

Turbulence Algorithm Intercomparison: Winter 2000 Results

**Barbara G. Brown¹, Jennifer L. Mahoney²,
Randy Bullock¹, Tressa L. Fowler¹,
Joan Hart^{2,3}, Judy Henderson^{2,3}, and Andy Loughe^{2,3}**

September 2000

¹ Research Applications Program, National Center for Atmospheric Research, Boulder, CO

² Forecast Systems Laboratory, Environmental Research Laboratories, National Oceanic and Atmospheric Administration, Boulder, CO

³ Joint collaboration with the Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO

Contents

Section	Page
Executive Summary	ii
1. Introduction	1
2. Approach	3
3. Algorithms	5
4. Data	7
5. Methods	9
5.1 Matching methods	9
5.2 Statistical verification methods	9
5.3 Stratifications	12
6. Real-time verification	13
6.1 Mechanics	13
6.2 Results	15
6.2.1 General comparisons	15
6.2.2 Variations with time	21
6.2.3 Variations in ITFA	24
6.2.4 Variations with height	27
7. Post-analysis	29
7.1 Mechanics	29
7.2 Results	30
7.2.1 Overall results for 3-h forecasts	30
7.2.2 Comparisons among lead times	36
7.2.3 Comparisons among valid times	49
7.2.4 Comparisons with TURB98-99 results	53
8. Summary, conclusions, and discussion	58
Acknowledgments	60
References	60

Executive Summary

This report summarizes basic results of a second 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 intercomparison, which took place during winter 1998-99, as well as two additional algorithms. The algorithm forecasts are based on output of the RUC-2 numerical weather prediction model during the period 10 January through 31 March 2000. Forecasts issued at 1200, 1500, 1800, and 2100 UTC, with 3-, 6-, 9, and 12-h lead times were included in the study. Turbulence AIRMETs, the operational turbulence forecast product that is issued by the NWS's Aviation Weather Center (AWC), also were included in the evaluation. The evaluation was limited to the continental United States and to altitudes above 20,000 ft.

The forecasts were verified using Yes and No turbulence observations from pilot reports (PIREPs), as well as No observations based on automated vertical accelerometer (AVAR) data that were obtained from a number of aircraft. 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 was 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/RTVS-project_des.html). 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.

Results of the intercomparison suggest that a few of the algorithms (e.g., DTF3, ITFA, Ellrod) have somewhat better overall forecasting performance than the others. In particular, these algorithms have somewhat larger values of the True Skill Statistic for comparable thresholds, and they have a slightly larger overall discrimination skill statistic. However, the best algorithms have very similar performance characteristics. In some (but not all) cases the algorithm performance is approximately the same as the performance of the AIRMETs. Results of the study are consistent with the results obtained for winter 1998-99.

<p

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 during the winter of 2000, and is the second in a series of evaluations of the algorithms' forecasting performance. The previous intercomparison took place during the winter of 1998-99; results of that evaluation are presented in Brown et al. (1999, 2000a). Both 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 2000 intercomparison (hereafter, denoted TURB2000) were to (i) develop a 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 intercomparison (hereafter, denoted TURB98-99). 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 two 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 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. Basic results from the real-time evaluation also are presented. Results of the post-analysis are presented in somewhat greater detail.

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. The verification methods are described in Section 5. Results of the real-time study are presented in Section 6, with results from the post-analysis presented in Section 7. Finally, Section 8 includes the conclusions and discussion.

2. Approach

A total of 11 CAT algorithms were included in TURB2000. Most of these algorithms also were included in TURB98-99; three algorithms that were evaluated in TURB98-99 were excluded from TURB2000, and two algorithms (including a “random” algorithm) were added. 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, 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 limited to the region of the atmosphere above 20,000 ft, as was the case in TURB98-99.

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

TURB2000 began on 10 January and ended on 31 March. Due to some data problems, however, only forecasts through 22 March are currently included in the post-analysis results.

The verification approach is identical to the approach taken in TURB98-99, except that a few additional metrics and graphics were added to the RTVS displays. In addition, a “random” algorithm was evaluated to provide assurance that the verification software was performing appropriately, and a method was developed to compute confidence intervals for the statistics in post-analysis. The algorithm forecasts and AIRMETs were verified using Yes and No PIREPs of turbulence. In addition, vertical accelerometer (AVAR) observations, which were systematically recorded from observations provided by certain United Airlines aircraft, were used as an indicator of No turbulence under certain conditions. 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.

A “forecaster evaluation” of algorithm performance also was included in TURB2000. In this subjective evaluation, five forecasters at the AWC examined forecasts produced by the turbulence algorithms and completed a questionnaire on a daily basis. The questionnaire concerned the synoptic meteorological conditions associated with observed turbulence events, as well as the forecasters’ perceptions of the relative performance of the various algorithms. Results of this study are presented in a separate report (Mahoney and Brown 2000).

3. Algorithms

The set of algorithms that was evaluated in TURB2000 differed slightly from the set that was considered in TURB98-99. Specifically, the differences include the following:

- (i) The Burke-Thompson, DTF4, and SCATR algorithms were not included in TURB2000 because the TURB98-99 results indicated that their performance either was much poorer than the performance of other algorithms (e.g., in the case of the SCATR index) or their performance was adequately represented by other algorithms (e.g., in the case of DTF4). However, SCATR was included in some of the post-analysis results, to provide confirmation of its performance in TURB98-99.
- (ii) ULTURB was included in TURB2000 (and TURB98-99), but its implementation was believed to have errors. Thus, ULTURB results will not be considered here. If possible, ULTURB output will be obtained directly from the AWC for future evaluations.
- (iii) A “random” algorithm was included to test the functioning of the verification software.

The algorithms that were included in TURB2000 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).

CCAT: The CCAT (Clark's Clear Air Turbulence) index has been used on a semi-operational basis by the US Navy's FNMOC for at least 2 decades. It was developed by Leo Clark in consultation with Hans Panofsky, by applying aerodynamicist Theodore Theodorsen's theory for the generation of vortices to clear air turbulence. There is no direct documentation on this index other than a definition and evaluation in an NRL verification study document (Vogel and Sampson 1996).

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-2: 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).

Endlich: The Endlich "index" is based on a paper by R. Endlich (1964) in which jet stream structures were compared to turbulence measurements. The best agreement of the observations was obtained for an empirical parameter that is simply the product of the local wind speed and the vertical gradient of the wind direction.

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.

Random: This index was created using a random number generator to randomly generate uniformly distributed random numbers over the range 0 to 1.

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).

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, the data that were used in TURB2000 include model output, PIREPs, AVAR observations, 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, 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.

In addition to the PIREPs, vertical accelerometer (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 value of the acceleration of gravity (9.8 ms^{-2}) were included as observations of No turbulence. Unfortunately, recent evaluations of the AVAR observations (Brown et al. 2000b) have suggested that the observations may not appropriately distinguish positive and negative turbulence conditions; thus, only limited results based on AVAR observations are presented in this report



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 Section 5.3).

5. Methods

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.

5.1 Matching methods

The same methods were used to connect PIREPs to forecasts as in TURB98-99. 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. AVAR observations are interpolated/matched to model gridpoints using the same approach. As in TURB98-99, a time window of ± 1 hour around the model valid time was used to evaluate both the algorithm forecasts and the AIRMETs.

5.2 Statistical verification methods

The statistical verification methods used to evaluate the TURB2000 results are the same as the methods used in TURB98-99, 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. Thus, verification methods described here generally are based on the two-by-two contingency table (Table 2). 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 3 lists the verification statistics used in both TURB98-99 and TURB2000. As shown in this table, POD_y and POD_n are the primary verification statistics based on the 2×2 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 2 : 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 3: 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)] x 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 x 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 3 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 3, 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 volume⁴ 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 extended to all of the statistics in Table 3; none of them should be considered alone – all should be examined in combination.

⁴ 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

As in TURB98-99, 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.

5.3 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 criteria was not applied in TURB2000. 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. 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.

All of the evaluations were limited to PIREPs, AVAR observations, and algorithm output above 20,000 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.

6. Real-time verification

Real-time verification was provided for TURB2000 to accomplish the following goals: (i) to provide near real-time statistical feedback to the algorithm developers, AWC forecasters, and other users through an interactive Web-based graphical user interface; (ii) to test the verification methods, evaluate whether realistic algorithm thresholds were applied to the algorithm output, and gather feedback on statistical displays so that adjustments could be made prior to the post analysis; and (iii) to generate statistics using only the forecasts and observations available in near real-time, much like the activities within an operational forecasting environment.

6.1 Mechanics

The real-time verification component of TURB2000 was provided by the RTVS (Mahoney et al., 1997). The system, which was developed by FSL with funding from the FAA, was enhanced to include the new turbulence algorithms considered in TURB2000, the ability to generate statistics for regions in the vicinity of mountains, an enhanced PIREP decoder, AREP data, and to include a number of new statistical displays, as well as a new graphical user interface. In addition, during the course of the evaluation, the system was upgraded to interact with a database, to allow creation of user-defined plots “on-the-fly.”

As in TURB98-99, model-based forecasts of turbulence, hourly turbulence observations from voice PIREPs, and automated AVAR reports were provided to RTVS through FSL's NIMBUS (Networked Information Management client-Based User System; Wahl et al. 1997). Scheduled processes were established within RTVS to access IDL (Interactive Data Language) routines for reading, writing, and stratifying data, bi-linearly interpolating algorithm output to observation locations, and generating statistical results. These processes ran continuously from 10 January - 1 March 2000. The algorithm thresholds used in the real-time verification are shown in Table 4. These thresholds were selected as an initial attempt to cover the range of possible forecasts. Based on the TURB98-99 results, additional thresholds for each algorithm were added to the system.

RTVS processed forecasts and observations that were available to the system at specified time periods. If data were missing or were late getting to the system, and/or the system processing or data transmission failed, results were not generated for that specific time period in near-real-time. However, after the evaluation was completed, attempts were made to fill in missing time periods and re-analyze the data.

In RTVS, the model output is connected to the PIREP and AVAR observations using the following process. First, the model-based output, available on the RUC-2 hybrid B coordinate system, is bi-linearly interpolated to flight levels to match the vertical resolution of the observations. Second, the four grid points surrounding the observation are interpolated horizontally to the observation location (e.g. PIREPs or AVARs), producing a forecast/observation pair as described in Section 5. If one of the grid points is missing or contains bad data, the forecast/observation pair is excluded from the statistical computations. A

±1-h time window around the model valid time is used to connect both the PIREP and AVAR observations to the forecasts.

Table 4: Algorithm thresholds used in RTVS analyses during TURB2000.

Algorithm	Threshold Values					
Brown-1	.00005	.00007	.00009	.00012	.00015	.0002
CCAT	1×10^9	3×10^9	7×10^9	1.5×10^8	3.5×10^8	5×10^8
DTF3	0.2	0.4	0.6	1.3	2.0	3.0
DTF5	0.08	0.1	0.15	0.25	0.5	0.9
Dutton	12.0	15.0	22.0	30.0	60.0	80.0
Ellrod-2	2×10^7	2.5×10^7	4×10^7	7×10^7	1.2×10^6	1.6×10^6
Endlich	0.07	0.14	0.17	0.24	0.38	0.5
ITFA	0.06	0.08	0.15	0.2	0.3	0.4
Random	0.25	0.5	0.6	0.75	0.9	0.95
Richardson	0.5	1.0	2.0	4.0	9.0	15.0
Shear	.004	.005	.006	.009	.015	.02

The Web-based graphical user interface (http://www-ad.fsl.noaa.gov/afra/rtvs/RTVS-project_des.html) utilized in TURB98-99 was enhanced for TURB2000. In addition, a new interface was developed, which connects the user to a database, where he/she can define particular combinations of algorithm, verification data, statistics, and displays to be created “on-the-fly”. Finally, the system also was enhanced to compute and display the TSS.

As in TURB98-99, web-based displays of the statistical results were presented through algorithm comparison and height series plots, as well as on scatter plots and contingency tables. In addition, the displays were enhanced to include time series plots of longer-term results. These time series plots are based on the accumulation of Yes/No counts and other data over 7-day periods. The plots were generated for each of the individual algorithms, issue and lead times, statistical measures, algorithm thresholds, and observation types. Plots were produced daily and for the overall evaluation period.

6.2 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.

6.2.1 General comparisons

Overall results for the 6-h lead from the 1200, 1500, and 1800 UTC issue times for PODy vs % Volume, PODy vs 1-PODn, and TSS vs % Volumes are presented in Figs. 2-10. Though all time periods presented in this Section indicate similar results, the plots are provided to serve as a reference guide to forecasters and algorithm developers. Each plot represents an individual model issue and lead time covering the period from 10 January – 31 March 2000. Each curve in the plots represent a single turbulence algorithm while each point on that curve represents a particular threshold used to create the Yes/No forecasts, with the AIRMETs represented by a single point. The thresholds were chosen to represent a range of turbulence forecasted over the specified

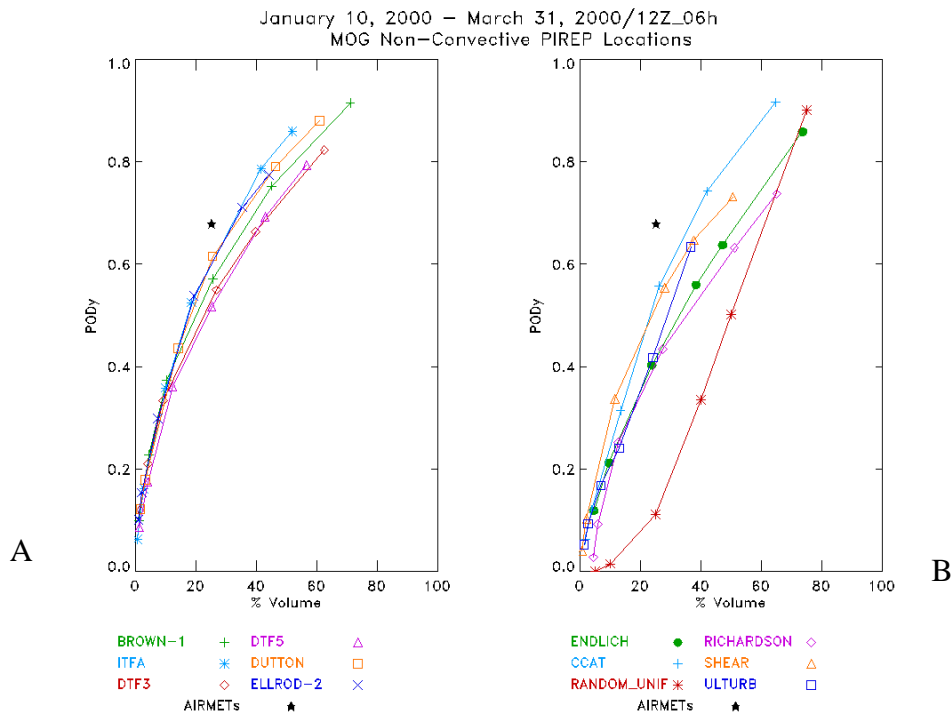


Figure 2. Two panels for 10 January – 31 March 2000 for the 1200 UTC issuance, 6-h lead, for MOG non-convective PIREPS are displayed for algorithm groups A and B, PODy vs. % Volume, with each plot containing 6 of the 12 algorithms. Each shape represents the PODy and % Volume for a particular algorithm. The line segments connect the results for different thresholds for a particular algorithm. The AIRMETs results are represented by a single point on the plots.

