuals within R-Rbar of the mean	Same as above, except for residual
SP	Percent of standard deviations within S-Sbar of the mean	Same as above, except for regression standard deviation
CP	Confidence probability	Overall confidence probability for the station observation on a given day. Currently, CP=RP

6. REFERENCES

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., 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. 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., T.G.F. Kittel, A. McNab, W.P. Gibson, J.A. Royle, D. Nychka, T. Parzybok, N. Rosenbloom, and G. Taylor. 2000. Development of a 103-year high-resolution climate data set for the conterminous United States. In: *Proc., 12th AMS Conf. on Applied Climatology,* Amer. Meteorological Soc., Asheville, NC, May 8-11, 249-252.

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,* Amer. Meteorological Soc., Seattle, WA, January 13-16, this volume.

Gibson, W.P., C. Daly, M. Doggett, J. Smith, and G. Taylor. Application of a probabilistic spatial quality control system to daily temperature observations in Oregon. *Proc., 14th AMS Conf. on Applied Climatology,* Amer. Meteorological Soc., Seattle, WA, January 13-16, this volume.

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.