P1.42

ON THE PERFORMANCE, IMPACT, AND LIABILITIES OF AUTOMATED PRECIPITATION GAUGE SCREENING ALGORITHMS

Edward I. Tollerud[1], Randall S. Collander[2], Ying Lin[3], and Andrew Loughe[4]

[1]NOAA Research - Forecast Systems Laboratory, Boulder, CO
[2]Cooperative Institute for Research in the Atmosphere (CIRA), Fort Collins, CO
[3]National Centers for Environmental Prediction, Washington, D.C.
[4]Cooperative Institute for Research in the Environmental Sciences, Boulder, CO

1. Introduction

Automated quality control (QC) procedures applied to real-time datastreams have the distinct advantages of reducing tedious and repetitious work, generalizing procedures across regions and personnel, and reducing many types of human error (transcription, for instance). When the datasets themselves are heavily automated and subject to inaccuracy, and the observations are difficult and wildly variable (for instance, precipitation measurements), automation of QC presents difficult problems. The bottom line is that regardless of how thorough and calibrated the automated procedures are, counter examples clear to a knowledgeable eye are almost certain to arise. A good example is the hourly gauge precipitation observations from the Hydrometeorological Automated Data System (HADS). To facilitate timely use as initialization data in numerical models, we have developed a system of algorithms written in Perl script specifically designed for the HADS and for the operational stations that are part of ASOS.

These data and the results of their QC provide a good opportunity to assess the impacts of such screening. In this paper, we describe factors that affect the performance of the QC procedures using CONUS rainfall data and diagnostic output designed into the system itself. The results from the individual algorithms of the QC system are first illustrated with recent examples. To assess and calibrate in a more general sense the overall impact of gauge screening procedures quantitatively, we also use a performance algorithm based on common verification scoring applied between stations and sets of their neighbors (usually these scores are applied between verification data and model results). From a user's perspective, interest in QC is primarily dependent on its impact on a particular application. Thus, we present diagnostic descriptions of the impacts on two fundamental gauge applications: analysis of precipitation observations to grids (specifically, the Stage II products of NCEP; (http://www.emc.ncep.noaa.gov/mmb/ylin/pcpanl/stage2/), and real-time verification of model predictions from the FSL-based Real-Time Verification System (RTVS) (Loughe et al. 2001; Mahoney et al. 2002).

2. The QC System

Three general types of checks are employed within the QC system. First, a set of standard screening algorithms for known problems are applied. In part, these checks include extreme values for 24-h totals and individual hours, nonphysical repeating patterns of hourly observations, and stuck-on gauges. Each of these checks involves the establishment of threshold values, which change periodically as experience with the system is improved and as new and different observations are introduced. Concerning extreme value thresholds, though they are currently set fairly large, on rare occasions they still will screen an accurate but very large observation. The tradeoff is the system's ability to find a larger number of inaccurate medium-range extreme values.

The second set of checks involves the neighbor checks with a set of nearby stations. Given the possibility of clustering of "reject" HADS sites around a target site, we have found it preferable to perform these checks using a high-quality set of daily total (1200–1200 UTC) observations that have

----

*Corresponding author address: Edward Tollerud, FSL/NOAA FS1, 325 Broadway, Boulder, Colorado, 80305.
email: edward.tollerud@noaa.gov

been processed (in many cases, qualitatively evaluated) by individual River Forecast Centers (RFCs) and routed through the National Centers for Environmental Prediction (NCEP), our principal source for station precipitation observations. These data in many cases include 24-h HADS totals computed by the RFCs and implicitly designated as accurate by inclusion in their daily observation datastream. At present each HADS target site is compared to 8 neighbors within a threshold distance from the target. To further ensure a good neighbor set, the verification sites from the daily total datastream are themselves first compared with neighboring daily sites. This use of a qualitatively screened RFC datastream introduces a nonautomated but important human element into the QC system.

The third set of checks are assessments of QC performance during a brief (at present, 30 days) historical period just prior to the present observation time. To evaluate this performance, the results of daily screening from each of the several algorithms at each station are imbedded in a long character string that is saved for an extended period. These station-specific character strings also provide critical information for retrospective system evaluation, tuning and revision.