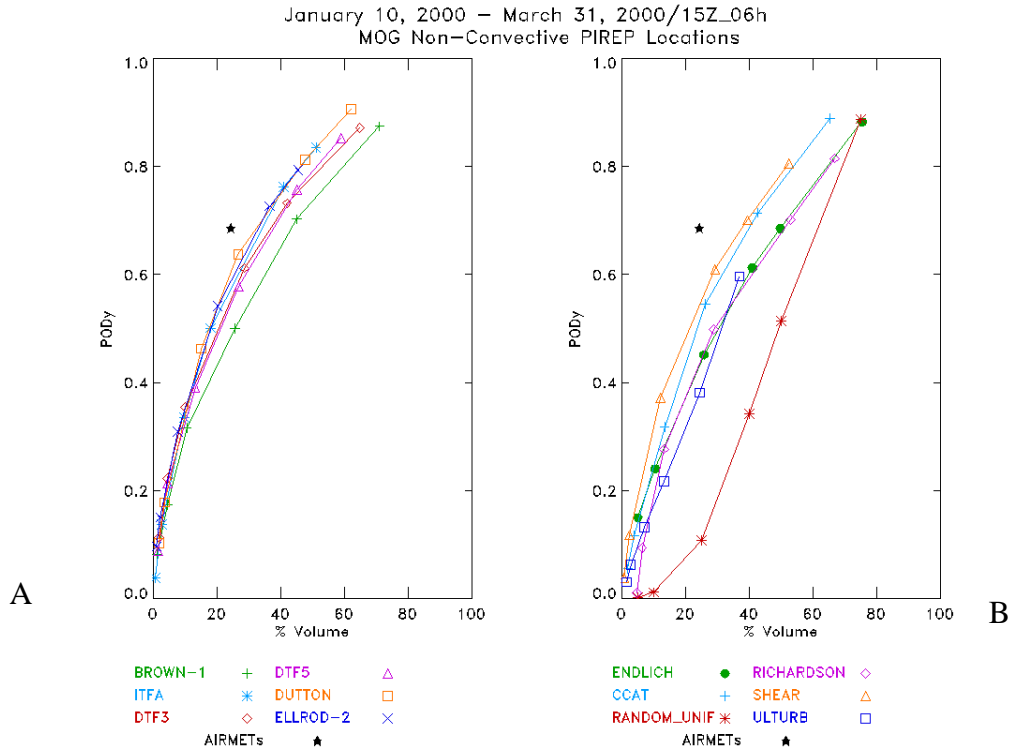


Figure 3. Same as Fig. 2, except for 1500 UTC, 6-h lead.

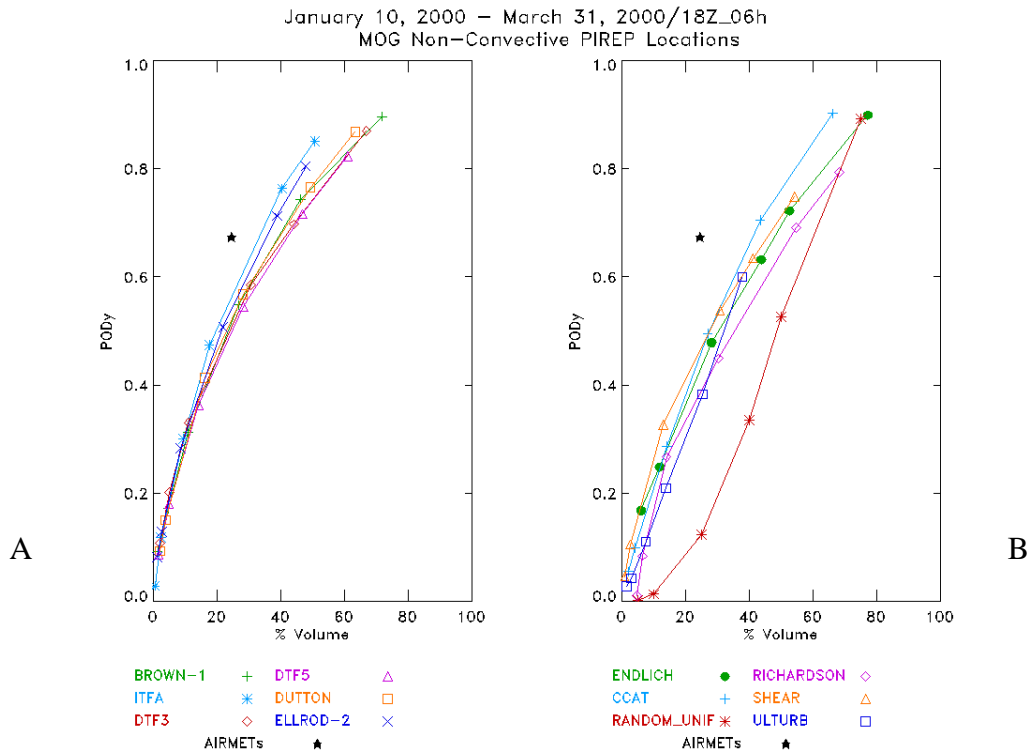


Figure 4. Same as Fig. 2, except for 1800 UTC, 6-h lead.

domain, where a low threshold may produce turbulence forecasts covering the entire domain, while higher values of the threshold limit turbulence to specific well-defined regions. As noted earlier, the ultimate goal for improved forecasting performance is to maintain a reasonable % Volume while improving the PODy, PODn, and TSS statistics, (i.e., moving closer to the upper left hand corner of the PODy vs. % Volume and PODy vs. 1-PODn plots).

Initial examination of the PODy vs % Volume plots (Figs. 2-4) indicates that differences in performance between the algorithms generally are small, particularly for those algorithms in Group A, as shown by the tight cluster of lines. The algorithms of Group A, which include Brown-1, ITFA, DTF3, DTF5, Dutton, Ellrod index, generally perform better than the algorithms of Group B, which include Endlich, CCAT, Random, Richardson Number, and Shear. This result is consistent with the results presented in the post-analysis. For instance, at a 20% volume, the ranges of values of MOG PODy for Groups A and B are 0.42 – 0.54 and 0.30 – 0.48, respectively; a difference of nearly 0.12 at the lower bound and 0.06 at the upper bound. These results are slightly better than those computed in TURB98-99. Nevertheless, all of the algorithms perform better than the Random “no-skill” forecast, indicating that each of the algorithms have some skill in forecasting turbulence. For all time periods, the AIRMETs include a somewhat smaller volume than the algorithms, for comparable MOG PODy values.

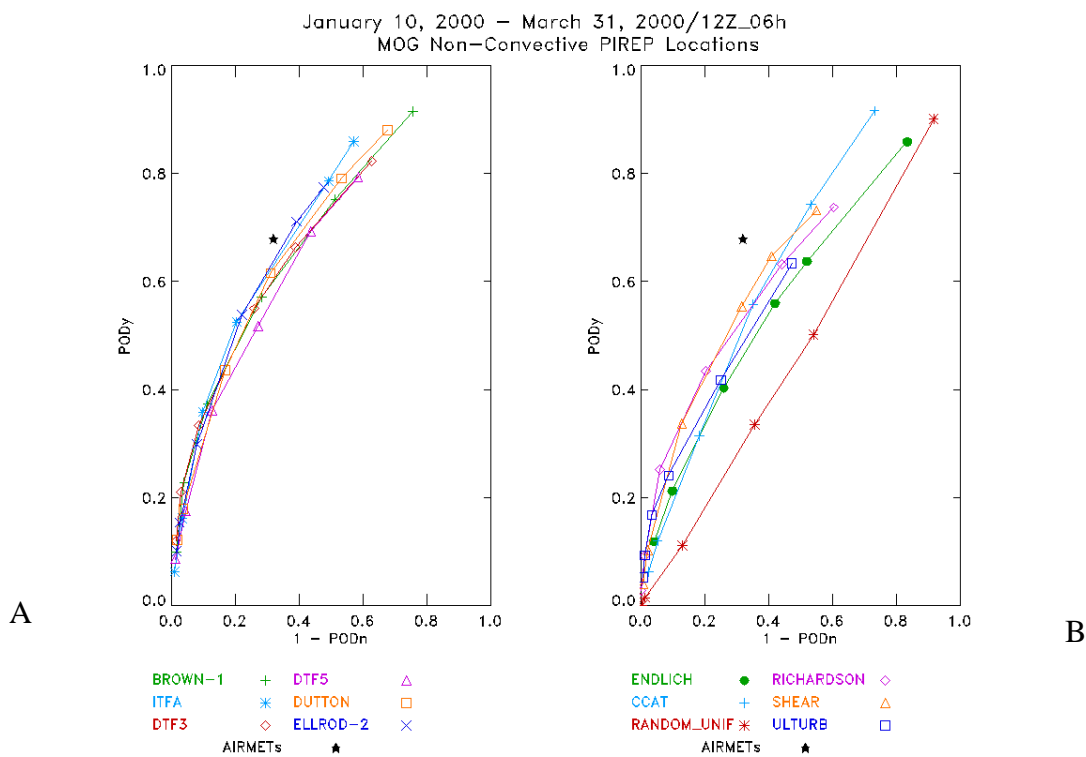


Figure 5. Same as Fig. 2, except for 1200 UTC, 6-h lead and for PODy vs. 1-PODn.

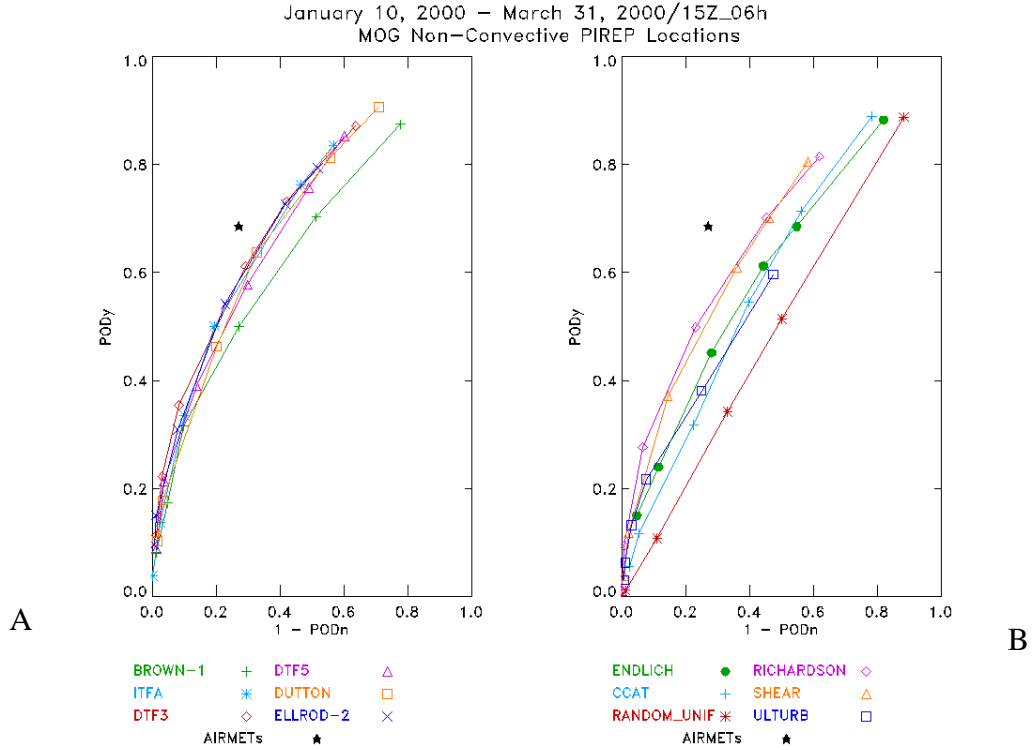


Figure 6. Same as Fig. 2, except for 1500 UTC, 6-h lead and for PODy vs. 1-PODn.

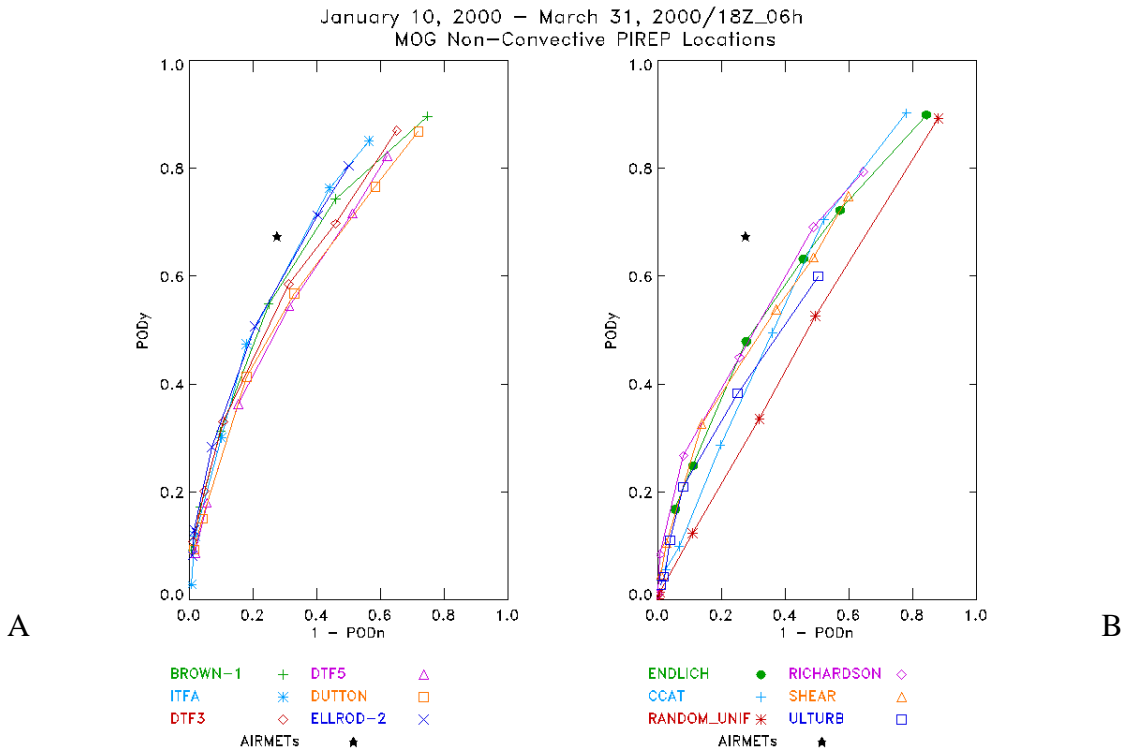


Figure 7. Same as Fig. 2, except for 1800 UTC, 6-h lead and for PODy vs. 1-PODn.

