9.8

THE REAL-TIME VERIFICATION SYSTEM (RTVS) AND ITS APPLICATION TO AVIATION WEATHER FORECASTS

Jennifer Luppens Mahoney[1]
NOAA Research-Forecast Systems Laboratory
Boulder, CO

Judy K. Henderson[2], Barbara G. Brown[3], Joan E. Hart[4], Andrew Loughe[4],
Christopher Fischer[4], and Beth Sigren[4]

1. INTRODUCTION

For aviation, nothing is more important than safety, and in order to make safe decisions, nothing is more important than having access to the best possible information. World-class scientists, engaged in government-sponsored research, have had enormous success in producing a new generation of weather products for a variety of users (e.g., airline meteorologists and private pilots). However, as these emerging technologies come into operational use, great care must be - and is - exercised to ensure that they truly are a step forward (Dave Pace, personal communication). Historically, the quality of these new aviation weather products was often tested through controlled studies on a sample of the data which were manually and subjectively analyzed. If successful in the sample, the product was deemed "good" and placed into operations; however, no sample was ever complete enough, nor any subjective analysis extensive enough to truly reveal the strength and weaknesses of a new technique.

The development and availability of the Real-Time Verification System (RTVS) has changed all that. Since 1997, a project team at NOAA/FSL and NCAR/RAP has been developing the Real-Time Verification System (RTVS) which is now an integral part of the Federal Aviation Administration (FAA) Aviation Weather Research Program (AWRP) and the National Weather Service Aviation Weather Center (AWC). The RTVS provides a new mechanism for establishing the level of quality of weather forecasts (Mahoney et al. 1997). The system allows for consistent, unbiased, objective verification statistics to be computed for a variety of forecasts in near real-time, generally with an emphasis on forecasts critical to aviation. The system has been designed to be accessed with an easy-to-use interface via the Web (*http://www-ad.fsl.noaa.gov/afra/rtvs*) so that local, as well as remote users, may obtain the information they need to support their decision-making process.

[1]Corresponding author: NOAA Forecast Systems Laboratory, 325 Broadway, Boulder, CO 80305; mahoney@fsl.noaa.gov
[2] Forecast Systems Laboratory, Boulder, CO
[3] Research Application Program, National Center for Atmospheric Research, Boulder, CO.
[4] Joint collaboration with Cooperative Institute for Research in the Environmental Sciences, University of Colorado Boulder, Boulder, CO

The RTVS has been developed to provide a statistical baseline for weather forecasts and model-based guidance products, and to support real-time forecast operations, model-based algorithm development, and case study assessments. To this end, the RTVS was designed to ingest weather forecasts and observations in near real time (as data become available) and store the relevant information in a relational database management system (RDBMS). A flexible, easy-to-use Web-based graphical user interface assures users quick and easy access to the data stored in the RDBMS. Users can compare various forecast lengths and issue times, over a user-defined time period and geographical area, for a variety of forecast models and algorithms.

The verification methods, underlying the system architecture, are developed from state-of-the-art techniques (Brown et al. 1997). These techniques often must be modified to accommodate the peculiarities of aviation forecasts.

This paper describes the architecture of the RTVS and briefly summarizes the verification methods used to evaluate the forecasts.

2. SYSTEM DESIGN

To alleviate the reprocessing of large amounts of data, the RTVS system is designed to allow the processing of forecasts and observations to occur as they become available in near real time. In addition, the system allows reprocessing of large amounts of data by emulating the real-time function so that consistent statistical baselines can be maintained.

The RTVS relies on forecasts and observations that are then processed into NetCDF files. This internal format is a self-describing format that allows easy access to specific variables within a file. The data files, mainly consisting of forecast/observation pairs, are stored in a data directory structure that identifies the pattern of "data-type" (e.g., icing), "model-type" (e.g., RUC), and "observation-type" (e.g., PIREPs). In addition to these data files, the forecast/observation pairs are stored in a RDBMS, which allows flexible generation of and access to the statistics. The RDBMS used for RTVS is MySQL, an off-the-shelf software package that is easily portable to other laboratories, such as the AWC. Users can access the statistical results through a Web-based graphical user

Interface that was developed in Java script to interact with SQL queries. Users can select from various forecast products, observation types, and regions for any period of time, with the results combined weekly, monthly, or yearly. Finally, users can define the plot type and statistic to display. The selections defined by the user through this interface are combined to produce the query that is used to access the data from the RDBMS.

