

Turbulence Algorithm Intercomparison: Winter 2001 Results

**Jennifer L. Mahoney¹, Barbara G. Brown³,
Randy Bullock³, Tressa L. Fowler³,
Chris Fischer^{1,2}, Judy Henderson¹, and Beth Sigren^{1,2}**

August 31, 2001

¹ Forecast Systems Laboratory, Environmental Research Laboratories, National Oceanic and Atmospheric Administration, Boulder, CO

² Joint collaboration with the Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO

³ Research Applications Program, National Center for Atmospheric Research, Boulder, CO

Turbulence Algorithm Intercomparison: Winter 2001 Results

Abstract. This report summarizes basic results of a third intercomparison of the capabilities of a number of clear-air turbulence (CAT) forecasting algorithms to predict the locations of CAT. The algorithms considered in the study include most of the algorithms that were included in the first two intercomparisons, which took place during winter 1998-99 and 2000, as well as several additional algorithms. The algorithm forecasts are based on output of the Rapid Update Cycle (RUC-2) numerical weather prediction model for the period from 8 February through 31 March 2001. Forecasts issued at 1200, 1500, 1800, and 2100 UTC, with 3-, 6-, 9-, and 12-h lead times were included in the study. The evaluation also includes the turbulence AIRMETs, the operational turbulence forecast product that is issued by the NWS Aviation Weather Center (AWC) and is limited to the continental United States and to two altitude bands: above 20,000 ft (as in TURB98-99 and TURB2000) and from 15,000 to 20,000 ft. In addition, the altitude band from 15-42,000 ft was considered in post-analysis.

The forecasts were verified using Yes and No turbulence observations from pilot reports (PIREPs). The algorithms were evaluated as Yes/No turbulence forecasts by applying a threshold to convert the output of each algorithm to a Yes or No value. A variety of thresholds were applied to each algorithm. The verification analyses were primarily based on the algorithms' ability to discriminate between Yes and No observations, as well as the extent of their coverage.

The study was comprised of two components. First, the algorithms were evaluated in near real time by the Real-Time Verification System (RTVS) of the NOAA Forecast Systems Laboratory (FSL), with results displayed through a graphical user interface on the World Wide Web (<http://www-ad.fsl.noaa.gov/afra/rtvs; link turbulence>). Second, the verification results were re-evaluated in greater depth in post-analysis, using a post-analysis verification system at the National Center for Atmospheric Research (NCAR), with additional thresholds applied to each algorithm to provide a more complete depiction of algorithm quality. Only initial basic results of the post-analysis are presented here.

The results of the TURB2001 intercomparison suggest that overall algorithm performance at the 6-h lead between BROWN-2, DTF3, DTF5, Dutton, Ellrod-1 and -2, and ITFA is similar. This finding is consistent with TURB98-99 and TURB2000 exercises. However, on-going statistics for ITFA show improvement over DTF3, Ellrod-1, and AIRMETs in the summer. In addition, turbulence generated by the algorithms usually covers 15 to 40 % of the forecast domain, and is independent of the algorithms ability to capture the turbulence events. The ability for the AIRMETs to capture the turbulence is related to the volume of the domain that is covered by the AIRMET. This result may be an artifact of the volumetric manner in which the AIRMETs are issued. The results by height for ITFA, DTF3, and Ellrod-1 were nearly identical. However, differences between the algorithms and the AIRMETs occurred at 30,000 and 40,00 ft.

1. Introduction

This report summarizes basic results of an intercomparison of the forecasting capability of various clear-air turbulence (CAT) forecasting algorithms. This intercomparison took place from 8 February – 31 March 2001 with some ongoing analysis from 10 January 2000 – 31 March 2001 provided by RTVS, and is the third in a series of evaluations of the algorithms' forecasting performance. The first intercomparison took place during the winter of 1998-99; results of that evaluation are presented in Brown et al. (1998, 1999, 2000a) with the second taking place during the winter of 2000; results of that evaluation are presented in Brown et al. (2000b and 2000c) and Mahoney and Brown (2000). Each of the turbulence algorithm intercomparisons were sponsored by the Turbulence Product Development Team (PDT) of the Federal Aviation Administration's (FAA's) Aviation Weather Research Program (AWRP).

Purposes of the winter 2001 intercomparison (hereafter, denoted TURB2001) were to (i) develop and monitor the baseline for the quality of current CAT forecasting algorithms; (ii) consider the consistency of the verification statistics from year to year; (iii) demonstrate to-date progress in the development of these forecasting tools; (iv) examine the strengths and weaknesses of the algorithms; and (v) perform an evaluation that is independent, consistent, comprehensive, and fair. Except for the second goal, all of these goals are the same as the goals for the winter 1998-99 and winter 2000 intercomparison (hereafter, denoted TURB98-99 and TURB2000). To meet the first goal, a number of different CAT algorithms were included in the study, as were the operational turbulence forecasts, or Airmen's Meteorological Advisories (AIRMETs), that are produced by the National Weather Service's (NWS's) Aviation Weather Center (AWC). The second goal will be met by comparing the results for the three winters. To meet the third goal, algorithms that have been developed over the last several years, with support of the AWRP, were included. The fourth goal will be met through the analyses presented in this report, as well as on-going studies of the results by the Quality Assessment Group (QAG) and by the algorithm developers. Finally, the fifth goal was met by pre-defining the verification methods and other features of the intercomparison, with approval by all members of the Turbulence PDT. In addition, the intercomparison and analyses of the results were the responsibility of the QAG, which includes the verification groups of the NOAA Forecast Systems Laboratory (FSL) and the National Center for Atmospheric Research Applications Program (NCAR/RAP), rather than the responsibility of the individual algorithm developers.

The study included two major facets: (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 through a graphical user interface on the World-Wide Web; and (ii) a post-analysis component in which the verification data were re-generated and examined in detail at NCAR and FSL. This report summarizes the displays and analyses that were presented by RTVS, including upgrades to that system that were implemented as a result of this project. Initial results from the post-analysis are presented here.

The report is organized as follows. The study approach is presented in Section 2. Section 3 briefly describes the algorithms that were included in the evaluation, and the data that were utilized are discussed in Section 4. Results of the real-time study are presented in Section 5, with initial results from the post-analysis presented in Section 6. Finally, Section 7 includes the summary and conclusions. The verification methods are described in the Appendix A.

2. Approach


A total of 14 CAT algorithms were included in TURB2001. Most of these algorithms also were included in TURB98-99 and TURB2000, but some changes occurred in TURB2001. For instance, Brown-2, Ellrod-2, Gravity Wave Breaking (GWB), Horizontal Shear, and Temperature Gradient were added to RTVS and post-analysis. CCAT, Endlich, Random and Brown-1 were excluded from both systems in the TURB2001 exercise.  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, 9, and 12 hours, out to a valid time of 0000 UTC, were included in the study, as shown in Table 1. In addition, turbulence AIRMETs, which are the operational turbulence forecasts issued by the National Weather Service's Aviation Weather Center (NWS/AWC) were included for comparison purposes. Due to the emphasis placed on forecasting upper-level CAT, the evaluation was mainly limited to the region of the atmosphere above 20,000 ft, as was the case in TURB98-99 and TURB2000. An extension was added in TURB2001 where the forecasts were also evaluated separately from 15,000 to 20,000 ft in RTVS and from 15,000 to 42,000 ft in post-analysis.

Table 1. Issue, lead, and valid times included in TURB2000.

Issue time (UTC)	Lead times (hr)	Valid times (UTC)
1200	3, 6, 9, 12	1500, 1800, 2100, 0000
1500	3, 6, 9	1800, 2100, 0000
1800	3, 6	2100, 0000
2100	3	0000

TURB2001 began on 8 February and ended on 31 March 2001. The verification approach is identical to the approach taken in TURB98-99 and TURB 2000, except that a few additional metrics and graphics that were added to the RTVS displays. A description of the verification methods is listed in Appendix A. The algorithm forecasts and AIRMETs were verified using Yes and No PIREPs of turbulence.



As in TURB2000, A “forecaster evaluation” of algorithm performance also was included in TURB2001. This subjective evaluation will be summarized in a separate report.

3. Algorithms

The set of algorithms that was evaluated in TURB2001 differed slightly from the set that was considered in TURB98-99 and TURB2000. Specifically, Brown-2, Ellrod-2, GWB, Horizontal Shear, and Temperature Gradient were added to RTVS and post-analysis. CCAT, Endlich, Random, and Brown-1 were excluded from the TURB2001 exercise due to poor performance as was identified in Turb2000. The algorithms that were included in the TURB2001 are described briefly in the following paragraphs. Further information about the algorithms and their development can be found in the references that are provided.

Brown-1: This index is a simplification of the Ri tendency equation originally derived by Roach (1970). The simplifications involve use of the thermal wind relation, the gradient wind as an approximation to the horizontal wind, and finally some empiricism (Brown 1973).

Brown-2: This is an extension of Brown-1 to provide a measure of turbulence intensity as expressed as an eddy dissipation rate (Brown 1973).

DTF3 and 5: The DTF (“Diagnostic Turbulence Formulation”) algorithms were developed to take into account several sources of turbulent kinetic energy in the atmosphere (e.g., upper fronts), with the output in terms of tke (Marroquin 1995, 1998). These algorithms are related to one another, with the algorithm associated with DTF5 incorporating greater complexity.

Dutton: This index is based on linear regression analyses of a pilot survey of turbulence reports over the North Atlantic and NW Europe during 1976 and various synoptic scale turbulence indices produced from the then-operational UK Met Office forecast model (Dutton 1980). The result of the analyses was the “best fit” of the turbulence reports to meteorological outputs for a combination of horizontal and vertical wind shears.

Ellrod-1: This index was derived from simplifications to the frontogenetic function. As such it depends mainly on the magnitudes of the potential temperature gradient, deformation and convergence (Ellrod and Knapp 1992).

Ellrod-2: Ellrod-2 is similar to Ellrod-1 except it also includes a term to account for convergence (Ellrod and Knapp 1992).

Gravity Wave Breaking (GWB): GWB is an abbreviation for the Gravity Wave Breaking algorithm. This uses a computation of divergence of Reynold’s stress over mountainous regions as an indicator of potential mountain-induced gravity wave breaking and therefore turbulence. It is an adaptation of the algorithm described in Palmer et al. (1986)

to account for the effect of gravity wave drag on the general circulation of the atmosphere.

