Preprints, *Symposium on Precipitation Extremes: Prediction, Impacts, and Reponses,* January 14-19 2001, Albuquerque, NM

P1.36

A VERIFICATION APPROACH SUITABLE FOR ASSESSING THE QUALITY OF MODEL-BASED PRECIPITATION FORECASTS DURING EXTREME PRECIPITATION EVENTS

Andrew F. Loughe *
Cooperative Institute for Research in Environmental Sciences (CIRES)
University of Colorado/NOAA Forecast Systems Laboratory (FSL)
Boulder, Colorado

Judy K. Henderson, Jennifer Luppens Mahoney, and Edward I. Tollerud
NOAA Forecast Systems Laboratory
Boulder, Colorado

1. INTRODUCTION

Accurate forecasts of precipitation are exceedingly important to the public. Emergency management teams rely on these forecasts to plan when and where disasters might strike, and the general public is concerned about whether precipitation will impact their many outdoor activities, particularly in situations when heavy precipitation is possible. In either case, accurate precipitation forecasts are largely dependent upon numerical model predictions. Therefore, a verification approach that assesses a model's ability to accurately predict precipitation at specific locations, and over relatively short (~3 hour) time periods, has been developed. Although numerical models are not prepared to directly address the problem of point-specific precipitation forecasting, this stringent verification approach will serve to track the progress of models over time as they evolve to meet this high expectation of the public.

2. THE REAL-TIME VERIFICATION SYSTEM (RTVS)

The Real-Time Verification System (Mahoney et al., 1997), developed at NOAA's Forecast Systems Laboratory (FSL) and funded by the Federal Aviation Administration's Aviation Weather Research Program (FAA/AWRP), is an automated system developed to baseline Aviation Weather Center (AWC) forecast products. RTVS is also utilized to assess the quality of aviation-related algorithms, and to provide helpful information to AWC forecasters in near real time, such as for the Collaborative Convective Forecast Product (CCFP). Most displays and reports generated by RTVS are accessible via the World-Wide Web (the Web) at: *www-ad.fsl.noaa.gov:80/afra/rtvs/RTVS-project_des.html.* Recently, the RTVS has been enhanced to include verification of precipitation forecasts. Currently within RTVS, precipitation forecasts from the National Centers for Environmental Prediction's (NCEP) Rapid Update Cycle (RUC-2) and Eta models are verified using hourly gauge data from ASOS stations and from the National Weather Service's Hydrometeorological Automated Data System (HADS).

There are five main steps carried out by the automated RTVS:

1. Observational data are acquired, tested for quality, and stored for eventual comparison with model forecasts.

2. Numerical modeling forecast data are acquired and stored for eventual verification against the observations.

3. At various model run times and forecast lengths, the two data sources are grouped together for accumulation periods ranging from 3 to 24 hours.

4. The forecast values are interpolated to observation points, tested versus the observations at thresholds ranging from .01 to 5 inches, and *scored* by tallying results into a standard 2 X 2 contingency table of YY, YN, NY, NN forecast/observation pairs. A variety of skill measures are then computed from these data.

5. The contingency data and skill measures are represented in graphical and tabular form for display over the Web.

* *Corresponding author address:* Andrew F. Loughe, NOAA/OAR/FSL/AD R/FS5, 325 Broadway, Boulder, CO 80305-3328; e-mail: loughe@fsl.noaa.gov

2.1 Ingest of Observational Data

Each day, hourly precipitation gauge measurements from approximately 4500 HADS and ASOS stations are gathered by the Climate Prediction Center (CPC) at NCEP. Along with these hourly observations, CPC collects 24-h precipitation data from approximately 7500 stations. These 24-h data are from the NWS' cooperative observing network, ASOS and HADS stations, various local government mesonets, and a few private spotter networks. CPC applies measures to these 24-h precipitation totals to flag questionable reporting sites. These measures include buddy checks, comparisons to climatology, and the use of WSR-88D radar estimates (Sid Katz of CPC, personal communication).