Several assumptions (“observational philosophies”, if you will) are implicit in the design of this system. First, our experience shows us that automated gauges such as the HADS are susceptible to errors that tend to repeat themselves from hour to hour and from day to day. Hence, a “proving period” is built into the system during which time a formerly rejected but improved gauge can be reinstated (moved off the reject list). A second assumption is that there are inherent limits on the periods of accumulations that can be addressed by QC algorithms. For instance, neighbor comparisons (now done on 24 h accumulations) perform poorly at hourly intervals because of the extreme temporal and spatial variability of hourly precipitation, especially during warm-season convective regimes. This means that QC cannot be performed instantaneously except for certain gross error checks (extreme values, for instance). We assume that an observation site screened during the daily QC cycle should be excluded from use for the entire 24 h. An area of planned research is assessment of the feasibility of 6h QC increments.

Details of thresholds, station history character string content, and output diagnostic files for the QC system can be found in the readme file at http://www-frd.fsl.noaa.gov/mab/sdb/diagnostic.cgi.
Note that a critical part of the system are daily updates to authoritative metadata sites.

3. Examples of System Performance

Figure 1 shows three examples of the kinds of screening described above in the context of typical HADS observation fields. In Fig. 1a, the observations at a site near Grand Junction in western Colorado would fail the neighbor check because it shows nonzero rainfall in a general field of zero observations. As the time series of hourly reports reveals, this site should also fail the check for recurrent nonphysical temporal patterns of hourly reports. In all likelihood, many of the examples of this kind of faulty reporting are due to telemetry or other kinds of communications problems. In Fig. 1b, the time series reveals a likely case of a stuck gauge. The neighborhood check for this day would also screen this station, but it is quite possible that such a report would escape neighbor screening if it occurred on a day of regional rainfall in eastern Montana. The observation at Nampa, Idaho, in Fig. 1c is clearly faulty because of its extreme size, and also because of nonmatching neighbors. This station is in fact an example of a recurrent instrument problem.

The chart in Fig. 2 provides a general review of the overall effect of the QC system. It lists counts of HADS stations in several categories of the QC system output produced during a short evaluation of the QC installation at FSL. Something like 10–15% of the HADS reports is screened by one or another of the QC algorithms on a typical day (it is possible for an individual report to be screened by several of the algorithms). Day-to-day variability is largely a result of precipitation reports; days with considerable rainfall in regions with denser station distributions will have a larger count. Of the screening types, neighbor checks have the largest effect. Stations thus screened are also predominantly clustered, located in regions of rainfall (see Fig. 3 for an example from another day). Days on which the neighbor screening results in “no rejected stations" are in fact days during which the datastream of daily stations was late. It might be preferable on those days to perform neighborhood screening based on the previous one day or more set of accepted HADS sites. Although small in number, the stations screened because of extreme values are potential sources of significant error in such applications as analysis to grids and computation of verification scores such as magnitude bias. It is encouraging that the quantity of stuck gauges has been improving during the course of several years and, at least during the period exhibited, is small.

4. Effects on Precipitation Analysis

The NCEP Stage II is a real-time, hourly, multi-sensor (radar+gauges) national precipitation analysis produced at NCEP and archived at NCAR (Lin and Mitchell 2005). Currently, the Stage II program uses a "gauge reject list" to screen out gauges that have previously been flagged as problematic. This gauge reject list is updated manually. Unfortunately, this approach has several limitations, principally because it is very labor intensive and there is no way to reinstate a rejected gauge. For this reason, updates to the gauge reject list have been infrequent because of the amount of work involved, and because of a concern about the ultimate thinning of the gauge population. As of June 2005, only 80 gauges had been placed on the reject list. For these reasons, the automated scheme described here is being considered to replace the existing scheme in the analyses procedure.

This begs the question: what is the impact of the new automated scheme on the actual analyses? Figure 4 shows an example of hourly precipitation analysis performed with the previous procedure (b) and for comparison a similar analysis that employs the newly quality-controlled stations (a). The difference between the two (a-b) is shown in c. Overall, the differences between them are small. For example, in eastern Kansas, the test analysis (new QC) has higher values than the operational analysis, a difference that appears to be caused by the excluded gauges (mostly by failing the hourly/daily neighbor checks) reporting lower values of precipitation amounts. The difference in South Dakota is an anomaly – the gauge at Pierre, SD, (PIES2) is listed in the operational HADS reject list, and has been reporting ~0.2 in per hour of rainfall on and off, for roughly half of each day, day after day. The fact that this (clearly faulty) station eludes the QC in the new system reveals a failure of its present algorithms.