3. VERIFICATION METHODS

The methodology used to verify the forecasts is fundamentally based on the statistical framework for verification developed by Murphy and Winkler (1987) and was later modified for aviation forecasts by Brown et al. (1997). In general for each variable type verified through RTVS, the forecasts are matched (or interpolated when using a grid from a numerical weather prediction (NWP) model) to the observation locations. For example, the icing and turbulence forecasts are matched to PIREPs (Brown et al. 2000), ceiling and visibility forecast interpolated to surface observations (Brown et al. 2001), and precipitation forecasts produced from NWP models are interpolated to the precipitation gauge observations (Loughe et al. 2001). Moreover, grids of convective forecasts are directly compared with grids of the National Convective Weather Detection (NCWD; Mueller et al. 1999) product, a convective product that combines radar and lightning observations. In each of these cases, the forecasts and observations are treated dichotomously (Yes/No) by applying thresholds to the data. The computation of the statistics is then based on the standard two-by-two contingency table (Brown et al. 1997).

Some of the statistics available for verification of dichotomous (Yes/No) forecasts are summarized in Table 1. It is noteworthy that not all statistics listed in the table can be used to evaluate all of the forecasts. In particular, the primary statistics used to evaluate icing and turbulence forecasts are PODy, PODn, % Volume and % Area (where the latter two statistics represent the areal and volumetric extent of the forecast). Furthermore, since the icing or turbulence forecast grid is not adequately sampled by the PIREPs, the FAR, Bias and other standard statistics (e.g., CSI, Heidke and Gilbert skill scores) are not and should not be computed (Brown and Young 2000).

3.1 Complexities

Creating a matched set of forecasts and observations is one of the most difficult aspects of forecast verification. Some of the difficulties encountered when developing verification methods follow.

First, scaling the observations to match the forecasts has been a particular problem when evaluating forecasts of convection. Mahoney et al. (2000) state that the statistical results for convective forecasts are influenced, in part, by both the scale at which the forecasts are produced and the grid size used to verify them. Therefore, within RTVS the grid used to map the convective observations is based on the scale at which the convective forecasts are issued and intended to be used. As one example, the observational data used to verify forecasts for large areas of convection, such as the Collaborative Convective Forecast Product (CCFP), are interpolated from their native 4-km grid to a 40-km grid in order to more accurately represent the scale of the forecast itself.

Table 1. Standard verification measures that can be computed from the 2x2 contingency table.

Statistic	Definition	Description
PODy	$\frac{YY}{YY+NY}$	*Probability of Detection* of "Yes" observations: Proportion of "Yes" obser vations that were forecasted correctly
PODn	$\frac{NN}{YN+NN}$	*Probability of Detection* of "No" observations: Proportion of "No" obser vations that were forecasted correctly
FAR	$\frac{YN}{YY+YN}$	*False Alarm Ratio*: Proportion of "Yes" fore casts that were incorrect
CSI	$\frac{YY}{YY+YN+NY}$	*Critical Success Index*: Number of correct "Yes" forecasts relative to number of "Yes" forecasts or observations
TSS	PODy +PODn-1	*True Skill Statistic* A measure of discrimination
PC	$\frac{YY+NN}{T}$	*Proportion Correct*: Proportion of "Yes" and "No" observations that were forecasted Correctly
Bias	$\frac{YY+YN}{YY+NY}$	Frequency of "Yes" forecasts relative to frequency of "Yes" observations
% Vol	$\frac{\text{Forecast Vol.}}{\text{Total Vol.}} \times 100$	% of the total airspace that is impacted by the forecast

Second, the lack of evenly distributed and consistently reported observations, such as PIREPs (Schwartz 1996) poses a problem particularly for forecasts of icing and turbulence. Therefore, since the PIREPs do not provide a representative sample of the forecast grid and pilots are often encouraged to avoid areas that contain the verifying information, standard verification methods, such as FAR, cannot be computed for these variables (Brown et al. 1997; Brown and Young 2001).