The RTVS selects for verification only those hourly stations that also exist within the same day's 24-h precipitation station list, and are not included in the list of *flagged* 24-h stations. RTVS then checks for stations that are regularly reporting, by only keeping those hourly stations that have reported at least 5 of the last 10 days.

By automatically testing hourly station data in this manner, the system includes, on average, 2250 of the 4500 hourly precipitation stations, or roughly half of the hourly station data assembled daily by the CPC. FSL scientists and programmers are currently developing in-house QC methods for including more good hourly data from the large list of 4500 stations (see section 4). Additionally, it has been encouraging to note the continued expansion of the HADS network of satellite telemetered precipitation data, as is evident from this Web site: *http://www.nws.noaa.gov/oh/hads/hadsinfo.*

2.2 Ingest of Model Data

Once the RTVS collects hourly observations, compares them to the 24-h data, throws out stations that have been flagged as questionable, and checks that each station is reporting regularly, the automated system stores the station data into archive files indexed for 24 individual time periods — one for each hour of the day. These files are then retrieved and matched with corresponding model forecast data. The two data streams are collected over a series of 3-h accumulation periods, which are then summed for accumulation studies ranging from 3 to 24 hours. The model data are then verified for accumulation thresholds of .01 to 5.0 inches.

2.3 Comparing Models with Observations

The permutation resulting from various combinations of run time (0000, 0600, 1200, 1800 UTC), forecast length (3, 6, 9, 12, 24 hours), and threshold amount (.01, .10, .25, .50, .75, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0 inches) allows for great flexibility in investigating the strengths and weaknesses of the numerical models. The Web interface also allows one to group data over run time, forecast hour, or threshold, in order to present a clearer view of pervasive trends in the data.

2.4 Scoring the Forecasts

RTVS is unique from many other verification efforts in that it:

1. Performs model verification for accumulation periods less than 24 hours, possible because it uses hourly rather than daily precipitation observations. This is similar to the approach of Schwartz and Benjamin, 2000.

2. Interpolates model forecast values to actual observation points, instead of performing a grid-to-grid comparison by extrapolating observational data across a model grid box.

Model data summed over a given accumulation period are interpolated to observation points, and during verification for a specific threshold value, tallies are kept in a 2 X 2 contingency table consisting of YY, YN, NY, NN dichotomous forecasting values. These contingency pairs are stored for each run time, forecast length, threshold, and accumulation period listed above. Once the four count totals are grouped and stored in this manner, various skill scores are computed including probability of detection, probability of non-detection, bias, false alarm ratio, and ESS (Doswell et al., 1990).

The approach of bilinearly interpolating forecast values to observation points to score a model, sets what some may consider an unreasonable standard for the models to achieve. Nevertheless, each day members of the public verify numerical model performance in exactly this manner. A typical scenario is that forecast customers wake in the morning, listen to a weather report (which is largely based upon output from a numerical model), and ascertain whether or not it will rain in their particular area. If they expect rain, and it does not rain, they make a YN entry, if they expect no rain, but it rains, they make an NY entry. All of these experiences of comparing forecasts to observations can

be assembled into a contingency table and skill measures computed, as is done by the RTVS. It is true that this method can severely penalize an otherwise *good* forecast for inaccuracies in the placement and timing of rain events, but some of these negative effects are often mitigated by viewing results over longer accumulation periods, from 6-12 hours, for example.

2.5 Publishing the Results

One of the joys of modern research is that results can be shared rather quickly via the Web. At present, there are only two models scored by the enhanced RTVS, and so the Web interface is designed to highlight differences between these two models. This interface visually provides plots of equitable skill score (ESS) and bias (Fig. 1), and serves as an aid in determining which model is scoring better than the other (higher ESS), but also whether a particular score is higher due to overforecasting (larger bias) — a generally effective method of achieving higher skill scores (Schwartz and Benjamin, 2000).

