7.6 **Quality Control of pre-1948 Cooperative Observer Network Data**

Kenneth E. Kunkel[1]
Illinois State Water Survey, Champaign, Illinois
David R. Easterling
National Climatic Data Center, Asheville, North Carolina
Kenneth Hubbard
University of Nebraska, Lincoln, Nebraska
Kelly Redmond
Desert Research Institute, Reno, Nevada
Karen Andsager, Michael Kruk, Michael Spinar
Illinois State Water Survey, Champaign, Illinois

1. Introduction

The National Weather Service's (NWS) cooperative observer network (COOP) is the core climate network of the U.S. In operation since the late 19th century, it consists primarily of volunteer observers using standard equipment provided by the NWS. The typical suite of elements, observed daily, include precipitation, maximum temperature, minimum temperature, snowfall, and snow depth. Some stations report only precipitation variables. A few stations observe other variables such as pan evaporation and soil temperature.

These observations were routinely digitized beginning with 1948. Although there have been occasional projects to retroactively digitize selected data (e.g. Kunkel, et.al. 1998), most pre-1948 observations remained available only on paper or microfiche. This has recently changed. The U.S. Congress has provided funding to the National Climatic Data Center (NCDC) for the Climate Database Modernization Program (CDMP 2001). The goal of CDMP is to convert data only available in hard-copy form to computerized formats. The pre-1948 COOP data was one of the first data sets chosen for this conversion.

The authors have undertaken a project to quality control this data set. This paper describes the QC procedures and discusses certain aspects of the data set.

2. Data Set Description

COOP observations are recorded on paper forms (1 sheet per month) and sent to NCDC at the end of each month. In the 1980s, NCDC copied all paper forms onto microfiche. The keying of these data in the CDMP was done from the images on the microfiche. Data were keyed manually by Image Entry, located in London, KY. All data were double-keyed and discrepancies between the two sets of keyed data resolved. This process minimizes the number of keying errors. Also, extremes tests were applied during post-processing, to help ensure accurate keying. Values that failed the extremes tests but verified with the source were retained. Estimated values were added to the database to fill in some missing values; in particular, if the daily precipitation values added up to the monthly total for a station, the other days were zero-filled. The total number of values keyed exceeded 300,000,000.

This data set was given the designation TD-3206 by NCDC. The digital COOP data beginning in 1948 is designated as TD-3200; this includes the routinely keyed COOP data plus the results of various state-based keying projects undertaken through the years. A recently developed data set of keyed COOP data done for nine central U.S. states (Kunkel et al. 1998) was designated as TD-3205. These three data sets were combined for this project to create the data set of all keyed COOP data.

3. Quality Control

There are a number of potential sources of errors in the data set. Some primary examples include observer errors in reading the instruments, observer errors in writing the observations on the form, liquid separation in the thermometers, and legibility of the forms. There are also a number of potential issues with continuity of the data for each station due to changes in instrumentation, observing practices, and exposure. The primary purpose of the QC for this project was to identify the largest errors in individual values, particularly those that might affect analyses of extreme events.

Automated procedures were used to identify unusual values ("outliers"). Outliers were then examined by experienced, trained climatologists to assess their validity. A basic set of procedures was applied to data for all precipitation stations and for all temperature stations with at least 5 years of data. A more detailed set of procedures was applied to long-term stations, defined as those with less than 10% missing data for the period 1895-2000. These stations will be heavily utilized to study climate trends; thus a

*Corresponding author address: Kenneth E. Kunkel, Illinois State Water Survey, 2204 Griffith Drive, Champaign, IL 61820; e-mail: kkunkel@uiuc.edu

greater allocation of quality control resources was justified.

3.1 Basic Procedures

The basic procedures identified the most extreme values in the dataset using either absolute thresholds or thresholds based on the station's own climatology. For precipitation, any value in the database that exceeded 10 inches was flagged as an outlier. For maximum and minimum temperature, a daily value T_i was flagged as an outlier if its standardized anomaly from the monthly mean exceeded 5.0 in absolute value, i.e.

$$\left| \frac{T_i - T_m}{\sigma_m} \right| > 5.0 \qquad (1)$$

where m is the month, T_m is the monthly mean maximum or minimum temperature, and σ_m is the standard deviation of daily maximum or minimum values.