Plots of PODy vs 1-PODn are presented in Figs. 4-6. As the issue time varies from 1200 to 1800 UTC, only small differences in algorithm performance are noted. For instance, in Group A, lower skill was identified for DTF5 and Brown while the Ellrod index and ITFA algorithms show the best skill. Of the algorithm in Group B, the Shear and Richardson Number show some skill above the rest for the 1200 and 1500 UTC issuances. The values for the AIRMETs remain above those computed for the algorithms with greater separation between the AIRMETs and the algorithms at the 1500 and 1800 UTC issuances.

The TSS vs % Volume plots are shown in Figs. 8-10. As described in Section 5.2, the TSS is a measure of the ability of the algorithms to discriminate between Yes and No reports of turbulence, and is a combination of the PODy and PODn statistics. As indicated by the MOG PODy plots, all of the algorithms show some skill in forecasting turbulence since the TSS values for all algorithms lie above the 0.0 “no-skill” line. ITFA and Ellrod index remain the best performers, by a slight margin, while DTF5 and Brown fall below the other algorithms in Figs. 8 and 9, respectively. Since AWC forecasters often rely on the Richardson Number and Shear as guidance to defining areas of turbulence (Mahoney and Brown 2000), it is worth noting that the skill statistics for the Shear and Richardson Number algorithms remain higher than the skill statistics for the other algorithms in Group B, particularly for the 1200 and 1500 UTC issue times. The results again indicate that the AIRMETs out-perform the algorithms at least somewhat, with TSS values around 0.4 and % Volume values of 25%. However, it is important to remember that the AIRMETs are human generated forecasts that benefit greatly from forecaster experience and the availability of an array of various model-based forecasts of turbulence, many of which include output from the algorithms evaluated in this exercise.

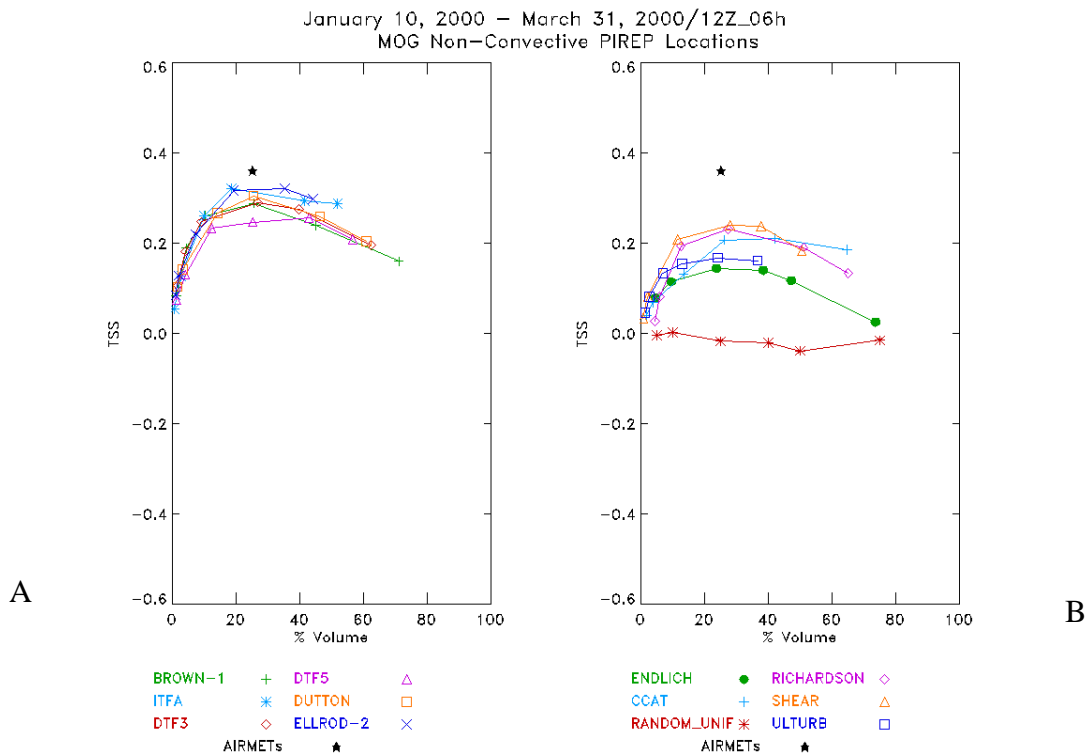


Figure 8. Same as Fig. 2, except for 1200 UTC, 6-h lead and for TSS vs. % Volume.

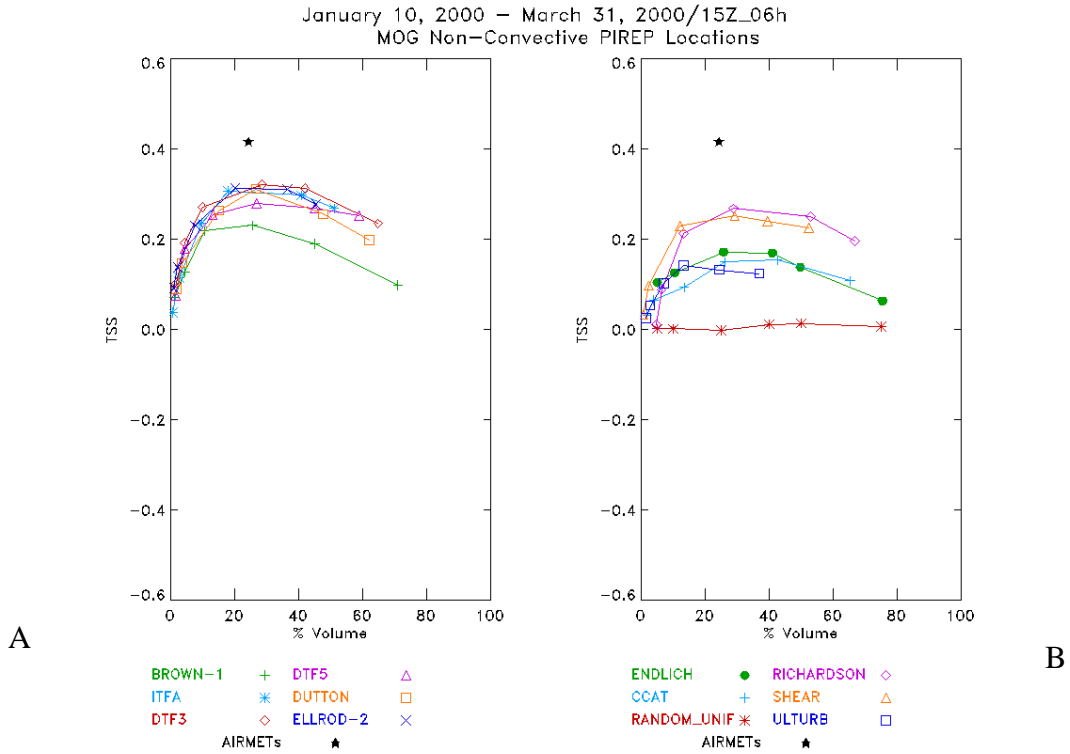


Figure 9. Same as Fig. 2, except for 1500 UTC, 6-h lead and for TSS vs. % Volume.

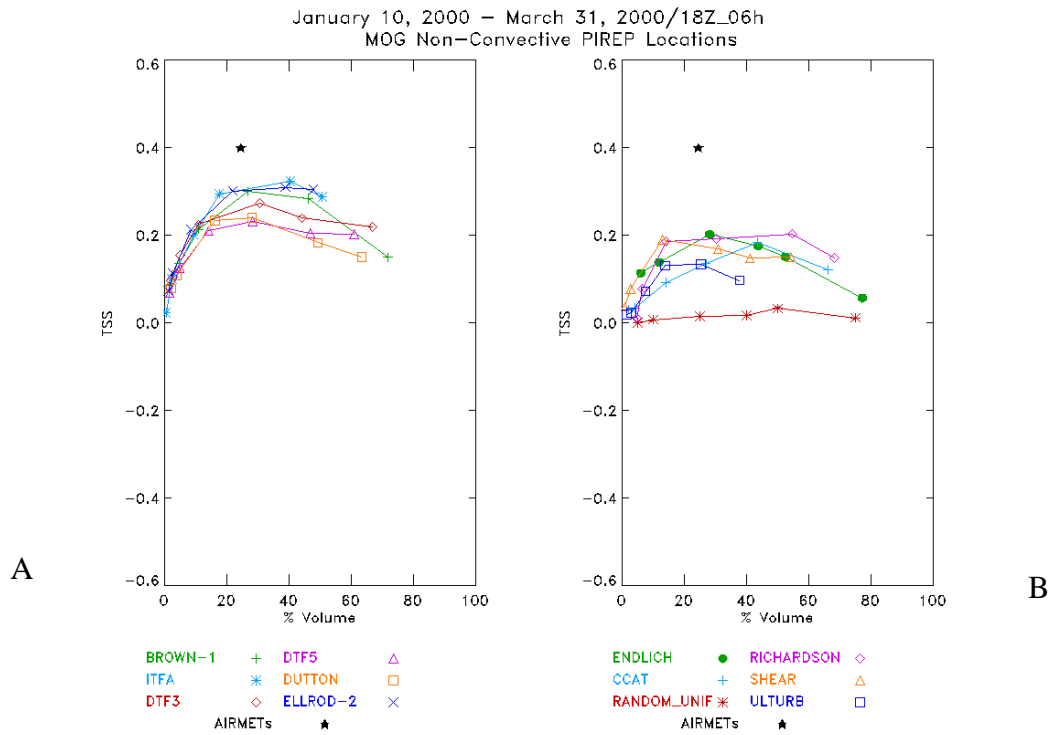


Figure 10. Same as Fig. 2, except for 1800 UTC, 6-h lead and for TSS vs. % Volume.

6.2.2 Variations with time

Figures 11-16 illustrate the variations in MOG PODy, and PODn by week for the 6-h lead time forecasts for the 1200, 1500, and 1800 UTC issue times. Each line on the plots represents results for one of the following algorithms: DTF3, Dutton, Ellrod index, and ITFA. The algorithm verification results were filtered to select the algorithm threshold for each algorithm that typically produced an overall MOG PODy value between 0.5 and 0.6. The algorithm thresholds are listed in the caption. Each symbol on a line represents a statistic generated from 7-day accumulations of Yes/No pairs.

Inspection of the MOG PODy statistics (Figs. 11-13) indicates a large variability in algorithm performance from week to week and over the various issue times. For instance, values of MOG PODy vary between values of 0.3 on “bad” days and 0.85 on “good” days. Possible explanations for these differences may be missing data during the period, inherent model biases, algorithm biases, or changing weather features. Further inspection of the figures indicates that the Dutton algorithm often has the largest MOG PODy and the smallest PODn. This result illustrates the trade off between obtaining a large MOG PODy and maintaining a large PODn. Several peaks in MOG PODy for the weeks ending on 26 January, 16 February, 1 March, and 22 March were identified. Investigation into the causes of these peaks will be performed in future analyses through comparisons of these results to those obtained from the Forecaster Assessment (Mahoney and Brown 2000). in hopes of determining the possible sources of the turbulence, which may provide clues regarding algorithm performance.

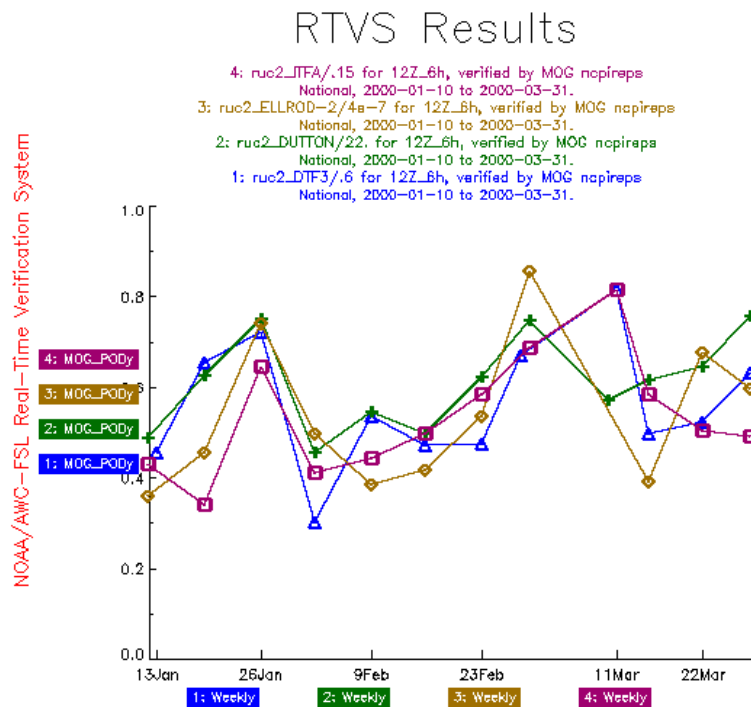


Figure 11. Time series plot for 1200 UTC issue 6-h lead for DTF3/0.6 (triangle), Dutton/0.22 (+), Ellrod index/4x10⁻⁷ (diamond), and ITFA/0.15 (square) by week for 10 January – 31 March 2000 for MOG PODy.

RTVS Results

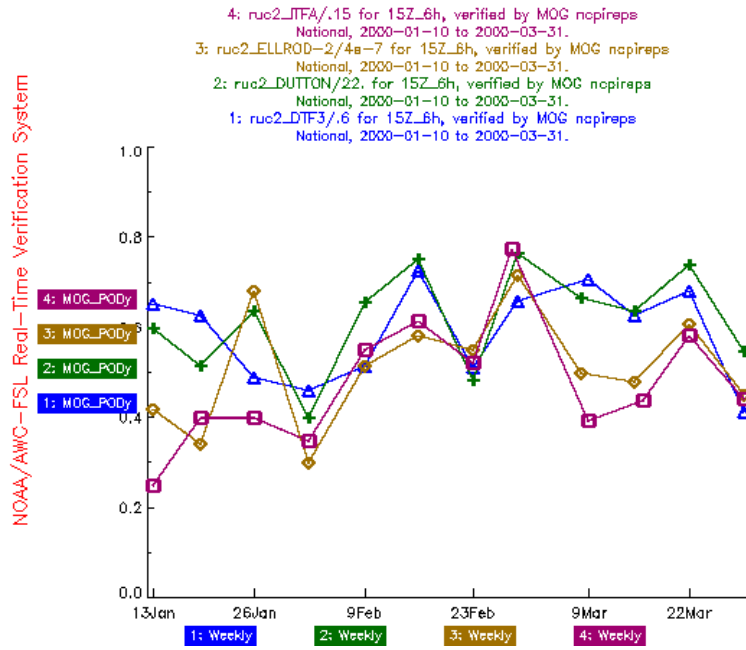


Figure 12. Same as Fig. 11, except for 1500 UTC, 6-h lead.