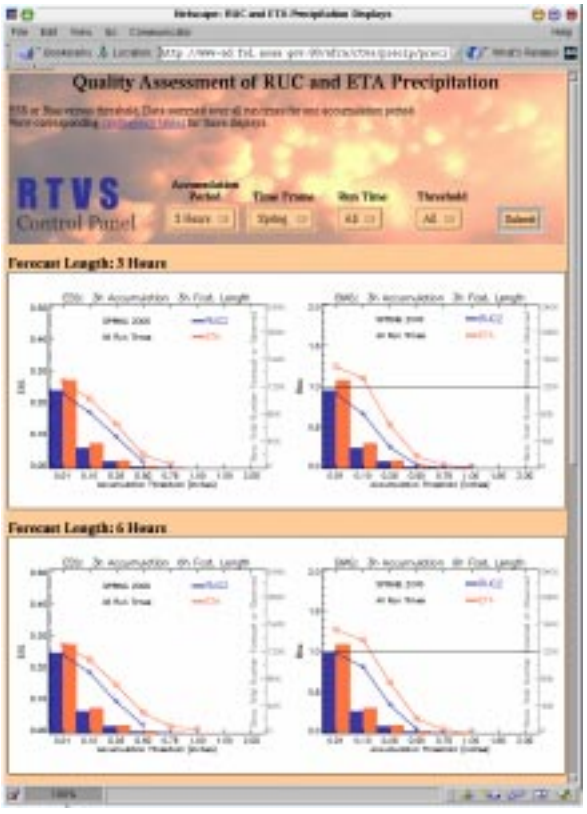

Fig. 1. Output from the RTVS precipitation Web page showing a typical view of equitable skill score (left) and bias (right) versus threshold (abscissas). The number of forecasts is represented in each plot by the bars and the axis on the right.

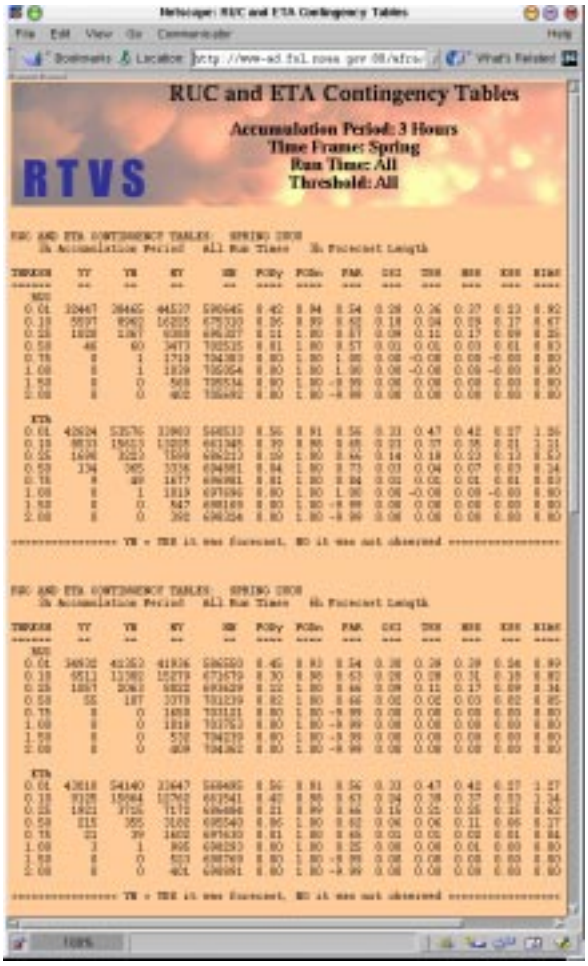

Fig. 2. Output from the RTVS precipitation Web page showing the Web-based contingency table corresponding to Fig. 1. This display is from verification of RUC-2 and Eta precipitation during spring 2000.

