7.2 STRATEGIES FOR EVALUATING QUALITY CONTROL PROCEDURES

Imke Durre *, Matthew J. Menne, and Russell S. Vose
NOAA National Climatic Data Center, Asheville, North Carolina

ABSTRACT

The evaluation strategies outlined in this paper constitute a set of tools essential to the development of robust automated quality control (QC) procedures. Traditionally, thresholds for the QC of climate data have been selected based on the target flag rate or statistical confidence limits. However, this approach by itself does not provide an indication of the procedures' effectiveness at detecting true errors in the data. This information is best obtained by means of a careful analysis of geographical and seasonal patterns of flag rate combined with a systematic manual inspection of samples of flagged values. During the development process, an iterative approach of pattern analysis and threshold selection aids in choosing the best possible combination of QC procedures and associated thresholds.

1. INTRODUCTION

Users of meteorological data may legitimately ask, "To what extent have quality-control (QC) procedures removed significant errors from the data, and at what cost"? In other words, users need to know what types of errors remain in a dataset and whether the QC procedures have inadvertently removed true climate extremes. Ideally, this information would be provided via a thorough evaluation of the type I and type II errors, i.e., the degree to which the QC process identified good observations as erroneous and the extent to which known errors remain undetected. An assessment of the circumstances under which the two types of error occur also benefits the data user.

This paper outlines three components of such an evaluation process. The approach relies on manual inspection as a tool for (1) the selection of appropriate thresholds for individual procedures, (2) the examination of patterns in flagged values, and (3) the determination of the type I and type II error rates. An "extremes check" for daily precipitation totals is used to illustrate the approach.

In brief, the philosophy behind the evaluation process is that a QC system should be tailored to the data to which it is applied and that empirical error rates should be documented for the end user. The reasons for this philosophy are discussed in Section 2. Section 3 contains a brief description of the precipitation extremes check. In Sections 4-6, the three evaluation strategies are explained and illustrated. Some concluding remarks are found in Section 7.

2. UNDERLYING PHILOSOPHY

A thorough evaluation of a QC system is of particular importance because obvious errors sometimes remain in quality-controlled datasets (e.g., see the appendix of Durre et al. 2006) and because interesting climatic features are occasionally identified as errors (Wolter 1997; Fiebrich and Crawford 2001; Graybeal et al. 2004a,b). For example, the Complex Quality Control applied to the Comprehensive Aerological Reference Data Set did not identify clearly erroneous pressures, including a surface level at 70 mb (Durre et al. 2006). Conversely, the QC system for Release 1 of the Comprehensive Atmosphere Ocean Data Set rejected a significant portion of the unusually warm sea surface temperatures in the central tropical Pacific prior to the 1982/83 El Niño event because the limits of its "trimming" check were set too tightly to accommodate both synoptic and interannual variability (Wolter 1997). These cases represent examples of type II and type I errors, respectively.

Typically a quality control procedure is treated as a hypothesis test in which the null hypothesis is that a datum is valid. The null hypothesis is rejected when the datum (or a parameter derived from it) exceeds a specified threshold. Approaches for establishing thresholds frequently employ statistical confidence limits (Collins, 2001; Hubbard et al., 2005), measures of deviation from the mean (Kahl et al., 1992; Wolter, 1997), or target flags rates (Graybeal et al., 2004b). The resulting thresholds often imply that an expected percentage of values will be flagged regardless of how many errors are actually in the data. If applied to error-free data, this percentage is equivalent to the type I error rate. Otherwise, the type I error rate is unknown because both data errors and valid values are likely to have been flagged.

Consider, as in Hubbard et al. (2005), a simple test in which a value is valid only when it lies within ± 3 standard deviations of the long-term mean. Assuming a normal distribution, the expectation is that 99.73% of all data values fall within these limits, yielding a flag rate of 0.27%. If no errant values exist, then all flagged values are type I errors. In this case, the type 1 error rate is also 0.27%. On the other hand, the percentage of flagged values that are false positives is 100. This “false positive rate” is of particular relevance to the data user.

* *Corresponding author address:* Imke Durre, NOAA National Climatic Data Center, 151 Patton Avenue, Asheville, NC 28801; e-mail: Imke.Durre@noaa.gov.

Unfortunately, when errors are present, the false positive rate is unknown unless flagged values are inspected (Kunkel et al. 1998; Graybeal et al., 2004b).