As noted above, the temperature tests were applied only to stations with at least five years of data while the precipitation test was applied to all stations regardless of their period of record.

3.2 Procedures Applied to Long-term Temperature Stations

The second set of procedures identified outliers by performing spatial comparisons using nearby stations, along with double-checks based on temporal continuity and extremes. Daily gridded fields of maximum and minimum temperature for the period of 1895-1948 were produced using the objective analysis scheme of Barnes (1964) as modified by Achtemeier (1987, 1989). For each station, each daily temperature value T_i was compared with an estimate E_i from the corresponding gridded field using a bi-linear interpolation from the four nearest grid points. A daily difference D_i was calculated as

$$D_i = (E_i - E_m) - (T_i - T_m) \qquad (2)$$

where E_m = the monthly mean of the gridded estimates. Next, 12 cumulative distribution functions, one for each month, were generated from the set of D_i values. An example is shown in Fig. 1 for the month of December for Grand Marais, Michigan. A daily value was considered an outlier if

$$D_i < D_{0.01} \qquad (3)$$

or

$$D_i > D_{0.99} \qquad (4)$$

where $D_{0.01}$ and $D_{0.99}$ are difference limits for fractional cumulative frequency values of 0.01 and 0.99, respectively. The application of this test results in flagging of 2% of all temperature values.

A subset of values identified as outliers by (3) and (4) were assessed manually and a "quality index", Q, was created to rank outliers in order of likely validity. For values exceeding $D_{0.99}$, this index was defined as:

$$Q_i = (D_{0.99} - M_D)/(D_i - M_D) \qquad (5)$$

where $M_D = 0.5(D_{0.99} + D_{0.01})$. A similar equation applies for $D_i < D_{0.01}$. Q_i values range from 0 to 1 with lower values representing more extreme outliers. All outliers with Q-rank less than 0.34 were validated.

For outliers with Q-rank 0.34 and higher, two double-checks were applied. One double-check was a temporal (spike) test, with the cutoff limits of 5% generated by each station's climatology. The other double-check was an extremes test, with the cutoff limits of 1% and 99% also generated by each station's climatology. If the outlier failed the extremes double-check, it's Q-rank was recalculated using the difference limits relaxed to 5% and 99%, which has the effect of decreasing the likely validity of the outlier.

3.3 Procedures Applied to Long-term Precipitation Stations

A similar methodology, using gridded estimates, was tested for daily precipitation. However, there were many valid precipitation values for which the calculated Q values were very low, thus requiring much unnecessary manual assessment. This was due to the high spatial variability of precipitation during convective situations. An alternate method was developed that proved to be superior at identifying invalid values. For each station, a set of nearest neighbor stations was identified based on geographical distance. All non-zero daily values were ranked from lowest to highest. Extreme values were defined as those exceeding the 95th percentile threshold and were subjected to further tests to identify outliers.

For each extreme value, P_i, two sets of Q values were calculated. The first set used actual precipitation amounts as follows:

$$Q_{amt}(i, n) = P_n / P_i \qquad (6)$$

Where Q_{amt} (i, n) is the Q-rank using precipitation "amounts" for day i and nearest neighbor station n and P_n is the precipitation amount for station n. The second set used percentile ranks as follows:

$$Q_{per}(i, n) = (100 - R_i) / (100 - R_n) \qquad (7)$$

Where Q_{per} (i, n) is the Q-rank using precipitation percentiles and R_n and R_i are the monthly percentile ranks for the nearest neighbor and validated stations, respectively.

The monthly percentiles were obtained by ranking all non-zero precipitation values for the month of day i.

The final Q-rank, Q_i, for P_i is the maximum individual value of the set of Q_{amt} and Q_{per} values. The key feature of the procedure is that a high Q-rank will be calculated if any single nearest neighbor station has a precipitation value that is seasonably high. Values with very low Q ranks only occur when no nearby station has a high precipitation value. Our tests indicated that this procedure was effective at identifying invalid values and maximizing use of personnel resources for manual assessment.