5. Effects on Precipitation Verification

Another frequent use for precipitation reports is as verification data for numerical models that predict precipitation. Although verification data are commonly considered as "truth," it is of course true that there is a range of uncertainty involved with them as well as with the model fields themselves. One source of this data variability is the quality of the data, hence the interest in QC issues in the verification community.

A first step in assessing the QC impact on verification is to determine the magnitude of the difference between verification with fields that have the advantage of thorough QC procedures and those that do not. We make this comparison in Figs. 5 and 6. In Fig. 5, we show a month of daily verification of the Eta model over the CONUS using two sources for truth fields: the HADS observations from this period with the advantage of only minimal QC as a red curve, and the higher-quality daily reports (the same as those we use as neighbors in the HADS screening algorithms) as a blue curve. Clearly, substantially better scores result from use of the better verification data. On rare days, however, the reverse is true (check April 6). We surmise that the geographical distribution of rainfall vis-à-vis the density variations in the distribution of stations in either set is the probable source of this anomalous result, emphasizing the impact of other data problems (data representativeness, for instance) The days with zero values (5, 11, and 20 April) are days when either the datastreams to FSL were delayed or processing hardware failed; these scores should be discounted.

The differences displayed in Fig. 5 result from extremities of data quality. In Fig. 7, on the other hand, a more useful comparison is shown. Both panels display equitable threat scores over a several-week period computed using the HADS dataset, with QC applied in the upper panel but withheld in the lower. As previously shown, improved scores result from the use of quality data; scores at all categories improve from 10% to 15%. Clearly, if small differences in verification scores are to be used to evaluate model performance, the magnitude of the differences due solely to verification data must also be considered.

Large day-to-day differences in scores are evident in Fig. 5, which also complicate its interpretation. To determine the extent to which these differences are driven by the distribution of verification stations ("representativeness" error) instead of reflecting different performance of the model, we present similar scores computed not from model fields but from observations at each station's nearest quality neighbor observing site. In some sense, these scores represent performance that is "as good as it gets." The two resulting curves shown in Fig. 6 clearly track each other, suggesting that the difference between these scores and the Eta model scores might actually be a better way to track model performance over time.

6. Future Plans

During examination of the QC results, several desirable revisions and improvements became evident. A more refined neighbor checking mechanism is necessary because our first attempt screens too many nonzero reports in situations of scattered rain reports. We are also considering ways to replace the linear yes-no nature of the set of algorithms with fuzzy logic decision tools, and ways to employ satellite or radar data directly when a station reporting zero rainfall is of ambiguous accuracy. Most importantly, we hope to determine the feasibility of reducing the analysis time increment from 1 day at present to 6 h.

