*Ninth Conference on Aviation, Range, and Aerospace Meteorology*
*11-15 September 2000, Orlando, FL*

# 7.4 THE TURBULENCE ALGORITHM INTERCOMPARISON EXERCISE: STATISTICAL VERIFICATION RESULTS

Barbara G. Brown[1], Jennifer Luppens Mahoney[2], Judy Henderson[2], Tressa L. Kane[3], Randy Bullock[3], and Joan E. Hart[4]

## 1. INTRODUCTION

This paper summarizes results of a winter 1998-99 intercomparison of the forecasting capability of a number of clear-air turbulence (CAT) forecasting algorithms. This study was undertaken by the Turbulence Product Development Team (PDT) of the Federal Aviation Administration's (FAA's) Aviation Weather Research Program (AWRP).

Purposes of the intercomparison were to (i) develop a baseline for the quality of current CAT forecasting algorithms; (ii) demonstrate to-date progress in the development of these forecasting tools; (iii) examine the strengths and weaknesses of the algorithms; and (iv) perform an evaluation that is independent, consistent, comprehensive, and fair. Implementation of the intercomparison and analyses of the results were the responsibility of the AWRP Quality Assessment Group, which includes the verification groups of the NOAA Forecast Systems Laboratory (FSL) and the National Center for Atmospheric Research, Research Applications Program (NCAR/RAP).

The study consisted of two major components: (i) a real-time component, in which the algorithms were evaluated in near-real-time by FSL's Real-Time Verification System (RTVS; Mahoney et al. 1997), with results displayed on the World-Wide Web; and (ii) a post-analysis component in which the forecasts were re-generated and examined in detail at NCAR and FSL. This paper primarily concerns some results of the post-analyses. However, example results from the RTVS evaluation for winter 2000 are shown in the final section.

The paper is organized as follows. The study approach is presented in Section 2. Section 3 briefly describes the algorithms that were included in the evaluation. The data that were utilized are discussed in Section 4, and the verification methods are described in Section 5. Some results of the study are presented in Section 6. Finally, Section 7 contains conclusions and future work.

## 2. APPROACH

A total of 14 CAT algorithms were included in the real-time portion of the study, and results for a total of 11 algorithms are presented here. The algorithms were applied to data from the RUC-2 (Rapid Update Cycle, Version 2) model (Benjamin et al. 1998), with model output obtained from the National Centers for Environmental Prediction. Model forecasts issued at 1200, 1500, 1800, and 2100 UTC, with lead times of 3, 6, and 9 hr were included in the real-time portion of the study, but the analyses presented here are limited to the combinations of issue and lead time shown in Table 1. Turbulence AIRMETs, which are the operational turbulence forecasts issued by the National Weather Service's Aviation Weather Center (NWS/AWC; NWS 1991) were included for comparison purposes. Because the emphasis of the study concerned forecasting upper-level CAT, the evaluation was limited to the portion of the atmosphere above 20,000 ft. The analyses included all forecasts available over the period 21 December 1998 through 31 March 1999.

**Table 1. Issue and lead times used in post-analyses.**

| *Issue time (UTC)* | *Lead times (hr)* | *Valid times (UTC)* |
|---|---|---|
| 1500 | 3, 6, 9, 12 | 1800, 2100, 0000, 0300 |
| 1800 | 3, 6, 9 | 2100, 0000, 0300 |
| 2100 | 3, 6 | 0000, 0300 |
| 0000 | 3 | 0300 |

The algorithm forecasts and AIRMETs were verified using Yes and No pilot reports (PIREPs) of turbulence. In addition, automated vertical accelerometer (AVAR) observations provided by certain United Airlines aircraft, were used as an indicator of No turbulence under certain conditions (described in Section 4). The algorithm forecasts were transformed into Yes/No turbulence forecasts by determining if the algorithm output at each model grid point exceeded or was less than a pre-specified threshold. A variety of thresholds was utilized for each algorithm. The Yes/No forecasts were evaluated using standard verification techniques available for Yes/No forecasts where observations are based on PIREPs, as described in Section 5 (Brown and Mahoney 1998).

