4.4   APPLICATION OF A PROBABILISTIC SPATIAL QUALITY CONTROL SYSTEM TO DAILY TEMPERATURE OBSERVATIONS IN OREGON

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

1. INTRODUCTION

Whether modeling in the near to real-time or climatic time scales, data is the backbone for our analyses. Traditionally for climate analyses, we have used data sets from observation networks such as COOP (Cooperative Observer), HCN (Historical Climate Network), and more recently SNOTEL (Snowpack Telemetry). For near to real time, the timely nature of the analyses has elevated the need to use hourly observation networks such as ASOS (Automated Surface Observing System), and RAWS (Remote Automated Weather Station). In all analyses, we rely on some form of quality control (QC) being applied to the data, and even then, the level and complexity of QC is insufficient. Traditional QC systems are applied on data sometimes as far back as six months after the observations have been recorded, not very useful for real time analyses. They also employ a series of quality checks that an observation must pass, and the observation is then considered either "valid" or "invalid". Examples of these checks include range checks and internal and external consistency checks, either in an automated or manual mode. This type of QC system can generally be termed categorical and deterministic; that is, they employ categorical validity checks, from which a determination of validity results.

A new generation of QC systems is being formulated, and one of these is referred as the PRISM Probabilistic-Spatial Quality Control (PSQC) System (Daly et al., 2004). The purpose of this new system is to provide a continuous, quantitative level of confidence for each observation, estimate a replacement value, and provide a measure of confidence in that replacement value. The PRISM PSQC system is an iterative system that systematically weeds out inconsistent observations from consistent ones. The system is based the Parameter-elevation Regression on Independent Slopes Model (PRISM). PRISM is a knowledge-based system that accounts for effects such as elevation, inversions, topographic facets, and coastal proximity (Daly et al, 1994, 2002).

This paper illustrates the use of that system by briefly discussing the pertinent statistics generated by the system, and examples to illustrate how the system operates.

2. PSQC SYSTEM

The PSQC system generates a variety of statistics that help in answering questions like: 1) how noisy is the

---
* *Corresponding author address*: Wayne Gibson, Spatial Climate Analysis Service, Oregon State University, 316 Strand Hall, Corvallis, OR 97331 email: gibson@coas.oregonstate.edu

prediction, 2) how unusual is the observation compared to the prediction, and 3) how much confidence can be placed in the prediction.

For prediction noise, the regression standard deviation (S), which describes the scatter around the regression line created by PRISM, is used.

Confidence placed on the prediction and the unusual nature of the observation is treated as one and the same statistic, the residual probability (RP). RP measures the relative success of the model in predicting the observation. A low probability indicates PRISM is not able to predict as well as it normally does for that station and time of year. That could be due to either an unusually poor prediction or a highly unusual observation. With highly varying residuals, which occurs when PRISM typically has a difficult time predicting for a station, the RP value becomes larger for a given deviation of the residual from its average value. This is illustrated in Figure 1. For overall confidence in the PRISM prediction (CP), RP is currently used since it represents the most stable estimate of confidence in the prediction and the observation. A more detailed discussion of the PSQC system is found in these proceedings (Daly et al., 2004).

3. CASE STUDIES

Four case studies follow. They represent a wide spectrum of issues encountered with current observation networks

3.1. Case 1: Drifting Observation, SNOTEL Station 21E07S, McKenzie Pass

As discussed in Daly et al. (2004), continuous drift exists in electronic measurement systems. Figure 2 illustrates the observations preceding and following such a period.

On February 7, 1996 the T_{max} observation increased from 6 to 35ºC, eventually reaching a value of 84ºC before returning to a more reasonable value on March 8. Figure 3 illustrates the Confidence Probability (CP) values created by the PSQC system. A CP of 0% indicates that the observations are truly outliers and lie on the tails of the CP distribution.

Figure 4 shows PRISM's jackknifed prediction for station 21E07S on February 8. Varying marker sizes illustrate the relative weight of each station's observation in the model prediction. It is obvious that the observation at station 21E07S was roughly 28ºC larger than it's predicted value of 8ºC, or 10 standard deviations from the value of S (2.8), yet the distribution of surrounding observations fit the linear model well. This appears to be a rather obvious case in which the observations were erroneous.

It is also worth mentioning the existence of a second station with a series of bad observations; RAWS station

COLGATE (32600728) appears to have flat-lined observations for this entire period. In the PRISM prediction, its weight in the regression calculation was correctly assigned a low value, therefore having little effect in the prediction process.

3.2. Case 2: Drifting Observation, SNOTEL Station 22F04S, Salt Creek Falls

In this case, a drifting T_{max} observation was again encountered but was not nearly as severe as in the previous example. Figure 5 illustrates the time series of observations and Figure 6 illustrates their anomalies from average values. For most of the period from July 7 through July 20, 2000, the anomalies of station 22F04S differed from those of neighboring stations, deviating by as much as +6ºC.

The PRISM prediction for one day in the period (July 14, 2000), as shown in Figure 7, clearly shows the unusual nature of the observation. The observed value is roughly 10ºC higher than the predicted value of 21ºC, or 14 standard deviations from S (.7). A low value of S also indicates a good fit to the regression model.

CP values created by the PSQC system indicate not only an inconsistent observation on the day in question, but a series of inconsistent observations for numerous days in the period (July 9-14, 17, and 19-21).

3.3. Case 3: Erratic Observations, COOP Station 356252, Odell Lake East

This case represents a situation in which the T_{max} observations are highly erratic over a two-week period. The time series of observations are illustrated in Figure 9 while the anomalies from average are found in Figure 10. In particular, the focus will be on the observation for July 14, 2000. The plots suggest that there could be a day-shift problem at 356252, but the shift is not consistent.

As shown in Figure 11, COOP station 356252 is located at the east end of Odell Lake, while SNOTEL station 22F03S is found at the west end. This site is located on the crest of the Cascades in Oregon. The erratic behavior may be partially related to an issue of site location adjacent to Odell Lake. However, on the day in question it appears that the prevailing winds are out of the southwest and the expectation might be cooler rather than warmer temperatures. Figure 12 further illustrates the proximity of the surrounding stations used in the prediction.

As shown in Figure 13, the observation for 356252 is approximately 2.5ºC higher than the prediction of 22ºC, or 5 standard deviations from the value of S (.5). The low value of S indicates a good fit to the regression model. The notable exception in the scatter plot is SNOTEL station 22F04S, which is the same station and date used in the previous case study. CP values for iteration 2, as illustrated in Figure 14, also reflect the erratic behavior and inconsistent nature of the station's observations compared to surrounding stations. While this plot illustrates the CP values after the first iteration in the PSQC system, Figure 15 illustrates decreasing CP values assigned with successive iterations for just the date in question. The PSQC system has assigned

low confidence that this observation is consistent with its surrounding neighbors.

3.4. Case 4: Time of Observation for COOP Station 358481 , Tidewater

The final case study deals with two issues: 1) afternoon time of observation for a COOP station, and 2) problems associated with combining an hourly observation network (ASOS, RAWS) with a daily observation network (COOP).

It is not uncommon in the summer for temperature observations to reach their maximum values anywhere from 12:00 to 17:00 local time. If temperature observations are recorded once a day during this period and the following day's maximum temperature is significantly lower than the previous day's, it is possible for the following day's temperature observation to reflect the previous day's warmer temperatures. Such a case existed for COOP site 358481 on August 1, 1995, when the observing time was 16:00. On July 31 there appeared to be a local maxima in the T_{max} observations for most stations, as shown in Figure 16. The accompanying anomalies are found in Figure 17. On the following day, the T_{max} values decreased 2-7ºC for most stations. ASOS stations KONP and KJNW represented coastal sites on or very near to the cool Pacific Ocean, so their temperature change was moderated accordingly. Both RAWS station 3260C236 and COOP station 358481 were located inland and were not nearly as affected by the cool ocean temperatures. Figure 18 is a 3D representation of the sites. Station 358481 tracked fairly well with RAWS station 3260C236 with some exceptions. The scatter plot in Figure 19 shows that the observation at 358481 on August 1 was 6ºC higher than the prediction of 26ºC, or 4.6 standard deviations from the value of S (1.3). A further time series plot of the hourly RAWS site, as found in Figure 20, shows it is likely that the T_{max} value on August 1 for the COOP site represented the warmer temperatures that occurred just after 16:00 on July 31, the previous day. This scenario could become more extreme with larger decreases in maximum temperature from one day to the next. Figure 21 shows CP values assigned at each of the 10 iterations.