Another unknown is the number of true errors that remain undetected (type II errors). One approach to estimating the type II error rate is to introduce erroneous values into a sample dataset in order to determine whether the "seeds" are detected by the QC process (Guttman et al., 1988; Graybeal et al. 2004b; Hubbard et al. 2005). The errors introduced either are chosen to reflect known types of errors (e.g., Graybeal et al. 2004b) or are generated randomly from a uniform or normal distribution (Hubbard et al. 2005). The results provide insight into the sensitivity of the check versus the magnitude of error. However, when there is little knowledge of both the type and distribution of true errors, the correspondence between the miss rate for seeded errors and the miss rate for true errors is unknown. Furthermore, since the values flagged are likely to be a combination of seeded errors, valid values, and true errors, error seeding is not well suited to determining the check's false positive rate (Graybeal et al., 2004b).

From the above discussion, it follows that neither seeding nor threshold selection based on expected error rates is sufficient for evaluating the performance of quality control procedures. Rather, a thorough evaluation should include an assessment of spatial and temporal patterns in the flag rate, of biases associated with particular meteorological conditions, and of the overall type I and type II error rates. An evaluation should further include the determination of whether any biases or unreasonable error rates are the result of inappropriate thresholds, an inadequate representation of the spatial or temporal variability, variations in data resolution, undocumented observing practices, or systematic errors in the data.

A critical component of such an evaluation is the manual inspection of a random sample of flagged values for false positives and a randomly selected sample of all values for obvious errors that are not detected by the procedure. This inspection process is similar to the practice of manual validation, which is often employed in semi-automatic quality assurance (Guttman et al., 1988; Loehrer et al. 1996; Wolter 1997; Shafer et al., 2000; Graybeal et al. 2004a). In both cases, inspection by a human expert is used to assess the validity of decisions made by automated procedures. However, in the approach proposed here, the purpose of the manual evaluation is to document the performance of the automatic procedures and to provide guidance for improvement rather than to override a system's automated decisions.

3. BRIEF DESCRIPTION OF THE PRECIPITATION EXTREMES CHECK

This section provides a brief description of the "precipitation extremes check" used as an illustrative example in subsequent sections. In brief, this check identifies erroneously large 24-hour precipitation totals by comparing the value in question to the overall distribution at a given station and time of year. The particular challenge in the case of precipitation is that the distribution of daily totals is positively skewed. The skewness results from the relatively large frequency of small totals versus the comparatively rare occurrence of heavy precipitation in conjunction with, for example, deep moist convection or tropical systems. The long tail of the distribution makes it difficult to distinguish between valid and erroneous extreme values, precluding the use of common measures of central tendency and dispersion (e.g., the mean and standard deviation).

The precipitation extremes check uses a reference value equal to the daily total that corresponds to the 95th percentile of all nonzero values for a given station and time of year. For each daily total a ratio is calculated as

$$ratio = \frac{x}{p_{95}}, \quad (1)$$

where x is the daily total and p_{95} represents the 95th percentile of all nonzero daily values for the period within a 29-day window centered on the day in question. A daily total is flagged when the ratio exceeds a specified threshold. This threshold must be selected such that the check identifies erroneous values without flagging a significant number of real extreme events.

For illustrative purposes, a threshold is developed here using observations from Cooperative Observer (Coop) Network stations in the contiguous U.S. Coop data are well suited to this task because several information sources are available to enhance the manual inspection process. For example, scanned images of original Coop observer forms can be used to identify digitizing errors. Likewise, tropical cyclone track data and qualitative comparisons with surrounding stations can assist in the verification of certain heavy precipitation totals, thus aiding the identification of false positives.

4. THRESHOLD SELECTION TECHNIQUE

The type 1 and type 2 error rates of a QC test are directly linked to the threshold chosen for that test. For instance, a high threshold usually implies a low false-positive rate, while a lower threshold tends to detect a larger number of errors, albeit at a greater risk of over flagging. Consequently, when implementing a particular QC procedure, the threshold ideally should strike a balance between the number of errors detected and the number of false positives.

A logical first step is to establish the range of parameter values within which the threshold is certain to fall. For the precipitation extremes check, this was accomplished by examining small number of observations with ratios exceeding 1. The initial

inspection suggested that all events with ratios greater than 15 were clearly erroneous and many events with ratios less than 6 were plausible because they coincided with heavy totals at neighboring stations.