## 3. ALGORITHMS

The 11 CAT algorithms that were included in this post-evaluation are listed in Table 2, along with references to the appropriate literature where they are described. The algorithms listed in Table 2 include standard turbulence metrics, such as Richardson number, as well as algorithms that have been recently developed (e.g., Ellrod Index, DTF3-5). The Integrated Turbulence Forecasting Algorithm (ITFA) is currently under development by the Turbulence PDT. ITFA uses fuzzy logic techniques to integrate turbulence observations with the forecasts provided by a suite of

[1]*Corresponding author address*: Barbara G. Brown, National Center for Atmospheric Research, Boulder, CO 80307; e-mail: bgb@ucar.edu
[2]Forecast Systems Laboratory, NOAA, Boulder CO
[3] NCAR, Boulder, CO
[4]Cooperative Institute for Research in the Environmental Sciences, University of Colorado; and Forecast Systems Laboratory

turbulence indices, such as the others included in this study. Four additional algorithms that were included in the real-time intercomparison (Burke-Thompson, CCAT, Richardson tendency, and ULTURB) were not included in this post-analysis. These algorithms are described and evaluated in Brown et al. (1999b). In addition, ULTURB (McCann 1997), which also was included in the real-time intercomparison, was found to have errors that are being corrected; this algorithm will be included in future evaluations. Additional algorithms, such as Brown-2 (Brown 1973), two algorithms developed by the NWS (Reap 1996), and a statistically-based integrated algorithm (Tebaldi et al. 2000), also will be included in future analyses.

One additional algorithm that was included in the analyses, but is not listed in Table 2 is a "Random" algorithm, which randomly assigns values in the range 0-1 to each grid point. The purposes of including the Random algorithm were to (i) determine that the verification software was functioning correctly and (ii) ascertain that the algorithms produce forecasts that are better than what could be attained through a random process.

**Table 2. Diagnostic algorithms included in the post-analysis.**

| ***Algorithm*** | ***References*** |
|---|---|
| Brown-1 | Brown (1973) |
| Colson-Panofsky | Colson and Panofsky (1965) |
| DTF3, DTF4, DTF5 | Marroquin (1985, 1988) |
| Empirical Dutton | Dutton (1980) |
| Ellrod-1,2 | Ellrod and Knapp (1992) |
| Endlich | Endlich (1964) |
| ITFA | Sharman et al. (2000) |
| Richardson number | Drazin and Reid (1981); Dutton and Panofsky (1970); Kronebach (1964) |

## 4. DATA

Output from the RUC-2 model was used as input for the various algorithms. This model is the operational version of the Mesoscale Analysis and Prediction System (MAPS), Version 2 model, developed at FSL (Benjamin et al. 1998). The model vertical coordinate system is based on a hybrid isentropic-sigma vertical coordinate, and the horizontal grid spacing is approximately 40 km. The turbulence algorithms were applied to the model output files to create algorithm output files. As part of this process, the algorithm output data were interpolated to flight levels (i.e., every 1,000 ft).

All available Yes and No turbulence PIREPs were included in the verification analyses. These reports include information about the severity of turbulence encountered, which was used to categorize the reports. In particular, reports of moderate to extreme turbulence were included in the "Moderate-or-Greater" (MOG) category. In addition to the PIREPs, AVAR data were obtained from certain United Airlines aircraft, through the Aircraft Communications, Addressing, and Reporting System (ACARS). These data are available every 10 minutes through the FSL Aircraft Data Web. The AVAR observations are a measure of the aircraft's vertical acceleration, which can be associated with either internal motions of the aircraft, or external forces such as turbulence. Due to the effects of aircraft motions on the value of the vertical acceleration, the AVAR data only can be used as an indicator of no turbulence. Thus, only AVAR observations that were within 20% of the acceleration of gravity were included as observations of No turbulence.