7. Acknowledgments

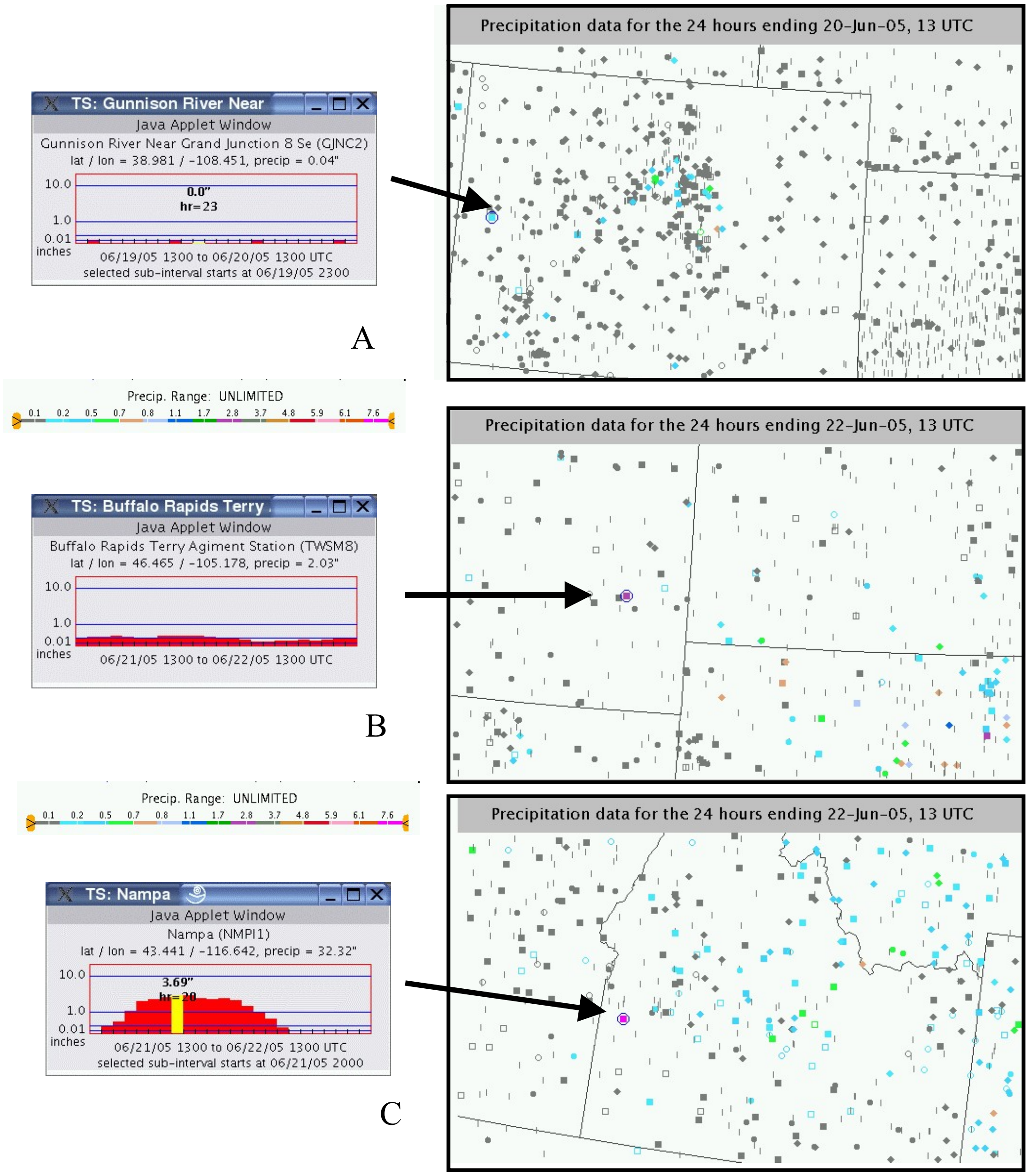
We thank Chungu Lu and Nita Fullerton for their reviews of this paper. Some of the work described here was supported by a grant from the United States Weather Research Program.

References

Lin, Y. and K. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: development and applications. Preprint, *AMS 19th Conference on Hydrology*, 9-13 January, San Diego, CA.

Loughe, A.F., J.K. Henderson, J.L. Mahoney, E.I. Tollerud, 2001: A verification approach suitable for assessing the quality of model-based precipitation forecasts during extreme precipitation events. Preprint, *Symposium on Precipitation Extremes: Prediction, Impacts, and Responses,* 14-19 January, Albuquerque, NM.

Mahoney, Jennifer Luppens, J. K. Henderson, B. G. Brown, J. E. Hart, A. Loughe, C. Fischer, and B. Sigren, 2002: The real-time verification system (RTVS) and its application to aviation weather forecasts. Preprints, *10th Conference on Aviation, Range, and Aerospace Meteorology,* 13-16 May, Portland, OR.