4. CONCLUSIONS AND FURTHER WORK

Initial inspections of the results from the PRISM PSQC system were promising. The system appears to have correctly isolated inconsistent observations from neighboring stations. When station density is low or surrounding stations are also inconsistent, an appropriate level of confidence is assigned. Some difficult situations during exceptional days along narrow coastal fog belts have been identified, however, and are being targeted for improvement.

While RP (residual probability) is currently being used as the overall CP value, there is most likely more information that can be extracted from the system. Some of the issues that need further investigation are: 1) should something other than a Gaussian distribution be assumed when calculating probabilities, 2) are skewness and kurtosis useful indicators about a stations

behavior (distribution), and 3) are the probabilities of a station observation, prediction, and standard deviation from normal values useful in fine tuning determination of CP values.

From the user's perspective, the question remains as to what is the lowest CP value for which an observation should be replaced with its predicted value. This lower bound will most likely be on a sliding scale, depending on an individual's application. Some applications are more sensitive to outliers than others and may thus require more stringent criteria.

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

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

Doggett, M., C. Daly, J. Smith, W. Gibson, G. Taylor, G. Johnson, and P. Pasteris. 2004. 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.
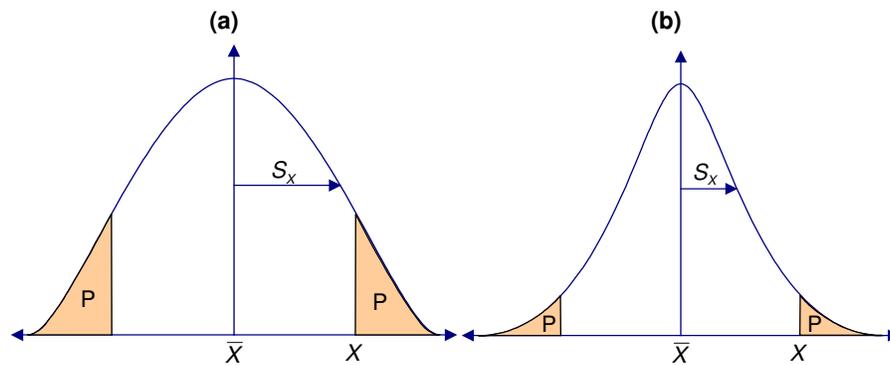
Figure 1. Two-tailed p-values (orange shaded areas) for a generic daily value (*X*) its mean ($\overline{X}$), and standard deviation (s_x) for: (a) a distribution with a large s_x, and (b) a distribution with a small s_x. Note that the p-value is much higher for a given deviation from $\overline{X}$ when s_x is large.
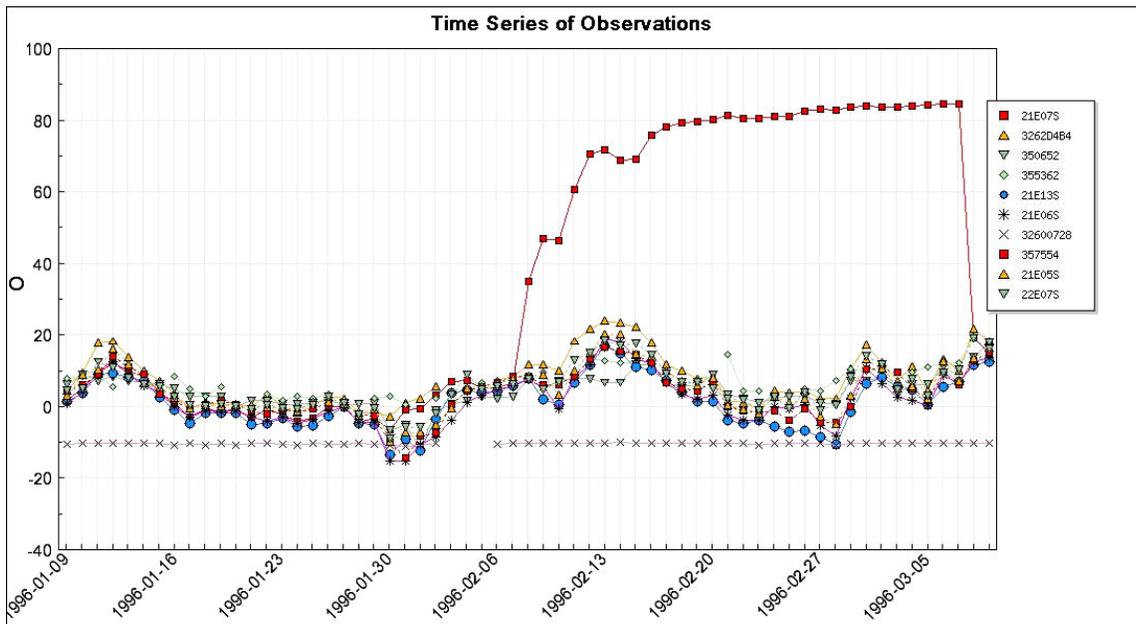
Figure 2. Time Series of Observations, January 9 – March 9, 1996. SNOTEL Station 21E07S exhibited severely drifting observations over a one-month period.
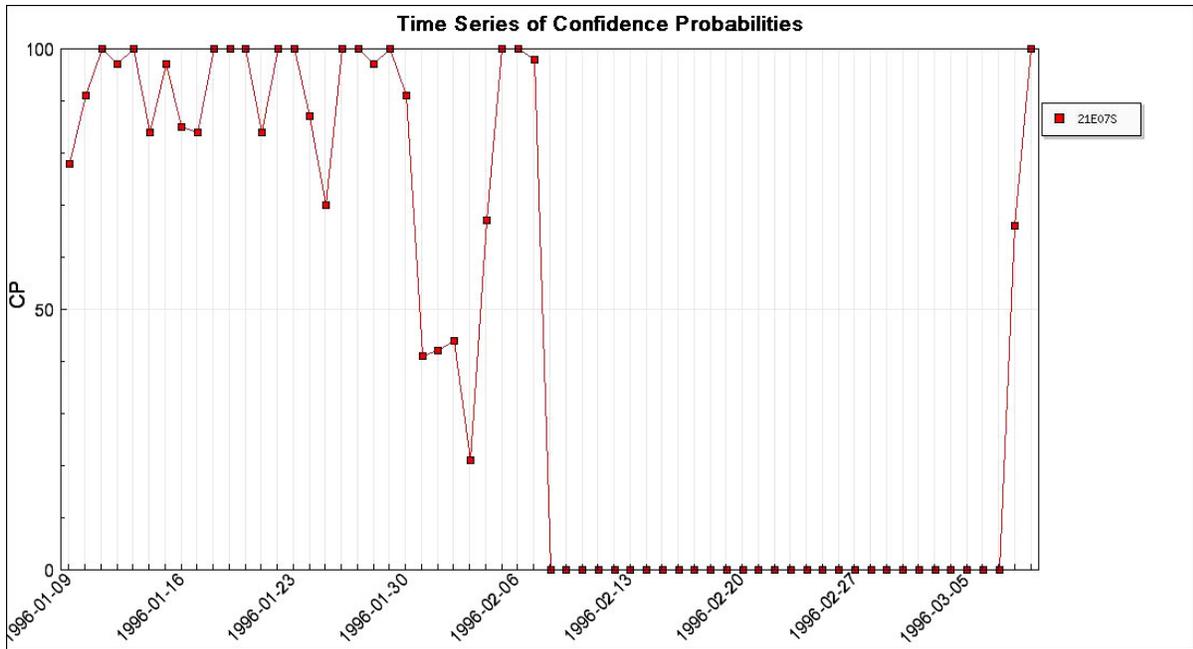
Figure 3. Time series of *CP* (Confidence Probability) values for SNOTEL station 21E07S, January 9 – March 9, 1996.