3.4. Manual Assessment

The manual assessment took a conservative approach. An outlier was assumed to be valid if there was any confirming evidence. Each outlier was assessed and assigned one of four flags described as follows:

"Valid"-there is some confirming evidence. Usually, this evidence consisted of values at one or more nearby stations that were also relatively extreme. Or, the observed spatial pattern of values is recognized as a usual one for the region and time of year.

"Plausible"-there may be no nearby stations with similar extreme values, but the assessor recognizes that such a pattern has occurred in the past with some regularity at the location and time of year.

"Questionable"-the assessor judges that the observed pattern is not a regularly occurring one and the value is unlikely to be valid, but cannot discount the physical possibility of the observed pattern.

"Invalid"-the assessor judges that the observed value is outside a physically possible range or that the observed spatial pattern is not likely to be physically possible.

4. Results

For the basic temperature test, a total of 4380 values were identified that met the eq. (1) criterion. The results of the manual assessment (Fig. 2) show a clear and expected relationship to the magnitude of the standardized anomaly. For standardized anomalies of greater than 7, more than 80% of the values were judged to be invalid. This percentage drops to about 20% for the 5.0-5.5 category.

For the basic precipitation test, a total of 498 precipitation values exceeded 10 inches and were manually assessed. The results of the manual assessment (Fig. 3) indicate that the percentage of invalid values increased with increasing amount, from about 20% in the 10-12 inch category to about 95% for values greater than 20 inches.

For the spatial tests applied to long-term temperature stations, a total of 6547 values with Q-values less than 0.35 were manually assessed. The results of the manual assessment (Fig. 4) indicate that the percentage of invalid values decreased with increasing Q-value, from 100% for Q < 0.10 to about 70% for the 0.30-0.35 category.

For the nearest neighbor tests applied to long-term precipitation stations, a total of 7421 values with Q-values less than 0.50 were manually assessed. The results of the manual assessment (Fig. 5) indicate that the percentage of invalid values decreased with increasing Q-value, from roughly 40% to less than 10% at a Q-rank of 0.50.

5. Conclusions

The newly keyed pre-1948 data represents a major enhancement to the COOP data set, which is widely used for analysis of climate variability and change. The quality control applied in this project increases its value by eliminating flagrant errors in individual values. Approximately 6000 values were flagged as invalid.

This effort was limited by available personnel resources. The results for temperature outliers from the spatial tests, summarized in Fig. 4, indicate that further manual assessment would likely result in a substantial number of additional invalid values.

The advanced objective quality control procedures developed here were found to be quite effective at identifying outliers that were likely to be invalid, but the manual assessment was critical to avoid removing valid values from the dataset.

6. References

Achtemeier, G. L., 1987: On the concept of varying influence radii for a successive corrections objective analysis. *Mon. Wea. Rev.*, 115, 1760–1772.

Achtemeier, G. L., 1989: Modification of a successive corrections objective analysis for improved derivative calculations. *Mon. Wea. Rev.*, 117, 78–86.

Barnes, S. L., 1964: A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteorol.*, 3, 396–409.

Climate Database Modernization Program (CDMP), Annual Report, available from the National Climatic Data Center, Asheville, NC, 8 pp, 2001.

Kunkel, K.E., K. Andsager, G. Conner, W.L. Decker, H.J. Hilaker Jr., P. Naber Knox, F.V. Nurnberger, J.C. Rogers, K. Scheeringa, W.M. Wendland, J. Zandlo, and J.R. Angel, 1998: An expanded digital daily database for climatic resources applications in the Midwestern United States. *Bull. Amer. Meteor. Soc.,* 79, 1357-1366.

