18.13 AUTOMATED VALIDATION FOR SUMMARY OF THE DAY PRECIPITATION DATA

Michael L. Urzen *, and Stephen A. Del Greco
NOAA National Climatic Data Center, Asheville, North Carolina

Steve Ansari
STG Inc, Asheville, North Carolina

1. INTRODUCTION

In recent years, the value of weather and climate observations has become ever more broadly recognized. Given that weather observations are susceptible to a number of potential data errors, their value is partially dependent on the degree to which errors can be identified and appropriately addressed. While many data problems are characteristic of an observing system or instrument type, no system is immune to at least occasional instrument malfunction or human error. At NOAA's National Climatic Data Center (NCDC), which is responsible for the acquisition, quality assurance and archival of U.S. weather and climate information, a variety of quality assurance techniques have been used to identify *in situ* temperature and precipitation observation errors during data ingest and review. Error detection approaches have ranged from subjective review by human "validators" to more objective and automated quality assurance algorithms.

From the standpoint of efficiency and reproducibility, fully automated and objective data reviews clearly are desirable. However, the establishment of threshold checks for data validation that work well under most circumstances is a constant challenge. Frequently, subjective interventions are necessary in automated data review procedures. Many quality assurance systems, including those in place at NCDC for NWS Cooperative Observer (Coop) Network observations, are at best nearly automated. Whatever the approach used, it is important to note that observations are not considered to be "error-free" when they pass the scrutiny of a review system, but rather to be reasonable or plausible as defined within the framework of the testing algorithms or review by data experts.

* *Corresponding author address: M*ichael L. Urzen, National Climatic Data Center, Asheville, NC 28801; e-mail: Michael.Urzen@noaa.gov.

For convenience, observational errors are often classified into two basic types: random and systematic. Some random errors may be identified by logical checks designed to expose internal consistency problems. In the Coop Network, internal inconsistencies include, for example, cases where a minimum temperature value for a given day is listed as greater than the corresponding maximum temperature or when frozen precipitation is recorded at temperatures well in excess of freezing. Random errors that pass logical checks may be identified by comparing a target observation to values from nearby stations. In this case, attempts are made to establish the plausibility of an observation given the characteristics of a local field of values from surrounding locations. This type of validation rests on the assumption that the correlation between observations is large at least within the distances defined by the spacing between stations. While this assumption may be generally true in temperature fields, in the case of daily precipitation totals, the expected nature of the field of observations can be difficult to establish under many (e.g., convective) circumstances. Consequently, the use of station-to-station difference thresholds to determine the limits of plausibility is problematic. In addition, daily precipitation totals in the Coop Network are recorded by volunteers according to a variety of observation schedules. Precipitation events in an area may therefore be differentially apportioned across more than one calendar day at various neighboring stations as an artifact of differences in the gauge reading schedule. The result is an additional dimension of complexity to the local precipitation field that further complicates data validation.

A potential way to overcome some of the problems associated with the identification of random errors in precipitation totals is to compare gauge totals to independent assessments. PrecipVal (*Precip*itation *Val*idation) is the first systematic attempt at NCDC to compare in situ precipitation totals to independent or quasi-independent estimates for the purpose of data

validation. Here we describe a system that can be used to evaluate the potential of this approach to data validation.

2. PRECIPITATION ESTIMATES

The PrecipVal system currently has the capability of comparing daily precipitation totals from the Coop Network to three different precipitation estimates. These include other (non Coop) in situ measurements, radar and gauge blended estimates, and a merged estimate from multiple sources. The in situ estimate is obtained by inverse distance-squared weighted averaging of observations from the nearest three non-Coop stations within 25 mi. (40 km) of the target Coop site. The radar/gauge estimate is the National Centers for Environmental Prediction's (NCEP) Stage IV product, which is a blend of radar data and sub-daily in situ measurements. The third value is derived by averaging multiple data sources that include satellite, radar, forecast model and gauge-based data (Fig. 1). More specifically, these sources consist of six satellite-based estimates, the WSR-88D Level III radar Digital Precipitation Array, Stage IV values, estimates from the Rapid Update Cycle (RUC) forecast model and gauge totals from other observation networks. Fig. 2 provides examples of daily precipitation estimates from two of these sources. The non-gauge-based data sources used in the merged estimate are screened for gross errors using various inter-comparison tests. At any particular location and time, the screening eliminates those sources considered to be of limited utility. The remaining values are used in the calculation of an average and various other summary statistics at each nominal 4.7 kilometer grid cell.

3. PRECIPITATION VALIDATION

The PrecipVal system was established to carry out a series of comparison tests between Coop precipitation totals and the various estimates. The first test attempts to identify multi-day accumulations of precipitation that can occur when the observer fails to read the gauge each observational day. When a gap is encountered in the Coop data, the daily values of each of the three estimates are accumulated throughout the period of missing Coop data to create three alternative accumulated precipitation estimates for the day on which Coop observations resume. If the first non-missing Coop value following the data gap more closely matches one of the three accumulated values than the corresponding daily estimate, the Coop value is considered to be a likely accumulation. The Coop accumulated value is then distributed over the missing period based on percentages derived from the daily estimates whose accumulation best matches the Coop value.