Third, grid vs. point verification presents a number of complexities in developing matched pairs, particularly in cases where the observations are nonstandard. For example, the ceiling and visibility AIRMET forecasts have been evaluated at both station locations (i.e., points) and at grid points, where the stations were put on a grid. Overall, the differences in the statistics were minor (Brown et al. 2001). In general, both the PODy and FAR values were slightly smaller when the gridded method was used to evaluate the AIRMETs, but no particular improvement of one method over another was demonstrated. Nevertheless, information can be gained by using both approaches to evaluate forecasts.

Finally, specific verification statistics associated with a forecasting system are less meaningful or valuable, if they cannot be compared to values associated with another forecasting system, or another appropriate standard of comparison (Brown et al. 2001). However, in general it is difficult to compare human-generated and automated forecasts; automated forecasts are often more precisely defined than those produced by humans while human-generated forecast often incorporate more detail. Considering that this extra detail provided by human-generated forecasts is often nonstandard and difficult to decode, the verification techniques are designed to treat these forecasts in a similar manner to the automated forecasts. Ideally, automated verification of forecast products by a system such as RTVS, should provide motivation for forecast formats to be standardized such that all available information may be verified resulting in improved forecasts as well as improved usability of the forecasts by users.

3.2 Applications

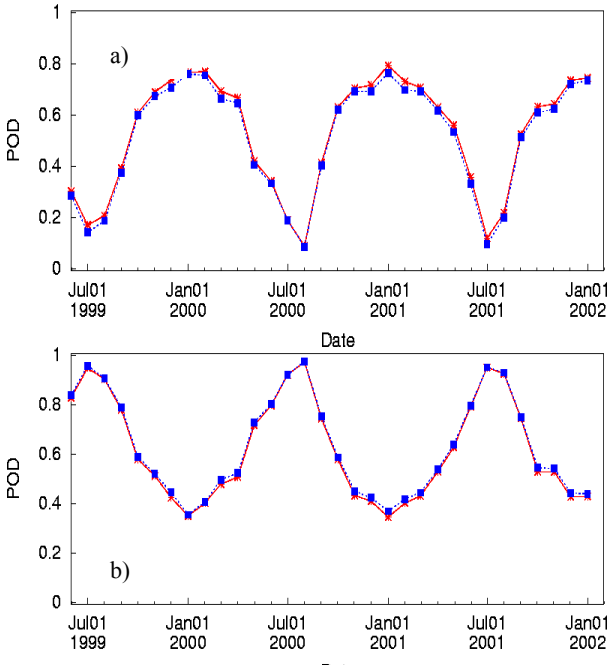
Using the verification methods described in Section 3, a selection of statistical results for icing, convection, and precipitation forecasts are presented. More information for each of these variables and others can be obtained through the RTVS web site (http://www-ad.fsl.noaa.gov/afra/rtvs).

3.2.1 Icing

Verification results from RTVS have been used to track the quality of icing AIRMETs since 1999 (Mahoney et al.1998). For instance, AIRMETs with and without amendments can be compared to determine the impact of amending the forecasts, as shown in Fig. 1. Each dot on the line represents a PODy (Fig. 1a) or PODn (Fig. 1b) value computed from forecast/observation pairs generated for each month from 1 January 1999 – 31 January 2002. As shown by the overlapping lines for both PODy (Fig. 1a) and PODn (Fig. 1b), little overall improvement occurs in the quality of the AIRMETs when they are amended.