For those interested in looking more closely at the raw contingency data, the automated RTVS also provides access to the count data via contingency tables of the form shown in Fig. 2. Although the RTVS performs forecast verification daily, these graphics and tables are only provided for an entire month, and are made available for the previous month by the third day of the present month.

Some *daily* contingency tables are available by selecting a particular month (not season) and a specific threshold amount which the user wishes to investigate. The ability to display daily contingency data is provided for those wishing to more clearly understand how the models perform during particularly heavy rain events within a given month.

3. PRELIMINARY FINDINGS

The precipitation portion of the RTVS has been actively verifying the RUC-2 model since September 1999, and the Eta model since March 2000. In that short period of overlap, a few interesting trends have been noted. Results indicate that the quality of precipitation forecasts produced by Eta (as measured by the ESS) are generally better than those produced by RUC-2 at all model initialization and lead times; however, Eta consistently overforecasts precipitation at smaller accumulation thresholds, while RUC-2 generally underforecasts compared to the Eta model. Unexpectedly, RUC-2 improves, relative to Eta, at forecasting all accumulation amounts when initialized at 0600 UTC. This result is likely due to the inclusion of initialization data not available to Eta, or possibly due to decreased convective activity for forecasts generated at this hour.

Statistical results from RTVS for RUC-2 and Eta will be presented. Highlights will include variations in the models due to model initialization and lead times, seasonal dependencies, diurnal effects, and possible model spinup problems.

Fig. 3. HADS observing sites for 11 September 2000. This figure displays (a) all operating sites, and (b) sites selected by RFCs for 24 h totals.

4. FURTHER ISSUES INVOLVING DATA QUALITY

Comparisons between real-time HADS data and better-quality retrospective data from the Hourly Precipitation Dataset (HPD) discussed in Tollerud (1997) suggest that data quality can have a dramatic effect on the representativeness of verification fields. Because hourly reporting sites are sparsely located in regions of the U.S., it is vital to include as many as possible in each verification period. There is thus a constant trade-off between quality control procedures that rigorously screen out questionable stations and those that allow most stations into the verification process at the risk of including poor observations.

The extent of this dilemma is demonstrated in Fig. 3, from a Website located at *http://precip.fsl.noaa.gov/hourly_precip.html,* showing HADS observing sites on 11 September 2000. A daily screening of HADS stations during preparation of 24-h precipitation totals at individual River Forecast Centers (RFCs) typically results in a station set much like that plotted in the bottom panel. While sites in the Southern Plains and Appalachians are dense, sites in many regions, particularly Virginia, Georgia, Michigan, and the northwestern states, are very sparse. In contrast, the full set of station reports available on this day (top), although still sparse in some regions, is considerably more representative of the U.S. as a whole. Many of the site removals are a result of obviously bad observations, but differences in procedures among RFCs and availability of time from day to day, results in many stations not being included in the final set of 24-h observations for a given day.

RTVS verification of 3-h to 24-h precipitation has relied so far on the set of HADS stations that survive the selection process at RFCs (e.g., Fig. 3b). To supplement this set, we are developing a process that introduces new stations into the data stream by comparing the previous month's daily total precipitation at excluded HADS sites with neighboring HADS observations that have been selected by RFCs, and with other available 24-h observations. The assumption is that, at automated sites in the HADS network, instruments that observe accurately will do so consistently, at least over a period of several weeks. Thus far, this screening process consists of qualitative examination of time series of observations at neighboring stations (Fig. 4). For example, the exclusion of stations NAPM1 (no. 32) and BIKN3 (no. 60) from the RFC set can be easily justified, while

FFDN3 (no. 58) and MLTM3 (no. 14), although excluded from the RFC set, appear to provide observations consistent with their neighbors. Based on this kind of examination of the HADS data from June 2000, it appears possible to add several hundred HADS stations in poorly represented regions to the verification station set.