The second comparison is designed to identify cases where the precipitation total is attributed to the previous calendar day rather than to the day on which the gauge was read. In an effort to detect the error, a lagged comparison between the Coop total and a selected estimate is performed. The Coop values for the current day and the following day are compared to the estimate shifted forward by one day. If the lagged values agree better than the concurrent values and the Coop values are increasing as estimated values are decreasing, or vice versa, then the Coop value on the current day is flagged as a shifter. Conversely, the "reverse shifter" check identifies cases in which the precipitation total of a particular day is mistakenly reported for the subsequent day. The estimate used is based on availability.

When either the Coop value or the estimate closest in magnitude to the observed value exceeds one inch, then the observation is further tested for a shift in the decimal point. Such decimal point displacement is deemed to be present when multiplication of the observation by a factor of 0.01, 0.1, 10, or 100 results in the closer match with the estimate.

Following these three checks, all observations except those identified as accumulations are subjected to a same day comparison test in which each Coop observation is compared to the estimate closest in magnitude. An observation is considered valid if the estimate falls within 50% or within 0.25 in. (6 mm) of the observed amount. If at least two of the estimates are available, and none falls within the validation threshold, the observation is considered "suspect". When only one estimate is present, and this estimate falls outside of the validation range, the observed value is neither validated nor flagged as suspect. When zero precipitation is reported at the Coop station and the estimate is greater than zero, but less than or equal to 0.25 in. (6 mm), it is noted that precipitation may have been present. In the case in which an observation is missing, estimates are available.

4. CONCLUSION

In addition to random error, systematic errors should be considered when comparing observed rainfall totals to estimates. Nevertheless, PrecipVal provides a convenient mechanism for testing the feasibility of validating precipitation gauge totals with independent data sources. The statistics produced by the approach when applied to several months of Coop observations serve as a starting point for evaluating its robustness, the value of each individual data source, and the utility of using estimates generated by spatial interpolation of independent station observations. A principal benefit of using independent data sources is that they provide hourly estimates and thus aid in establishing the timing of precipitation events. Further, the system lends itself to the addition of other sources as they become available.

5. ACKNOWLEDGEMENTS

The authors thank Imke Durre and Matthew Menne of NCDC for guidance and review of this paper.