TS: Gunnison River Near

Gunnison River Near Grand Junction 8 Se (GJNC2)
lat / lon = 38.981 / -108.451, precip = 0.04"

06/19/05 1300 to 06/20/05 1300 UTC
selected sub-interval starts at 06/19/05 2300

Precipitation data for the 24 hours ending 20-Jun-05, 13 UTC

Precip. Range: UNLIMITED

TS: Buffalo Rapids Terry

Buffalo Rapids Terry Agiment Station (TWSM8)
lat / lon = 46.465 / -105.178, precip = 2.03"

06/21/05 1300 to 06/22/05 1300 UTC

Precipitation data for the 24 hours ending 22-Jun-05, 13 UTC

Precip. Range: UNLIMITED

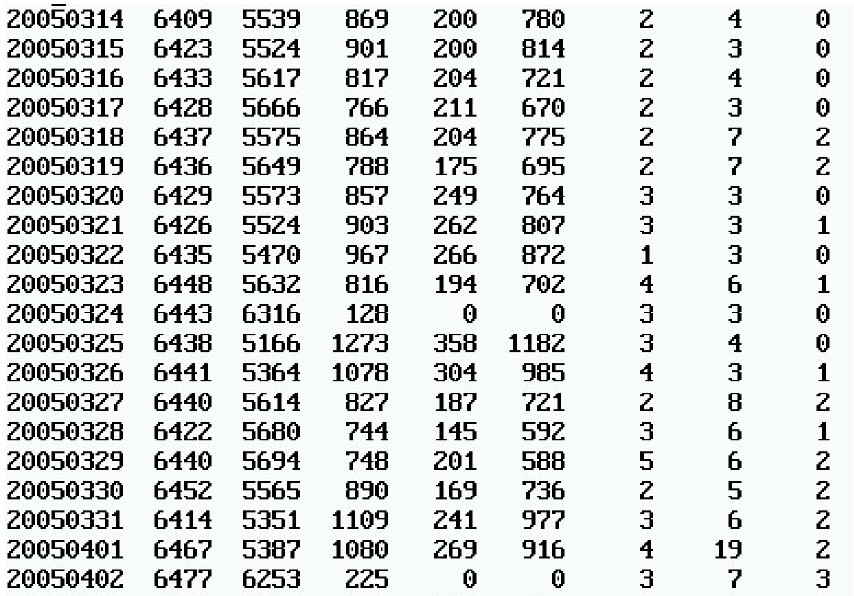
TS: Nampa

Nampa (NMPI1)
lat / lon = 43.441 / -116.642, precip = 32.32"

06/21/05 1300 to 06/22/05 1300 UTC
selected sub-interval starts at 06/21/05 2000

Precipitation data for the 24 hours ending 22-Jun-05, 13 UTC

Fig. 1. Time series of hourly precipitation and geographical plots of precipitation gagesites for three anomalous gage observations. See text for explanations. Note that color scale is inches.



(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(I)
20050314	6409	5539	869	200	780	2	4	0
20050315	6423	5524	901	200	814	2	3	0
20050316	6433	5617	817	204	721	2	4	0
20050317	6428	5666	766	211	670	2	3	0
20050318	6437	5575	864	204	775	2	7	2
20050319	6436	5649	788	175	695	2	7	2
20050320	6429	5573	857	249	764	3	3	0
20050321	6426	5524	903	262	807	3	3	1
20050322	6435	5470	967	266	872	1	3	0
20050323	6448	5632	816	194	702	4	6	1
20050324	6443	6316	128	0	0	3	3	0
20050325	6438	5166	1273	358	1182	3	4	0
20050326	6441	5364	1078	304	985	4	3	1
20050327	6440	5614	827	187	721	2	8	2
20050328	6422	5680	744	145	592	3	6	1
20050329	6440	5694	748	201	588	5	6	2
20050330	6452	5565	890	169	736	2	5	2
20050331	6414	5351	1109	241	977	3	6	2
20050401	6467	5387	1080	269	916	4	19	2
20050402	6477	6253	225	0	0	3	7	3

Fig. 2. Daily HADS QC statistics for the days indicated. Columns are: (a) dates; (b) number of HADS observations; (c) observations that passed the QC; (d) hourly HADS observations screened by the QC algorithms; (e) daily observations screened by neighbor checks; (f) HADS observations screened by neighbor checks; (g) observations screened by stuck gage algorithm; (h) observations screened by extreme daily total check; and (I) observations screened by extreme hourly total(s) check.