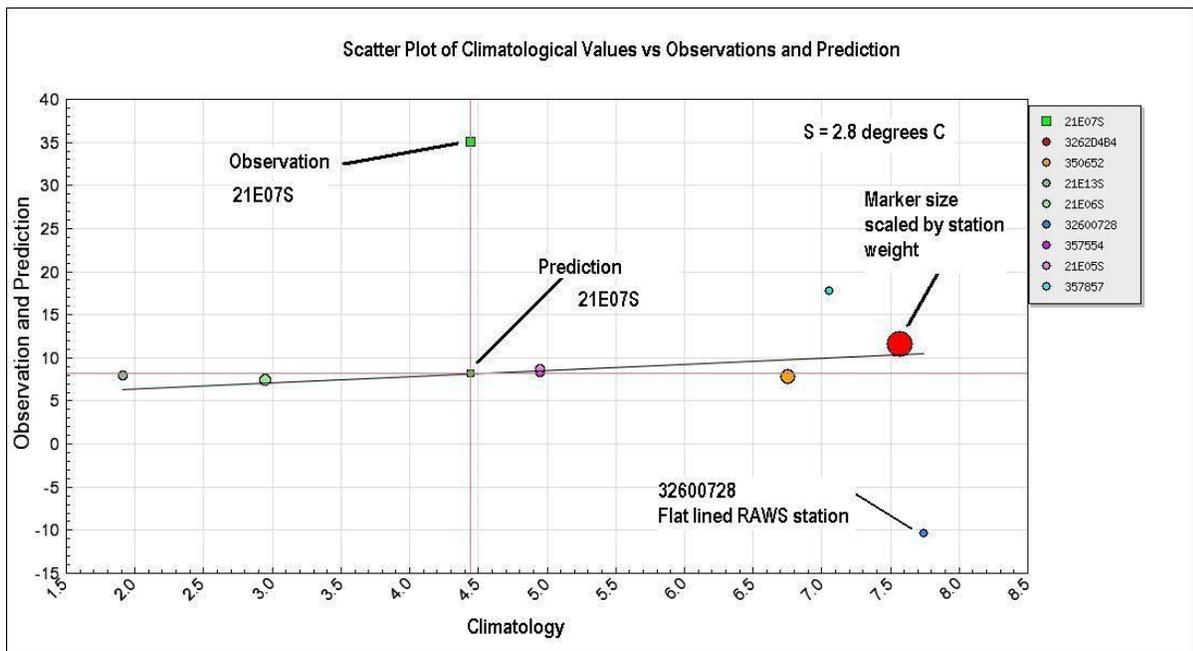
Figure 4. Scatter plot of PRISM regression function for SNOTEL station 21E07S on February 8, 1996.
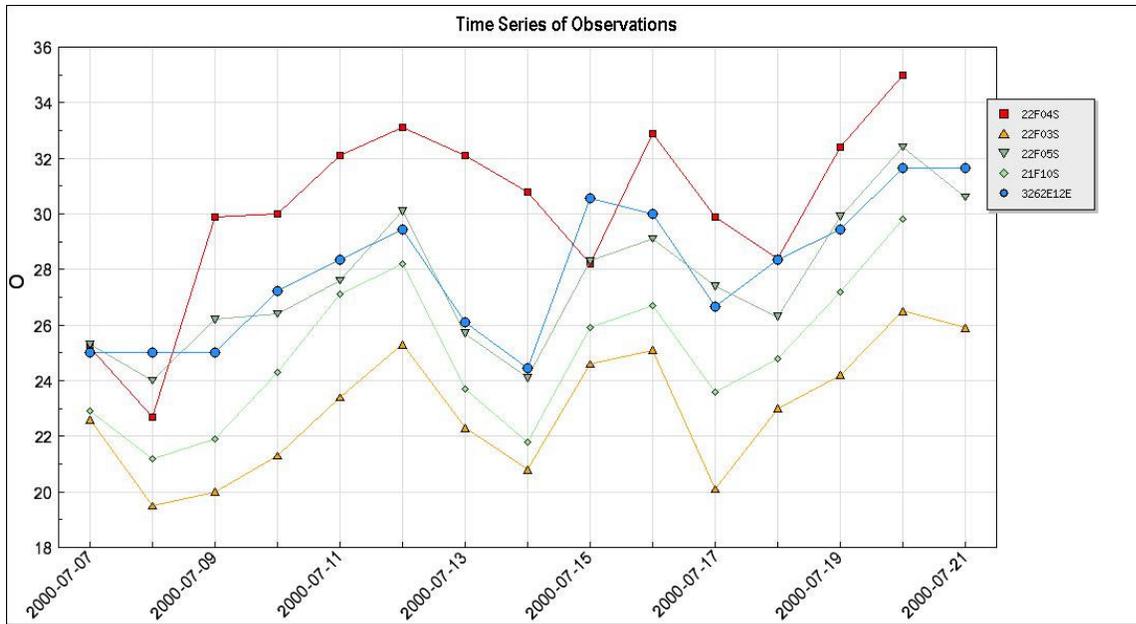
Figure 5. Time series of observations for SNOTEL Station 22F04S and surrounding stations, July 7-20, 2000. Station exhibited a warm bias for much of the period.
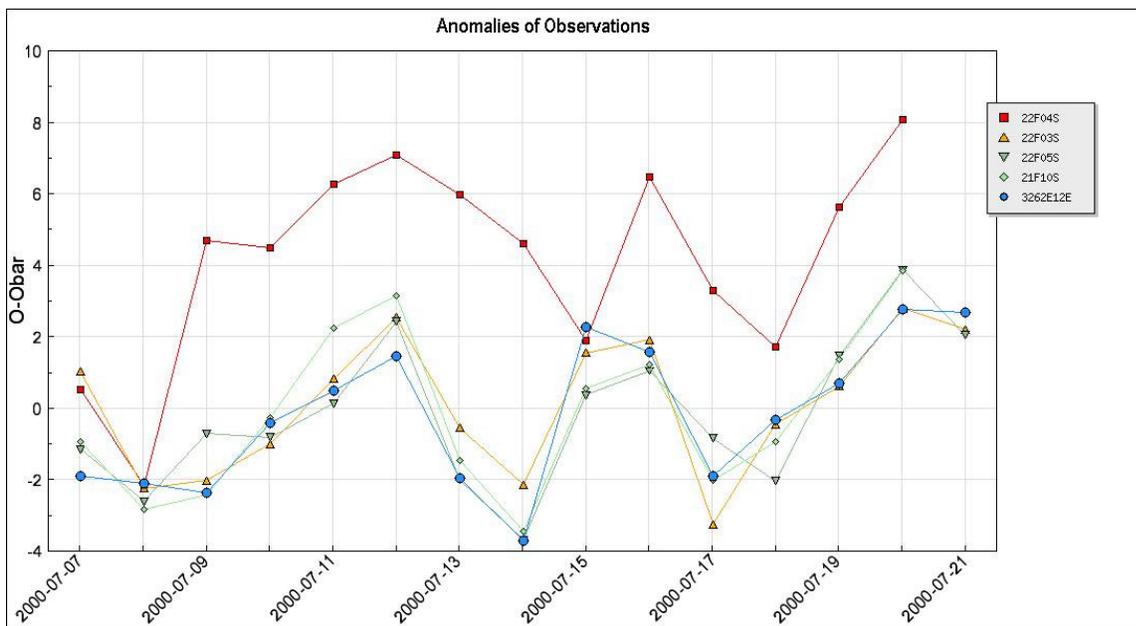
Figure 6. Time series of anomalies from average for SNOTEL Station 22F04S and surrounding stations, July 7-20, 2000. Station exhibited a warm bias for much of the period.
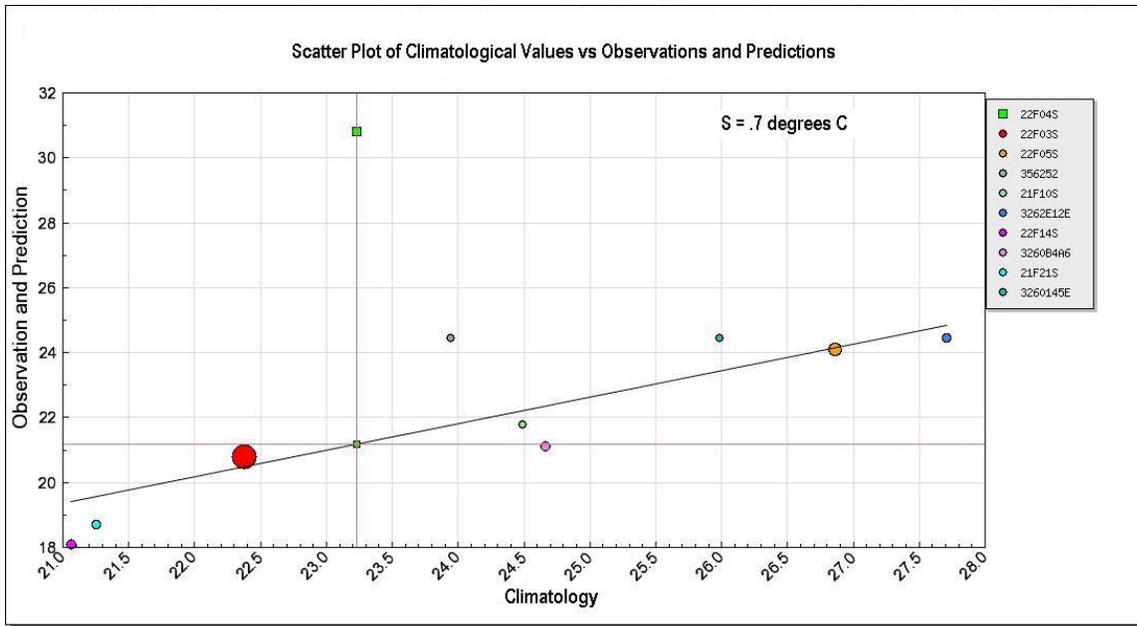
Figure 7. Scatter plot of PRISM regression function for SNOTEL station 22F04S on July 14, 2000.
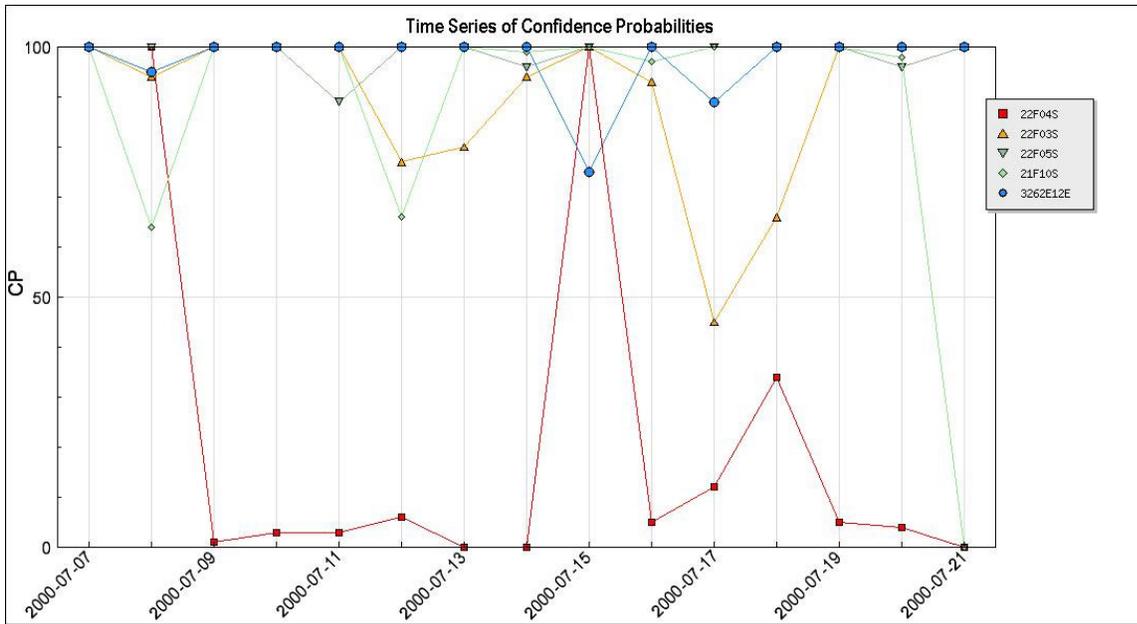
Figure 8. Time Series of Confidence Probabilities, SNOTEL Station 22F04S
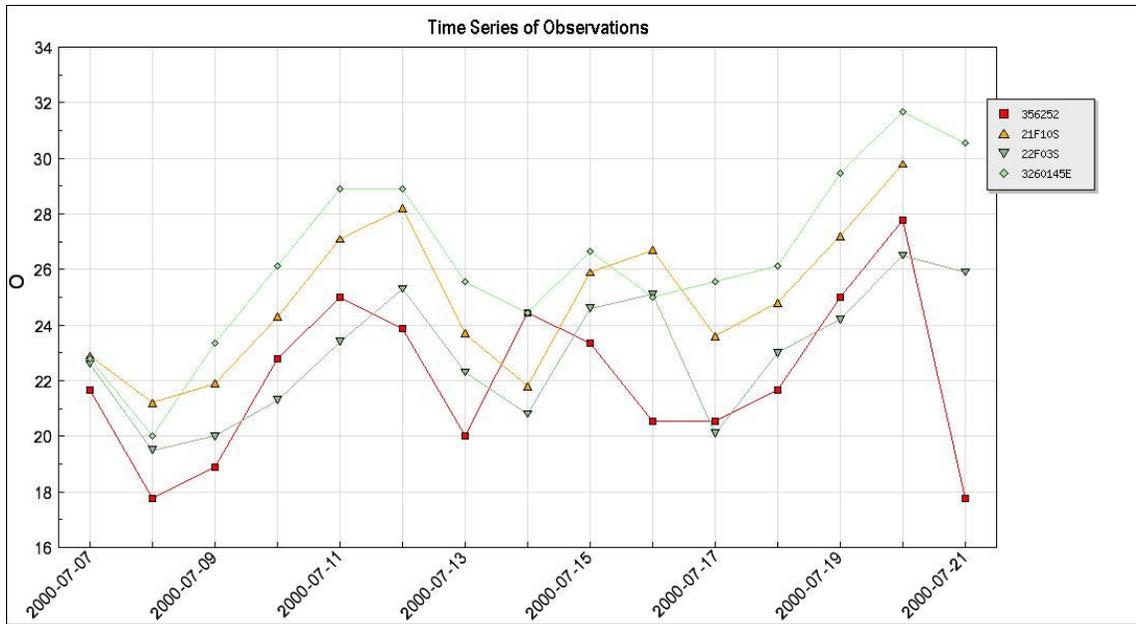
Figure 9. Time series of observations for COOP Station 356252 and surrounding stations for the period July 7-21, 2000.
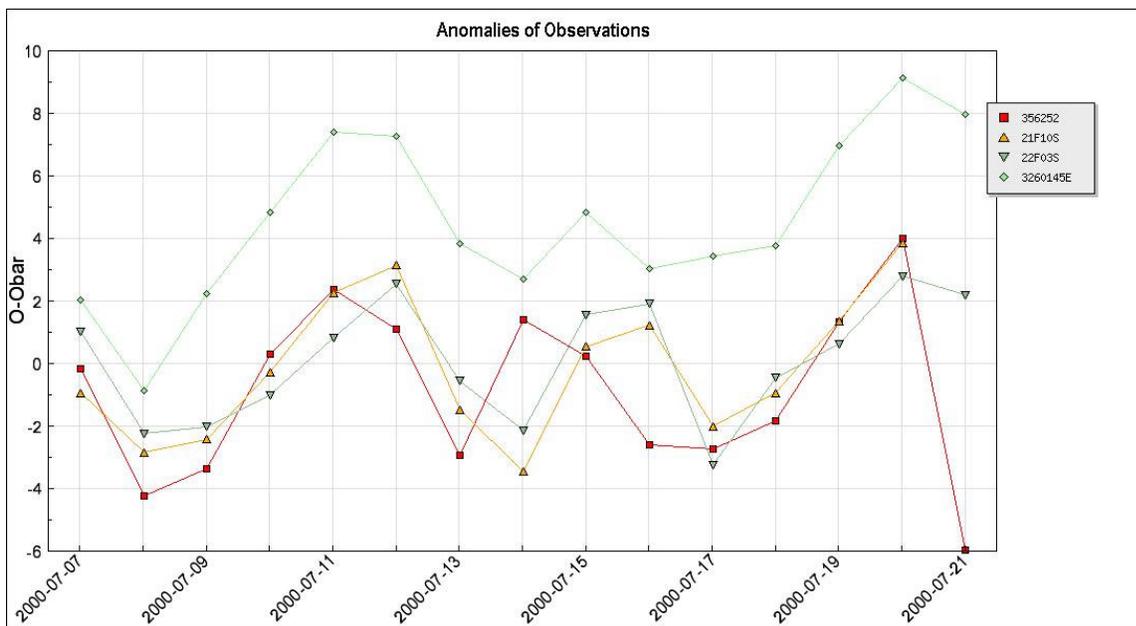
Figure 10. Time series of anomalies from average for COOP Station 356252 and surrounding stations for the period July 7-21, 2000.
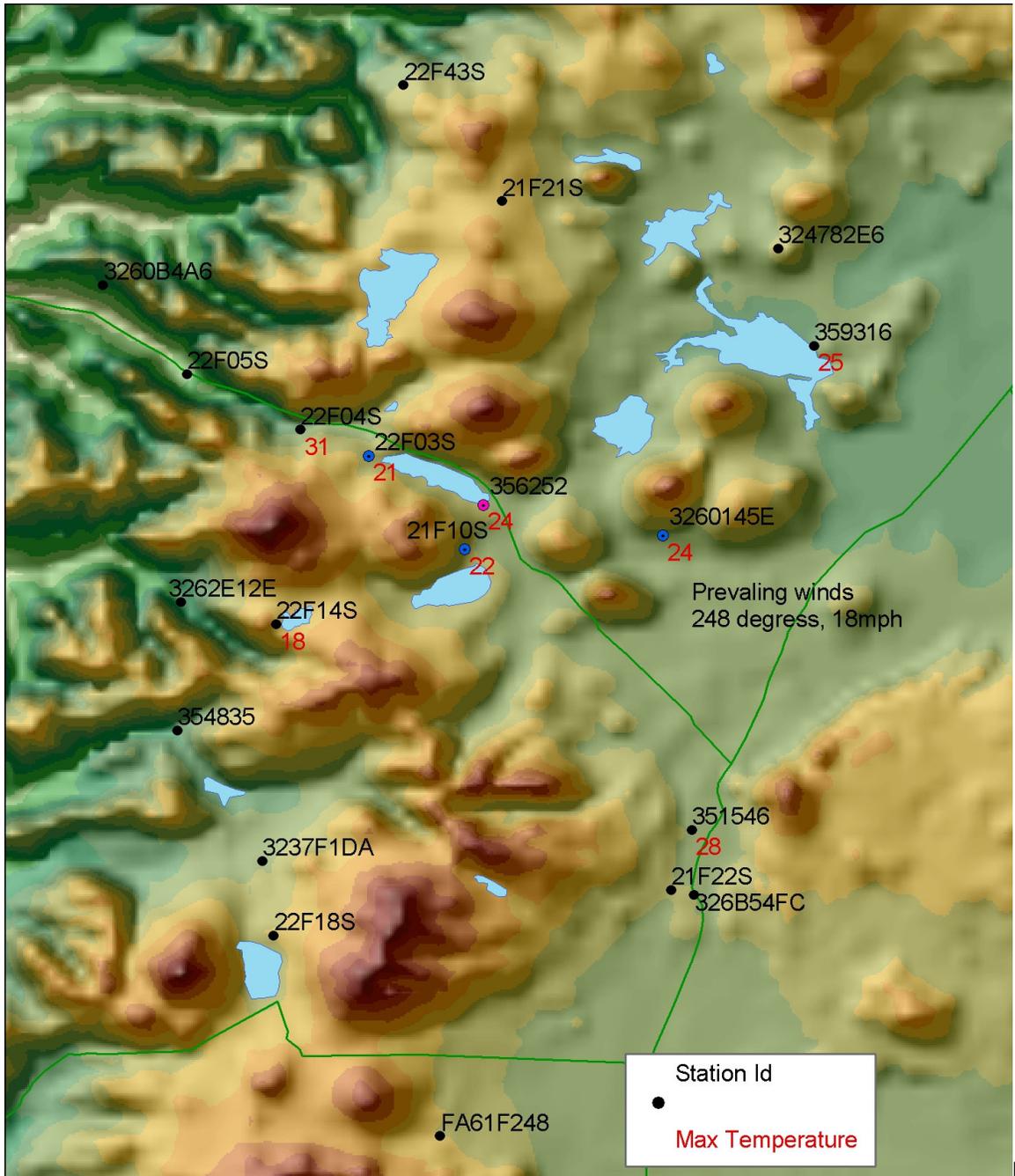
Figure