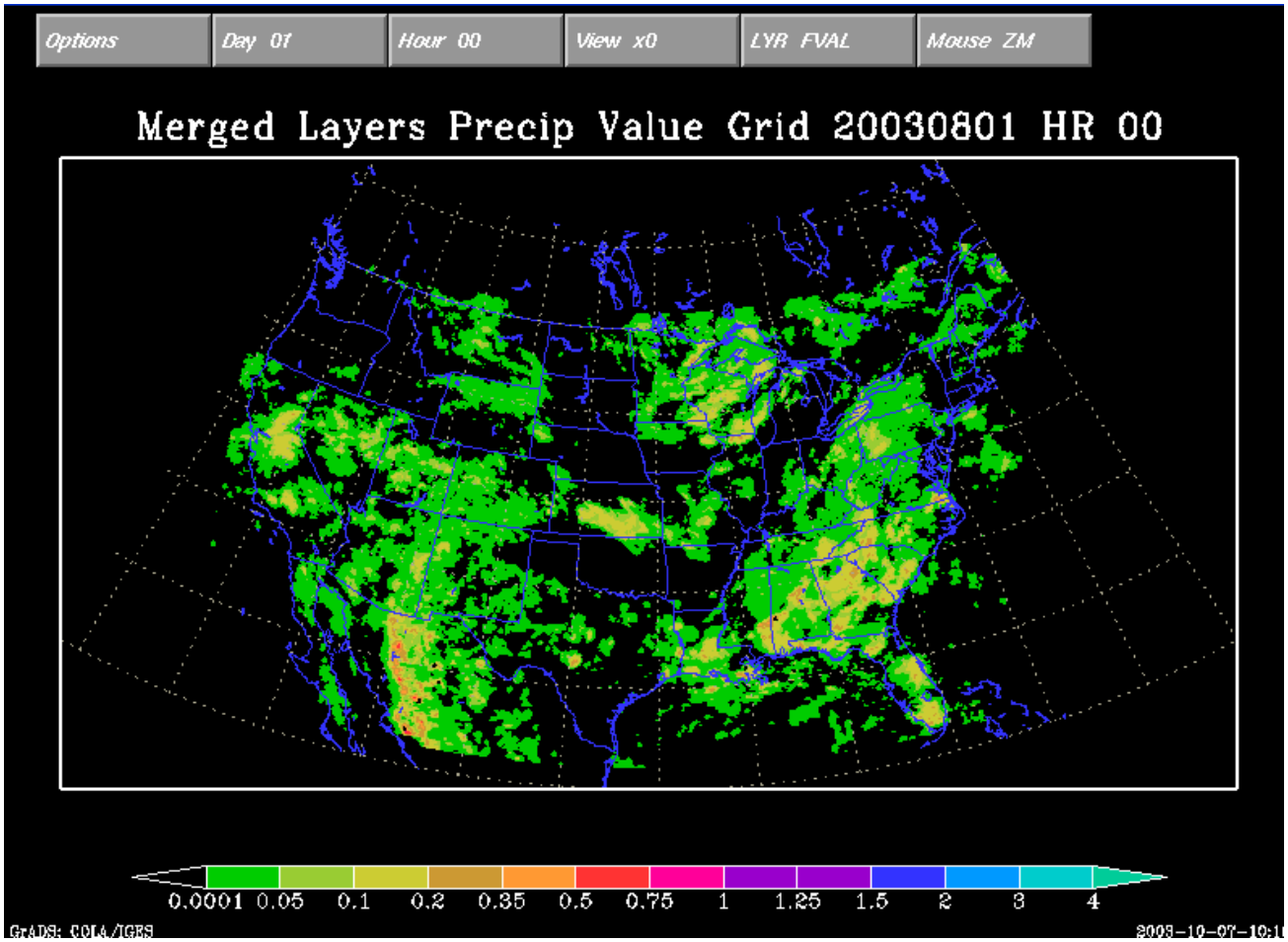

Figure 1. Example of multiple data source precipitation estimates.

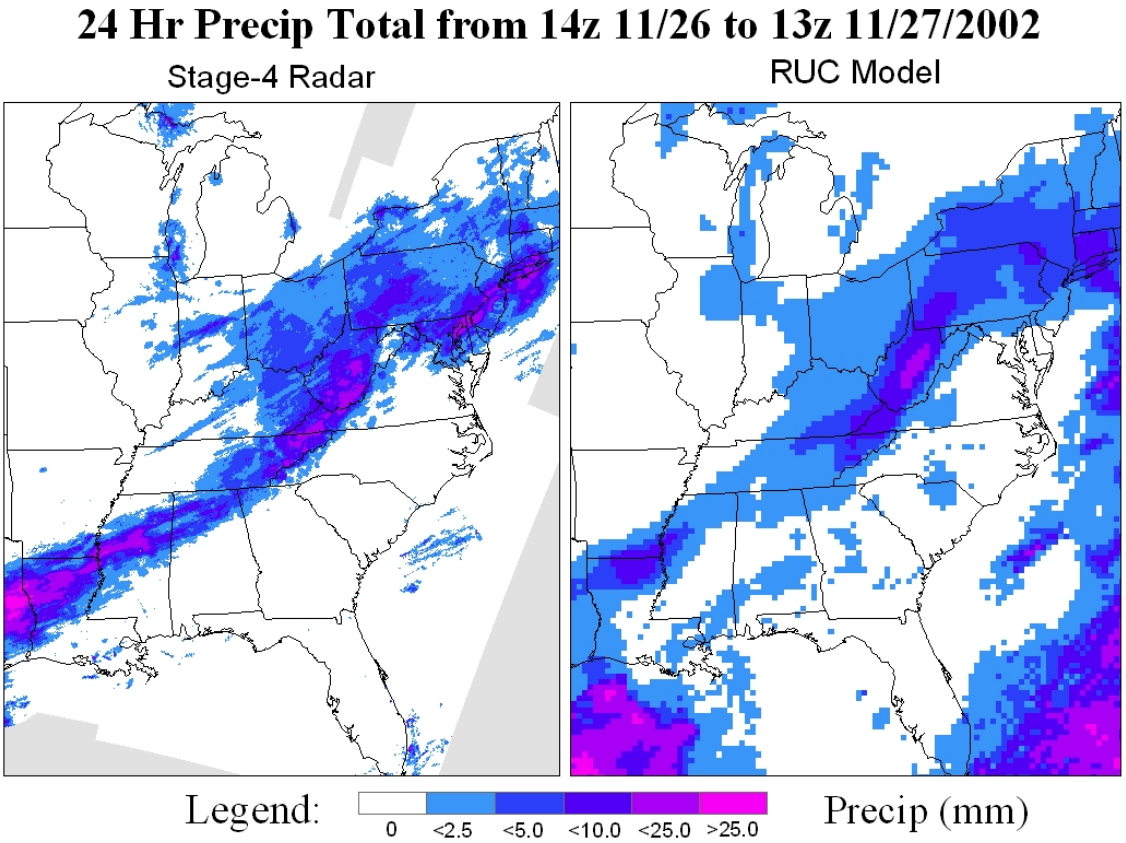

Figure 2. Comparison of daily precipitation estimates from two data sources.