Horizontal Shear: This is the horizontal gradient of temperature on a constant theta surface. It is a measure of deformation and also vertical wind shear from the thermal wind relation. This technique was recommended by Delta Airlines forecasters as a good indicator for turbulence locations (Dutton 1980)

ITFA : The ITFA (Integrated Turbulence Detection and Forecasting Algorithm) forecasting technique uses fuzzy logic to integrate available turbulence observations (in the form of PIREPs and AVAR data) together with a suite of turbulence diagnostic algorithms (a superset of algorithms used in the verification exercise and others) to obtain the forecast (Sharman et al. 1999, 2000). This algorithm is under development by the Turbulence PDT; the version included in this exercise is an early version of the algorithm.

Mwave: MWAVE is a mountain wave diagnostic developed by the Experimental Forecast Facility (EFF) at the Aviation Weather Center. MWAVE computes two diagnostics. First is the strength of the wave which MWAVE estimates as the drag the mountain wave exerts on the atmosphere. Second is the breaking potential which MWAVE estimates as a non-dimensional wave amplitude. Mountain waves may be strong but non-breaking, as evidenced by an aircraft experiencing a smooth ride but significant up-and-downdrafts. They may also be breaking but weak with barely noticeable turbulence. Additional information is available at <http://www.awc-kc.noaa.gov/awc/help/mwaveinfo.html>.

Richardson Number: Theory and observations have shown that at least in some situations patches of CAT are produced by what is known as Kelvin-Helmholtz (KH) instabilities. This occurs when the Richardson number (Ri), the ratio of the local static stability to the local shears, becomes small. Therefore, theoretically, regions of small Ri should be favored regions of turbulence (Drazin and Reid 1981; Dutton and Panofsky 1970; Kronebach 1964).

SCATR: This index is based on attempts by several investigators to forecast turbulence by using a time tendency (i.e., prognostic) equation for the Richardson number (Roach 1970). The version used in this study was based on a formulation of this equation in isentropic coordinates by John Keller, who dubbed the algorithm “SCATR” (Specific CAT Risk; Keller 1990).

Temperature Gradient: This is the horizontal gradient of temperature on a constant theta surface. It is a measure of deformation and also vertical wind shear from the thermal wind relation. This technique was recommended by Delta Airlines forecasters as a good indicator for turbulence locations.

Ulturb: The ULTurb (Upper-Level Turbulence) forecasting index was developed by Don McCann (1997) from the AWC. It attempts to correlate unbalanced (ie. nongeostrophic) flow to regions of clear-air turbulence. Three different measures of this imbalance are

computed, and the maximum of these relates to turbulence potential. The correlation between unbalanced flows and turbulence is supported at least qualitatively from numerous field experiments, both over the continental U.S. and the N. Pacific (Knox 1997).

Vertical Wind Shear: Wind shear has been known to be a destabilizing force from the time of Helmholtz. This can be seen from its inverse relation to Richardson's number: large values favor small Ri , which in turn produce turbulence in stratified fluids (Drazin and Reid 1981; Dutton and Panofsky 1970).

4. Data

As in TURB98-99 and TURB2000, the data that were used in TURB2001 include model output, PIREPs, and lightning. These data were obtained and used in near-real-time by the RTVS, and they were obtained and archived for use in post-analysis at NCAR.

Model output was obtained from the RUC-2 model, which is run operationally at NOAA's National Centers for Environmental Prediction, Environmental Modeling Center. 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 RUC-2 assimilates data from commercial aircraft, wind profilers, rawinsondes and dropsondes, surface reporting stations, and numerous other data sources. The model produces forecasts on an hourly basis; however, only the forecast and lead time combinations listed in Table 1 were used in this study. Fig. 1 depicts the RUC-2 domain and horizontal resolution. The verification analyses were limited to the domain covered by the AIRMETs, which also is shown in Fig. 1.

Algorithms were applied to the model output files to create algorithm output files. This part of the process was undertaken by the algorithm developers – the DTF forecasts were computed at FSL, Mwave and Ulturb were computed at AWC, and all of the other forecasts were computed at NCAR. As part of this process, the algorithm output data were interpolated to flight levels (i.e., every 1,000 ft) rather than the raw model levels.

All available Yes and No turbulence PIREPs were included in the study. 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. Information about turbulence type (e.g., "Chop," "CAT") frequently is missing, and was ignored.

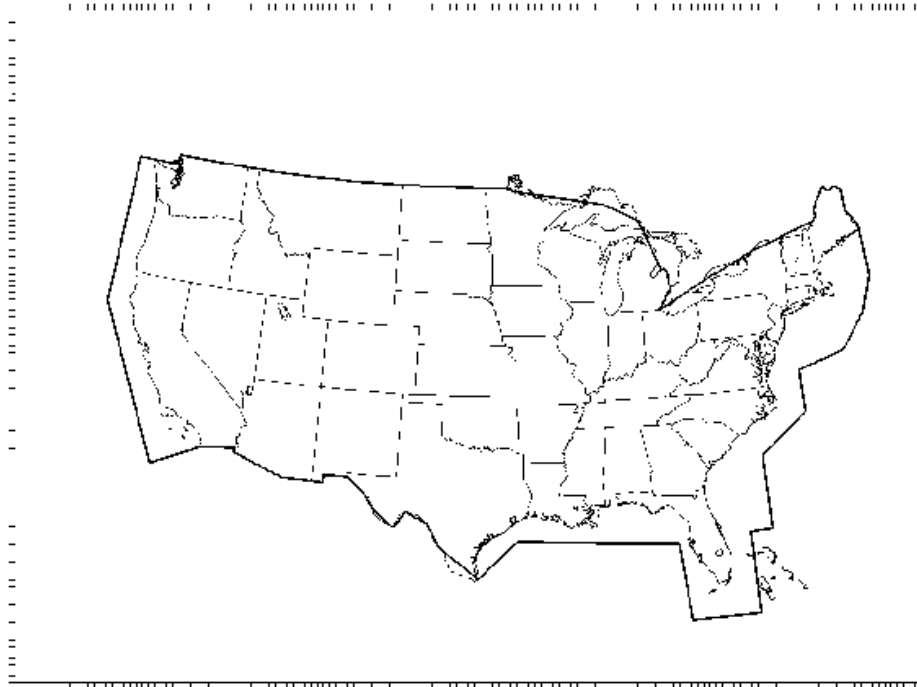


Figure 1. RUC-2 domain. Tics on the edges of the frame identify the model grid lines; dark outline around continental U.S. denotes the total domain of the AIRMETS.

Finally, lightning data were obtained from the National Lightning Data Network (Orville 1991). These data were used to identify PIREPs that were likely to be associated with convection (see Appendix A).

5. Results

The results presented in this Section for the non-convective PIREP category reporting moderate-or-greater (MOG) severities (hereafter MOG PODy), are limited to the 6-h lead from the 1200, 1500 and 1800 UTC issue times. These periods were chosen to correspond to those times used most often as forecast guidance by the forecasters at AWC (Mahoney and Brown 2000). Refer to the Web for results from the other time periods.

5.1. Overview results

Overall results for all turbulence algorithms and the AIRMETS included in the TURB2001 evaluation are presented in the comparisons plots shown in Figs. 2 - 6. Panel A of the Figs. 2 - 6 include the following algorithms: Brown-2, DTF3, DTF5, Dutton, Ellrod-1, Ellrod-2, and ITFA. Panel B of Figs. 2 - 6 include the following algorithms: Mwave, Richardson Number, Shear, Gravity Wave Breaking (GWB), Ulturb, Horizontal Shear, and Temperature Gradient. The AIRMETS are included on both panels.

The statistics were computed for the National domain as in TURB 1998-99 and TURB2000, but were also generated on several smaller domains (i.e., East, Central, West, and Mountain region) to determine if geographic location impacts algorithm performance. Each line on the comparison plots shown in Figs. 2 - 6, represents a particular MOG PODy value with respect to 1- PODn for one particular algorithm. Each symbol on the line represents the statistic at a particular algorithm threshold. Typically, a low threshold will produce turbulence forecasts covering the entire domain while higher values of the threshold limit turbulence to specific well-defined regions. The ultimate goal for improved forecasting performance is to maintain a reasonable 1-PODn while improving the PODy, (i.e., moving closer to the upper left hand corner of the PODy vs. 1-PODn plots). The AIRMETs are represented in the algorithm comparison plots by a single value. Comparison plots of MOG PODy and % Volume can be obtained from the Web-based displays and from the post-analysis results presented in Section 6.

As shown in each Figs. 2 - 6., overall, the algorithms in Panel A perform better than those in Panel B. For instance, the algorithms presented in Panel A, are tightly grouped with very little change in performance between the algorithms. This result is consistent with the results provided in TURB98-99 and TURB2000 and for the smaller domains. For instance, the differences between the A-Panel of algorithms on the National domain (Fig. 2) are similar to the statistics computed for those algorithms for the East (Fig. 3) or West (Fig. 5) domains. The arc of the lines on the East domain (Fig. 3), at lower values of 1- PODn, for Panel A is slightly higher than the other domains.

Although the algorithms presented in Panel A are overall better performers than the Panel B algorithms, some interesting differences are evident in the B-Panel of algorithms. For instance, the performance of the B-Panel of algorithms changes dramatically depending on region. Specifically, GWB is the best performer of Panel B (and a top performer when compared to the algorithms in the A-Panel) in the Central region (Fig. 4), but is the worst performer in all other regions (Fig. 2, 3, 5 and 6). The top performers to capture mountain wave turbulence (Fig. 6) include Ellrod-1, ITFA, and Mwave (with other Panel-A algorithms close behind).

In each domain, the AIRMETs out perform the algorithms where at a particular value of 1- PODn, higher values of MOG PODy are recorded (Fig. 2- 6). The largest difference between the statistical values for AIRMETs and the algorithms occurs on the Central domain (Fig. 4) where the MOG PODy value for GWB is closest to the AIRMETs. Interestingly however, the performance of the AIRMETs differs from domain to domain. The largest values of MOG PODy for the AIRMETs are computed for the National (Fig. 2) and Central (Fig. 4) domains where the lowest 1- PODn values are computed for the Mountain Wave (Fig. 6) domain.