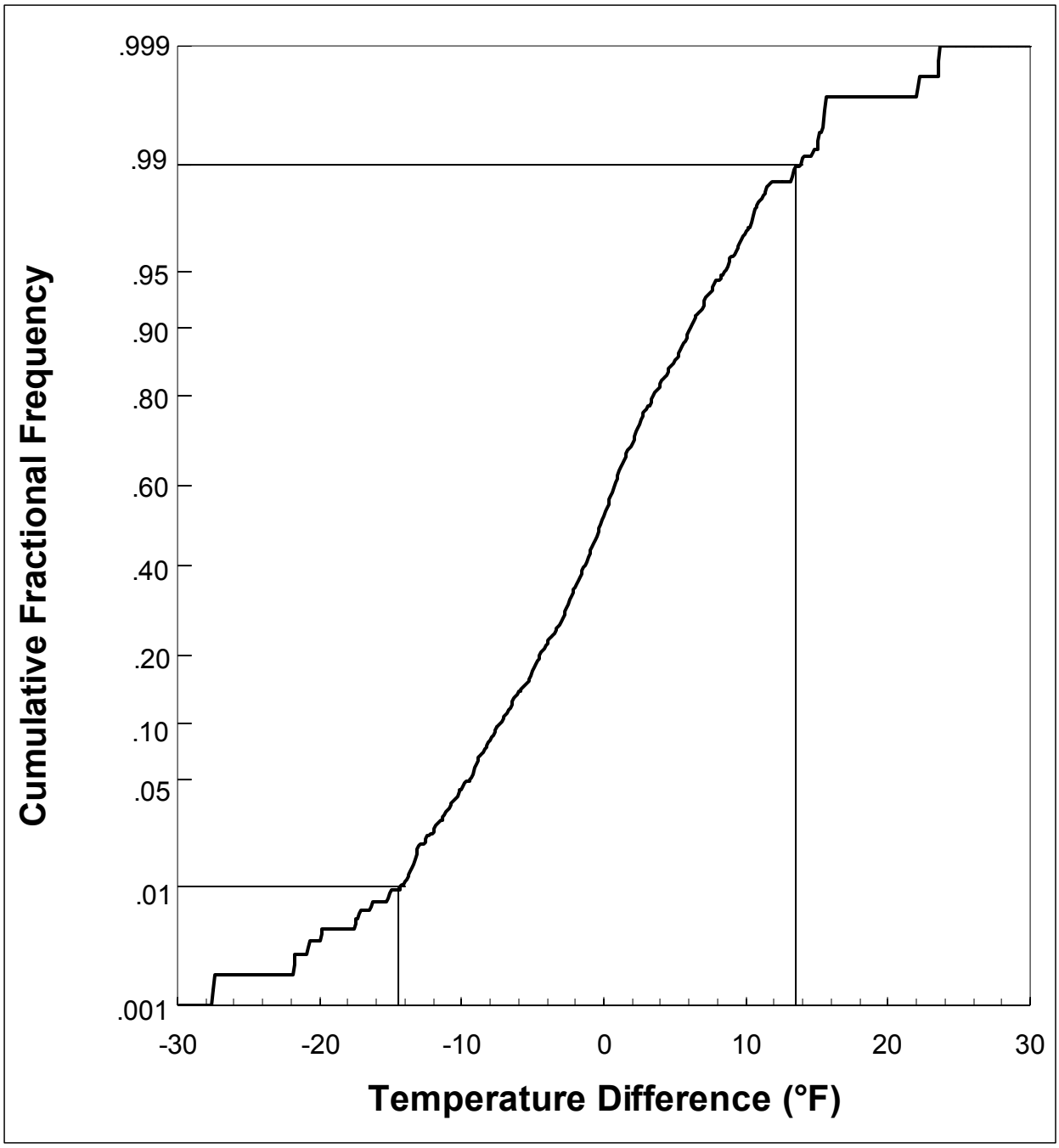

Figure 1. Cumulative frequency (expressed as a fraction) as a function of difference between station temperature anomalies and estimated temperature anomalies from the gridded data for Grand Marais, Michigan in December. The light lines show the temperature difference values at cumulative frequencies of 0.01 and 0.99.