Observations with ratios between 6 and 15 were then examined in detail. Specifically, ten values were chosen at random from each “bin” in Table 1, the validity of each value being assessed by manually examining observations at surrounding stations as well as by consulting the original observer forms when available. Each sample value that was judged to be erroneous was counted in the "number of errors in sample" column in Table 1. When a value was found to be questionable but not clearly erroneous, half an error was counted.

The total number of errors in each bin is estimated from the error rate in the corresponding sample of inspected values. For a particular threshold, the cumulative false positive rate is obtained from the total number of false positives in all bins above the threshold relative to the total number values above the threshold. For example, for ratios 12 to 15, 8.5 out of the 10 sample values inspected were considered erroneous. Assuming that this 85% error rate applies to the entire bin, 43 out of the 51 bin values are errors. For a ratio threshold of 12, the cumulate number of errors is 268, while the total number of values with ratios greater than 12 is 276. Therefore, the cumulative false-positive rate is equal to 8 out of 276, or approximately 3%. Note that this percentage is relative to the total number of values flagged, not to the total number of data values processed.

In general, the false-positive rate increases significantly for ratios below 9. Furthermore, half of the values are false positives when the threshold is between 8 and 9, implying that the probability that a value is an error is equal to the probability that it is valid. An example of a false positive occurred at Benevides, Texas where 488 mm of rain was reported on September 11, 1971, yielding a ratio of 8.7. This total occurred in conjunction with the landfall of Hurricane Fern and is corroborated by similarly heavy totals at several nearby stations. A developer interested in preserving this type of extreme value could therefore set the ratio threshold to 9, leaving errors in lower ratio categories undetected by this check. If the number of true errors in the lower bins were considered excessive, additional checks could be developed to explicitly target those undetected errors.

5. ANALYSIS OF PATTERNS IN FLAG RATE

The evaluation of a QC check should also include the examination of spatial and temporal patterns in the flagged values. In theory, such patterns may be caused by concentrations of true data errors in specific regions or periods (Collins, 2001; Graybeal et al., 2004a,b). On the other hand, a pattern may be indicative of procedural deficiencies such as systematic flagging of particular climatic conditions or the failure to adequately account for different observing practices (Wolter 1997; Fiebrich and Crawford 2001). Patterns may also arise as an artifact of variations in the temporal or spatial resolution of the data that limit a procedure's applicability at specific places and times (e.g., a lack of neighbors for a “spatial” consistency check).

Pattern analysis involves the generation and interpretation of summary statistics. Typical examples include histograms of the overall percentage of values flagged during each calendar month and maps of stations with flagged values. An example of the latter is the spatial distribution of flags set by the precipitation extremes check at stations across the contiguous United States. For a threshold of 9, flags are concentrated in the northern and interior western United States during the cold season months of November through March (Fig. 1a). In contrast, no apparent bias exists in the corresponding map for the months of May through September (Fig. 1b). However, when the threshold is lowered to 6, a small bias towards the South and the Eastern Seaboard appears during the warm season (Fig. 1c) while the overall cold-season bias (not shown) remains essentially the same.

Based on the examination of observation forms, the concentration of flags in the north during winter months (Fig. 1a) appears to be related to the practice of inadvertently recording a snowfall total in the water equivalent field. The areas with a relatively high concentration of warm season flags for a threshold of 6 (Fig 1c) coincide with regions most likely to be affected by tropical cyclones, which is consistent with the evaluation of specific cases during the threshold selection process. Thus, the wintertime pattern in Fig 1a reflects a systematic error in the data, while the spatial bias in Fig. 1c is likely an indication of overflagging.

6. ANALYSIS OF QC SYSTEM PERFORMANCE

Most QC systems consist of a suite of checks that are applied in succession (e.g., an extremes check followed by a spatial check). Once a threshold has been set for each check individually, the final step in the evaluation process should be an analysis of overall system performance (e.g, Lorenc and Hammon, 1988). The effectiveness of the system is best characterized by its false-positive rate relative to all flagged values and its miss rate relative to known errors. On the other hand, the Type I and Type II error rates are useful for assessing the impact of the QC system on the dataset as a whole. All of these quantities can be determined through the manual inspection of a representative sample of the processed data. In general, this involves 1) applying the QC system to the entire dataset, 2) manually inspecting random samples of the data for invalid values, and 3) computing the performance metrics for the system as a whole.