6hr National
MOG Non-Convective PIREP Locations
010208 010331
21,000-41,000 ft

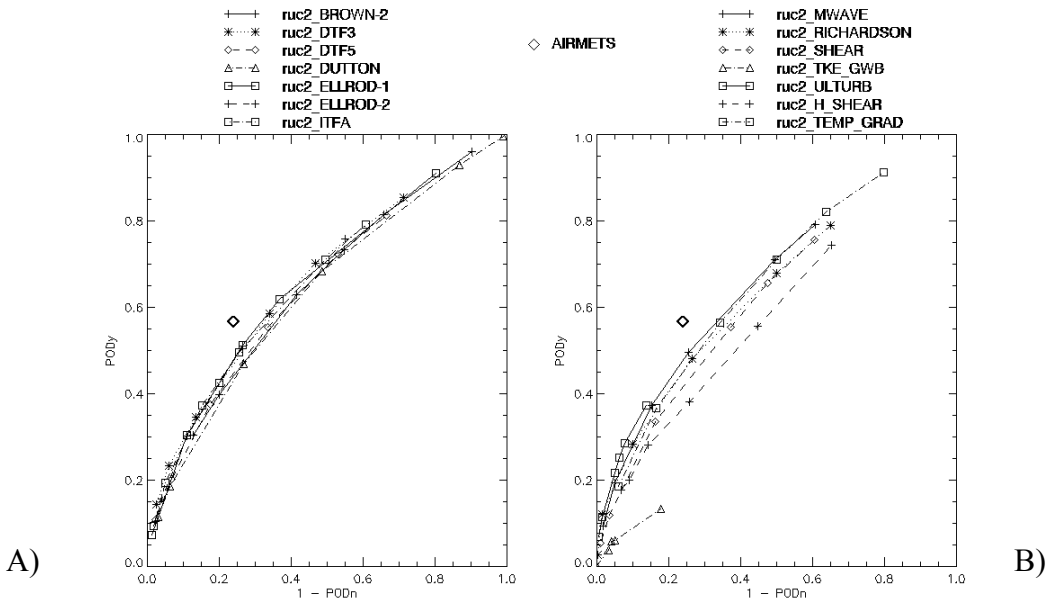


Figure 2. Two panels for 8 February – 31 March 2001 for 6-h lead all issue times combined, for MOG non-convective PIREPS are displayed for algorithm panel A and B, MOG PODy vs. 1- PODn, with each panel containing 7 of the 14 algorithms for the National domain. Each shape represents the MOG PODy and 1-PODn for a particular algorithm. The line segments connect the results for different thresholds for a particular algorithm. The AIRMETs results are represented by a single point (‘diamond’) on the plots.

6hr East
MOG Non-Convective PIREP Locations
010208 010331
21,000-41,000 ft

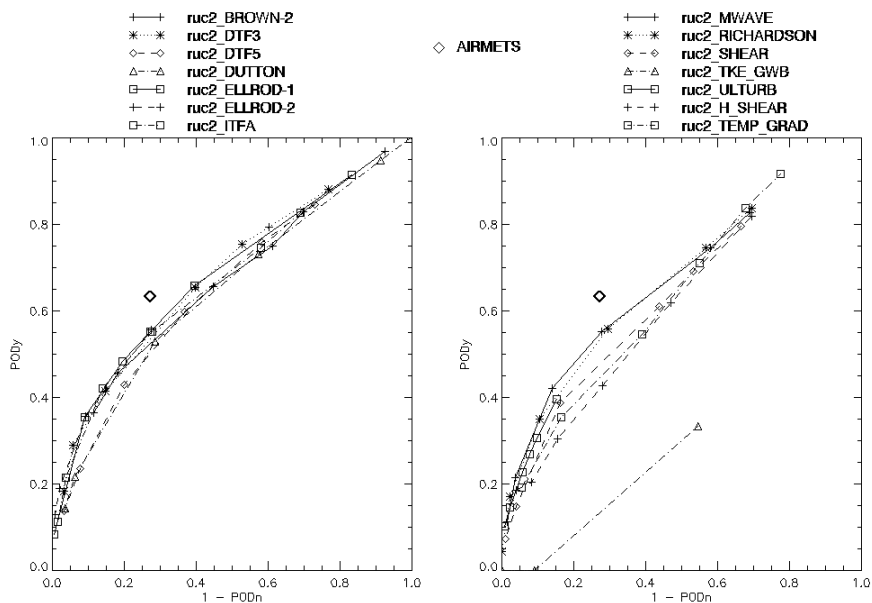


Figure 3. As in Fig. 2, except for East domain.

6hr Central
 MOG Non-Convective PIREP Locations
 010208 010331
 21,000-41,000 ft

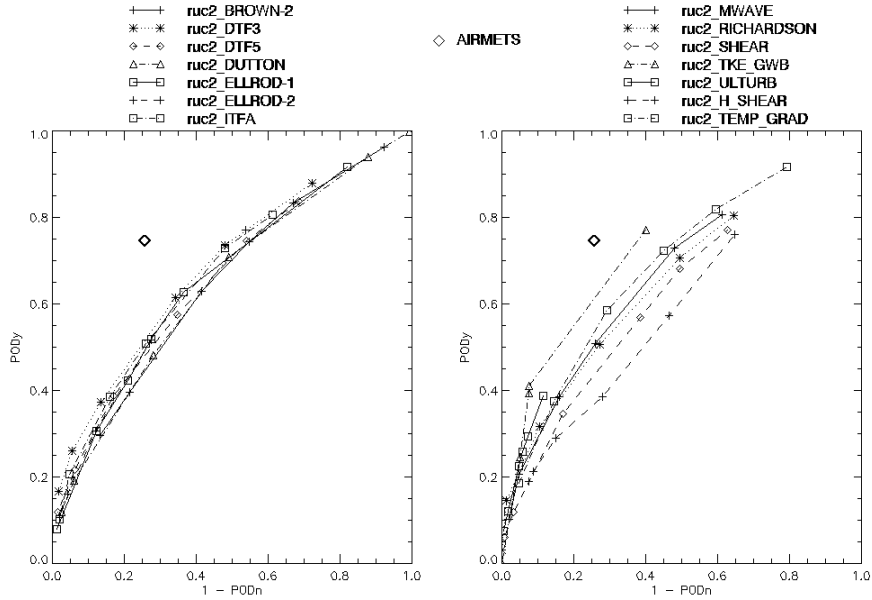


Figure 4. As in Fig. 2, except for Central domain.

6hr West
 MOG Non-Convective PIREP Locations
 010208 010331
 21,000-41,000 ft

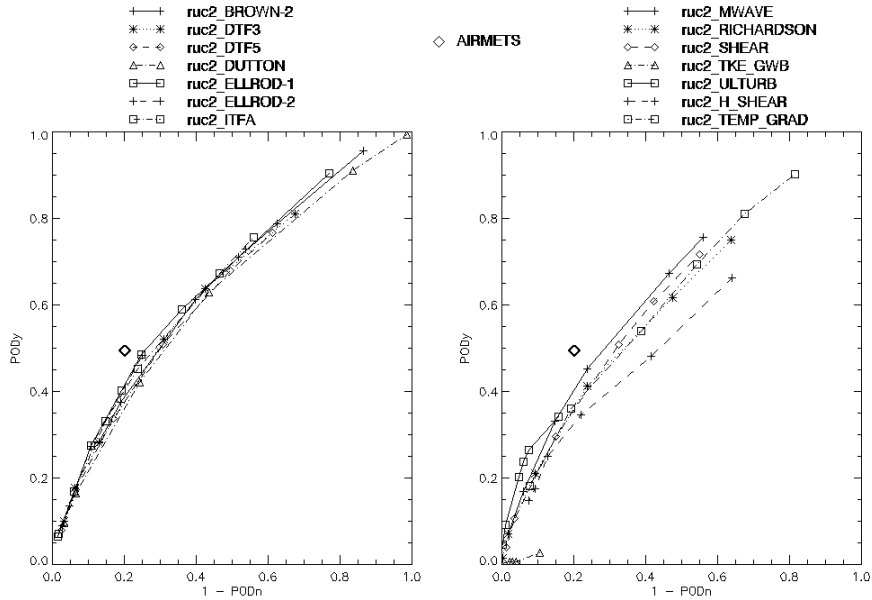


Figure 5. As in Fig. 2, except for West domain.

6hr Mountain
MOG Non-Convective PIREP Locations
010208 010331
21,000-41,000 ft

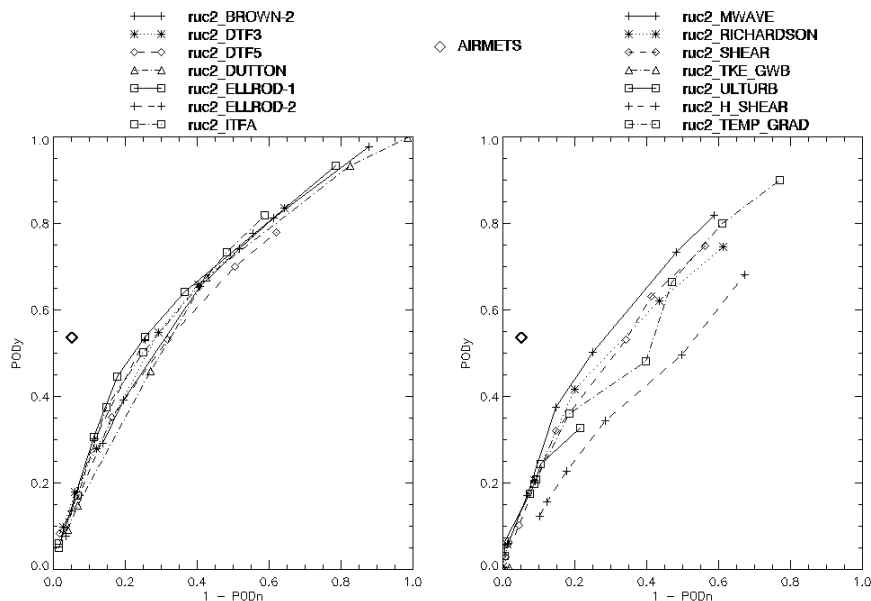


Figure 6. As in Fig. 2, except for Mountain domain.