## 5. METHODS

### 5.1 *Matching methods*

For the post-analyses, PIREPs and AVAR observations were matched to the largest algorithm value (or smallest value in the case of Richardson number) among the surrounding eight gridpoints (i.e., the four surrounding gridpoints horizontally, over two levels vertically). AIRMETs were matched to the PIREPs in the same way, by overlaying the RUC-2 grid on the AIRMET areas and assigning a Yes forecast to all gridpoints located inside an AIRMET. This approach is somewhat different from the bi-linear interpolation method used by the RTVS, which estimates the algorithm values at the location of each PIREP. However, consistency between the RTVS and post-analysis results suggest that the verification results are robust to this choice.

Previous evaluations of time windows for matching PIREPs to the forecasts have suggested that ±1 hour is an appropriate time length to allow fair representativeness of the model valid time and to obtain an adequate number of PIREPs (Mahoney 1998). Thus, this time window was applied in these analyses, both in real time and in post analysis. A time window of ±1 hour around the model valid time also was used to evaluate the AIRMETs, so that the AIRMET verification results are comparable to the algorithm results.

### 5.2 *Statistical verification methods*

Turbulence forecasts and observations are treated here as dichotomous (i.e., Yes/No) values. In particular, the algorithm forecasts are converted to a variety of Yes/No forecasts by application of various thresholds for the occurrence of turbulence. Thus, the verification statistics generally are based on methods for dichotomoous variables. However, some special methodological concerns arise due to the use of PIREPs as the verification data. For example, it is not appropriate to compute verification statistics such as the False Alarm Ratio, Bias, and Critical Success Index, due to the non-systematic nature of PIREPs (Brown and Young 2000). Moreover, other verification statistics based on PIREPs should not be interpreted in an absolute sense, but can be used in a comparative sense, for comparisons between algorithms and forecasts.

Table 3 lists the verification statistics used in this evaluation. As shown in this table, PODy and PODn are the primary verification statistics based on the 2x2

verification table. These statistics should be interpreted as the proportions of Yes and No PIREPs that are correctly forecast. Together, PODy and PODn measure the ability of the forecasts to discriminate between Yes and No turbulence observations. This discrimination ability is summarized by the True Skill Statistic (TSS), which frequently is called the Hanssen-Kuipers discrimination statistic (Wilks 1995). Note that it is possible to obtain the same value of TSS for a variety of combinations of PODy and PODn. Thus, it always is important to consider PODy and PODn, as well as TSS. PODn is computed in two ways in this study – (i) using the negative PIREP observations and (ii) using the negative AVAR observations.

**Table 3. Verification statistics used in this study.**

| *Statistic* | *Definition* | *Description* |
|---|---|---|
| PODy | Probability of Detection of "Yes" PIREPs | Proportion of Yes PIREPs that were correctly forecasted |
| PODn | Probability of Detection of "No" PIREPs or AVARs | Proportion of No PIREPs or AVARs that were correctly forecasted |
| True Skill Statistic (TSS) | PODy + PODn –1 | A measure of discrimination |
| Curve Area | Area under the ROC curve relating PODy and 1-PODn | A measure of overall skill |
| % Volume | Forecast Volume divided by Total Volume, x 100 | % of the total airspace that is impacted by the forecast |
| Volume Efficiency (VE) | PODy divided by % Volume, x 100 | PODy (x 100) per unit % Volume |

The relationship between PODy and 1-PODn for different algorithm thresholds is the basis for the verification approach known as "Signal Detection Theory" (SDT). This relationship can be represented for a given algorithm by the curve joining the (1-PODn, PODy) points for different algorithm thresholds. The resulting curve is known as the "Receiver Operating Characteristics" (ROC) curve in SDT. The area under this curve is a measure of overall forecast skill (e.g., Mason 1982), another measure that can be compared among the algorithms.