To obtain the overall false positive and Type I error rates, one should choose a random sample of the flagged values and then determine, via manual

inspection, the number of false positives in the sample. As an example, suppose that a hypothetical system flags 10,000 values in a dataset of 10 million. If 100 flagged values are examined manually and 20 are considered valid, then the system's false-positive rate is estimated to be 20%. By implication, the corresponding Type I error rate is approximately 2000 out of 10 million or 0.02%.

Although the actual total number of true errors in the dataset is rarely, if ever, known, manual inspection can also be used to estimate the miss and Type II error rates. To obtain such estimates, one might randomly choose a certain number of values from the entire dataset and manually determine their validity. Suppose, for example, that two of 100 values selected are manually identified as errors, but only one is detected by the automated QC. In this case, the miss rate if 1 out of 2 or 50%, while the Type II error rate is 1 out of 100 or 1%. In many circumstances, the sample may not include any obvious errors. In that case, the miss rate is undefined and the Type II cannot be quantified precisely, but is known to be less than 1%.

7. CONCLUSION

The strategies outlined in this paper constitute a set of tools for evaluating automated QC procedures. These strategies, which rely heavily on manual review, are essential to quantifying the performance of QC checks and should be used to ensure a robust QC system. If each test is thoroughly evaluated as it is developed, the system developer has the luxury of continually adapting the QC strategy thereby maximizing the effectiveness of the overall system.

In general, QC system development should include the following:

- the design of tests to detect known data problems
- the use of manual evaluation and pattern analysis of flagged values to select test thresholds such that each check has a low false positive rate
- the quantification of the overall type I error, type II error, false positive and miss rates for the combination of checks
- when necessary, the development of additional checks that target undetected gross errors and re-evaluation of system performance.

The question arises, when is it necessary to develop additional checks? Although the answer may depend on the particular application for which the quality-controlled data are to be used, the following general considerations may serve as guidance. Firstly, while each individual procedure may be designed to detect only a certain kind of error (e.g., unrealistically extreme values), one would expect the entire QC system to be able to identify the vast majority of egregious errors, i.e., those erroneous data points whose presence in the output data would damage the credibility of the quality-control effort. If the miss rate is unacceptably high, the development of additional checks may be a more appropriate remedy than the lowering of parameter thresholds. In fact, in our experience, the best results are obtained when the developer maintains an open mind towards the possibility of abandoning ineffective procedures and adding new checks as insights are gained about the error characteristics of the data and the efficiency of different procedures.

Following QC system development, both the QC procedures and the evaluation results should be documented. Such documentation should include a description of each check and its false positive rate, the type I and type II errors rates for the overall system as well as the percentage of values by the system. Information regarding the types of errors being detected, the types of errors that might remain in the data, and the conditions under which valid values might be misidentified as errors should also be provided. This kind of comprehensive documentation enables users to make informed decisions about the suitability of the data given their particular application.

8. REFERENCES

Collins, W.G., 2001: The Operational Complex Quality Control of Radiosonde Heights and Temperatures at the National Centers for Environmental Prediction. Part I: Description of the Method. *J. Appl. Meteor.,* **40**, 137–151.

Durre, I., R. S. Vose, and D. B. Wuertz, 2006: Overview of the Integrated Global Radiosonde Archive. *J. Climate*, *19*, 53-68.

Eischeid, J. K., C. B. Baker, T. Karl, and H. F. Diaz, 1995: The quality control of long-term climatological data using objective data analysis. *J. Appl. Meteor.,* **34**, 2787-2795.

Fiebrich, C. A., and K. C. Crawford, 2001: The impact of unique meteorological phenomena detected by the Oklahoma Mesonet and ARS Micronet on automated quality control. *Bull. Amer. Meteor. Soc*, **82**, 2173-2187.

Graybeal, D. Y., A. T. DeGaetano, and K. L. Eggleston, 2004a: Complex quality assurance of historical hourly surface airways meteorological data. *J. Atmos. Oceanic Tech.*, **21**, 1156-1169.

_____, _____, and _____, 2004b: Improved quality assurance for historical hourly temperature and humidity: Development and application to environmental analysis. *J. Appl. Meteor*., **43**, 1722-1735.

Guttman, N. B., C. Karl, T. Reek, and V. Shuler, 1988: Measuring the performance of data validators. *Bull. Amer. Meteor. Soc.,* **69**, 1448-1452.