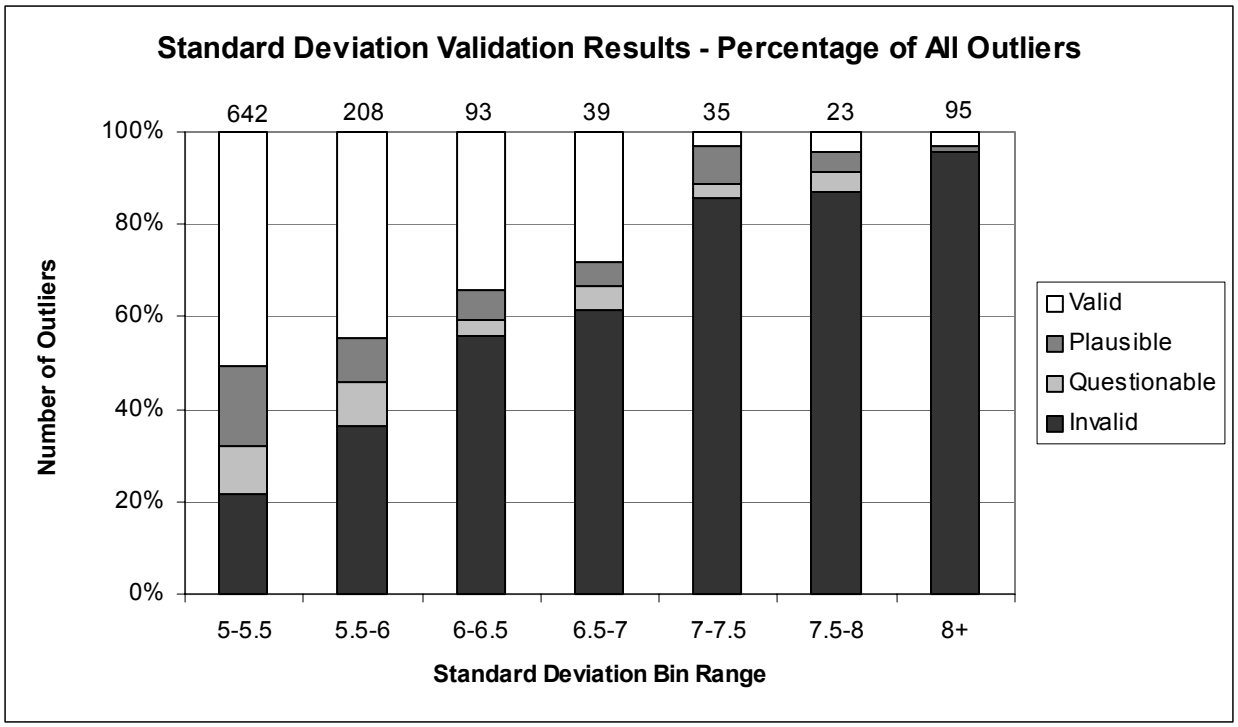

Figure 2. The percentage of manual outlier assessments in each category (valid, plausible, questionable, and invalid) as a function of the outlier standard deviation for temperature outliers. The total number of outliers in each standard deviation bin is shown at the top of the bar.