As shown in Table 3, two other statistics are utilized for verification of the turbulence forecasts: Impacted Volume and Volume Efficiency (VE). Impacted Volume measures (in three dimensions) the amount of airspace that is forecast to have turbulence, and is computed by summing all grid volumes with a Yes forecast. In general, Impacted Volume is expressed as % Volume, relative to the maximum volume possible. Because the analyses are limited to 20,000 ft and above, the total possible volume is about 64 million $km^3$. VE represents the % PODy per unit % Volume. While this statistic is useful for comparing algorithms, it also cannot be used alone. In particular, it is easy to obtain a large efficiency value when the Impacted Volume is small, even if PODy is also very small. An appropriate use of this statistic is to compare the efficiencies of forecasting systems with nearly equivalent values of PODy (e.g., see Brown et al. 1999a).

Emphasis will be placed on PODy, PODn, and % Volume. Use of this combination of statistics implies that the underlying goal of the algorithm development is to include most Yes PIREPs in the forecast "Yes turbulence" region, and most No PIREPs in the forecast "No turbulence" region (i.e., to increase PODy and PODn), while minimizing the extent of the forecast region, as represented by % Volume. All of these statistics must be used together, because each by itself can provide misleading results. Percent Volume and VE can be used to compare algorithms with similar PODy and PODn values.

### 5.3 *Stratifications*

Two categories of reported severity are considered: (i) reports of any turbulence severity (light and greater) and (ii) reports of MOG severity. The results also are stratified by forecast lead time. During the real-time portion of the study and in the initial post-analyses, the PIREPs also were stratified by aircraft weight (to make the observations more homogeneous) and lightning data (to eliminate reports that may be related to convective activity). However, these stratifications vastly reduced the PIREP dataset, but had little impact on the verification results. Thus, these stratifications are not considered in the analyses presented here.

## 6. RESULTS

The verification analyses were limited to dates and times when algorithm output and verification data were available for all algorithms. As a result, a total of 235 3-hr forecasts, 161 6-hr forecasts, and 106 9-hr forecasts were included.

Overall results for MOG PIREPs are shown in Fig. 1 for three groups of algorithms, for all forecasts with a lead time of 3 hr. The algorithms are divided into three groups to make the diagrams more clear. Group A includes ITFA, DTF3, DTF4, and DTF5; Group B includes Brown-1, Ellrod-1, and Ellrod-2; and Group C includes Dutton, Colson-Panofsky, Endlich, and Richardson number. The plots in Fig. 1 were created by combining all of the 3-hr forecasts (i.e., for all issue times). The figures include plots of PODy (MOG PIREPs) versus 1-PODn. Individual points on the algorithm curves represent the various thresholds used to create Yes/No forecasts. Curves for better forecasts are located closer to the upper lefthand corner of the diagrams.

The first impression from Fig. 1 is that, in general, the forecasting performance is very similar among the algorithms. However, some differences are apparent even in these crowded plots. In these figures, results for the AIRMETs can be used as a separator for the algorithm curves. Curves that approximately cross or lie just below the AIRMET point in Fig. 1 include Ellrod-1, Ellrod-2, DTF3, ITFA, and Richardson number.