5.2. General comparisons between ITFA, Ellrod-1, DTF3, and AIRMETs

A continuous baseline of statistical results was obtained through RTVS over a 14-month period from 10 January 2000 – 31 March 2001 for ITFA, DTF3, Ellrod-1, and AIRMETs. During this period, only minor changes in ITFA occurred, while the other forecasts remained the same. Statistical results from the four forecasts presented in this Section were chosen because they were identified in TURB2000 as the “best” overall performers (Mahoney and Brown 2000) and a long consistent baseline of statistical results was available.

The overall results for ITFA, DTF3, AIRMETs, and Ellrod-1 are summarized in the time series plots shown in Figs. 7 –9. The statistics were generated by combining the pairs for all issue times (1200, 1500, and 1800 UTC) at the 6-h lead. The algorithm verification results were filtered to select the threshold for each algorithm that typically produced an overall MOG PODy value between 0.5 and 0.6. Each line, shown in Figs. 7 – 9, represent a single turbulence algorithm at that specific threshold. Each symbol on the line represents a statistic generated from a monthly accumulation of Yes/No pairs.

Turb2000–2001 Results

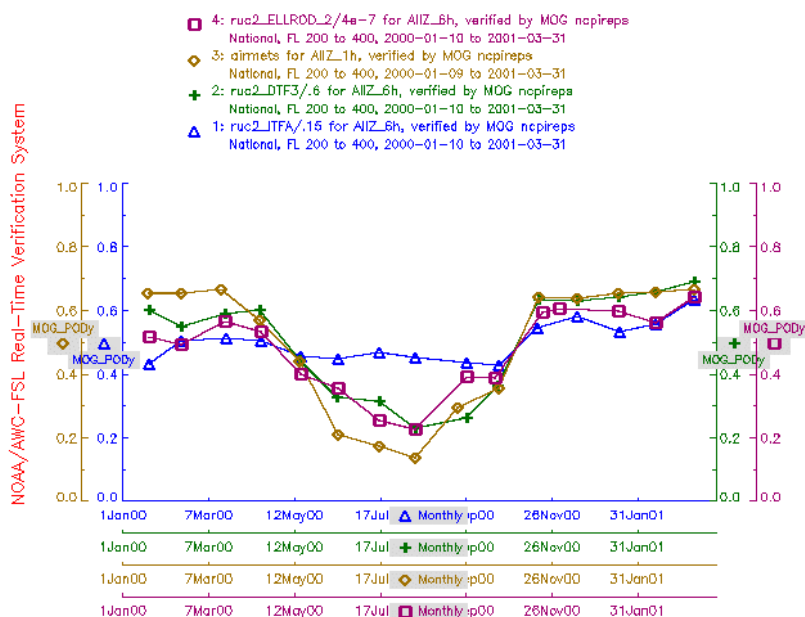
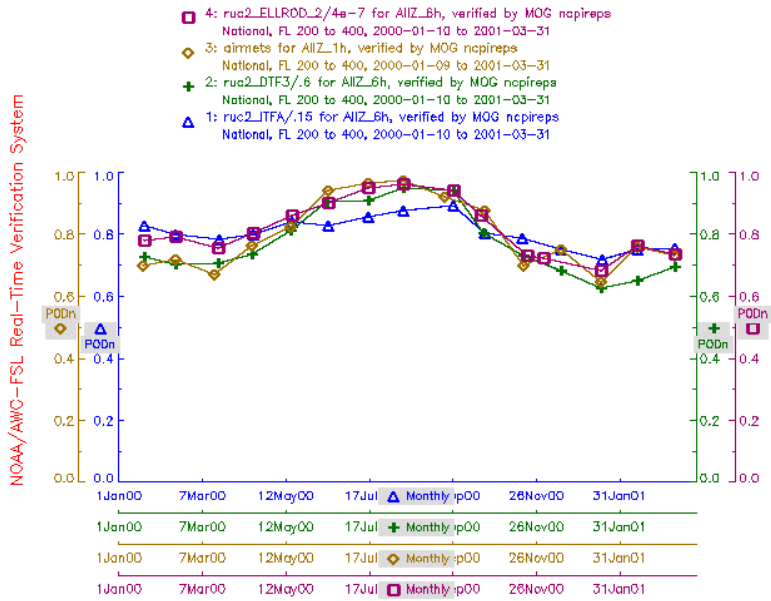


Figure 7. Time series plot for 6-h lead, all issue times combined for ITFA/.15 (triangle), DTF3/.6 (+), AIRMETs (diamond), and Ellrod-1/4e-7 (square) by month for 10 January 2000 – 31 March 2001 for MOG PODy.

Inspection of Fig 7 indicates that larger values of MOG PODy for the forecasts are present during the winter months (January – March) than during the summer months (April – September). For example, the MOG PODy for all forecasts during the winter range from 0.40 to 0.70, while in the summer the MOG PODy values range from 0.2 to 0.45. The performance in all of the forecasts improved slightly in the winter of 2001 as compared to 2000. However, this improvement could be due to changes in the weather from season to season. During the winter, the AIRMETs seem to perform better than the other algorithms with larger MOG PODy values. Surprisingly, during the summer however, MOG PODy values for ITFA are nearly 25% larger than those computed for the AIRMETs and 18 to 20% larger than those computed for Ellrod-1 and DTF3.

The PODn values indicate that the forecasts issued in the summer (Fig. 8) are generally better in identifying areas with no turbulence than during the winter. During the winter, the algorithms are better than the AIRMETs at detecting areas clear of turbulence because the algorithms have the ability to pinpoint turbulence in narrow layers of the atmosphere and the AIRMETs are restricted to capturing turbulence in volumetric shapes.

Turb2000–2001 Results



TSS values (Fig. 9), which measure the trade-off between MOG PODy and PODn, indicate that the performance between ITFA, DTF3, AIRMETs, and Ellrod-1 are similar in both the winter and summer months. The TSS values for ITFA during the summer are larger than for the other algorithms and for the AIRMETs. This result indicates that although there is some trade-off between large values of MOG PODy and

Turb2000–2001 Results

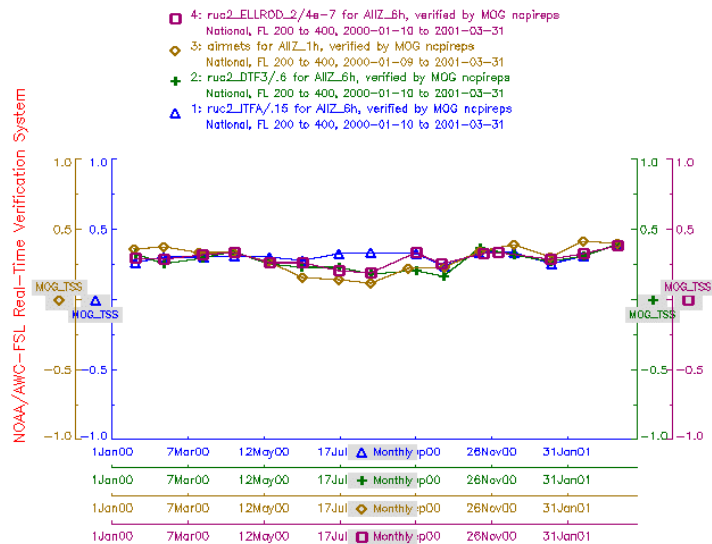


Figure 9. As in Fig. 7, except for TSS

small values of POD_n, the increase in MOG POD_y was larger than the decrease in POD_n for ITFA. This result suggests that ITFA could be used to improve the skill statistics of the other algorithms and AIRMETs during the summer.

Scatterplots of daily MOG POD_y with respect to % Volume for ITFA, DTF3, AIRMETs, and Ellrod-1 for the TURB2001 evaluation period (i.e., 8 February – 31 March 2001) are shown in Figs. 10 - 14. A comparison between the algorithms and AIRMETs is shown in Fig. 10 with scatterplots for the individual algorithms shown in Figs. 11 - 14. Each symbol on the plot represents a value for MOG POD_y computed from pairs accumulated over a one-day period. The % Volume was averaged over all issue times for the 6-h leads over the one-day period. The straight black line on the plot is the one-to-one correspondence line indicating that for an increase in MOG POD_y there is an equivalent increase in % Volume. The dotted black line is the regression line indicating the extent to which one of the variables decreases as the other increase.

Inspection of Fig. 10 shows that for ITFA, DTF3, Ellrod-1, and AIRMETs, the values of MOG POD_y cluster above the one-to-one correspondence line indicating that for a given MOG POD_y value, the % Volume is smaller than if the values fell along the one-to-one correspondence line. Of the entire volume computed for the forecast domain between 20,000 and 40,000 ft, the % Volume for the algorithms and AIRMETs is less than 40%.

When analyzing the scatterplots individually for each algorithm and for the AIRMETs (Figs. 11 –14), the character of the distribution is quite different. For instance, the regression line for ITFA (Fig. 11), DTF3 (Fig. 12), and Ellrod-1 (Fig. 13) are nearly vertical, while for the AIRMETs (Fig. 14), the regression line is slanted from the lower left corner of the plot to the upper right hand corner of the plot. These results indicate that for the algorithms (ITFA, DTF3, and Ellrod-1) a large variation in MOG POD_y occurs over a small range of volumes. Of the 3 algorithms, the smallest change in % Volume occurs for ITFA with a range from 15 to 28 %. For the AIRMETs, however, larger values of MOG POD_y occur with an increase in % Volume. This result may be an artifact of the volumetric manner with which the AIRMETs are required to be produced.