RTVS Results

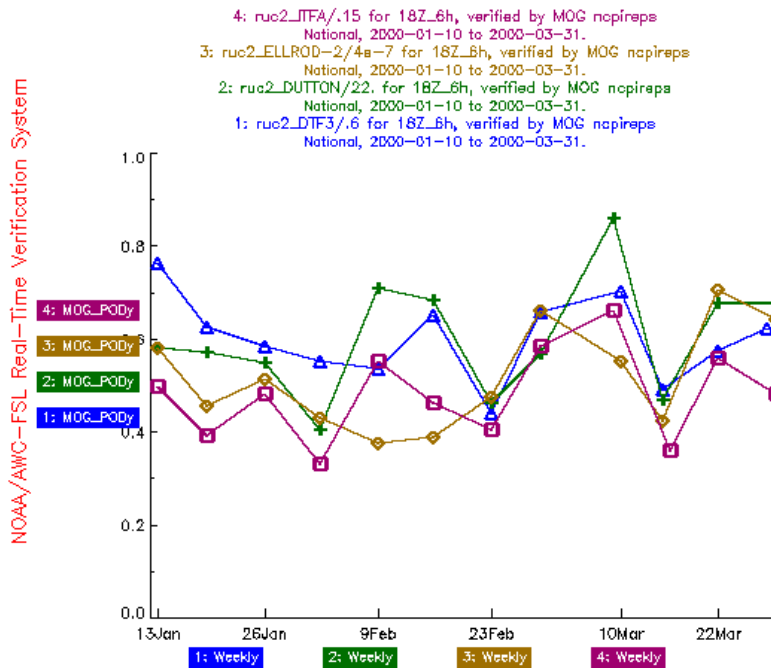


Figure 13. Same as Fig. 10, except for 1800 UTC, 6-h lead.

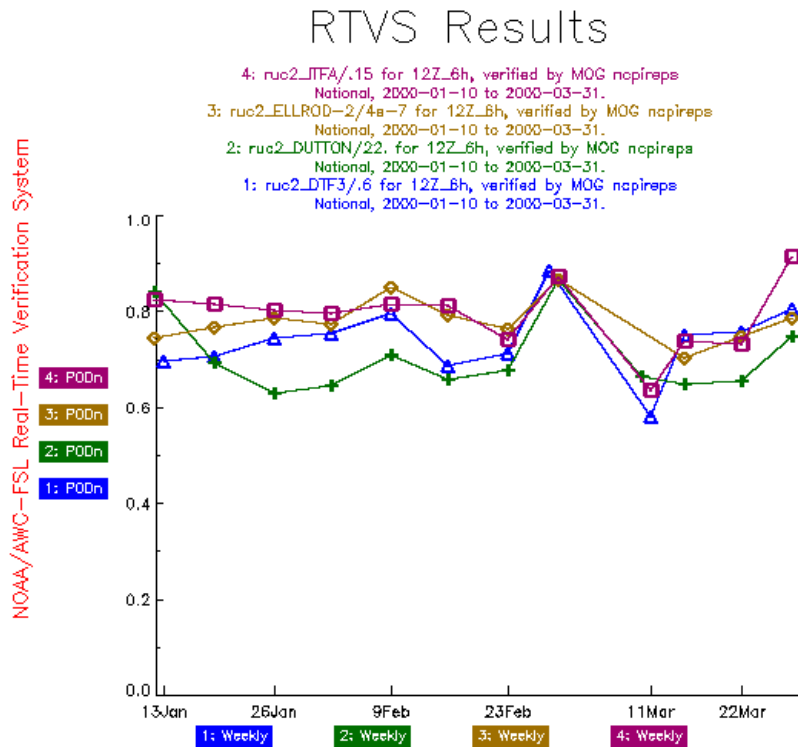


Figure 14. Same as Fig. 11, except for 1200 UTC, 6-h lead and for PODn.

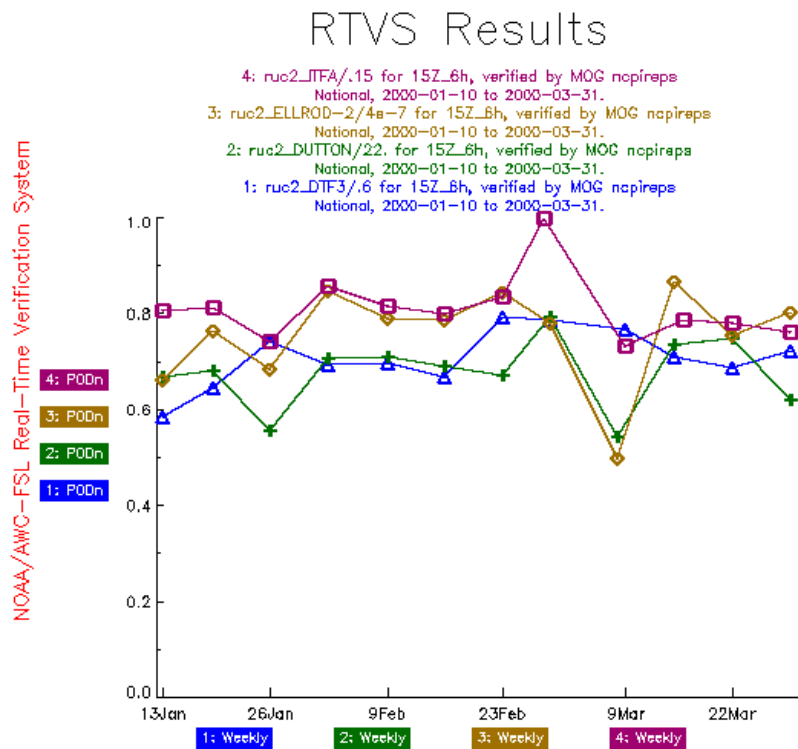


Figure 15. As in Fig. 11, except for 1500 UTC, 6-h lead and for PODn.

RTVS Results

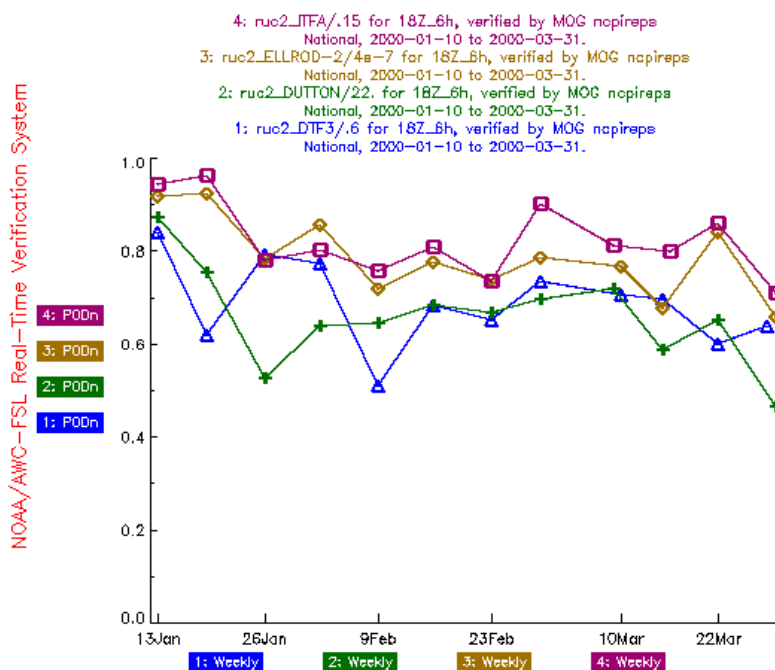


Figure 16. Same as Fig. 11, except for 1800 UTC, 6-h lead and for PODn.

6.2.3 Variations in ITFA

Figs. 17-20 are time series plots organized by issue time showing MOG PODy and PODn for ITFA at the 0.15 threshold. Each line on a plot represents a different forecast lead-time. Each symbol on the line is a statistic generated from 7-day accumulations of forecast/observation pairs. These results are of particular importance because ITFA is under development by the Turbulence PDT. Statistics such as these should be helpful in facilitating the development of ITFA from an experimental algorithm into a guidance product.

Initial inspection of Figs. 17-20 reveals larger variability in the MOG PODy values than in the PODn values with varying forecast lead time. As is generally expected, the statistical results for the longer forecast lead times are smaller than for the shorter leads. However, during some specific weeks the MOG PODy was larger for the longer lead time. For instance, the weeks ending 1 and 8 March have larger MOG PODy values for the 9- and 6-h leads than for the 3-h lead. MOG PODy values continue to be smaller for the 12-h lead than for the other lead times at the 1200 UTC (Fig. 17) and 1500 UTC (Fig. 19) issue time. For both the 1200 and 1500 UTC issuances, the statistics for MOG PODy at the 6-h lead time are nearly as large as those at the 3-h lead times. Thus, forecasters who use the ITFA forecasts with a 6-h lead time (Mahoney and Brown, 2000) are using one of the “best” ITFA forecasts.

RTVS Results

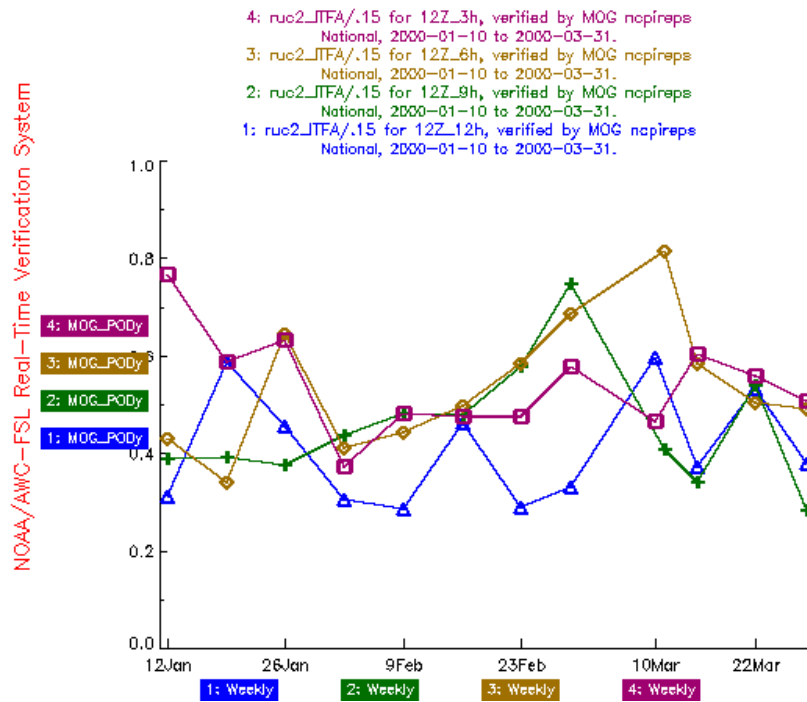


Figure 17. Time series plot for ITFA/.15 at 1200 UTC with 12-h lead (triangle), 9-h lead (diamond), and 3-h lead (square) by week for 10 January – 31 March 2000 for MOG PODy.

RTVS Results

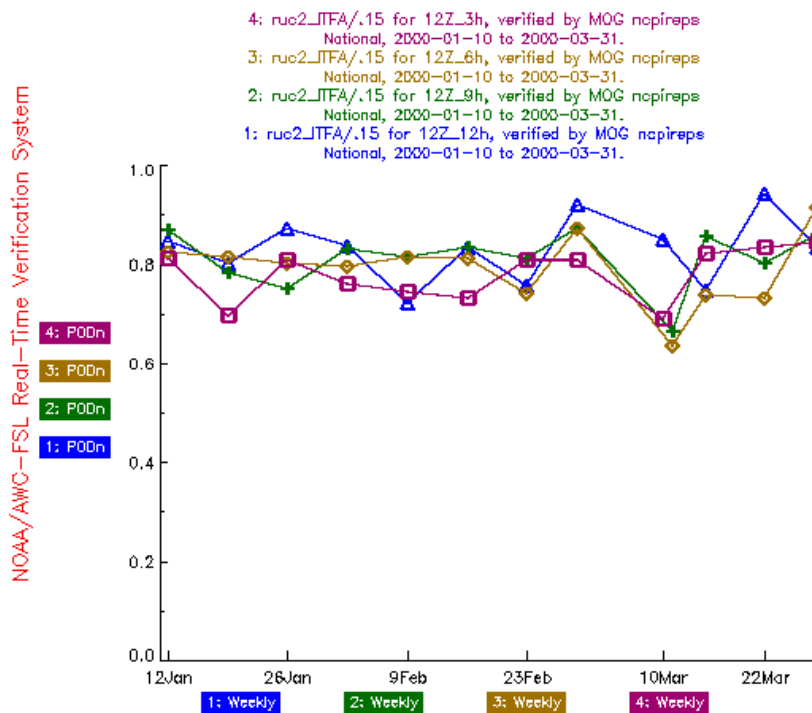


Figure 18. Same as Fig. 17, except for PODn.

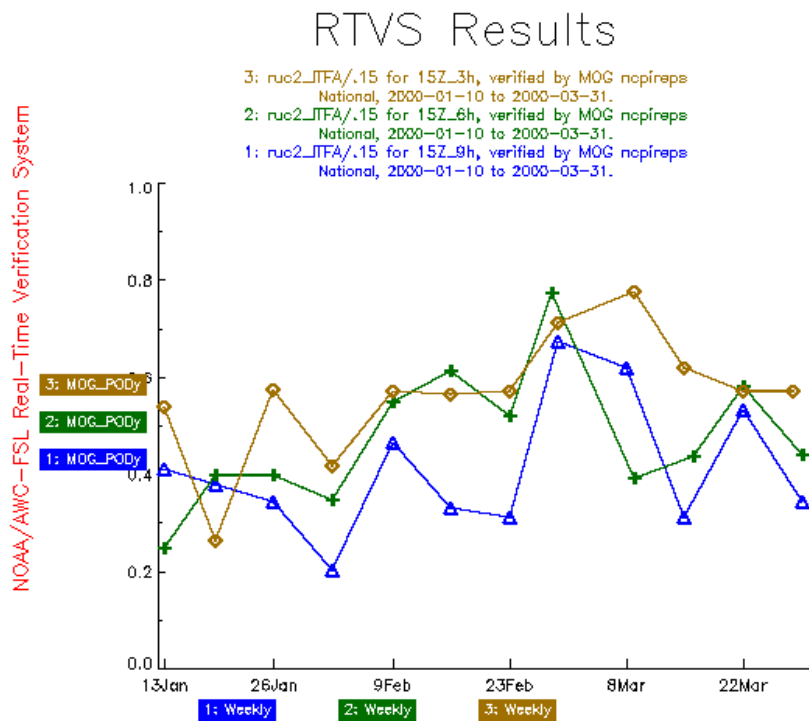


Figure 19. Time series plot for ITFA/.15 at 1500 UTC with 9-h lead (triangle), 6-h lead (+), and 3-h lead (diamond) by week for 10 January - 31 March 2000 for MOG PODy.