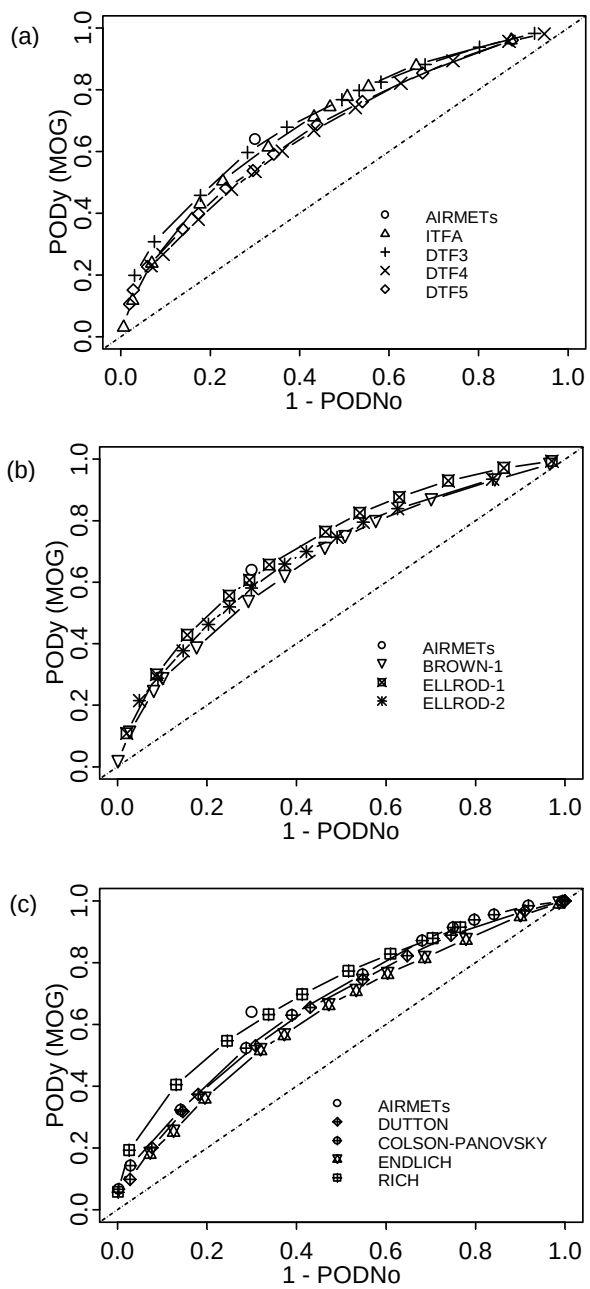


**Figure 1. ROC diagrams relating PODy to 1-PODn, for three groups of algorithms: (a) Group A; (b) Group B; and (c) Group C.**

All of the Group C algorithm curves lie below the AIRMET point.

The 3-hr results can be examined in greater depth by selecting appropriate, comparable thresholds for each algorithm and comparing the individual statistics among the algorithms. One rationale for this process is to select thresholds that lead to a PODy value that is approximately the same as the value attained by the AIRMETs. Table 4 shows the results of this exercise for the 3-hr forecasts. This table includes a variety of statistics associated with the specified thresholds.

Two values of PODy are included in Table 4 – one for All severities and one for MOG severities. In all cases, PODy (MOG) is slightly larger than PODy (All), which suggests that the MOG PIREPs are somewhat easier to capture than are PIREPs associated with less severe conditions. Two values of PODn also are included in Table 4 – one based on negative PIREPs, and the other based on AVAR data. Surprisingly, these two values of PODn are quite similar, even though the sources of the data are so different. For some algorithms (e.g., Brown-1), the value of PODn for the PIREPs is slightly larger, and in other cases (e.g., DTF5) the value for the AVARS data is slightly larger. However, the differences are always fairly small. The PODn values do, however, vary among the algorithms, with the largest values achieved by the AIRMETs, DTF5, Ellrod-1, ITFA, and Richardson number. Among the different forecasts and algorithms, the largest values of TSS are achieved by the AIRMETs, DTF3, Ellrod-1, Ellrod-2, ITFA, and Richardson number. In terms of VE, the best performance is achieved by the AIRMETs, Ellrod-1, Ellrod-2, and ITFA. The Richardson number has a relatively large % Volume value, and hence, a relatively small Volume Efficiency.

Thus, the results in Table 4 suggest that there are some discernible differences in the results among the algorithms, with the apparently best, all-around, algorithm performance associated with Ellrod-1 and -2, ITFA, and DTF3. Confidence intervals were computed for some of the statistics showing the largest differences, using methods identified by Kane and Brown (2000). These intervals indicate that none of the comparable statistics are significantly different from each other. For example, the 95% confidence intervals for the ITFA, Ellrod-1, and DTF3 values of PODn (based on PIREPs) in Table 4 are (0.63, 0.71) for ITFA, (0.60, 0.66) for DTF3, and (0.62, 0.70) for Ellrod-1. Because the intervals overlap, the apparent differences are not statistically significant.