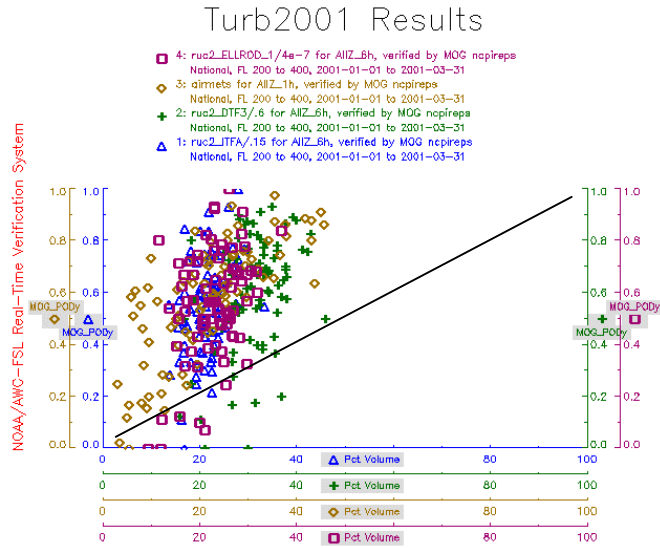


Figure 10. Scatterplot for 6-h leads, combined over all issue times for ITFA/.15 (triangle), DTF3/.6 (+), AIRMETs (diamond), and Ellrod-1/4e-7 (square) for TURB2001 evaluation period for MOG PODy vs. % Volume. Solid black line is one-to-one correspondence line. Dotted line is regression line indicating the extent to which one variable increases and the other decreases.

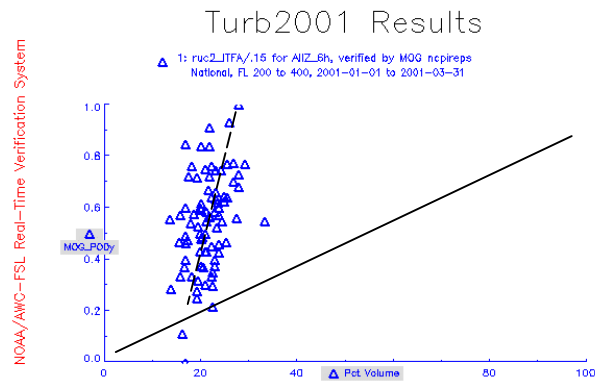


Figure 11. As in Fig. 10, except for ITFA/.15.

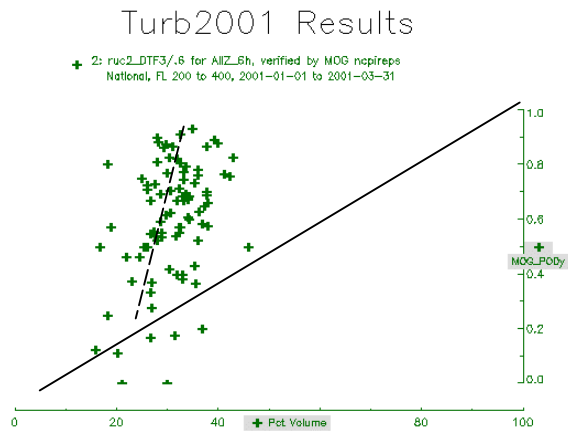


Figure 12. As in Fig. 10, except for DTF3/6.

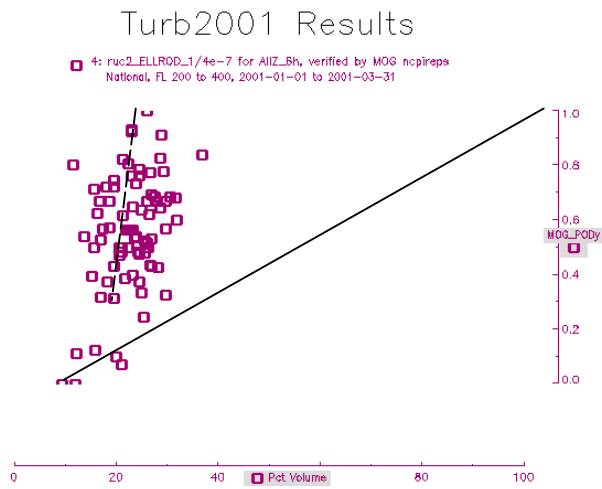


Figure 13. As in Fig. 10, except for Ellrod-1/4e-7.

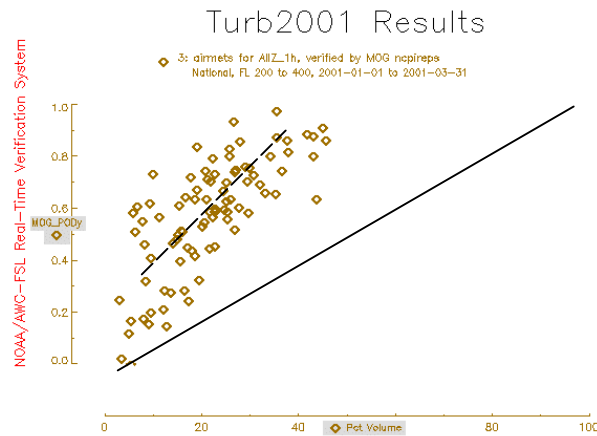


Figure 14. As in Fig. 10, except for AIRMETs.

Scatterplots of daily MOG PODy with respect to 1-PODn for ITFA, DTF3, AIRMETs, and Ellrod-1 for TURB2001 are shown in Figs. 15 - 19. First, a comparison between the algorithms and AIRMETs is shown in Fig. 15 with scatterplots for the individual algorithms shown next in Figs. 16 - 19. Each symbol on the plot represents a value for MOG PODy and 1-PODn computed from pairs accumulated over a one-day period.

Inspection of Fig. 15 for ITFA, DTF3, Ellrod-1, and AIRMETs shows larger scatter in the values for MOG PODy with respect to 1-PODn than with % Volume. Generally, the values of MOG PODy vs. 1-PODn on the plot lie above the one-to-one correspondence line indicating a tendency for the distribution of points to approach the desired upper left hand corner of the plot. Overall, the 1-PODn values are less than 0.5 with a large range in MOG PODy.

When analyzing the scatterplots individually for each algorithm and for the AIRMETs (Figs. 16 - 19), the scatter in the distribution of points is generally similar between the forecasts and AIRMETs, but with some noteworthy differences. For instance, the center of the cluster for ITFA (Fig. 16) is lower than DTF3 (Fig. 17) and Ellrod-1 (Fig. 18), but DTF3 (Fig. 17) and Ellrod-1 (Fig. 18) have larger 1-PODn values. Interestingly, the distribution of 1-PODn values for the AIRMETs (Fig. 19) is evenly spread from 0.0 to 0.5. On the other hand, the MOG PODy values for the AIRMETs increase with larger values of 1-PODn.

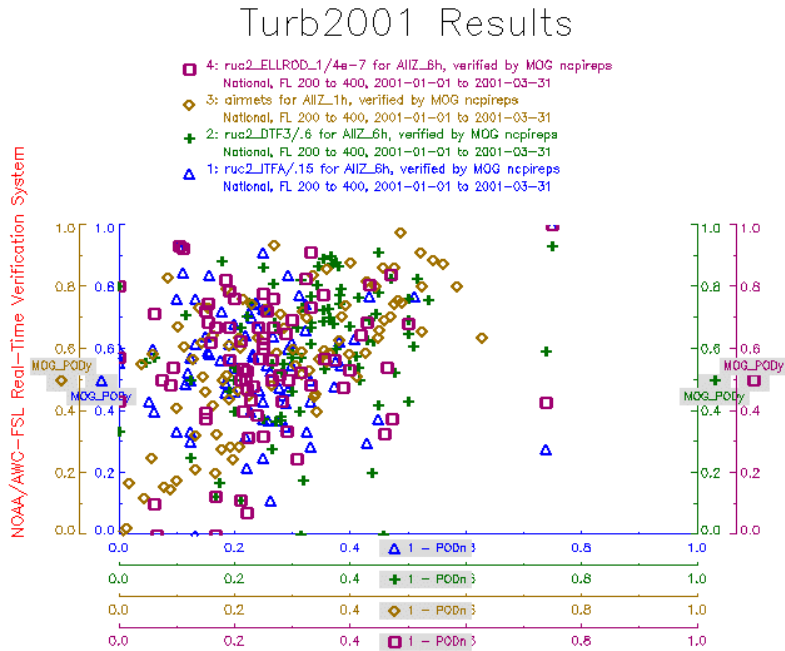


Figure 15. Scatterplot for 6-h leads, combined over all issue times for ITFA/.15 (triangle), DTF3/6 (+), AIRMETs (diamond), and Ellrod-1/4e-7 (square) for TURB2001 evaluation period for MOG PODy vs. 1- PODn. Solid black line indicates the one-to-one correspondence line.

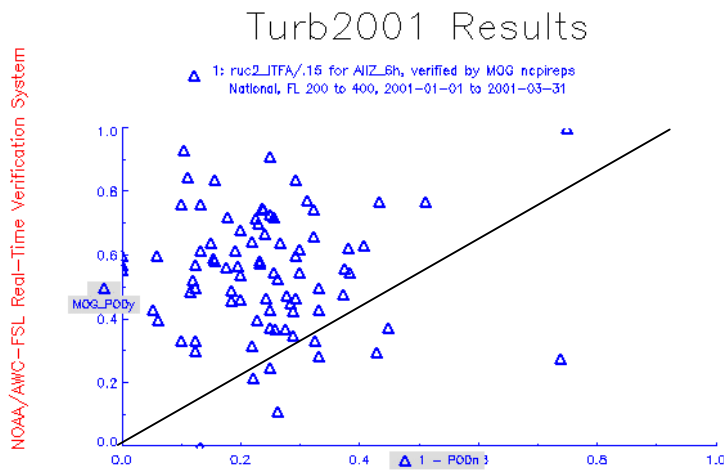


Figure 16. As in Fig. 15, except for ITFA/.15.

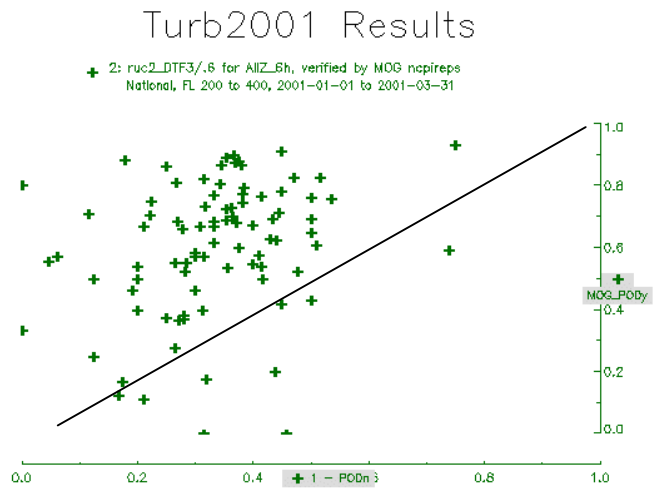


Figure 17. As in Fig. 15, except for DTF3/.6.