3.2.2 Convective

Displays are produced through RTVS that provide direct feedback to the forecasters. For instance, evaluations of independent CCFP forecasts are presented through displays that include the forecasts, verifying observations, and the statistics. An example is shown in Fig. 2 for the CCFP 6-h forecast issued at 1500 UTC on 4 June 2000. The light and dark gray areas in Fig. 2 represent the CCFP forecasts and the smaller square-like areas represent the verifying NCWD observations. The forecasts are colored (not available here) to represent a particular coverage of convection within the forecast area. The statistics computed for the example are shown along the left margin, and coverage of the NCWD within the CCFP is shown for each forecast area on the figure. These figures are generated for each forecast issue- and lead-time and are available to forecasters before the next forecast cycle. Using these displays which combine the graphical forecast information along with the verification statistics, AWC forecasters have been able to create smaller forecast areas resulting in improved forecasts for the aviation community (AWC forecasters, personal communication).



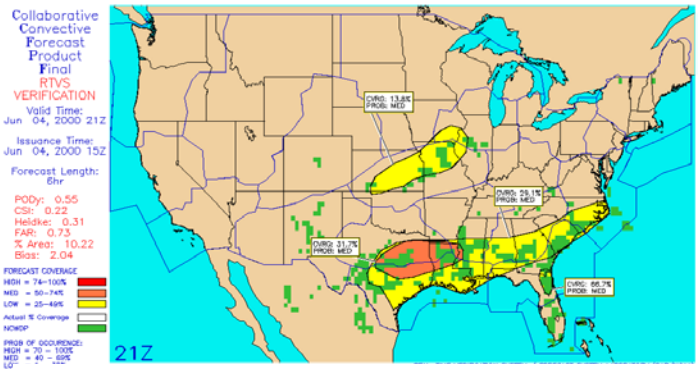


Figure 2. Map of the CCFP (large gray areas) and the NCWD (small dark gray square-like areas). Actual coverage computed from the NCWD is shown in white boxes on display. Statistics are shown on left margin.

3.2.3 Precipitation

In addition to turbulence, icing, ceiling, visibility, and convection, precipitation forecasts from NWP models are continuously being evaluated through the RTVS. An example of this evaluation is shown in Fig. 3 by plots of bias and equitable threat score (ETS) for 5 NWP models for several threshold values. The results shown in Fig. 3 were computed, from 1 June 2001–31 August 2001, for the 3-h lead times and by accumulating the precipitation

forecast/observation pairs over all runtimes. The pairs were computed by interpolating the model output to the precipitation gauge locations. Although the trend in ETS and bias is the same for nearly all models, there are some slight differences between them. For instance, the precipitation 3-h forecast from the Advanced Regional Prediction System (ARPS; '◇') has one of the largest bias values at smaller precipitation thresholds and the smallest ETS at all thresholds. It is interesting that the bias for the Mesoscale Modeling System version 5 (MM5) model remains between 1.0 and 1.5 indicating a slight tendency to overforecast precipitation at all thresholds. However, the other models considerably underforecast precipitation at thresholds larger than 0.5 in.

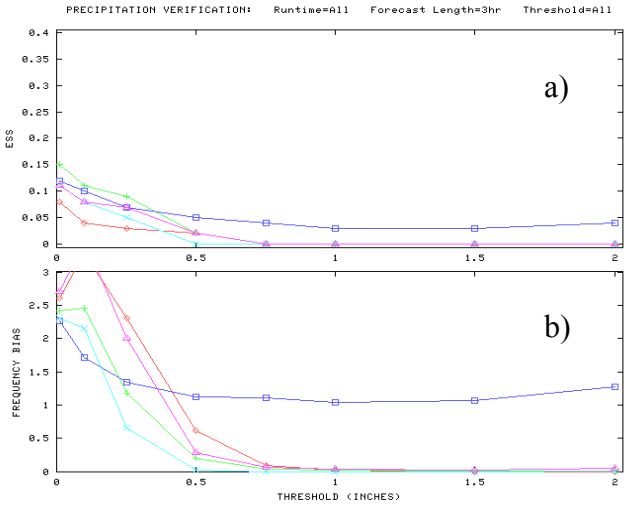


Figure 3. NWP precipitation statistics: (a) ETS and (b) bias computed for Eta40 ('+'), ARPS ('◇'), RUC40 ('*'), WRF22 ('△'),and MM5 ('□'); hourly observations are accumulated into 3-hourly periods for 1 June – 31 August 2001 for all forecasts with a 3-h lead time.