Estimates of the area under the curve relating PODy (MOG PIREPs) to 1-PODn (i.e., the ROC curves) for each algorithm, for all three lead times, are shown in Table 5. Note that this statistic is not included for the AIRMETs since only one (1-PODn, PODy) point is associated with the AIRMETs. The results in Table 5 indicate that the curve areas are quite similar for all of the algorithms, with slightly larger values for DTF3, Ellrod-1, Ellrod-2, ITFA, and Richardson number. The results in Table 5 also suggest only a very slight decrease in skill with increased lead time. Finally, the curve area values for the Random Index all are approximately equal to 0.50; although not shown in Fig. 1, the PODy vs. 1-PODn curve for the random forecasts lies along the diagonal line in the ROC diagrams. These results indicate that the verification software was functioning correctly, and that the algorithms do have positive skill relative to a random forecast.

**Table 4. Verification statistics for all 3-hr forecasts (all issue times combined), for thresholds with PODy (MOG PIREPs) about the same as the PODy for AIRMETs. Larger values of TSS and VE are in boldface.**

| *Algorithm* | *Threshold* | *PODy (All)* | *PODy (MOG)* | *PODn (PIREPs)* | *PODn (AVARs)* | *TSS* | *VE* |
|---|---|---|---|---|---|---|---|
| **AIRMETs** | -- | 0.57 | 0.64 | 0.70 | 0.66 | **0.34** | **2.8** |
| **Brown-1** | $9x10^{-5}$ | 0.58 | 0.62 | 0.63 | 0.58 | 0.25 | 2.3 |
| **Colson-Panofsky** | -1000 | 0.58 | 0.63 | 0.61 | 0.53 | 0.24 | 1.9 |
| **DTF3** | 0.7 | 0.63 | 0.68 | 0.63 | 0.63 | **0.31** | 2.3 |
| **DTF4** | 2.5 | 0.63 | 0.67 | 0.57 | 0.60 | 0.24 | 2.1 |
| **DTF5** | 0.15 | 0.56 | 0.59 | 0.66 | 0.68 | 0.25 | 2.5 |
| **Empirical Dutton** | 25 | 0.62 | 0.66 | 0.57 | 0.57 | 0.23 | 2.2 |
| **Ellrod-1** | $4x10^{-7}$ | 0.61 | 0.66 | 0.66 | 0.64 | **0.32** | **2.9** |
| **Ellrod-2** | $4x\ 10^{-7}$ | 0.61 | 0.66 | 0.63 | 0.59 | **0.29** | **2.8** |
| **Endlich** | 0.225 | 0.63 | 0.66 | 0.53 | 0.55 | 0.19 | 2.1 |
| **ITFA** | 0.13 | 0.57 | 0.62 | 0.67 | 0.65 | **0.29** | **2.6** |
| **Richardson #** | 4 | 0.58 | 0.63 | 0.66 | 0.64 | **0.29** | 2.4 |

**Table 5. Areas under the ROC curves for all algorithms, for three lead times (all issue times combined).**

| | *Lead time (hr)* | | |
|---|---|---|---|
| *Algorithm* | *3 hr* | *6 hr* | *9 hr* |
| Brown-1 | 0.67 | 0.67 | **0.69** |
| Colson-Panofsky | 0.67 | 0.66 | 0.66 |
| DTF3 | **0.71** | **0.70** | **0.70** |
| DTF4 | 0.67 | 0.67 | 0.66 |
| DTF5 | 0.67 | 0.67 | 0.67 |
| Empirical Dutton | 0.65 | 0.65 | 0.64 |
| Ellrod-1 | **0.72** | **0.71** | **0.71** |
| Ellrod-2 | **0.69** | **0.69** | **0.69** |
| Endlich | 0.63 | 0.62 | 0.60 |
| ITFA | **0.70** | **0.69** | **0.68** |
| Random | 0.50 | 0.50 | 0.49 |
| Richardson number | **0.70** | **0.69** | **0.68** |

## 7. CONCLUSIONS AND FUTURE WORK

In general, differences found thus far among the performance characteristics of the various algorithms are relatively small, except for certain differences that stand out. The Brown-1, Colson-Panofsky, Dutton, and Endlich algorithms generally exhibited poorer performance than the other algorithms. Algorithms that performed the best overall include DTF3, Ellrod-1 and -2, and ITFA. However, differences between the PODy and PODn values for different algorithms were, in general, not statistically significant. These results will provide a baseline for future development of turbulence forecasting algorithms.