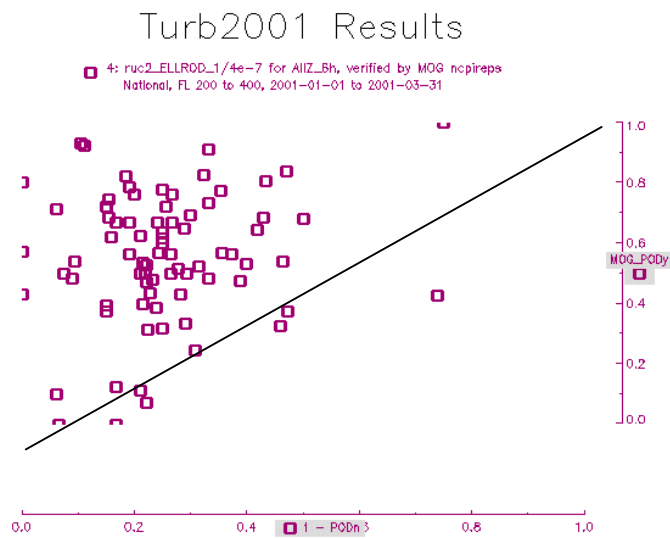
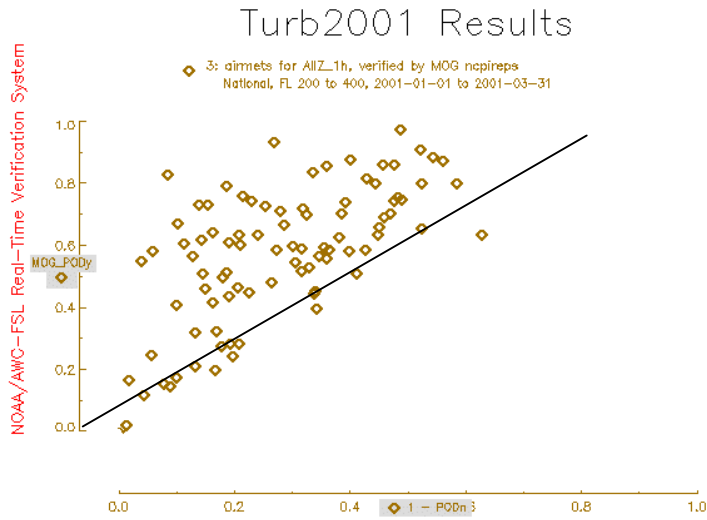


Figure 18. As in Fig. 15, except for Ellrod/4e-7.



5.3. Variations with height

Fig. 20 is a height series plot for the 6-h lead combined over all issue times where MOG TSS values are generated for ITFA (0.15), DTF3 (0.6), AIRMETs, and Ellrod-1 (4×10^{-7}) and separated into 5,000 ft intervals from 20,000 ft to 40,000 ft. Each line on a plot represents one of the algorithms or AIRMETs. Each symbol on the line is a statistic generated over the TURB2001 exercise period at a specific height. In place of a height series plot of MOD PODy and PODn, the MOG TSS is shown. The TSS statistic is a measure of the trade-off between the PODy and PODn statistics and, in this case, better represents the quality of the algorithms and AIRMETs than is represented by the MOG PODy and PODn plots.

Immediately apparent are the similarities with height between ITFA, DTF3, and Ellrod-1. The overall profile of the height series for the algorithms is interesting where the MOG TSS value for the algorithms decreases from a maximum at 20,000 ft to a minimum at 30,000 and then gradually increases from 30,000 ft to 40,000 ft. The AIRMETs are slightly different. Although the profile for the AIRMETs is similar to the algorithms below 30,000 ft, at 30,000 ft the MOG TSS value for the AIRMETs increases from 0.2 computed for the algorithms to nearly 0.3. Surprisingly, above 35,000 ft, the MOG TSS value for the AIRMETs decreases to a value of 0.2: a difference of nearly 0.5 between the AIRMETs and the algorithms.

Turb2001 Results

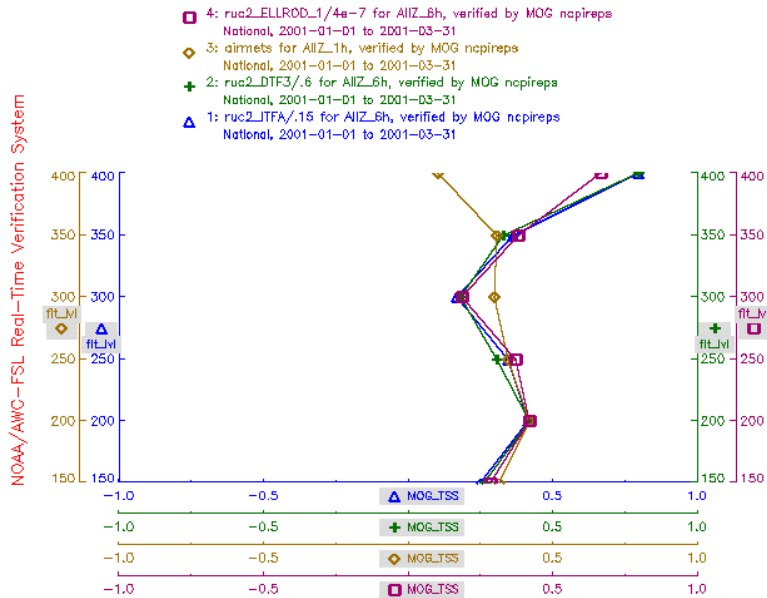


Figure 20. Height series plot for 6-h lead, all issue times combined for ITFA/.15 (triangle), DTF3/.6 (+), AIRMETs (diamond), and Ellrod-1/4e-7 (square) accumulated for TURB2001 for MOG TSS. Heights are every 5000 ft.

6. Basic post-analysis results

Algorithm comparison plots created at the start of the post-analysis are presented in Figs. 21 - 24, for 6-h forecasts. The results in these plots are consistent with the results obtained by RTVS, although these results are based on all altitudes between 15,000 and 42,000 ft. In particular, the same group of algorithms seems to perform well, while another subset performs less well. The statistics for the AIRMETs are somewhat better than the statistics for any of the algorithms.

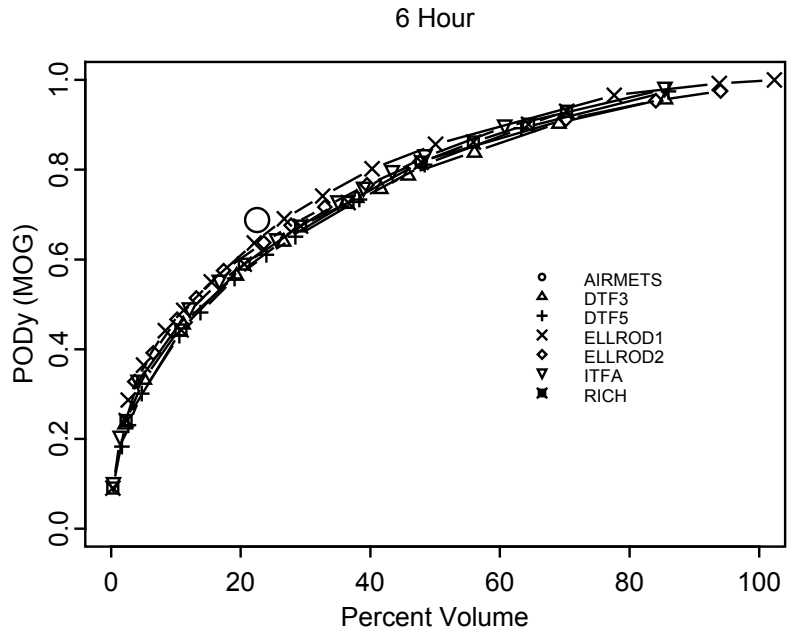


Figure 21. Algorithm comparison plots for 6-h forecasts, evaluated over 15-42,000 ft: PODY vs. % Volume.

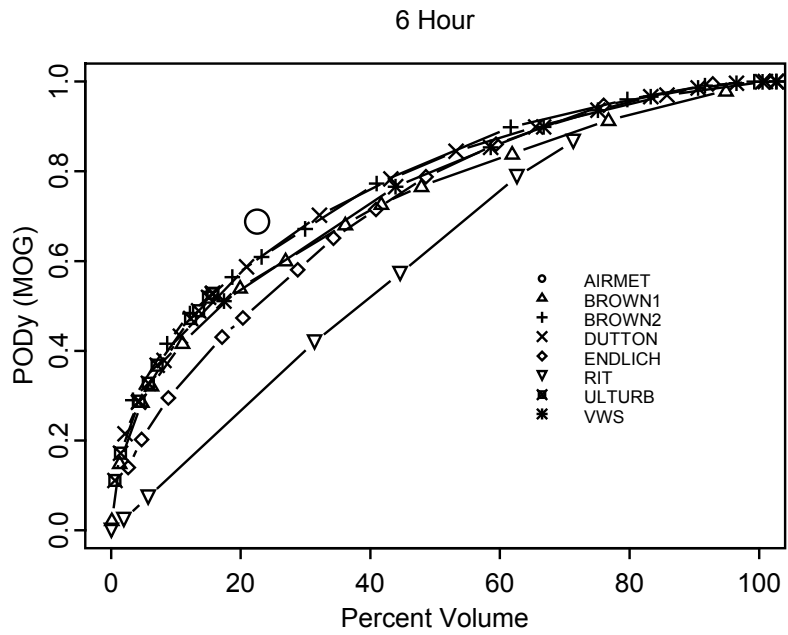


Figure 22. As in Fig. 21, for 2nd group of algorithms.