5. CONCLUDING REMARKS

Future enhancement of the RTVS precipitation program centers around these three objectives:

1. Increasing the number of quality-controlled hourly observations that are ingest into the system. This includes automation of the QC methods highlighted in section 4.

2. Expanding the number of models which are verified, and developing Web-based tools that provide greater flexibility in comparing the results. The ability to view regional summaries will also be investigated.

3. Enhancing statistical measures to include significance testing, and comparison of results with traditional grid-based verification methods.

6. REFERENCES

Doswell, C. A. III, R. Davies-Jones and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576-585.

Mahoney, J. L., J. K. Henderson and P. A. Miller, 1997: A description of the Forecast Systems Laboratory's real-time verification system (RTVS). Preprints, *7th Conference on Aviation, Range, and Aerospace Meteorology*, Long Beach, CA, Amer. Meteor. Soc., J26-J31.

Schwartz, B., and S. Benjamin, 2000: Verification of RUC-2 precipitation forecasts using the NCEP multisensor analysis. Preprints, *4th Symposium on Integrated Observing Systems*, Long Beach, CA, Amer. Meteor. Soc., 182-185.

Tollerud, E. I., 1997: The impact of data quality and internetwork consistency on central United States precipitation analyses using multiple gage networks. Preprints, *13th Conference on Hydrology*, Long Beach, CA, Amer. Meteor. Soc., 271-274.

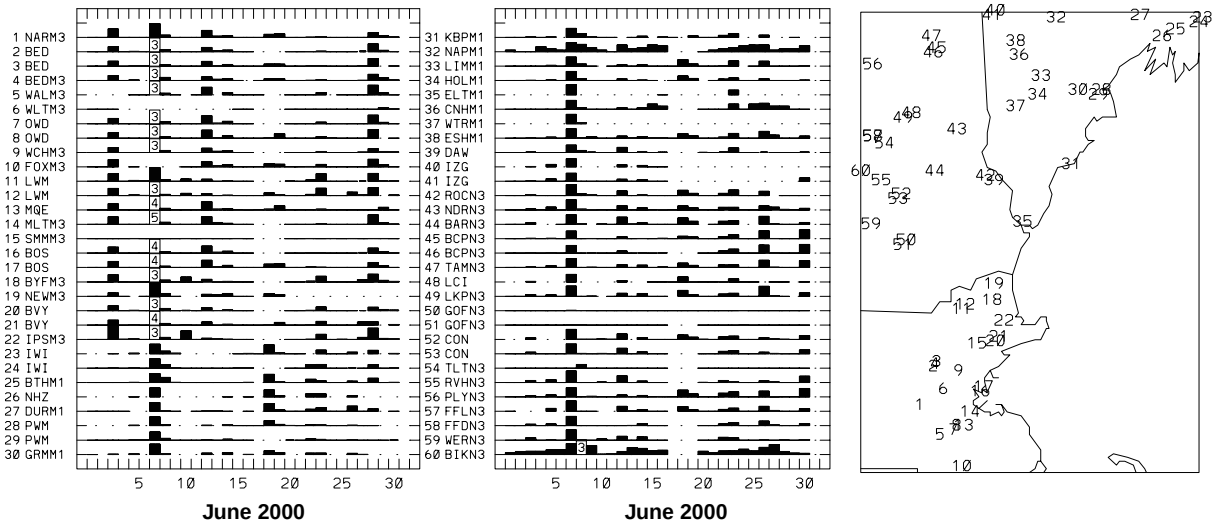

Fig. 4. Time series of daily 1200 UTC — 1200 UTC precipitation totals at HADS and daily reporting stations during June 2000. Largest bars are approximately 10 cm. Open bars display daily totals greater than 2 inches. Duplicate stations are HADS observations totalled at FSL followed by same sites totalled at RFCs. Dotted lines indicate missing data. The geographical location of these stations is indicated on the map to the right.