Kahl, J.D., M.C. Serreze, S. Shiotani, S.M. Skony, and R.C. Schnell, 1992: In-situ meteorological sounding archives for Arctic studies. *Bull. Amer. Meteor. Soc.,* **73**, 1824-1830.

Lorenc, A.C., and O. Hammon, 1988: Objective quality-control of observations using Bayesian metholds - theory, and a practical implementation. *Quart. J.*

Roy. Meteor. Soc., **114**, 515-543.

Loehrer, S.M., T.A. Edmands, and J.A. Moore, 1996: TOGA COARE Upper-Air Sounding Data Archive: Development and quality control procedures. *Bull. Amer. Meteor. Soc.*, **77**, 2651–2672.

Shafer, M. A., C. A. Fiebrich, D. S. Arndt, S. E. Frederickson, and T. W. Hughes, 2000: Quality assurance procedures in the Oklahoma Mesonetwork. *J. Atmos. Oceanic Technol,* **17**, 474 494.

Wolter, K., 1997: Trimming Problems and Remedies in COADS. *J. Climate*, **10**, 1980-1997.

Table 1. QC evaluation table for the precipitation extremes check. Results are based on the manual inspection of sample values for different ratio thresholds.

Ratio	Number of values in bin	Sample size	Number errors in sample	Number errors in bin	Cumulative false positive rate
>30	114	10	10	114	0%
20 to 30	56	10	10	56	0%
15 to 20	55	10	10	55	0%
12 to 15	51	10	8.5	43	3%
10 to 12	75	10	8	60	7%
9 to 10	52	10	7	36	10%
8 to 9	121	10	5	61	19%
7 to 8	185	10	4	74	30%
6 to 7	481	10	2	96	50%

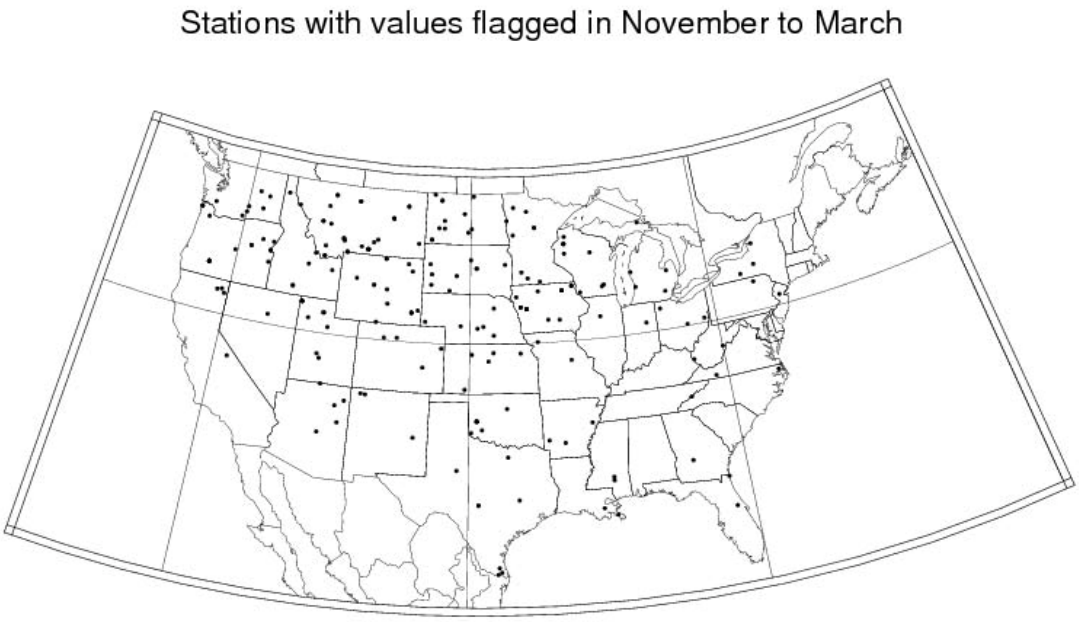

Figure 1a. Geographic distributions of stations with daily precipitation values flagged by the precipitation extremes check using a ratio of 9.0 during the months November through March.