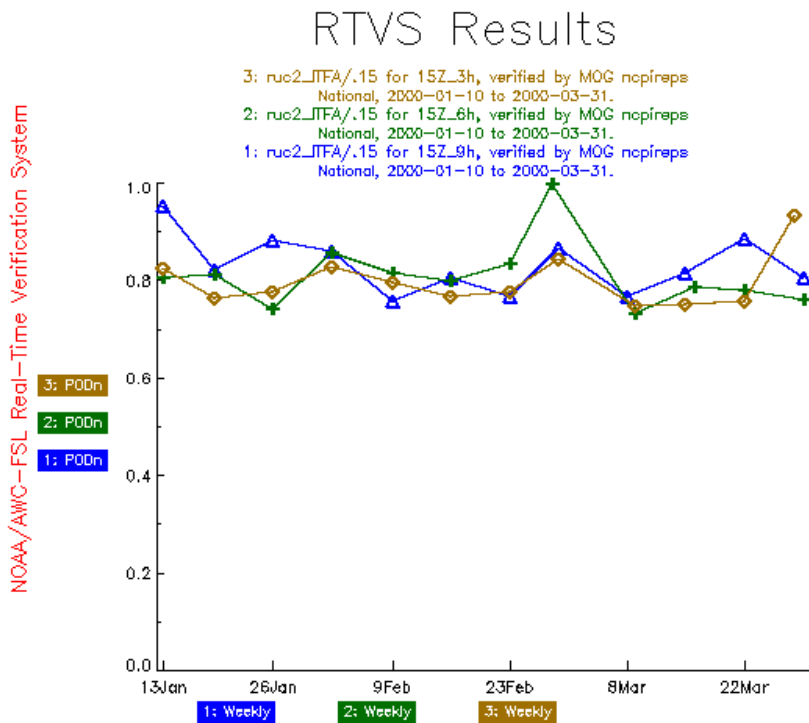


Figure 20. Same as Fig. 19, except for PODn.

6.2.4 Variations with height

Figs. 21-22 are height series plots for the 1500 UTC issue time, where MOG PODy and PODn are generated for ITFA at the 0.15 threshold and separated into 5,000 ft intervals from 20,000 ft to 40,000 ft. Each line on a plot represents a different forecast lead-time. Each symbol on the line is a statistic generated from 10 January – 31 March 2000.

As is generally expected, the 3-h lead forecasts have the largest MOG PODy; the smallest MOG PODy value was computed for the 9-h lead. Moreover, the difference in MOG PODy (Fig. 21) between the 3- and 9-h lead times below 35,000 ft is nearly 0.25. Interestingly, the MOG PODy value for the 6-h lead below 35,000 ft is similar to the 9-h value, with an approximate value of 0.35, but the 6-h MOG PODy value improves dramatically to 0.5 above 35,000 ft, nearly matching the value for the 3-h lead. The PODn (Fig. 22) for the 3-h lead is largest below 35,000 ft while above 35,000 ft, the 3-h value is smaller than the values for the other lead times. Therefore, at all levels, largest MOG PODy and PODn values were obtained by the 3-h lead, however, above 35,000 ft, the 6-h MOG PODy and PODn values approach those of the 3-h forecast. Variations in these statistics with issue time may occur. Those results can be found on the Web.

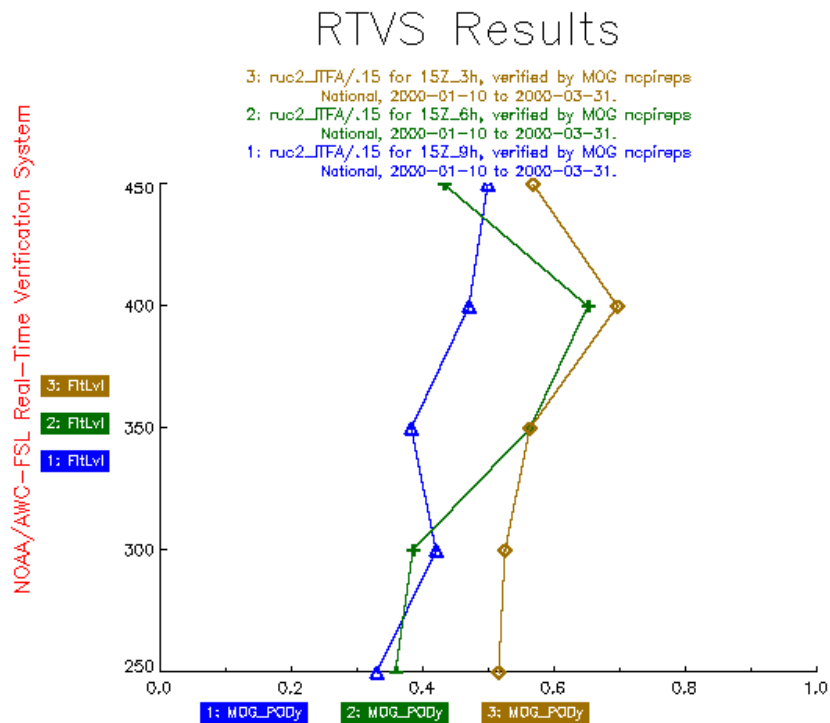


Figure 21. Height series plot of MOG PODy for ITFA at 0.15 threshold for the 1500 UTC, 9-h lead (triangle), 6-h lead (+), and 3-h lead (diamond) lead forecasts for 10 January – 31 March 2000 for MOG PODy. The MOG PODy values are computed for the range of 5,000 ft levels and are plotted on the y-axis at the upper limit of the range.

RTVS Results

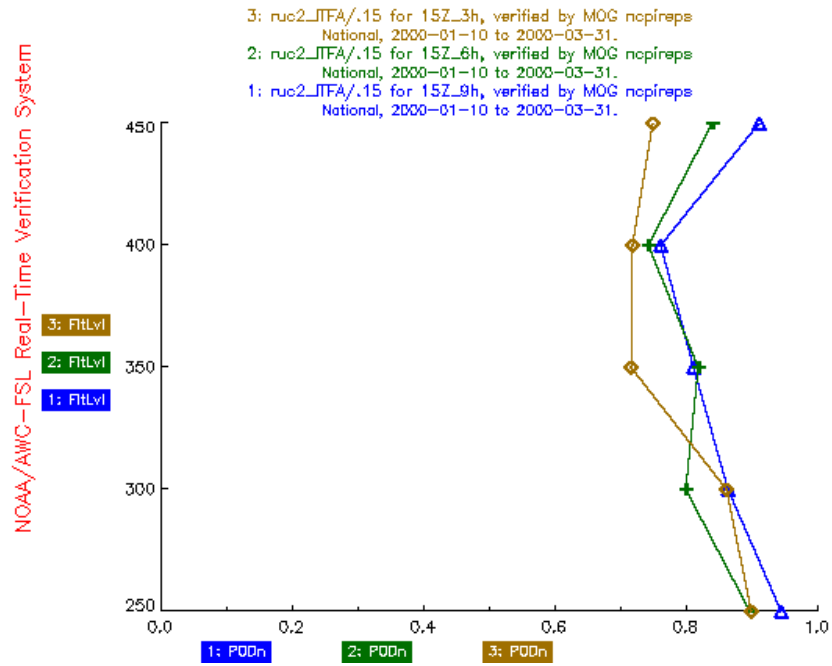


Figure 22. Same as Fig. 21, except for PODn.

7. Post-analysis

Basic results of the TURB2000 post-analyses are described in this section. As was the case for TURB98-99, this effort has included numerous steps. These steps included cataloging available data and making efforts to fill in the missing pieces, selecting additional algorithm thresholds to provide a more complete picture of algorithm performance, implementing some additional statistical methods, and re-evaluating the algorithm output using the additional data and techniques. The verification analyses were limited to dates and times when algorithm output was available for all algorithms, so all results would be comparable. A total of 142 3-h forecasts, 144 6-h forecasts, 93 9-h forecasts, and 48 12-h forecasts were included.

The mechanics of the verification analyses applied in the NCAR verification system are somewhat different than the methods used in RTVS. These methods are described briefly in Section 7.1. Some results of the post-analyses are presented in Section 7.2.

7.1 Mechanics

As in TURB98-99, the NCAR verification system used a matching approach to connect algorithm output to PIREPS. With this method, a PIREP is first matched to all of the model levels (i.e., flight levels) in the range of altitudes reported in the PIREP. Then, at each level, the four surrounding model grid points are compared to the PIREP. If any one of the four grid points has a Yes forecast, then a Yes forecast is assigned to the PIREP. If none of the four grid points has a Yes forecast, then a No forecast is assigned to the PIREP. The same procedure is applied to the AVAR observations. Essentially, this approach amounts to using the largest value of the algorithm output at the four surrounding grid points as the forecast assigned to the PIREP⁵.

To mimic this system, the AIRMETs also are treated somewhat differently by the NCAR verification system than by RTVS. In particular, the RUC-2 grid is overlaid on the AIRMETs and PIREPs. If any of the four RUC-2 grid points surrounding a PIREP is inside an AIRMET, then the PIREP is assigned a Yes AIRMET forecast; if none of the grid points are inside an AIRMET, then the PIREP is assigned a No AIRMET forecast.

Additional thresholds were included in the analyses for all algorithms. These thresholds were selected by examining the real-time results to identify regions where there was a large jump in PODy and/or PODn between the original thresholds. Additional thresholds also were added after examining some of the initial post-analysis results. Table 5 shows the algorithm thresholds that were used in most of the post-analyses.

⁵ Note that in the case of Richardson number, the minimum value is assigned.

Table 5: Algorithm thresholds used in post-analysis.

Algorithm	Thresholds
Brown-1	3.5×10^{-5} , 5×10^{-5} , 6×10^{-5} , 7×10^{-5} , 7.5×10^{-5} , 8×10^{-5} , 9×10^{-5} , 10×10^{-5} , 12×10^{-5} , 14×10^{-5} , 15×10^{-5} , 20×10^{-5} , 30×10^{-5}
CCAT	3×10^{-10} , 5×10^{-10} , 10^{-9} , 3×10^{-9} , 4×10^{-9} , 5×10^{-9} , 7×10^{-9} , 9×10^{-9} , 1.3×10^{-8} , 2×10^{-8} , 4×10^{-8}
DTF3	0.10, 0.20, 0.30, 0.40, 0.45, 0.50, 0.70, 0.90, 1.30, 2.00, 3.00
DTF5	0.06, 0.08, 0.10, 0.12, 0.15, 0.17, 0.20, 0.25, 0.30, 0.34, 0.50, 0.70, 0.90, 1.40, 1.80
Dutton	7, 10, 12, 15, 18, 20, 22, 25, 30, 40, 45, 60, 80
Ellrod-2	10^{-7} , 2×10^{-7} , 2.5×10^{-7} , 3×10^{-7} , 3.5×10^{-7} , 4×10^{-7} , 5×10^{-7} , 6×10^{-7} , 7×10^{-7} , 9×10^{-7} , 12×10^{-7} , 16×10^{-7}
Endlich	0.040, 0.080, 0.120, 0.150, 0.175, 0.20, 0.225, 0.275, 0.300, 0.400, 0.500, 0.600
ITFA	0.02, 0.05, 0.07, 0.08, 0.09, 0.10, 0.13, 0.17, 0.20, 0.30, 0.40, 0.50
Richardson	0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 7.0, 9.0, 12.0, 15.0
SCATR	10^{-7} , 1.5×10^{-5} , 2×10^{-5} , 10^{-4} , 2×10^{-4} , 3×10^{-4} , 4×10^{-4} , 6×10^{-4} , 8×10^{-4} , 0.001, 0.002, 0.003
Shear	0.001, 0.002, 0.0025, 0.003, 0.0035, 0.004, 0.005, 0.006, 0.007, 0.008, 0.010, 0.015

7.2 Results

In most cases, results are presented only for MOG PIREPs. Generally, the results for All PIREPs are similar to those for MOG PIREPs, with somewhat smaller values of PODy when all PIREPs were included. The statistics also are broken down by lead time and valid time. The analyses were limited to only include forecasts when data were available from all algorithms, and when AIRMETs and PIREPs also were available.

7.2.1 Overall results for 3-h forecasts

Overall results are shown in Figs. 23 and 24, for the 3-h lead time. The plots in Figs. 23 and 24 were created by combining the counts for all 3-h forecasts together. Figure 23 is a plot of PODy (MOG PIREPs) versus 1-PODn, whereas Fig. 24 shows plots of PODy versus % Volume. As in Figs. 2-7, the individual points on the algorithm curves represent individual thresholds used to create Yes/No forecasts. Results for better forecasts are located closer the upper left corner of the diagrams.

Two groups of algorithms are shown for each combination of statistics, in order to make the diagrams more clear. Group A includes Brown-1, DTF3, Ellrod-2, ITFA, and Richardson

number, while Group B includes CCAT, DTF5, Dutton, Endlich, SCATR, and Shear. Each plot also includes a point representing the AIRMETs. Recall that the PODy vs. 1-PODn plots represent the forecasts' ability to discriminate between Yes and No observations, whereas the PODy vs. % Volume plots consider the tradeoffs between capturing regions of turbulence and the size of the warned region.

The plots in Figs.23 suggest that there are fairly large differences in forecasting performance between the Group A and most of the Group B algorithms. In particular, the Group A algorithms generally lie quite a bit further above the 45° line, which defines no skill. In fact, most of the lines for the Group A algorithms are close to – just below – the point defining the AIRMET skill. In terms of this measure of performance, ITFA, DTF3, Richardson number, and the Ellrod index appear to have the most skill.