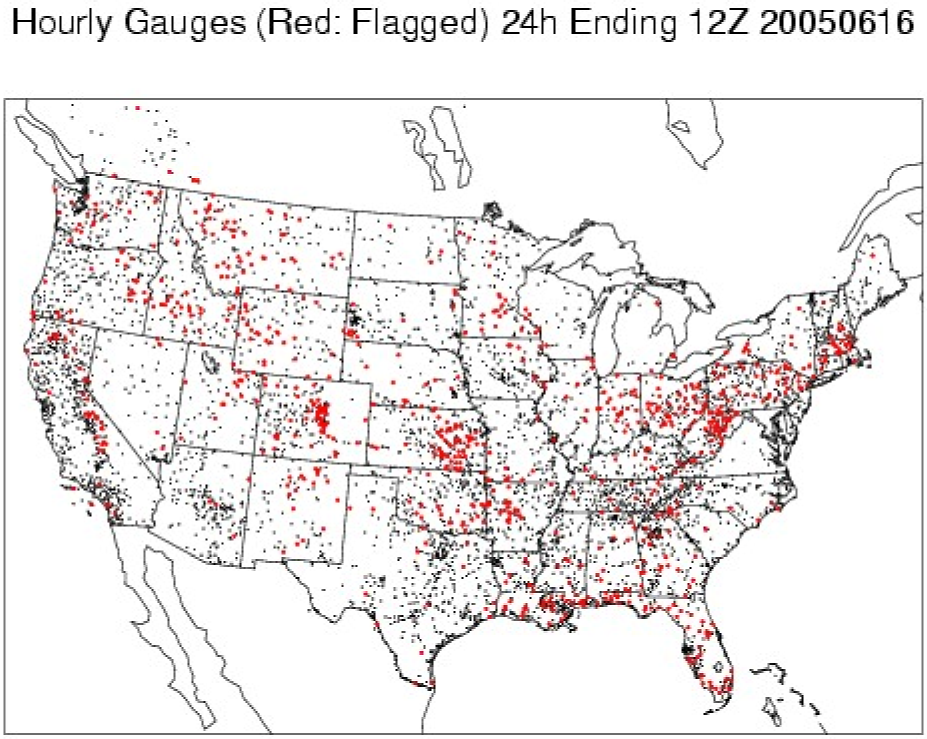


Fig. 3. Good (black) and bad (red) hourly gages determined by the automated QC system for 16 June 2005.

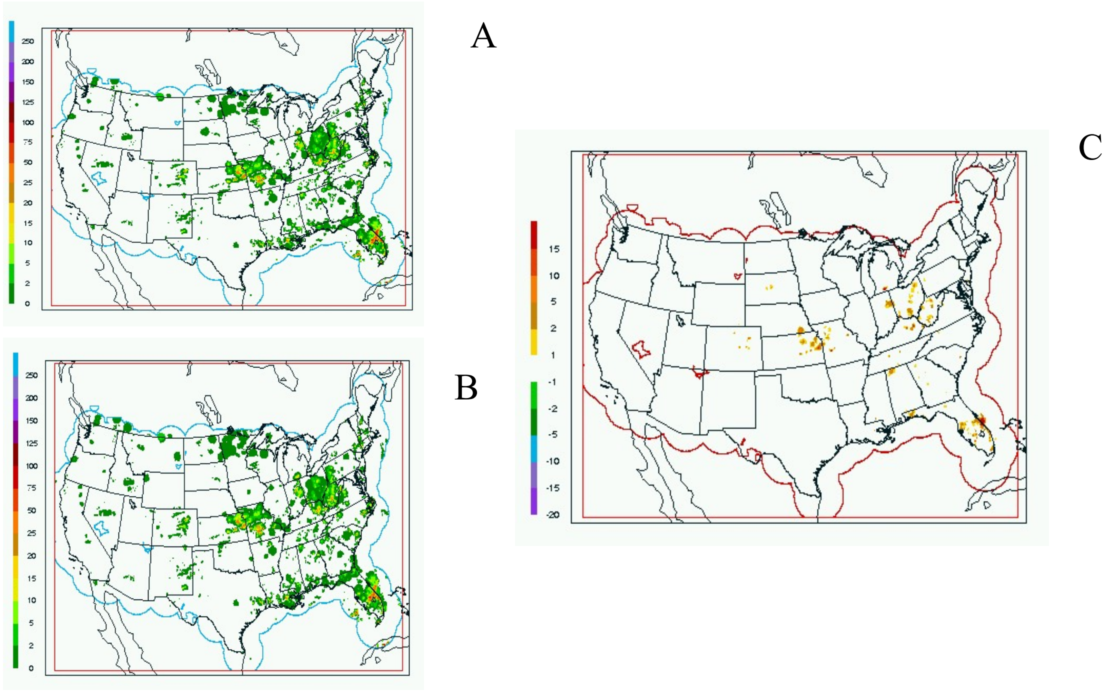


Fig. 4. Stage II analyses of hourly precipitation for 2300-0000 UTC 26-27 June 2005. Panel A is the analysis using QC'ed station, Panel B is the present operational analysis without additional new QC, and Panel C is the difference (A – B).