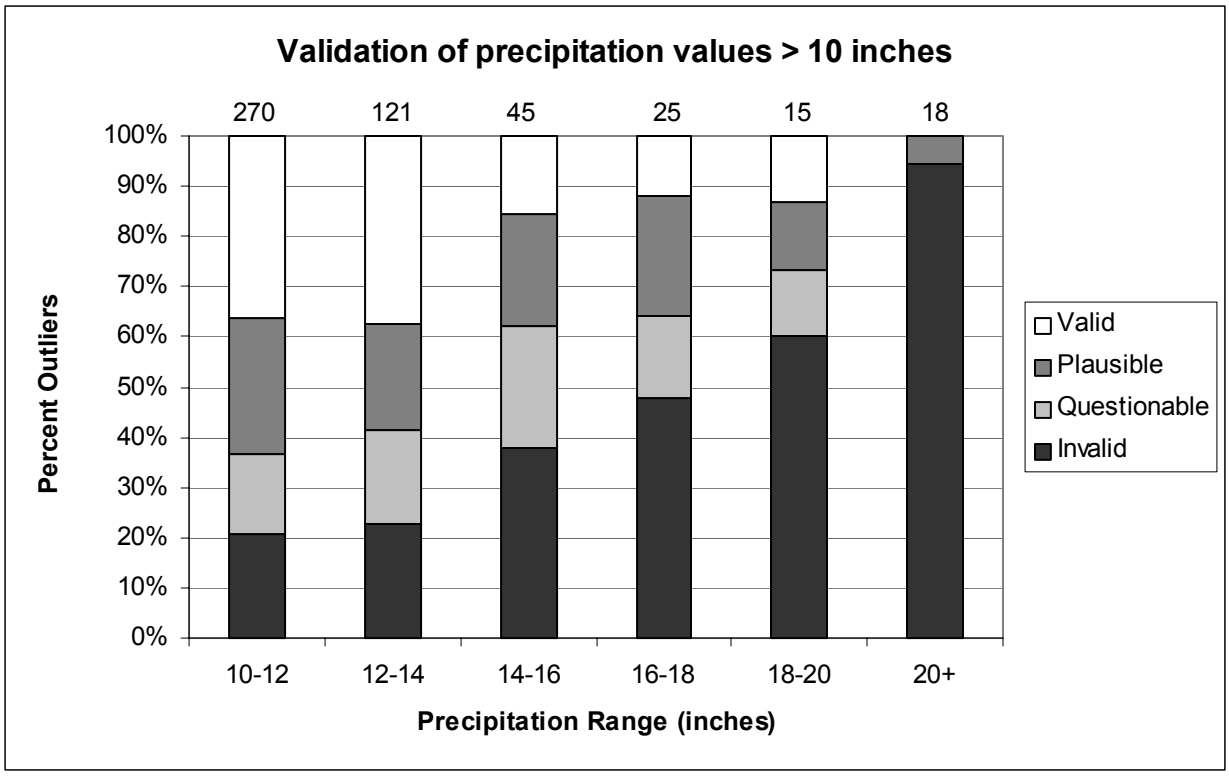

Figure 3. The percentage of manual outlier assessments in each category (valid, plausible, questionable, and invalid) as a function of amount for precipitation outliers. The total number of outliers in each precipitation bin is shown at the top of the bar.