Figure 11. Shaded relief map of terrain, locations of COOP Station 356252 and surrounding stations, and observed maximum temperatures for July 14, 2000.
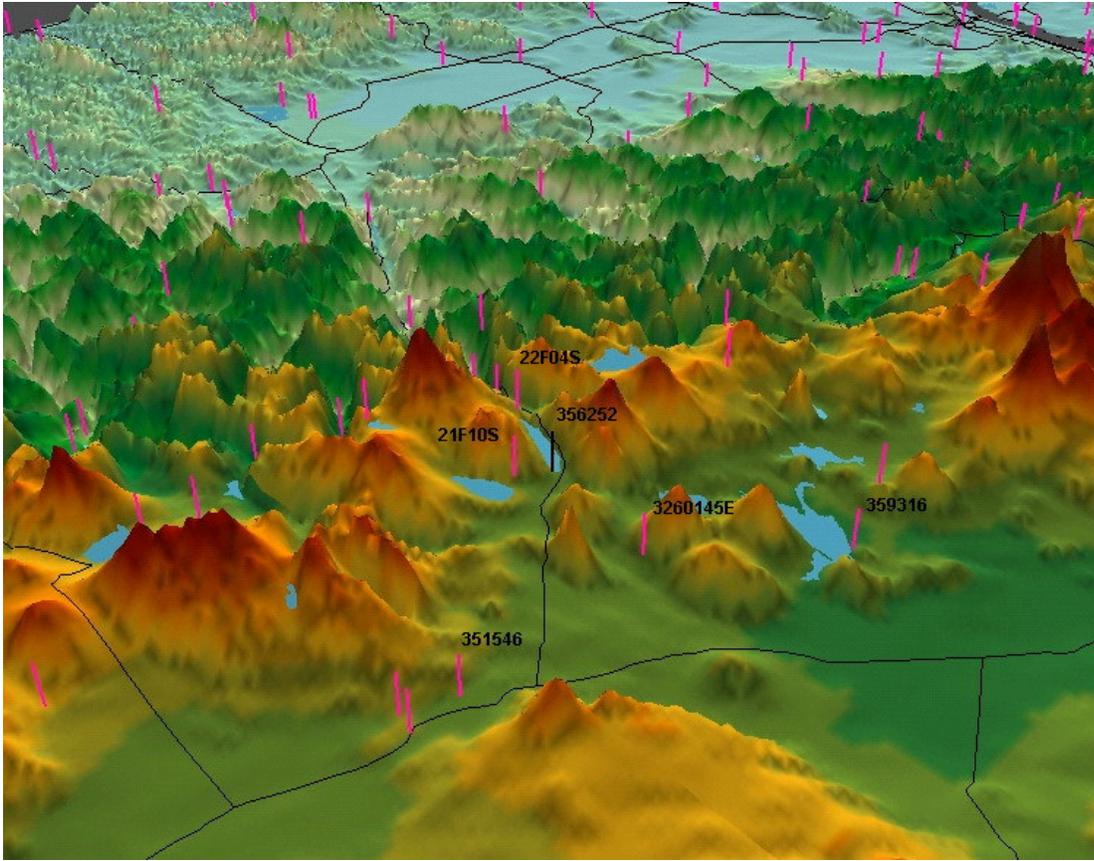
Figure 12. 3D rendering of shaded relief terrain map, and locations of COOP Station 356252 and surrounding stations, looking northwestward.
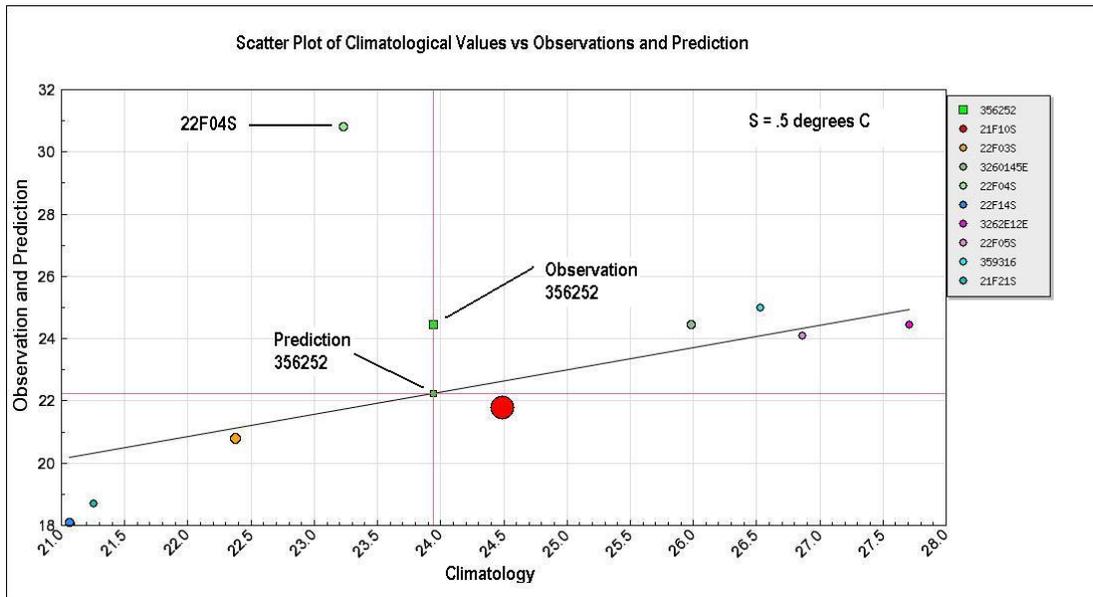
Figure 13. Scatter plot of PRISM regression function for COOP station 356252 on July 14, 2000.
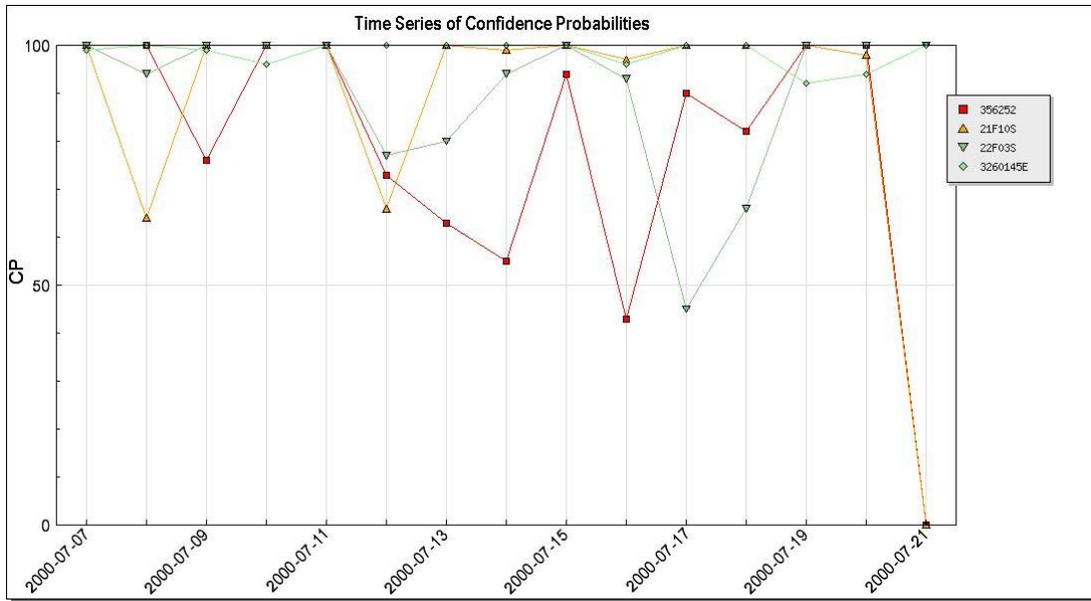