It is important to remember that the results presented here are based on a single season. A second turbulence algorithm intercomparison exercise took place during the winter of 2000. The winter 2000 exercise again included a real-time component through the RTVS. As part of this exercise, the RTVS capabilities were enhanced. For example, RTVS now includes time series plots (which can be specified by the user) showing results accumulated across several days. Fig. 2 shows an example of a time series plot from the winter 2000 real-time exercise. These types of plots allow a closer look at short-term variations in the statistics and will enable identification of situations in which one algorithm performs better than another.

The winter 1998-99 and winter 2000 intercomparison analyses will be combined and extended in the near future. Additional algorithms will be evaluated, including a statistically-based version of ITFA, as well as other basic algorithms. PIREPs that were recently obtained from Northwest Airlines also will be used to enhance the PIREP dataset. Finally, available feature detectors (e.g., jet stream, and possibly mountain wave) will be applied to the forecasts to determine the effects of these features on the verification results.

This intercomparison exercise not only developed a baseline for turbulence algorithm development, but also tested the robustness of the verification methods. Comparisons of the statistical results generated by the RTVS and the post-analysis indicate that the results are somewhat sensitive to the method used to match turbulence forecasts to the observations. However, despite the methodological and data differences between the systems, the basic conclusions are consistent between the real-time and post-analysis results. Thus, the combination of real-time and post-analysis provides a robust evaluation of the quality of the turbulence forecasts.

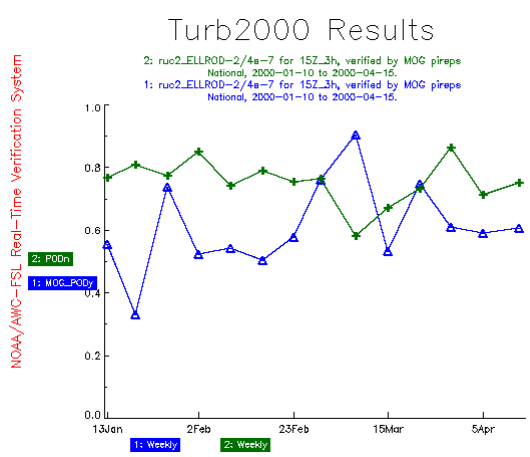


**Figure 2. Time series of weekly values of PODy (triangles) and PODn (+'s), for Ellrod-2 index with a threshold of $4x10^{-7}$, for the period 10 January to 15 April 2000, provided by the RTVS (http://www-ad.fsl.noaa.gov/afra/rtvs/RTVS-project_des.html).**

## ACKNOWLEDGMENTS

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). 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 thank the algorithm developers – particularly Bob Sharman of NCAR and Adrian Marroquin and Cecilia Girz of FSL – for their support of this effort and for their help in getting the project underway. We also greatly appreciate the efforts of Gerry Wiener (NCAR), Sue Dettling (NCAR), Missy Petty (NCAR), and Denise Walker (FSL) for making the algorithm output available during the real-time portion of the project and for the on-going re-computation of some of the fields.

## REFERENCES

Benjamin, S.G., J.M. Brown, K.J. Brundage, B.E. Schwartz, T.G. Smirnova, and T.L. Smith, 1998: The operational RUC-2. Preprints, 16th Conference on Weather Analysis and Forecasting, American Meteorological Society, Phoenix, 249-252..

Brown, B.G., T.L. Kane, R. Bullock, and M.K. Politovich, 1999a: Evidence of improvements in the quality of in-flight icing algorithms. *Preprints, 8th Conference on Aviation, Range, and Aerospace Meteorlogy*, Dallas, TX, 10-15 January, American Meteorological Society (Boston), 48-52.