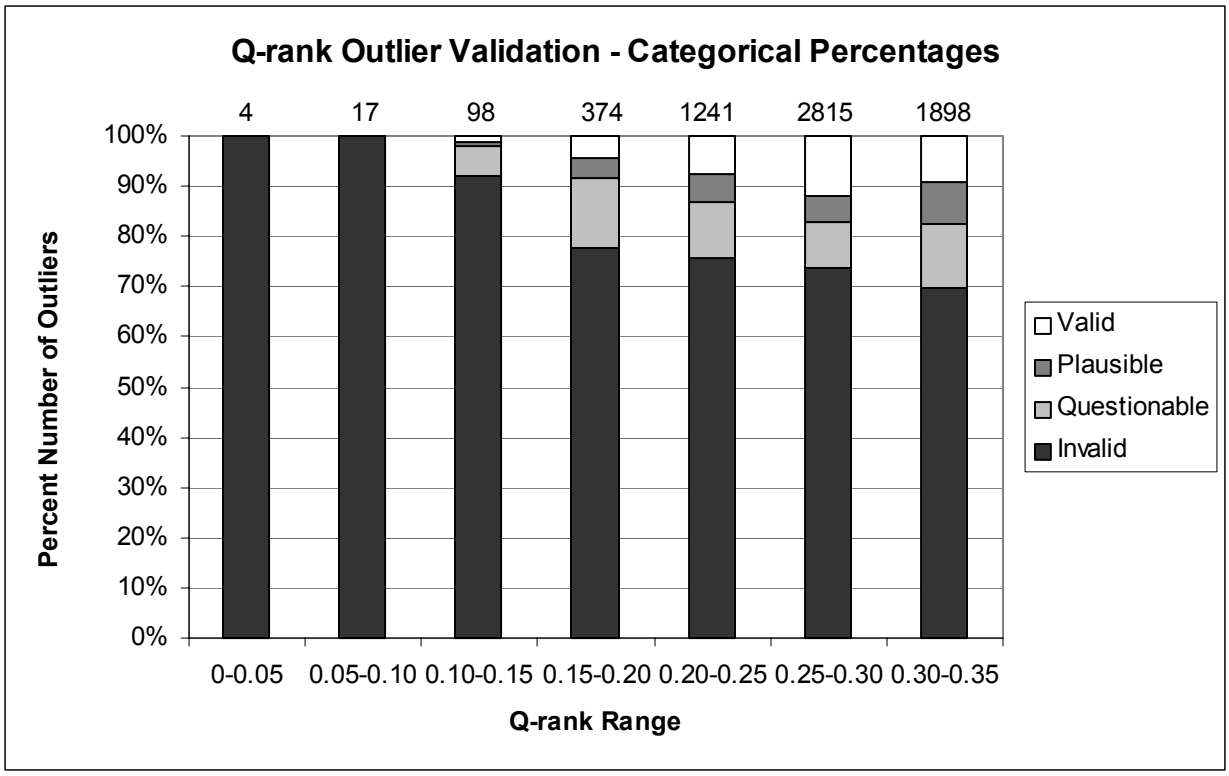

Figure 4. The percentage of manual outlier assessments in each category (valid, plausible, questionable, and invalid) as a function of the Q-value for temperature outliers. The total number of temperature outliers in each Q-value bin is shown at the top of the bar

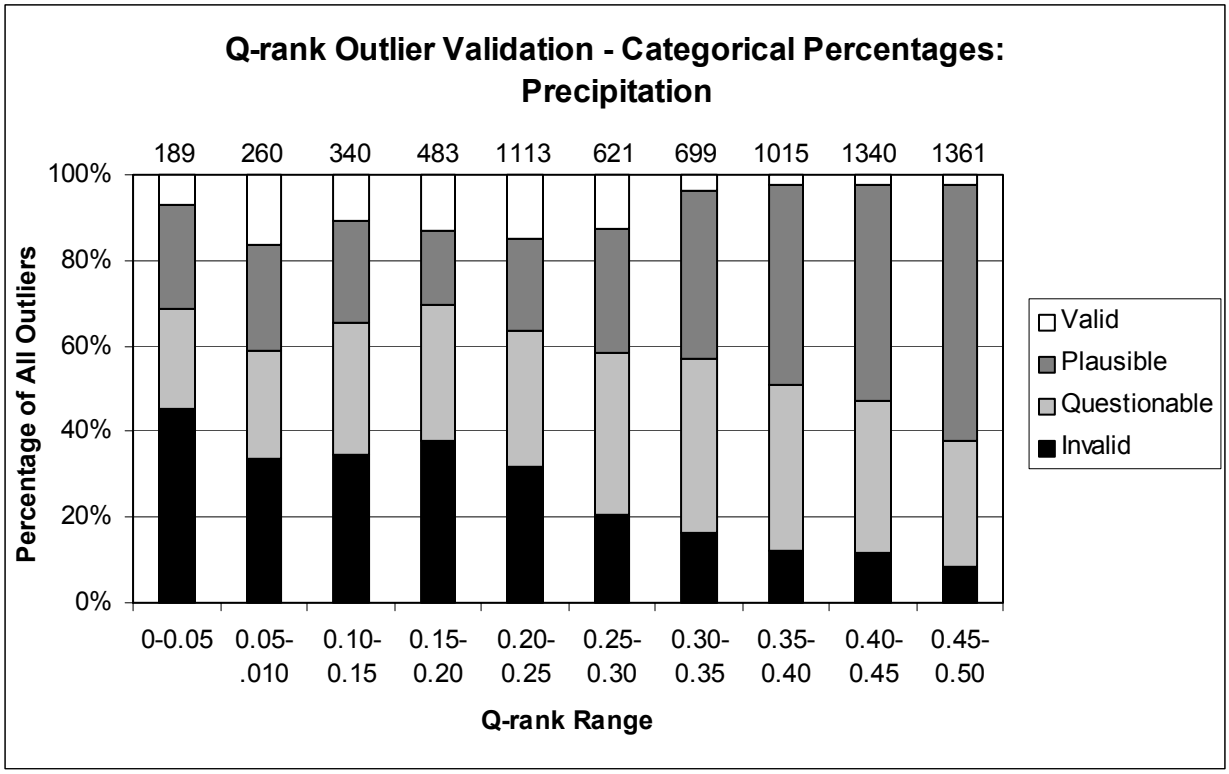

Figure 5. The percentage of manual outlier assessments in each category (valid, plausible, questionable, and invalid) as a function of the Q-rank for precipitation outliers from the long-term stations. The total number of outliers in each Q-rank bin is shown at the top of the bar.