Figure 14. Time series of *CP* (confidence probability) for COOP Station 356252 and surrounding stations for the period July 7-21, 2000, QC iteration 2.
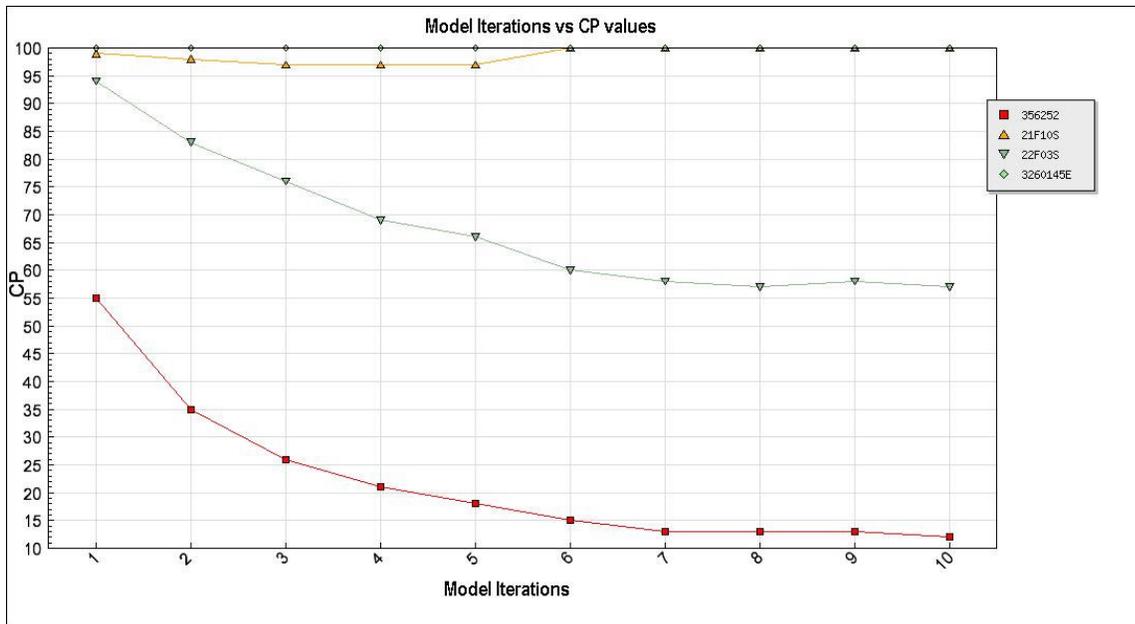
Figure 15. *CP* (confidence probability) as a function of QC iteration, COOP station 356252 and nearby stations, July 14, 2000.
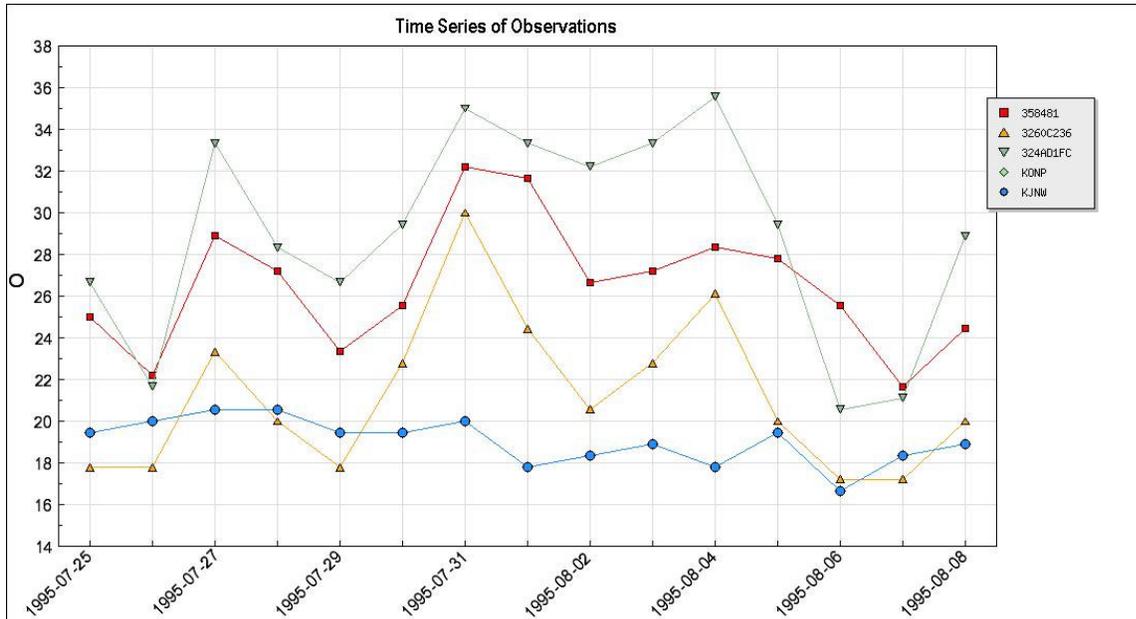
Figure 16. Time series of observations from COOP Station 358481 and surrounding stations, July 25 – August 8, 1995.
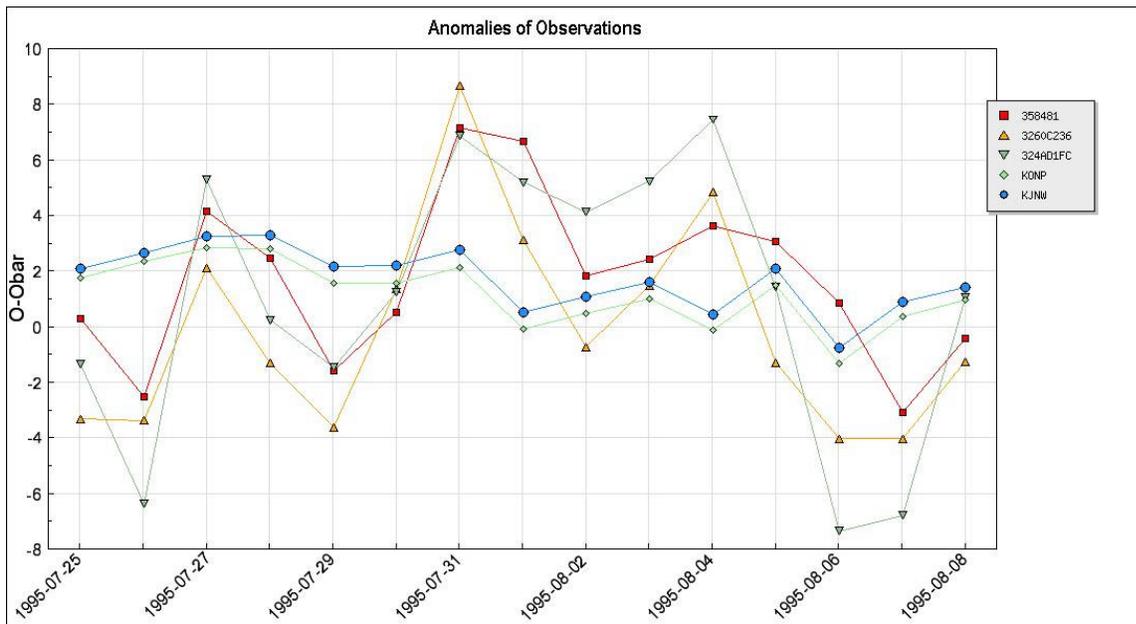