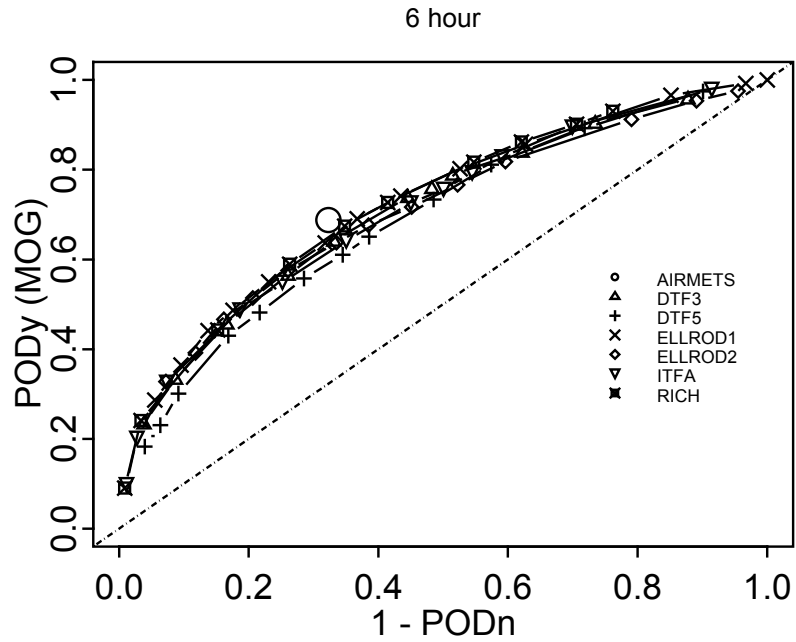


Figure 23. As in Fig. 21, for POD_y vs $1-POD_n$.

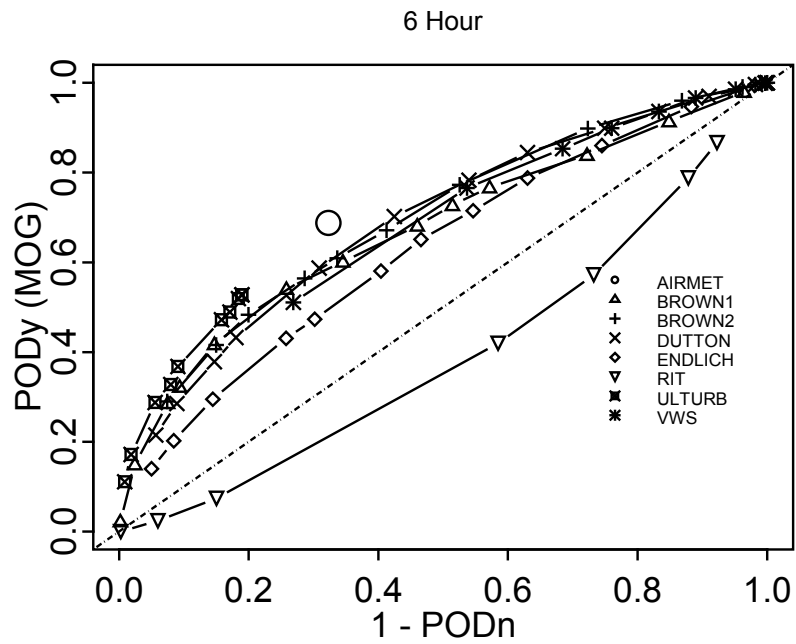


Figure 24. As in Fig. 23, for second group of algorithms.

7. Summary and conclusions

The basic results from the third turbulence intercomparison exercise (TURB2001) are summarized in this report. The exercise took place from 8 February – 31 March 2001 where fourteen turbulence algorithms were evaluated both objectively by RTVS and post-analysis and subjectively through input provided by AWC and Delta Airline forecasters. The subjective analysis will be summarized in a separate report. Only the basic statistical results were presented in the report. Additional displays and statistics can be obtained through the RTVS Web-based interface at <http://www-ad.fsl.noaa.gov/afra/rtvs>. An in depth analyses will be provided by post-analysis in a second report.

In addition to developing and maintaining a statistical baseline of the quality of the turbulence forecasts, one of our primary goals is to evaluate the differences and similarities between the various algorithms and AIRMETs so that the quality of the algorithms and forecasts can improve. The verification methods used to compare the algorithms and forecasts were developed to the best of our ability and with input provided by the Turbulence PDT and the AWC staff so that the comparisons are fair, independent, and consistent.

Several interesting similarities and differences between the algorithms and forecasts were revealed through this exercise. Overall, algorithm performance at the 6-h lead was similar between the seven main algorithms (Brown-2, DTF3, DTF5, Dutton, Ellrod-1 and -2, ITFA), and is consistent with the results presented in TURB98-99 and TURB2000. However, the on-going statistics for ITFA showed improvement over the other algorithms and AIRMETs during the summer months, although further testing is need to determine whether this improvement is dependent upon the weather or is consistent from year to year. Other differences were identified when the overall results were categorized by domain. For example, the AIRMETs were best at capturing turbulence in the Central region and best at capturing No turbulence over the Mountain domain. Furthermore for the smaller domains (i.e., Central, West, etc.), greater variation in the statistics between Mwave, Richardson Number, Shear, GWB, Ulturb, Horizontal Shear, and Temperature Gradient was apparent.

The scatterplots revealed that algorithms generally produce turbulence over 15 to 40 % of the National domain, with ITFA producing the smallest amount of turbulence, covering 15 to 28 % of the domain. However, ability for the forecasts to correctly capture the turbulence, as denoted by MOG PODy, was highly variable. The character of the statistics for the AIRMETs was different from those computed for the algorithms. For instance, an increase in performance for the AIRMETs was somewhat tied to an increase in volume where an increase in % Volume generated an increase in MOG PODy. These characteristic differences are likely a result of the algorithm's ability to capture turbulence in narrow layers, while the AIRMETs are limited to defining turbulence in volumetric shapes.

The height series plots showed that the performance of ITFA, DTF3, and Ellrod-1 was nearly identical from 20,000 to 40,000 ft. The AIRMETs showed improvements in

skill over the algorithms at 30,000 ft. However, the skill of the AIRMETs decreased at 40,000 ft when compared to the algorithms.

Initial results from the post-analysis are consistent with those shown from RTVS. An in depth analysis will be summarized in a second report.

Future work includes examining the statistics for other forecast lead times, correlating these results with those produced by the subjective analysis, further investigating the trends in the ITFA summer results, and analyzing the regional statistics to determine how that information can be used to improve ITFA.

Acknowledgements

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 members of the Turbulence Product Development Team for their support of this effort. We also thank Gerry Wiener (NCAR), Sue Dettling (NCAR), Missy Petty (NCAR), 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. We would like to thank Joan Hart and for her work on RTVS and the FSL Facilities Division for making the forecasts and algorithm output available to in real time to RTVS.

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., G. Thompson, R.T. Bruitjes, R. Bullock, and T. Kane, 1997: Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Weather and Forecasting*, **12**, 890-914.

Brown, B.G. and J.L. Mahoney, 1998: Verification of Turbulence Algorithms. Report, Available from B.G. Brown, NCAR, PO Box 3000 Boulder CO 80307-3000, 9 pp.

Brown, B.G., J.L. Mahoney, R. Bullock, J. Henderson, and T.L. Kane, 1999: Turbulence Algorithm Intercomparison: 1998-99 Initial Results. Report to the Aviation Weather Research Program, Federal Aviation Administration, U.S. Department of Transportation. Available from B.G. Brown, NCAR, PO Box 3000 Boulder CO 80307-3000, 64 pp.

Brown, B.G., J.L. Mahoney, J. Henderson, T.L. Kane, R. Bullock, and J.E. Hart, 2000a: The turbulence algorithm intercomparison exercise: Statistical verification results. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 Sept., American Meteorological Society (Boston), 466-471.

Brown, B.G., J.L. Mahoney, R. Sharman, J. Vogt, and J. Henderson, 2000b: Use of automated observations for verification of turbulence forecasts. Report to the Aviation Weather Research Program, Federal Aviation Administration, U.S. Department of Transportation. Available from B.G. Brown, NCAR, PO Box 3000 Boulder CO 80307-3000.

Brown, B.G., J.L. Mahoney, R. Bullock, T.L. Fowler, J. Hart, J. Henderson, A. Loughe, 2000c. Turbulence algorithm intercomparison: Winter 2000 results. NOAA Technical Memorandum OAR FSL-26, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, Forecast Systems Laboratory, 62 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*, Orlando, FL, 11-15 Sept., American Meteorological Society (Boston), 393-398.

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

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. *Meteorol. Mag.*, **109**, 293-310.

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

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

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*, Asheville, NC, 8-11 May, American Meteorological Society (Boston), 46-49.

Keller, J. L., 1990: Clear Air Turbulence as a response to meso- and synoptic-scale dynamic processes. *Mon. Wea. Rev.*, **118**, 2228-2242.

- Knox, J. A., 1997: Possible mechanism of clear-air turbulence in strongly anticyclonic flows. *Mon. Wea. Rev.*, **125**, 1251-1259.
- Kronebach, G. W., 1964: An automated procedure for forecasting clear-air turbulence. *J. App. Met.*, **3**, 119-125.
- Mahoney, J.L., J.K. Henderson, and P.A. Miller, 1997: A description of the Forecast System's Laboratory's Real-Time Verification System (RTVS). *Preprints, 7th Conference on Aviation, Range, and Aerospace Meteorology*, Long Beach, CA, American Meteorological Society (Boston), J26-J31.
- Mahoney, J.L., and B.G. Brown, 2000: Forecaster assessment of turbulence algorithms: A summary of results for the winter 2000 study. NOAA Technical Memorandum OAR FSL-27, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, Forecast Systems Laboratory, 134 pp.
- 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, 7th Conf. on Aviation, Range and Aerospace Meteorology*, American Meteorological Society, Long Beach, CA, 158-163.
- Orville, R.E., 1991: Lightning ground flash density in the contiguous United States – 1989. *Monthly Weather Review*, **119**, 573-577.
- Palmer, T.N., Shutts, G.J., and Swinbank, 1986: Alleviation of a systematic bias in general circulation and numerical weather prediction models through an orographic gravity wave drag parameterization. *Quart. J. R. Met. Soc.*, **112**, 1001-1039.
- Roach, W.T., 1970. On the influence of synoptic development on the production of high level turbulence. *Quart. J. R. Met. Soc.*, **96**, 413-429.
- Sharman, R, C. Tebaldi, and B. Brown, 1999: An integrated approach to clear-air turbulence forecasting. *Preprints, 8th Conference on Aviation, Range, and Aerospace Meteorology*, Dallas, TX, 10-15 January, American Meteorological Society, 68-71.