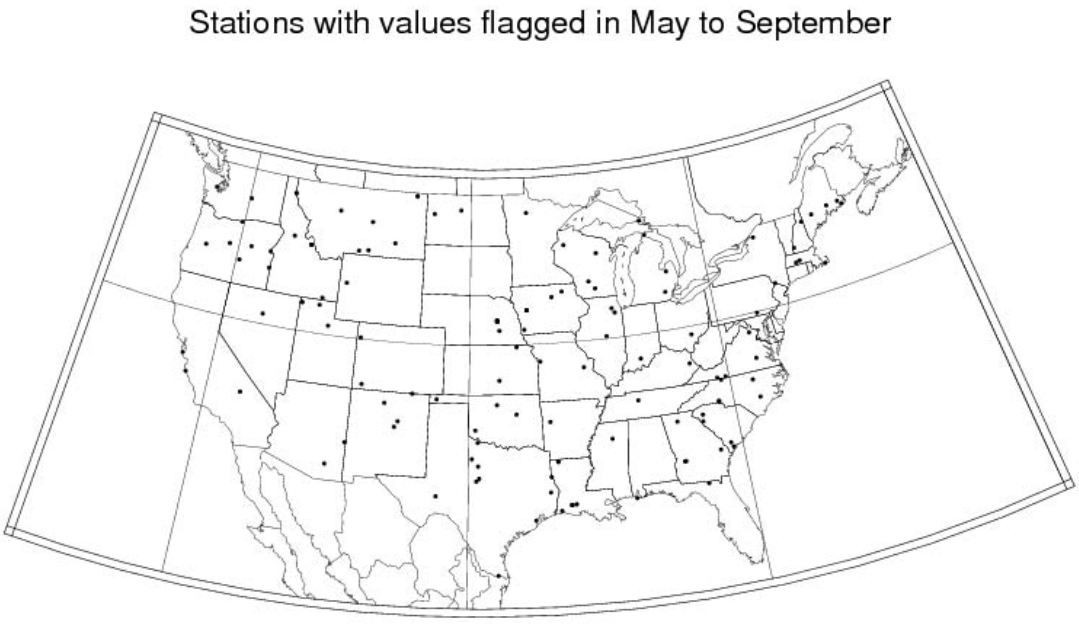

Figure 1b. Geographic distributions of stations with daily precipitation values flagged by the precipitation extremes check using a ratio of 9.0 during the months May through September.

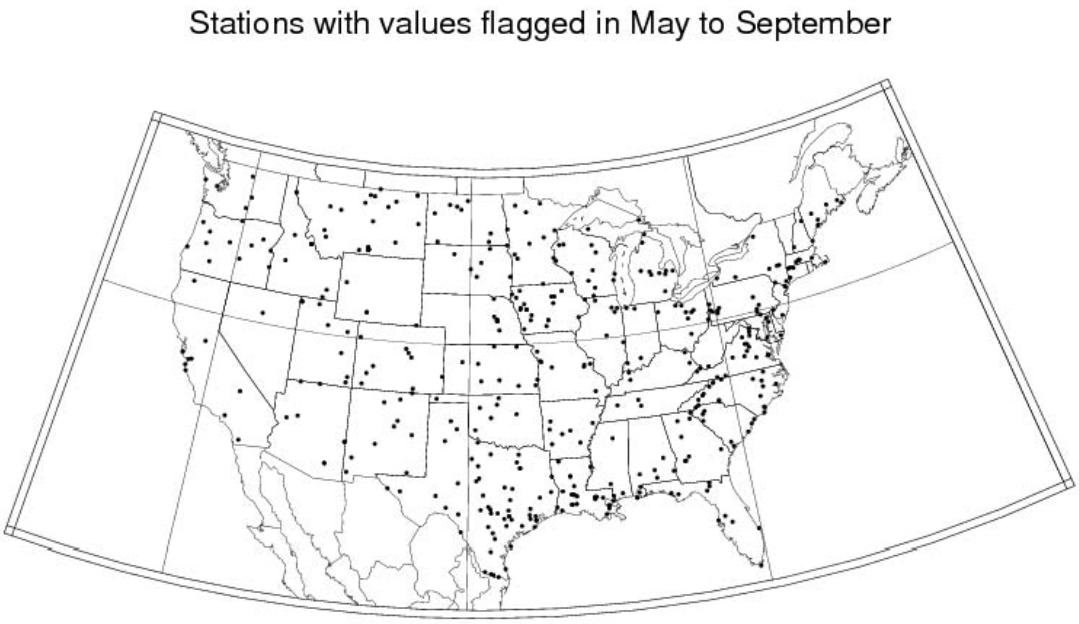

Figure 1c. Geographic distributions of stations with daily precipitation values flagged by the precipitation extremes check using a ratio of 6.0 during the months May through September.