In contrast, the PODy vs. 1-PODn plots for the Group B algorithms indicate that the SCATR index has negative skill, since all of the points on the SCATR curve lie below the 45° line. Moreover, the curve for the Endlich algorithm is located just above the 45° line. Only the curve for DTF5 is relatively close to the AIRMET point. The curves for other algorithms are located well below the AIRMET point. It is worth noting that the PODy vs. 1-PODn curve for the random algorithm (not shown) fell *on* the 45° line, as it should, since this line marks the boundary between positive and negative skill relative to chance.

Similar differences in performance between the Group A and Group B algorithms are apparent in the plots of PODy vs. % Volume (Fig. 24). In general, for Group A, the Ellrod-2 index has the best overall performance in terms of this pair of statistics, and the Brown-1 algorithm has the worst performance, although the differences among results for this group of algorithms are quite small. In most cases the lines for the Group A algorithms are just below the point depicting the performance of the AIRMETs. For the Group B algorithms, the curves are further below the point associated with the AIRMETs (with the exception of DTF5), and the curve for the SCATR index is well below the other curves, for most of the range of % Volume.

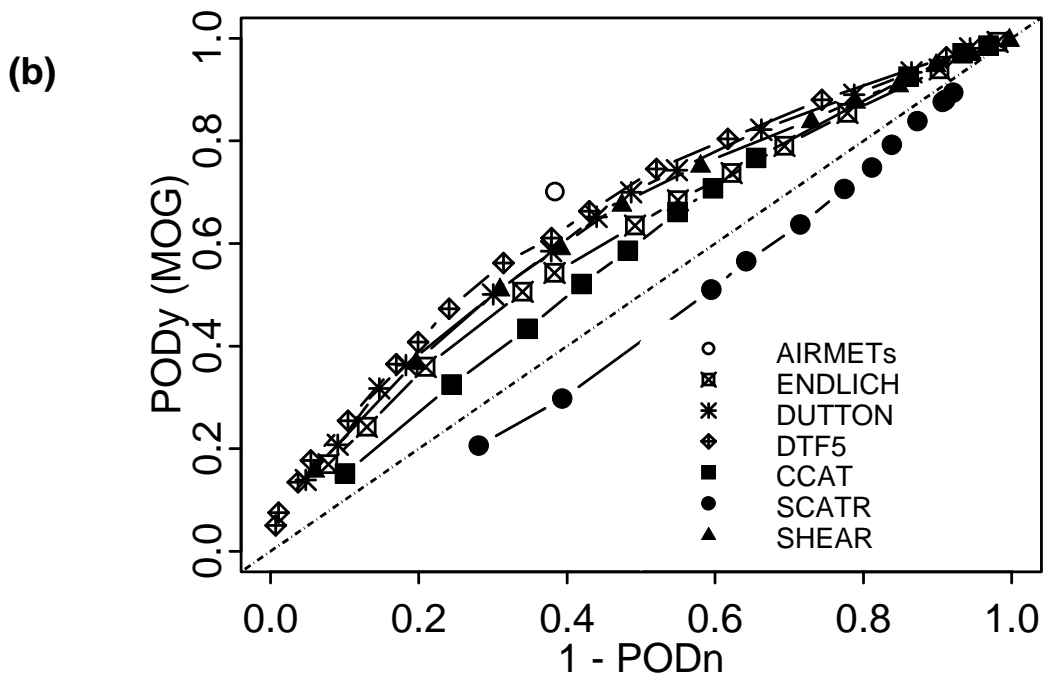
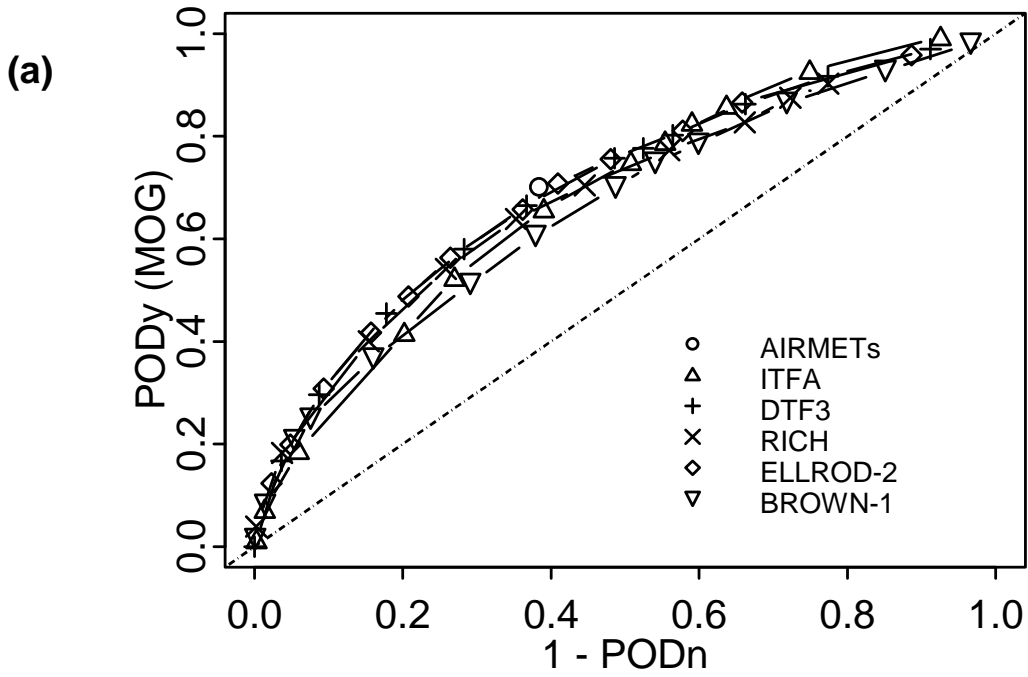


Figure 23. Plots of PODY vs. 1-PODn for 3-h forecasts: (a) algorithm Group A; and (b) algorithm Group B.

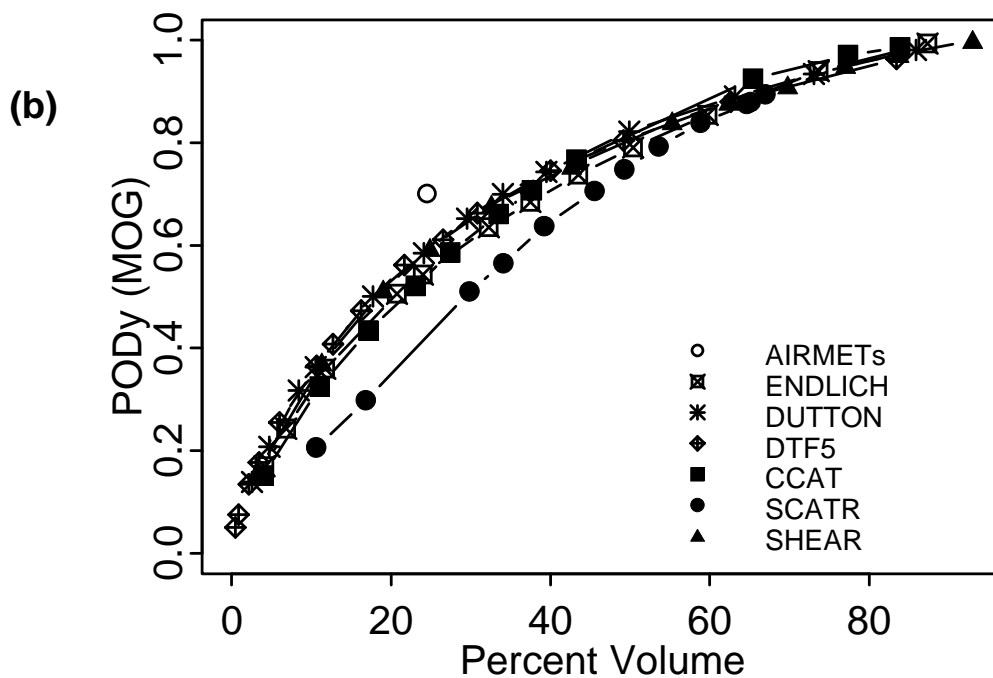
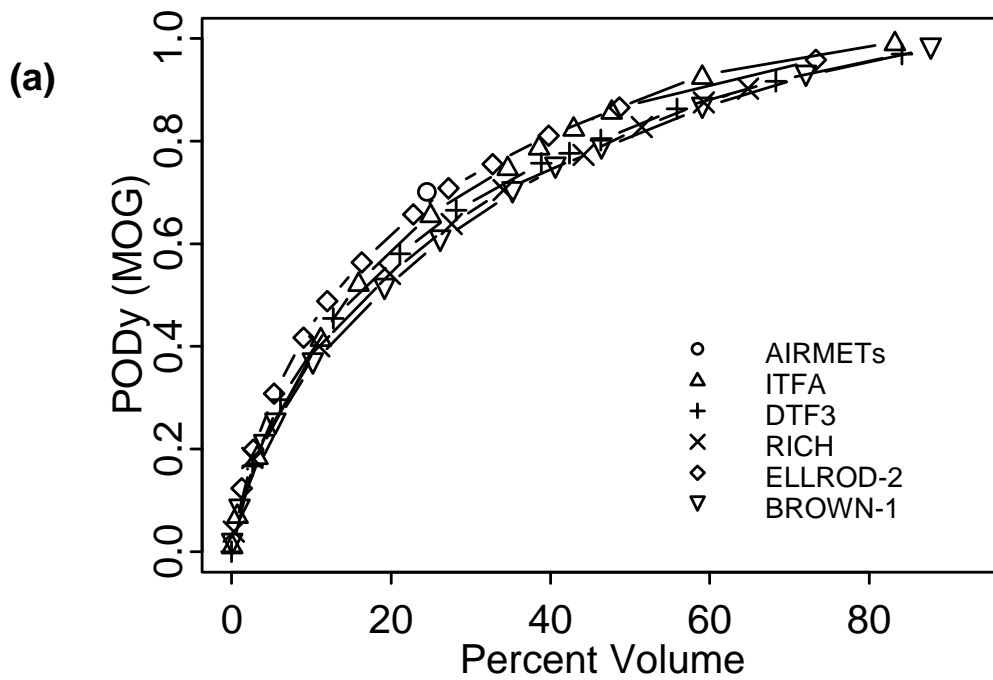


Figure 24. Plots of PODy vs. % Volume for 3-h forecasts: (a) algorithm Group A; and (b) algorithm Group B.

As in TURB98-99, the overall 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 6 shows the results of this exercise for the 3-h forecasts. This table includes a variety of statistics associated with the specified thresholds. In addition, Table 6 includes (in the last column) estimates of the total area under the curves relating PODy (for MOG PIREPs) to 1-PODn (i.e., the curves in Fig. 23). This statistic is not included for the AIRMETs since only one point is associated with the AIRMETs, whereas many points are associated with each of the algorithms; the area estimate for the AIRMETs would be smaller than the estimates for most of the algorithms simply due to this difference.

Two values of PODy are included in Table 6 – one for All severities and one for MOG severities. In almost all cases (except for SCATR), PODy (MOG) is somewhat larger than PODy (All). This result, which is consistent with results from TURB98-99, suggests that the MOG PIREPs are somewhat easier for the forecasts to capture than are PIREPs associated with less severe conditions. The PODn values vary among the algorithms, with the largest values of PODn achieved by the AIRMETs, DTF3, DTF5, Ellrod-2, ITFA, and Richardson number. Very poor values of PODn are achieved by CCAT, Endlich, and SCATR. Table 7 shows 95% confidence intervals for the PODn values in Table 6. These intervals reveal that the PODn values for CCAT, Endlich, and SCATR are significantly smaller than the PODn values for the other algorithms and for the AIRMETs. However, the PODn values for the other algorithms and for the AIRMETs are not significantly different from each other, since all of their confidence intervals overlap each other.

The TSS values in Table 6 provide a somewhat clearer comparison of the forecasting performance among the algorithms (i.e., because some of the variations in PODn are likely due to the minor variations in PODy among the algorithms). Among the different forecasts and algorithms, the largest TSS values are achieved by the AIRMETs, DTF3, DTF5, Ellrod-2, and ITFA. With regard to the ROC curve area, the best algorithm results are attained by DTF3, DTF5, Ellrod-2, ITFA, and Richardson number. Moreover, the ROC areas achieved by CCAT, Endlich, and SCATR are notably smaller than the values associated with the other algorithms and the AIRMETs. These results are not surprising, based on the curves in Figure 23.

In terms of the % Volume values in Table 6, the smallest (best) values are achieved by DTF3, Ellrod-2, and ITFA, with the smallest value associated with the AIRMETs. Because % Volume is strongly related to PODy, the small variations in PODy in Table 6 may have had some impact on these results. Thus, it is more appropriate to consider the Volume Efficiency (VE) values. The best (largest) VE values in Table 6 were achieved by Ellrod-2 and ITFA. The Richardson number has a relatively large % Volume value, and hence, a relatively small VE value.

Thus, the results in Table 6 suggest that there are some discernible differences in the results among the algorithms, with the apparently best, all-around, algorithm performance associated with Ellrod-2, ITFA, and DTF3.

Table 6: Verification statistics for all 3-h forecasts (all issue times combined), for thresholds with POD_y (MOG PIREPs) about the same as the POD_y for AIRMETs.

Algorithm	Threshold	POD _y (All)	POD _y (MOG)	POD _n	TSS	ROC Curve Area	Average % Vol	VE
AIRMETs	--	0.65	0.70	0.62	0.32	--	24.5	2.9
Brown-1	8x10 ⁻⁵	0.66	0.70	0.51	0.21	0.66	35.3	2.0
CCAT	4x10 ⁻⁹	0.68	0.71	0.40	0.11	0.58	37.7	1.9
DTF3	0.70	0.63	0.67	0.63	0.30	0.70	28.2	2.4
DTF5	0.15	0.63	0.66	0.57	0.29	0.66	30.1	2.2
Dutton	20.00	0.67	0.70	0.51	0.21	0.64	34.1	2.0
Ellrod-2	3.5x10 ⁻⁷	0.66	0.71	0.59	0.30	0.70	27.2	2.6
Endlich	0.20	0.66	0.68	0.45	0.13	0.60	37.5	1.8
ITFA	0.13	0.61	0.66	0.61	0.27	0.68	25.0	2.6
Richardson	5.00	0.67	0.70	0.55	0.26	0.68	34.1	2.0
SCATR	0.0004	0.71	0.71	0.22	-0.07	0.44	45.5	1.6
Shear	0.006	0.65	0.68	0.53	0.20	0.63	32.6	2.1

Table 7. 95% confidence intervals for the PODn values shown in Table 6.