Sharman, R, B. 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*, Orlando, FL, 11-15 Sept., American Meteorological Society (Boston), 460-465.

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

APPENDIX A

Verification Methodology

This section summarizes methods that were used to match forecasts and observations, as well as the various verification statistics that were computed to evaluate the CAT forecasts.

A1. Matching methods

The same methods were used in TURB2001 to connect PIREPs to forecasts as in TURB98-99 and in TURB2000. In particular, both the NCAR/RAP and RTVS systems connect each PIREP to the forecasts at the nearest 8 grid points (four surrounding grid points; two levels vertically). However, the RTVS uses bi-linear interpolation to compute the appropriate forecast value, whereas the RAP system matches the PIREP to the most extreme (largest, except in the case of Richardson number) forecast value among the four surrounding gridpoints. As in TURB98-99 and TURB2000, a time window of ± 1 hour around the model valid time was used to evaluate both the algorithm forecasts and the AIRMETs.

A2. Statistical verification methods

The statistical verification methods used to evaluate the TURB2001 results are the same as the methods used in TURB98-99 and TURB2000, with a few relatively minor extensions. More detail on the general concepts underlying verification of turbulence forecasts can be found in Brown and Mahoney (1998). These methods are described briefly here.

Turbulence forecasts and observations are treated here as dichotomous (i.e., Yes/No) values. AIRMETs essentially are dichotomous (i.e., a location is either inside or outside the defined AIRMET region). The algorithm forecasts are converted to a variety of Yes/No forecasts by application of various thresholds for the occurrence of turbulence. The thresholds used in RTVS are summarized in Table A1. Verification methods described here generally are based on the two-by-two contingency table (Table A2). In this table, the forecasts are represented by the rows, and the columns represent the observations. The entries in the table represent the joint distribution of forecasts and observations.

Table A3 lists the verification statistics used in TURB98-99, TURB2000 and TURB2001. As shown in this table, POD_y and POD_n are the primary verification statistics based on the 2x2 verification table. It is important to recognize that POD_y and POD_n are estimates of the conditional distributions that underlie the joint distribution of forecasts and observations, or they are functions of these distributions. For example, POD_y is an estimate of the conditional probability of a Yes forecast given a Yes

observation, $p(f=Yes|x=Yes)$, where f represents the forecasts and x represents the observations.

Table A1. Algorithm thresholds used in RTVS analyses.

Algorithm	Threshold Values					
Brown-2	.03	.1	.15	.25	.65	1.0
DTF3	.2	.4	.6	1.3	2.0	3.0
DTF5	.08	.1	.15	.25	.5	.9
Dutton	12.	15.	22.	30.	60.	80.
Ellrod-1	1×10^{-7}	3×10^{-7}	4×10^{-7}	5×10^{-7}	7×10^{-7}	2×10^{-6}
Ellrod-2	2×10^{-7}	2.5×10^{-7}	4×10^{-7}	7×10^{-7}	1.2×10^{-6}	1.6×10^{-6}
GWB	1.5	2.5	3.	4.5	8.5	12.
Horizontal Shear	2×10^{-5}	3.2×10^{-5}	4.8×10^{-5}	6.1×10^{-5}	7.5×10^{-5}	8.1×10^{-5}
ITFA	.06	.08	.15	.2	.3	.4
Mwave	.001	.1	10.	50.	100.	400.
Richardson	.5	1.	2.	4.	9.	15.
SCATR	1×10^{-6}		1×10^{-4}		1×10^{-3}	1×10^{-2}
Temp Gradient	1×10^{-5}	1.5×10^{-5}	2×10^{-5}	2.7×10^{-5}	3.8×10^{-5}	5.4×10^{-5}
ULTURB	.001	.005	.007	.01	.025	.04
VW Shear	.004	.005	.006	.009	.015	.02

Table A2. Contingency table for evaluation of dichotomous (Yes/No) forecasts. Elements in the cells are the counts of forecast-observation pairs.

<i>Forecast</i>	<i>Observation</i>		<i>Total</i>
	<i>Yes</i>	<i>No</i>	
<i>Yes</i>	YY	YN	YY+YN
<i>No</i>	NY	NN	NY+NN
<i>Total</i>	YY+NY	YN+NN	YY+YN+NY+NN

Table A3. Verification statistics used in this study.

<i>Statistic</i>	<i>Definition</i>	<i>Description</i>	<i>Interpretation</i>	<i>Range</i>
POD_y	$YY/(YY+NY)$	Probability of Detection of Yes observations	Proportion of Yes observations that were correctly forecasted	0-1 Best: 1 Worst: 0
POD_n	$NN/(YN+NN)$	Probability of Detection of No observations	Proportion of No observations that were correctly forecasted	0-1 Best: 1 Worst: 0
TSS	$POD_y + POD_n - 1$	True Skill Statistic	Level of discrimination between Yes and No observations	-1 to 1 Best: 1 No skill: 0
Curve Area	Area under the curve relating POD _y and 1-POD _n	Area under the curve relating POD _y and 1-POD _n (i.e., the ROC curve)	Overall skill (related to discrimination between Yes and No observations)	0 to 1 Best: 1 No skill: 0.5
% Volume	$[(Forecast\ Vol) / (Total\ Vol)] \times 100$	% of the total air space volume that is impacted by the forecast	% of the total air space volume that is impacted by the forecast	0-100 Smaller is better
Volume Efficiency (VE)	$(POD_y \times 100) / \% Volume$	POD _y (x 100) per unit % Volume	POD _y relative to airspace coverage	0-infinity Larger is better

It also will be noted that Table A3 does not include the False Alarm Ratio (FAR), a statistic that is commonly computed from the 2x2 table. As described in Brown et al. (1997) and applied in TURB98-99, it is not possible to compute FAR using only PIREPs (or PIREPs and AVARs). This conclusion, which also applies to other statistics such as the Critical Success Index and Bias, is documented analytically and by example in Brown and Young (2000). In addition, due to the limited numbers of PIREPs and other characteristics of the PIREPs, other verification statistics (e.g., PODy and PODn) should not be interpreted in an absolute sense, but can be used in a comparative sense, for comparisons between algorithms and forecasts. Moreover, PODy and PODn should not be interpreted as probabilities, but rather as *proportions of 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 both PODy and PODn, as well as TSS. PODn can be computed in two ways for turbulence forecasts – (i) using the negative PIREP observations and (ii) using the negative AVAR observations. However, results based on the AVAR observations are not presented in this report.

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 “Relative Operating Characteristic” (ROC) curve in SDT. The area under this curve is a measure of overall forecast skill (e.g., Mason 1982), and provides another measure that can be compared among the algorithms. These area values were computed only in the post-analysis.

As shown in Table A3, two other variables are utilized for verification of the turbulence forecasts: % Volume and Volume Efficiency (VE). The % Volume statistic is the percent of the total possible airspace volume² that has a Yes forecast. VE considers PODy relative to the volume covered by the forecast, and can be thought of as the POD per unit volume. The VE statistic must be used with some caution, however, and should not be used by itself as a measure of forecast quality. For example, it sometimes is easy to obtain a large VE value when PODy is very small. An appropriate use of VE is to compare the efficiencies of forecasting systems with nearly equivalent values of PODy.

Use of these statistics is considered in somewhat greater detail in Brown et al. (1999). In general, however, the argument presented in the previous paragraph can be

² The total possible area (limiting coverage to the area of the continental United States that can be included in AIRMETs) is 9.5 million km². Because the analyses are limited to 20,000 ft and above, the total possible volume thus is about 64 million km³

extended to all of the statistics in Table A3; none of them should be considered alone – all should be examined in combination.

As in TURB98-99 and TURB2000, emphasis in this report 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. ROC curve areas also will be considered as a measure of the overall skill of the forecasts at discriminating between Yes and No observations.

Quantification of the uncertainty in verification statistics is an important aspect of forecast verification that often is ignored. Confidence intervals provide a useful way of approaching this quantification. However, most standard confidence interval approaches require various distributional and independence assumptions, which generally are not satisfied by forecast verification data. As a result, the QAG has developed an alternative confidence interval method based on re-sampling statistics, which are appropriate for turbulence forecast verification data (Kane and Brown 2000). This approach is applied to some of the statistics considered in this report.

A3. Stratifications

In TURB98-99, the verification results were stratified and limited using a variety of criteria applied to the PIREPs. These criteria included aircraft weight and proximity to lightning (Brown et al. 1999). Results of the TURB98-99 analyses indicated that the aircraft weight criteria had little effect on the verification results, except that it vastly reduced the number of PIREPs available for the analysis. Thus, this criterion was not applied in TURB2000 or TURB2001. However, the lightning criterion was used by RTVS to eliminate reports that may have been located in convective regions, using the same approach as in TURB98-99 and TURB2000. In particular, this stratification considered the locations of lightning observations. If a PIREP was located within a 20-km radius of an area where there had been at least 4 lightning strikes during the previous 20 minutes, the observation was assigned a convective flag and was excluded from some analyses. Because the impacts of this stratification were also found to be relatively minimal, the lightning criterion has not been applied in post-analyses for TURB2000 or TURB2001.

The statistics were generated by RTVS on several domains that included the National, East, Central, West, and Mountain domains where the areas of those boundaries were defined using AWC criteria. A fifth domain, which defined the mountain regions, was also applied to RTVS and post-analysis. In this report, statistical results are briefly summarized for the smaller domains, but additional information is provided from on the Web.

All of the evaluations were limited to PIREPs, algorithm output above 20,000 ft and algorithm output from 15,000 – 20,000 ft. In post-analysis, heights were evaluated from 15,000 to 42,000 ft. Two categories of reported severity are considered: (i) reports

of any turbulence severity (light and greater) and (ii) reports of MOG severity. Most results are presented for the MOG category.