4. SUMMARY

The RTVS is a flexible, easy-to-use Web-based system that contains a wealth of statistical information for human-generated, automated algorithms, and NWP forecasts. Forecast and observations are processed in near real-time, which allows the statistical database to continuously build and enables rapid access to current information. The statistics generated through the RTVS are used to provide baseline statistics and track the quality of forecasts over time, and to support real-time forecast operations, model-based algorithm development, and case study assessment. The verification methods follow a well-developed framework and are adjusted to account for the complexities inherent in the forecasts. Future enhancements to the RTVS include verifying other forecasts, adding flexible tools that allow for interrogation observations and forecasts, and introducing new verification methods for diagnosing spatial and temporal forecast errors.

Acknowledgments

This research is in response to requirements and funding provided by the Federal Aviation Administration. The views expressed are those of the authors and do not necessarily represent the official policy and position of the U.S. Government. We would like to express our appreciation to those at NCAR and AWC who have provided assistance with developing verification methods and helped to define and develop the RTVS and to Nita Fullerton and Mike Kay for their helpful review of this work.

References

Brown, B.G., J.L. Mahoney, T.L Fowler, and J. Henderson, 2001: Approaches for verification of ceiling and visibility diagnosis and forecasts (available from the author, bgb@ucar.edu).

Brown, B.G., J.L. Mahoney, J. Henderson, T.L. Kane, R. Bullock, and J.E. Hart, 2000: The turbulence algorithm intercomparison exercise: Statistical verification results. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology,* Orlando, Fl. Amer. Meteor.Soc., 466-471.

Brown, B.G. and G.S. Young, 2000: Verification of icing and turbulence forecasts: Why some verification statistics can't be computed using PIREPs. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, Fl. Amer. Meteor. Soc., 393-398.

Brown, B.G., G. Thompson, R.T. Bruintjes, R. Bullock, and T. Kane, 1997: Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Wea. and Forec.*, **12**, 890-914.

Loughe, A.F., J.K. Henderson, J.L. Mahoney, and E.I. Tollerud, 2001: A verification approach suitable for assessing the quality of model-based precipitation forecasts during extreme precipitation events. *Preprints, Symposium on Precipitation Extremes: Prediction, Impacts, and Responses*, Albuquerque, NM, Amer. Meteor. Soc., 77-81.

Mahoney, J.L., B.G. Brown, J.E. Hart, and C.Fischer, 2002: Using verification techniques to evaluate the differences among convective forecasts. *Preprints, 16th Conference on Probability and Statistics in the Atmospheric Sciences*, Orlando, FL, Amer. Meteor. Soc, 12-19.

Mahoney, J.L., B.G. Brown, C. Mueller, and J.E. Hart, 2000: Convective intercomparison exercise: Baseline statistical results. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, Amer. Meteor. Soc., 403-408.

Mahoney, J.L., B.G. Brown, D. Mathews, F. Mosher, 1998: Verification of the Aviation Weather Center's in-flight aviation weather advisories: The methods, complexities, and limitations. *Preprints, 14th Conference on Probability and Statistics in the Atmospheric Sciences*, Phoenix, AZ, Amer. Meteor. Soc., J28-J32.

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.

Mueller, C.K., C.B. Fidalego. D.W. McCann, D. Meganhart, N.Rehak, and T. Carty, 1999: National Convective Weather Forecast Product. *Preprints, 8th Conference on Aviation, Range, and Aerospace Meteorology,* Amer. Meteor. Soc., 230- 234.

Murphy, A.H., and R.L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115,** 1330-1338.

Schwartz, B., 1996: The quantitative use of PIREPs in developing aviation weather guidance products. *Wea. Forec.*, **11,** 372-384.