Algorithm	Threshold	PODn	95% Confidence Interval (Lower bound, Upper bound)
AIRMETS	--	0.62	0.59, 0.65
Brown-1	8×10^{-5}	0.51	0.48, 0.57
CCAT	4×10^{-9}	0.40	0.37, 0.45
DTF3	0.70	0.63	0.60, 0.66
DTF5	0.15	0.57	0.54, 0.60
Dutton	20.00	0.51	0.48, 0.56
Ellrod-2	3.5×10^{-7}	0.59	0.56, 0.63
Endlich	0.20	0.45	0.42, 0.48
ITFA	0.13	0.61	0.58, 0.65
Richardson	5.00	0.55	0.53, 0.59
SCATR	0.0004	0.22	0.20, 0.25
Shear	0.006	0.53	0.50, 0.56

7.2.2 Comparisons among lead times

The algorithm curves for the 6-, 9-, and 12-h lead times (Figs. 25-30) are qualitatively similar to the 3-h results in Figs. 23 and 24, although the quality of the forecasts does seem to degrade some by the 9-h lead time. In particular, all of the curves in Figs. 27 and 28 lie below the AIRMET point, whereas several comparable curves lie above the AIRMET point in Fig. 23 and 24.

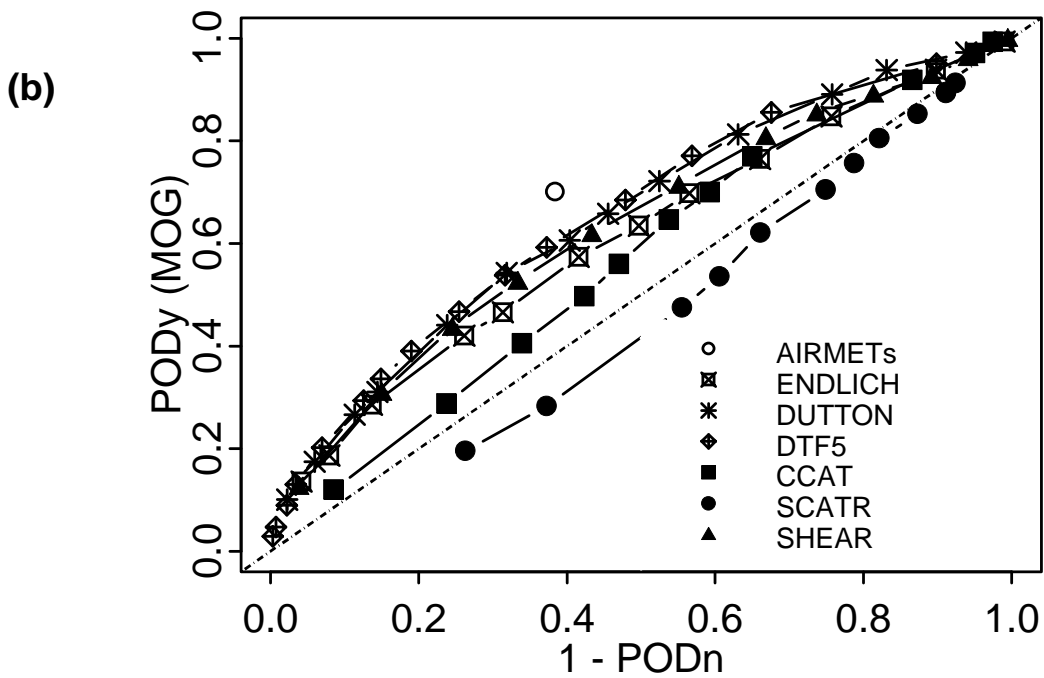
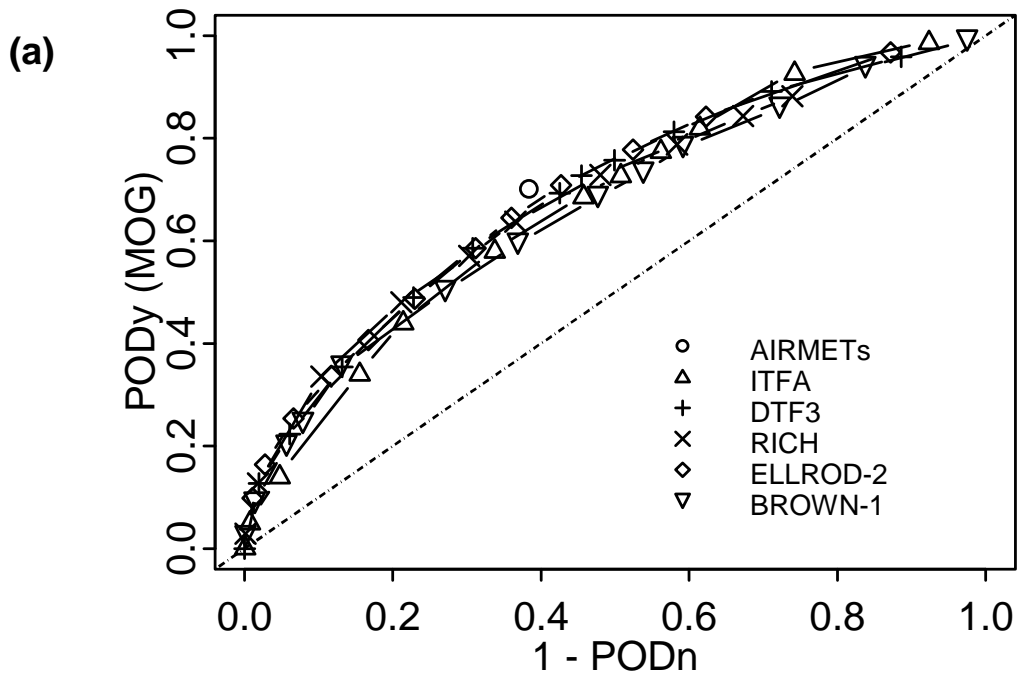


Figure 25. As in Fig. 23, for 6-h forecasts.

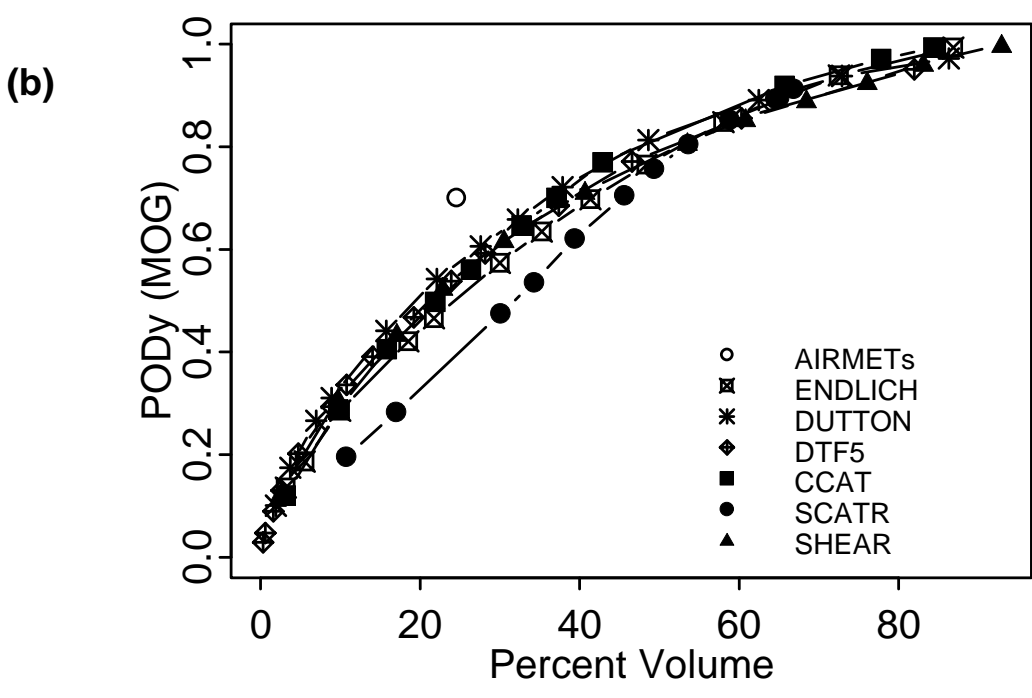
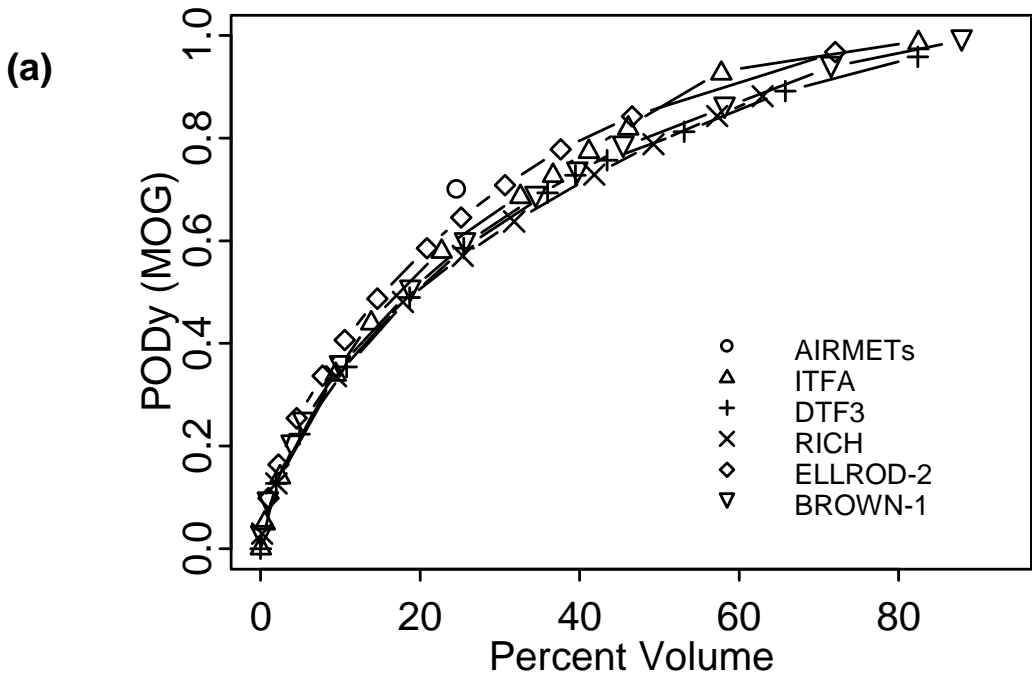


Figure 26. As in Fig. 24, but for 6-h forecasts.

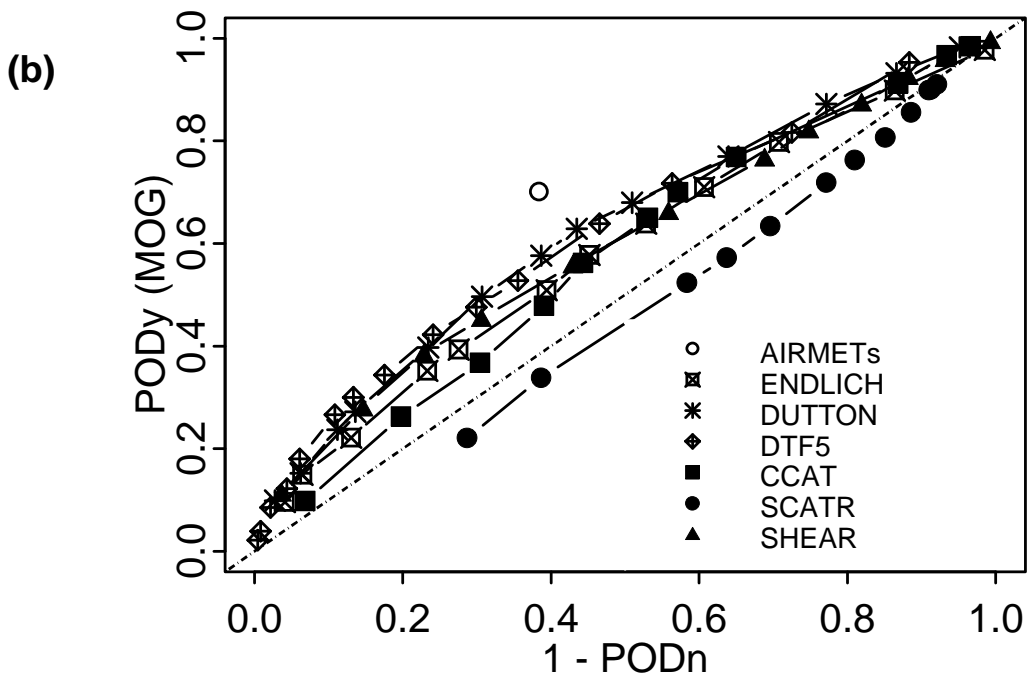
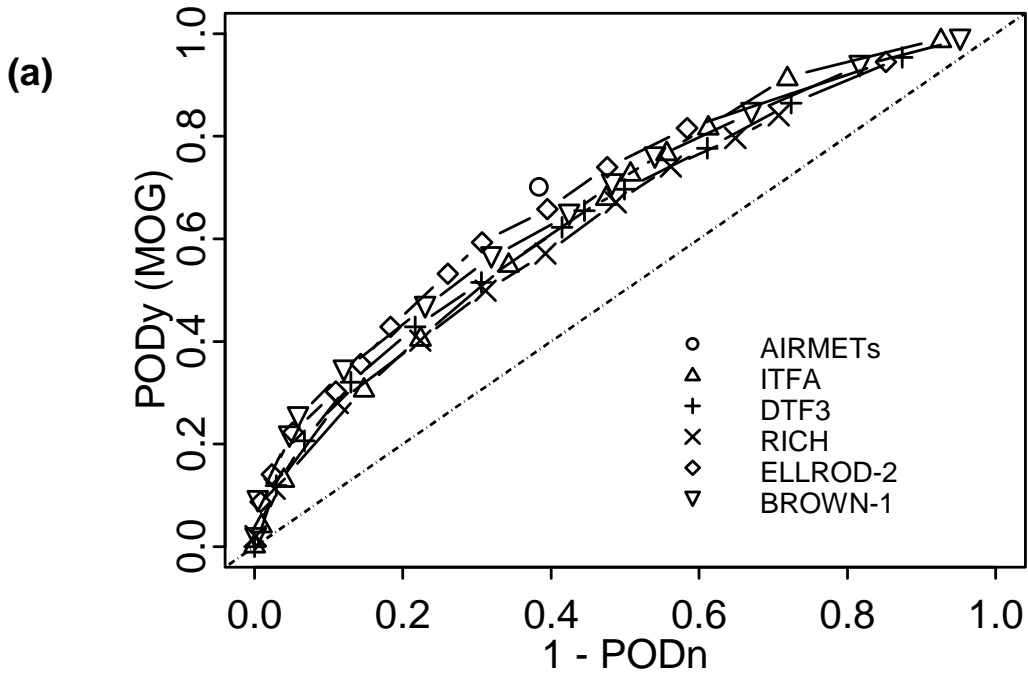


Figure 27. As in Fig. 23, for 9-h forecasts.