Figure 17. Time series of anomalies from average from COOP Station 358481 and surrounding stations, July 25 – August 8, 1995.
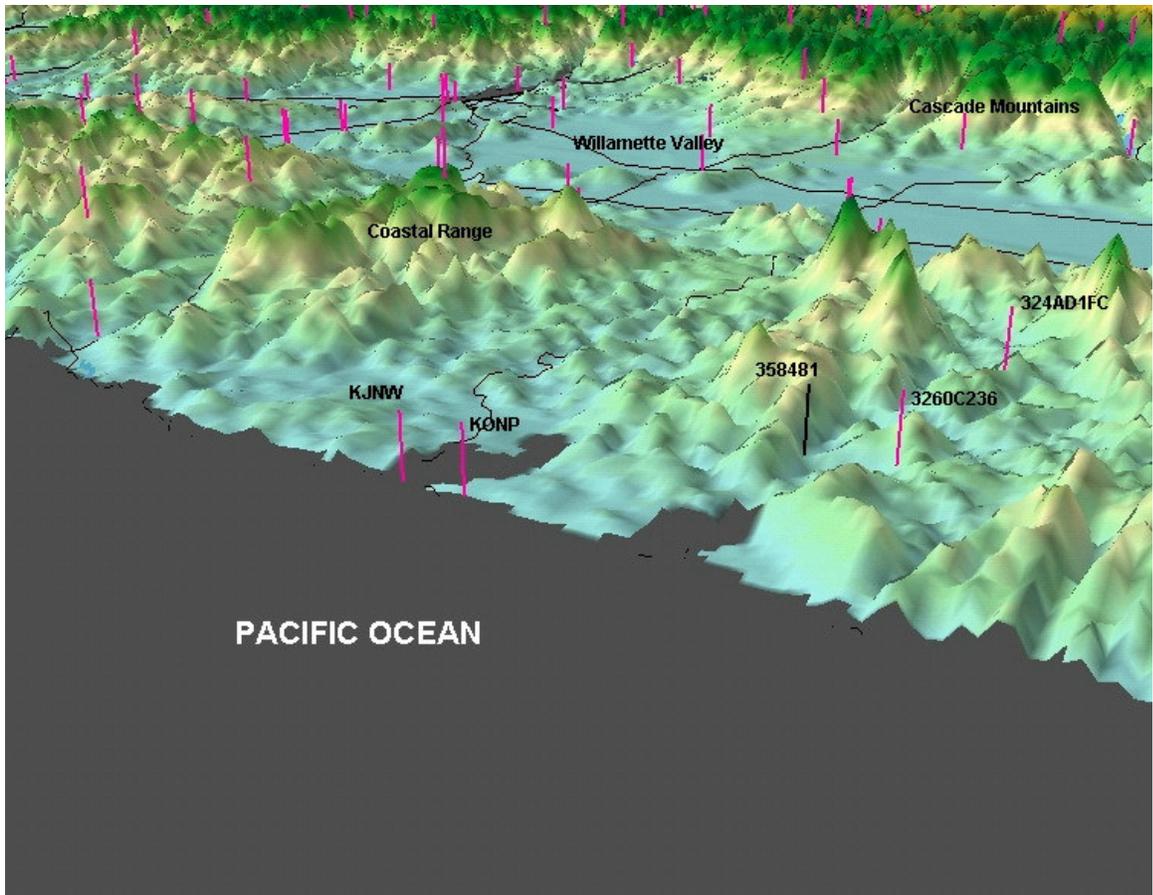
Figure 18. 3D rendering of shaded relief terrain map, and locations of COOP Station 358481and surrounding stations, looking northeastward.
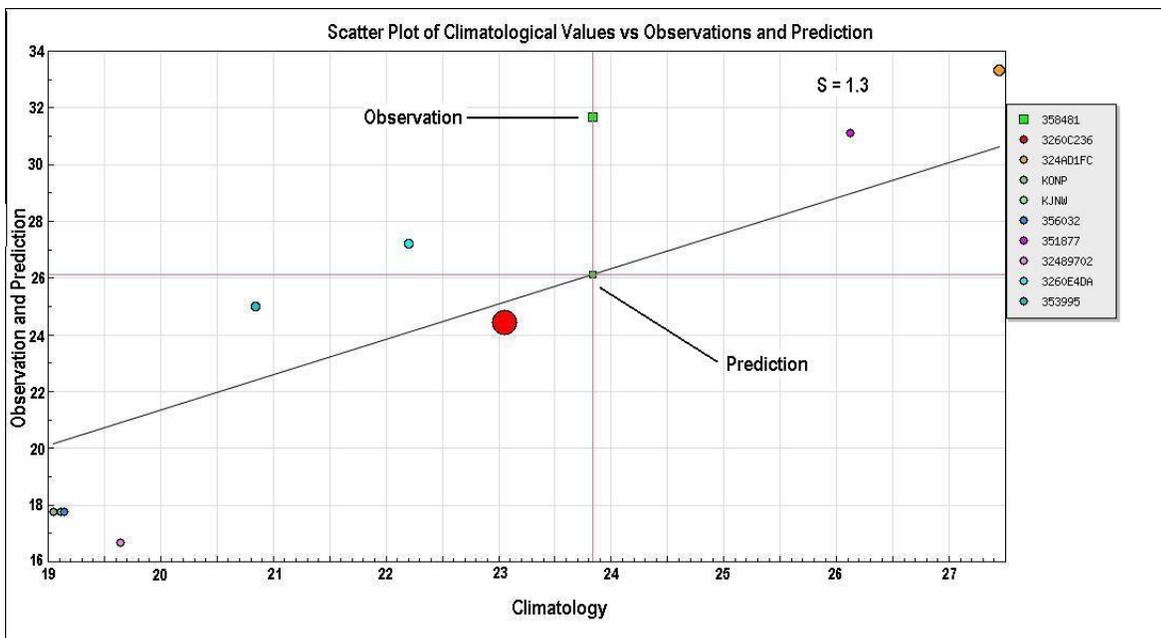