Brown, B.G. and J.L. Mahoney, 1998: Verification of Turbulence Algorithms. Report, Research Applications Program, National Center for Atmospheric Research, and Forecast Systems Laboratory, Environmental Research Laboratories, NOAA, 9 pp.

Brown, B.G., J.L. Mahoney, R. Bullock, J. Henderson, and T.L. Kane, 1999b: Turbulence Algorithm Intercomparison: 1998-99 Initial Results. Report, available from B. Brown, bgb@ucar.edu, 64 pp.

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*, 11-15 Sept., Orlando, FL, American Meteorological Society (Boston).

Brown, R., 1973: New indices to locate clear-air turbulence. *Meterol. Mag*., **102**, 347-361.

Colson, D. and H. A. Panofsky, 1965: An index of clear air turbulence. *Quart. J.Roy. Meteor. Soc*., **91**, 507-513.

Drazin, P.G. and W.H. Reid, 1981: **Hydrodynamic Stability**. Cambridge, 527 pp.

Dutton, J. and H. A. Panofsky, 1970: Clear Air Turbulence: A mystery may be unfolding. *Science*, **167**, 937-944.

Dutton, M.J.O., 1980: Probability forecasts of clear-air turbulence based on numerical model output. *Meterol. Mag.,* **109**, 293-310.

Ellrod, G.P. and D.I.Knapp, 1992: An objective clear-air turbulence forecasting technique: verification and operational use. *Wea. Forecasting*, **7**, 150-165.

Endlich, R. M., 1964: The mesoscale structure of some regions of clear-air turbulence. *Mon. Wea. Rev*., **92**, 261-276.

Kane, T.L., and B.G. Brown, 2000: Confidence intervals for some verification measures – a survey of several methods. *Preprints, 15th Conference on Probability and Statistics in the Atmospheric Sciences*, 8-11 May, Asheville, NC, American Meteorological Society (Boston), 46-49.

Kronebach, G. W., 1964: An automated procedure for forecasting clear-air turbulence. *J. App. Met.,* **3,** 119-125.

Mahoney, J.L., 1998: Statistical comparisons between four different time windows. Report, Forecast Systems Laboratory, Environmental Research Laboratories, NOAA, 4 pp.

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, American Meteorological Society, J26-J31.

Marroquin, A., 1995: An integrated algorithm to forecast CAT from gravity wave breaking, upper fronts and other atmospheric deformation regions. *Preprints, 6th Conference on Aviation Weather Systems*, Dallas, TX, American Meteorological Society, 509-514.

Marroquin, A., 1998: An advanced algorithm to diagnose atmospheric turbulence using numerical model output. *Preprints, 16th Conference on Weather Analysis and Forecasting*, Phoenix, AZ, 11-16 January, American Meteorological Society.

Mason., I., 1982: A model for assessment of weather forecasts. *Australian Meteorological Magazine*, **30**, 291-303.

McCann, D. W., 1997: A "novel" approach to turbulence forecasting. Preprints, *Seventh Conf. On Aviation, Range and Aerospace Meteorology*, 158-163. American Meteorological Society, Long Beach, CA.

NWS, 1991: National Weather Service Operations Manual, D-22. National Weather Service. (Available at Website http://www.nws.noaa.gov).

Reap, R.M., 1996: Probability forecasts of clear-air-turbulence for the contiguous U.S. *Technical Procedures Bulletin No. 430*, NWS Office of Meteorology 15 pp.

Sharman, R.C., B.G. Brown, and S. Dettling, 2000: Preliminary results of the NCAR Integrated Turbulence Forecasting Algorithm (ITFA) to forecast CAT. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, 11-15 Sept., Orlando, FL , American Meteorological Society (Boston).

Tebaldi, C., D. Nychka, B.G. Brown, and R. Sharman, 2000: Forecasting clear-air turbulence. Submitted to *Technometrics*.

Wilks, D.S., 1995: **Statistical Methods in the Atmospheric Sciences**, Academic Press, 467 pp