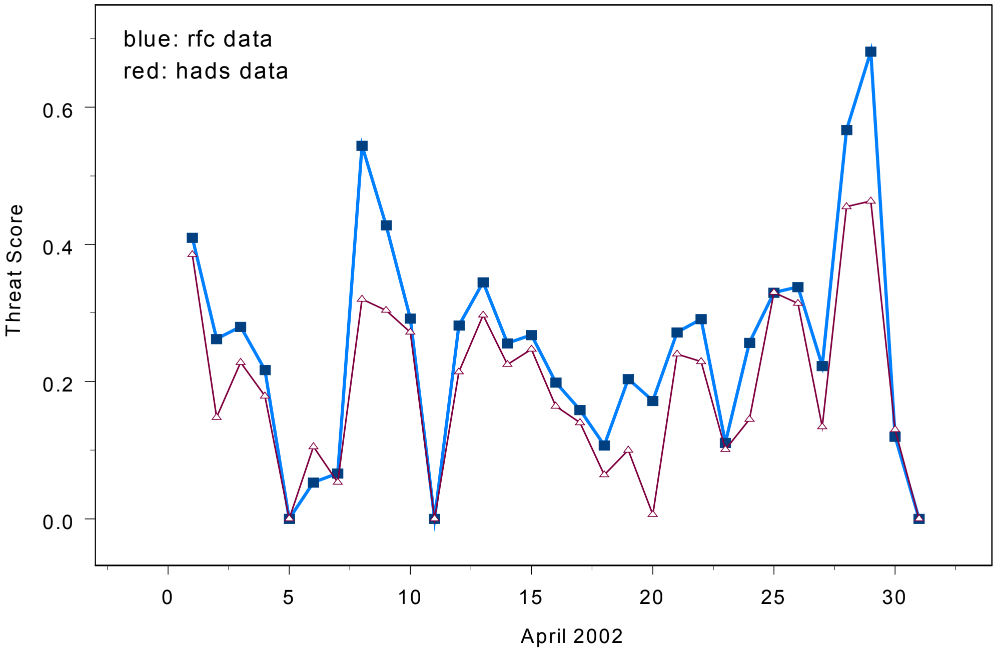


Fig. 5. Daily CONUS equitable threat scores for the eta model for April 2002. The scores represented by the blue curve were computed using high-quality daily (1200-1200 UTC) observations assembled by river Forecast Centers as verification data; scores represented by the red curve were computed using lower-quality lightly-QC'ed HADS daily totals as verification data.