Figure 19. Scatter plot of PRISM regression function for COOP station 358481 on August 1, 1995.
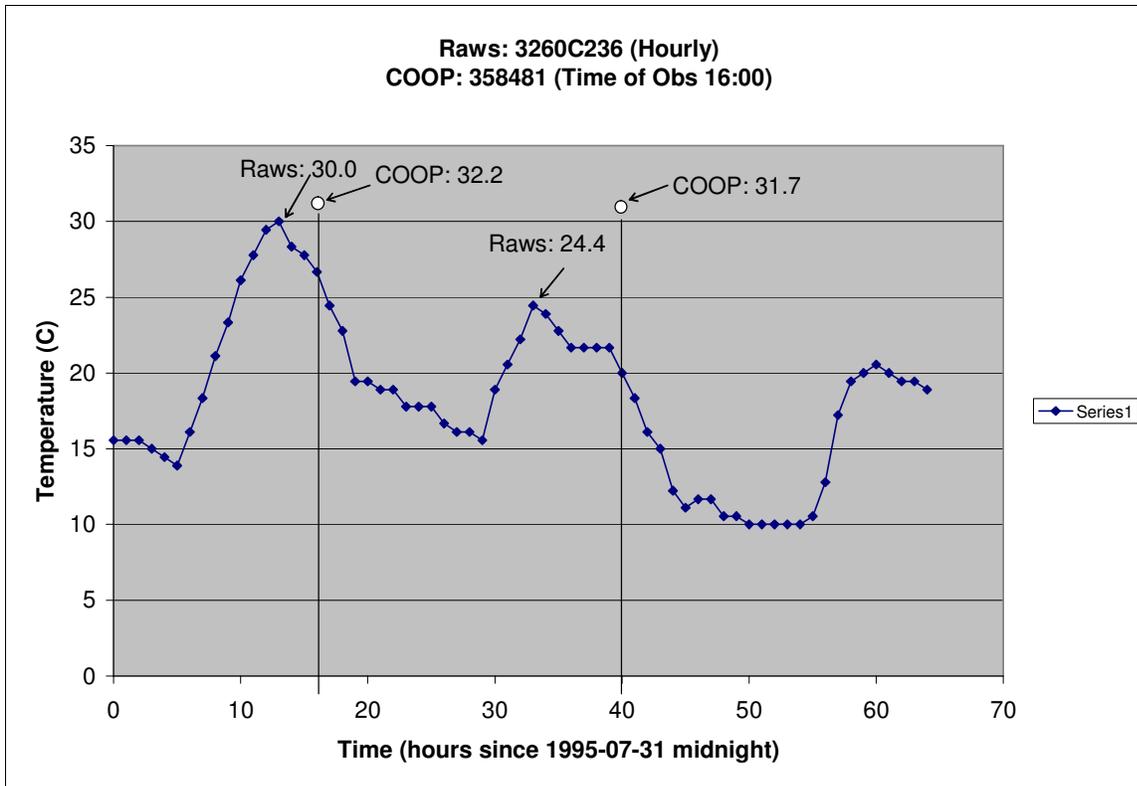
Figure 20. Hourly time series of RAWS station 3260C236 for the period July 31 – August 2, 1995. COOP values are the maximum temperatures at station 358481, observed at 16:00. The increase in maximum temperature at 358481 from July to August 1, despite a significant cool-down at the RAWS station, suggest that the COOP maximum temperature on August 1 actually reflects the previous day.
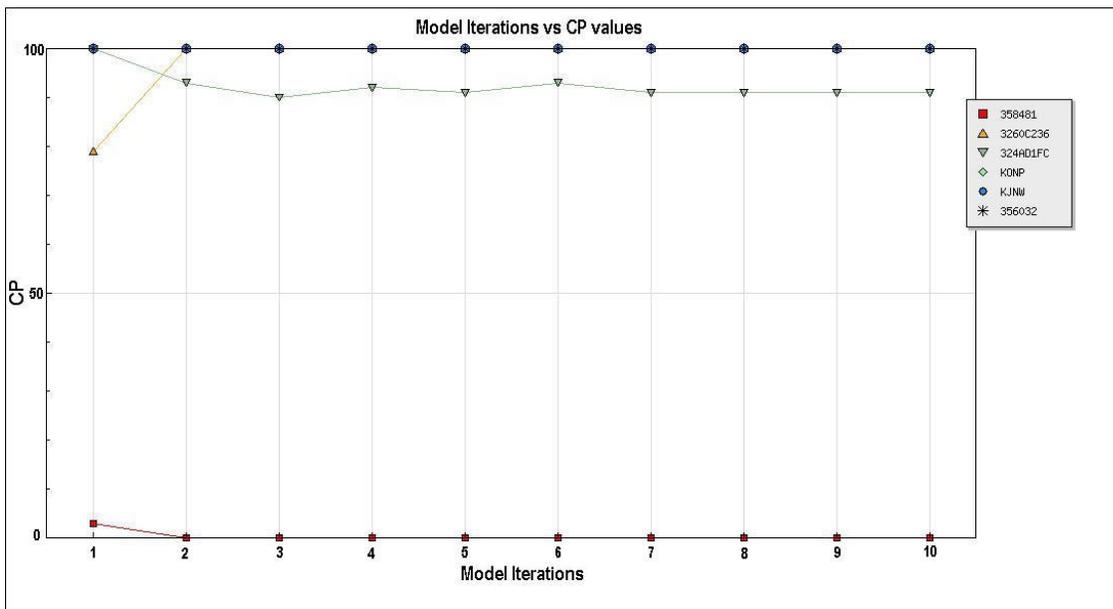

Figure 21. *CP* (confidence probability) as a function of QC iteration, COOP station 358481 and nearby stations, August 1, 1995.