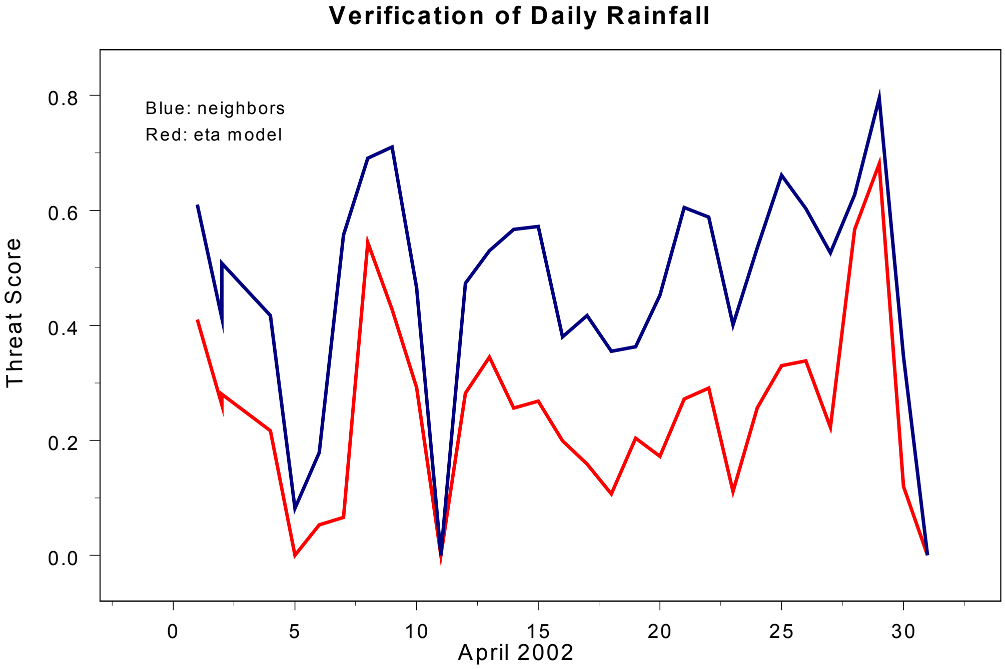


Fig 6. As in Fig. 5, except for equitable threat scores for the eta model (red curve; same as blue curve in Fig. 4) compared to scores computed between individual daily stations and their nearest neighbors (blue curve).

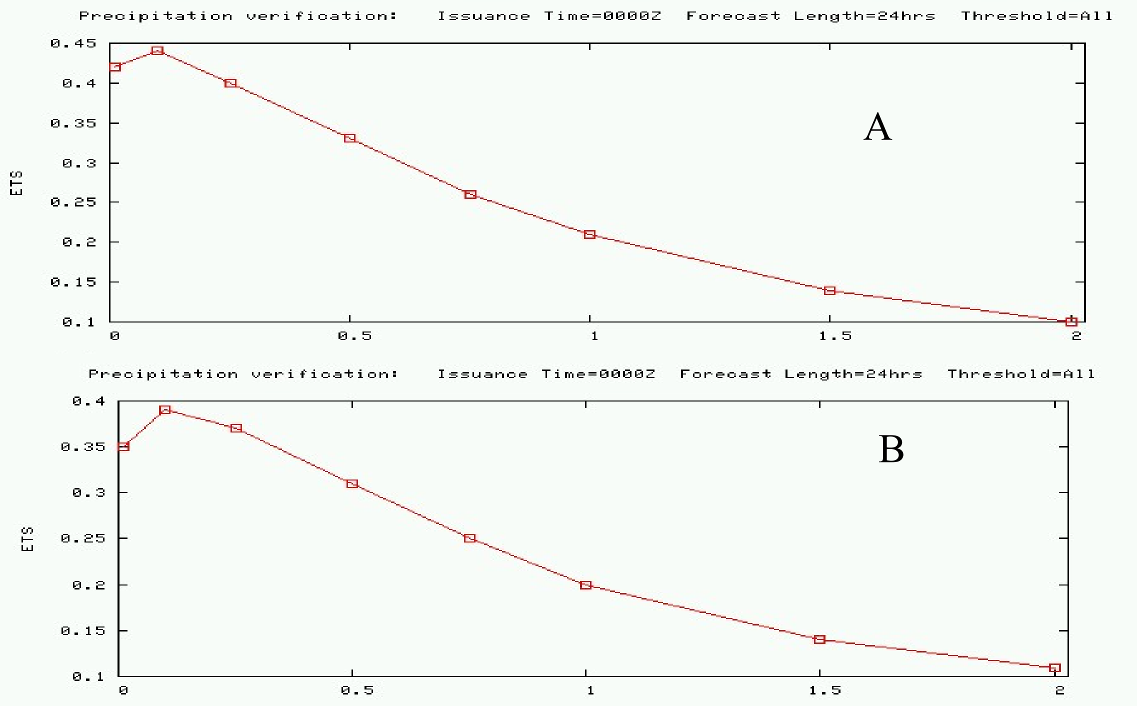


Fig. 7. Average daily (0000-0000 UTC) equitable threat scores over the CONUS for the WRF NNM model during the Development Testbed Center Winter Forecast Experiment (DWFE) from 15 January – 31 March 2005. Panel A represents scores computed using QC’ed HADS gage observations as verification data; Panel B represents scores computed using non-QC’ed HADS observations for verification.