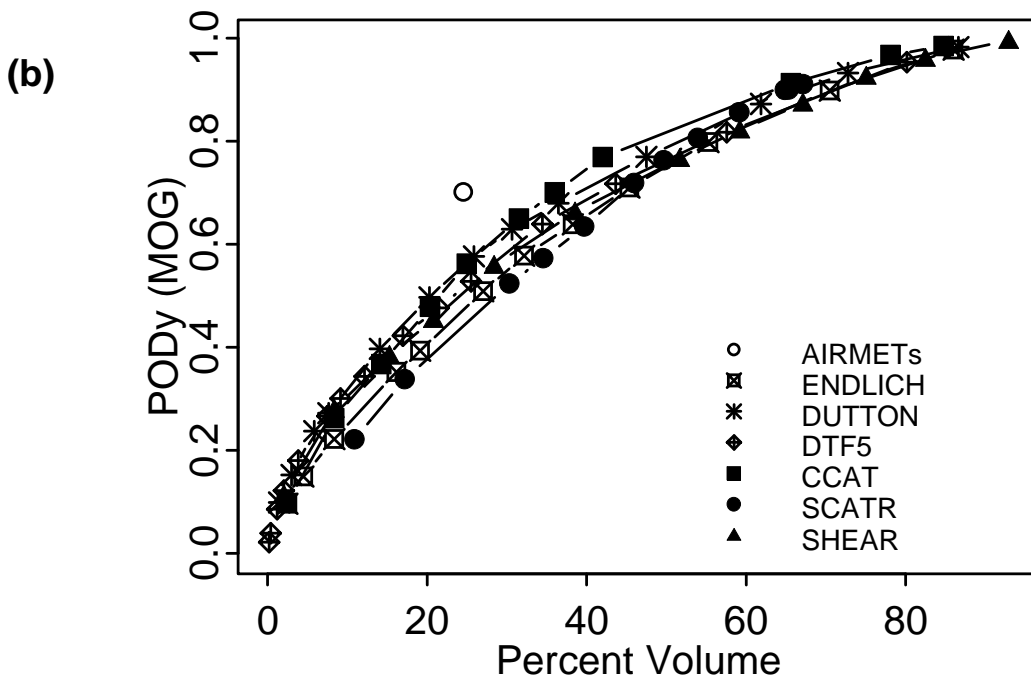
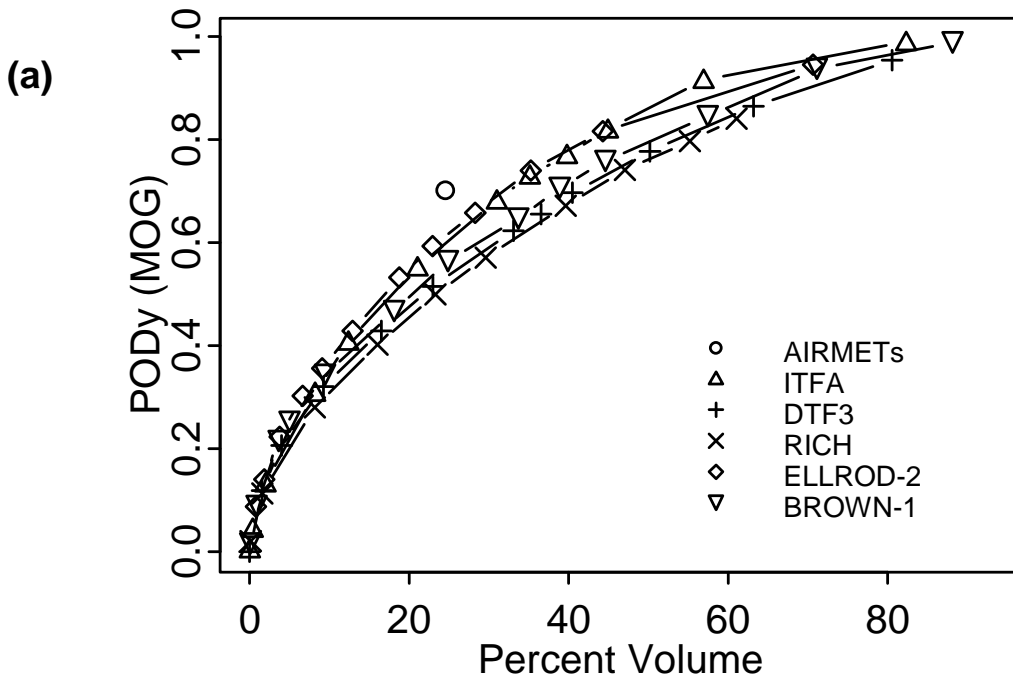


Figure 28. As in Fig. 24, for 9-h forecasts.

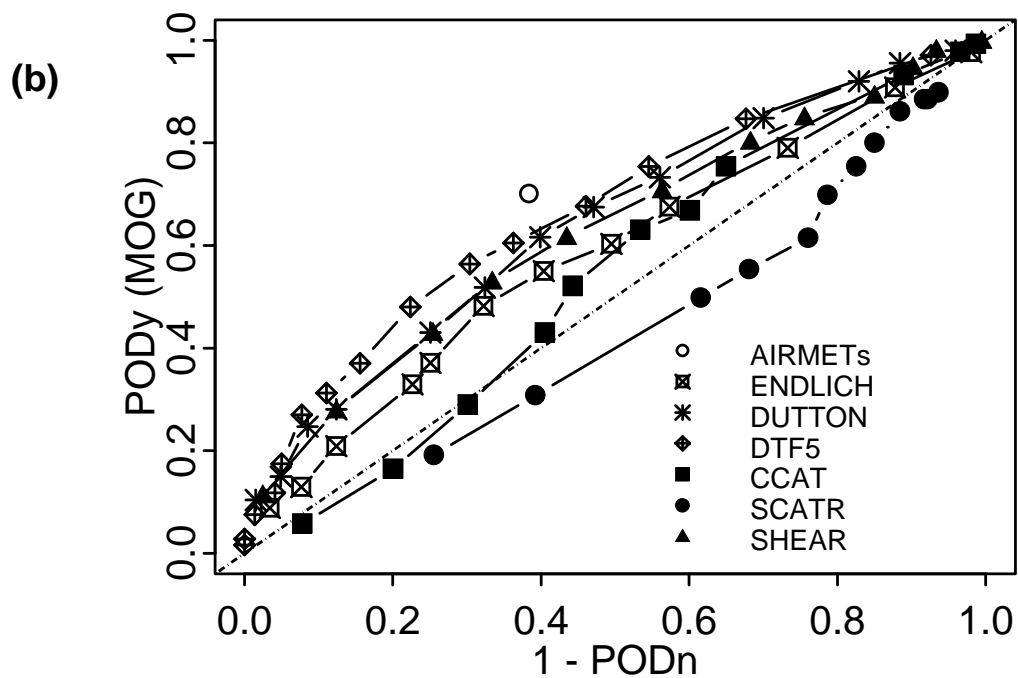
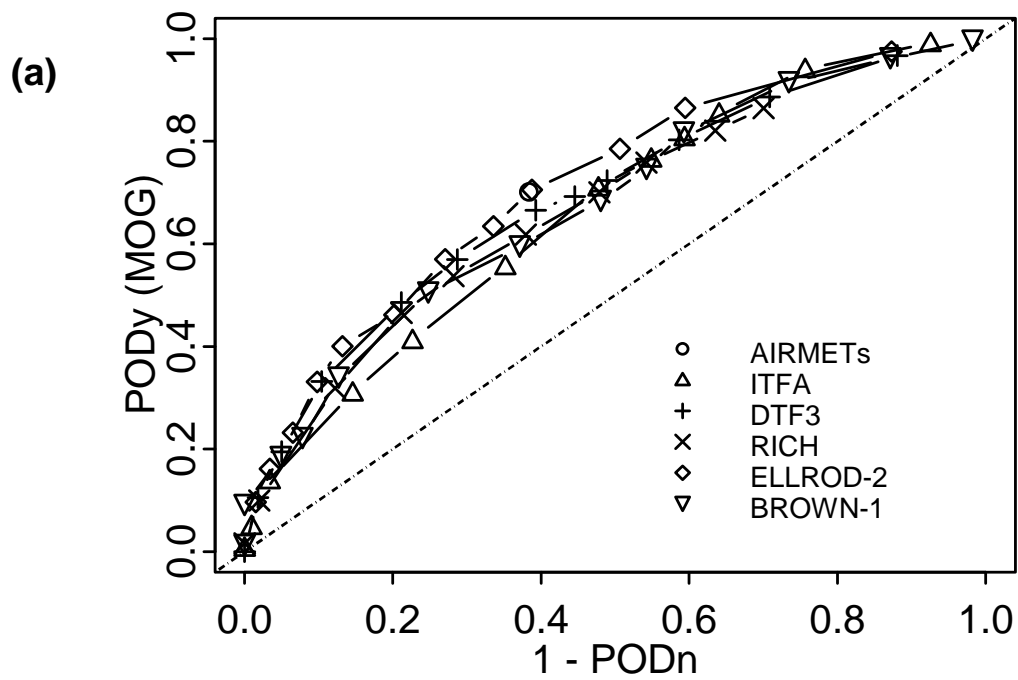


Figure 29. As in Fig. 23, but for 12-h forecasts.

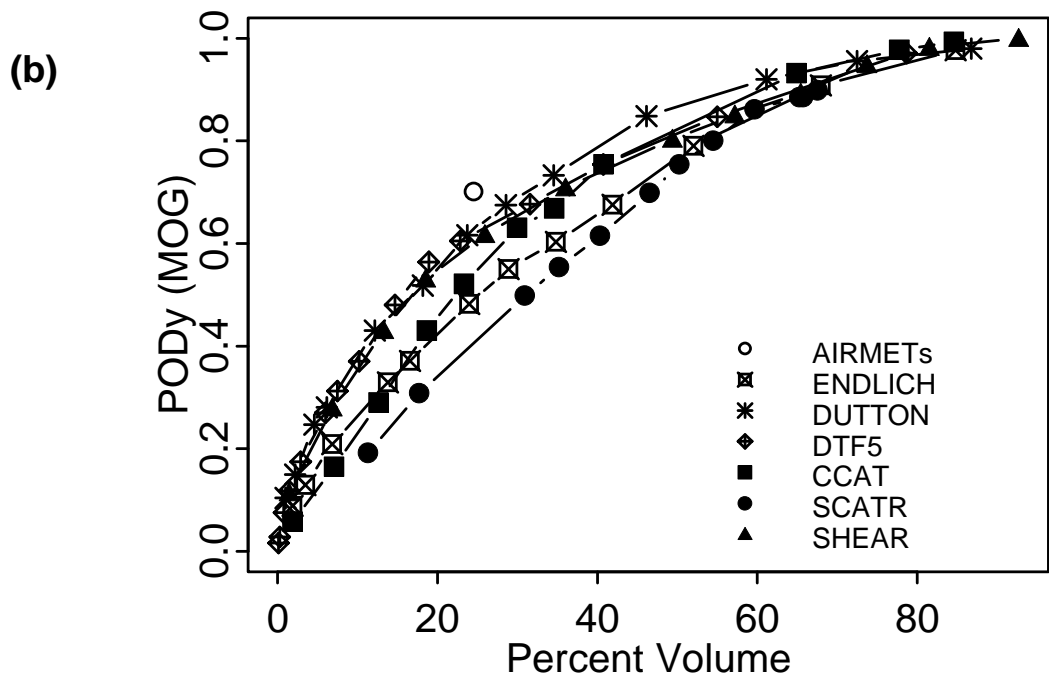
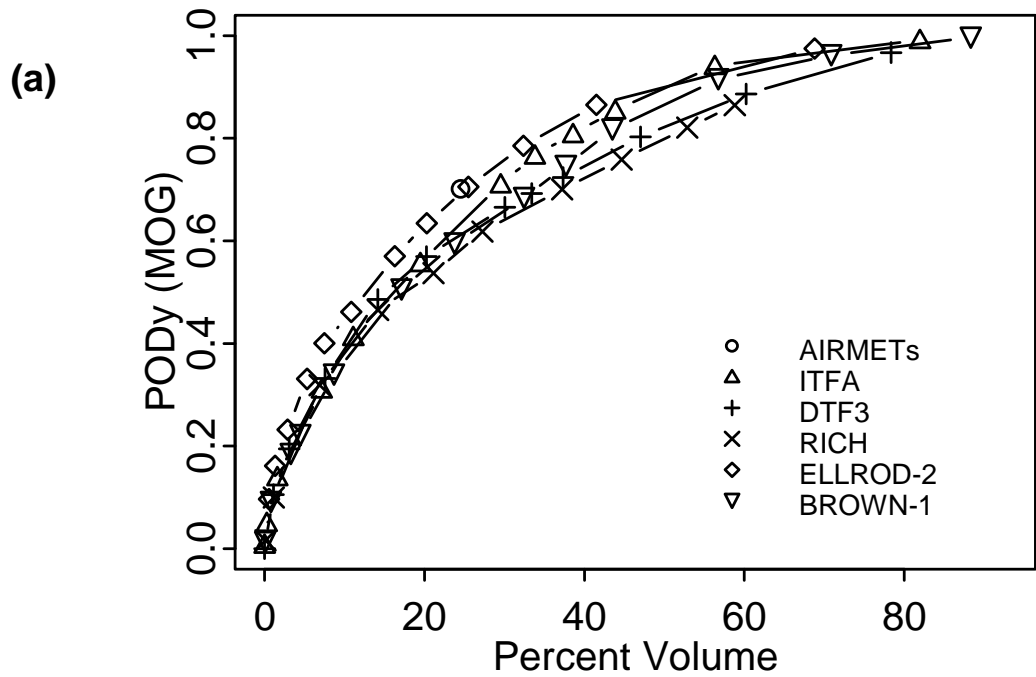


Figure 30. As in Fig. 24, but for 12-h forecasts.

Tables 8-10 show the same analysis as Table 6, but for the 6-, 9-, and 12-h forecasts. The results in these tables suggest that forecasting performance is approximately the same for 6-h forecasts as for 3-h forecasts, but that the statistics degrade somewhat by 9 hr. In particular, the TSS and VE values are somewhat smaller for most algorithms at the 9-h lead time than at the 3-h lead time. Similarly, the ROC curve areas are somewhat smaller for the 9-h forecasts than for the 3-h forecasts. In contrast, however, the statistics for the 12-h forecasts suggest that these longer-lead forecasts have almost as much skill as the 3-h forecasts. In fact, the TSS and VE values for the 12-hr Ellrod-2 forecasts are slightly larger than the comparable values for the 3-h forecasts. However, these differences may be related to the fact that only one forecast issue time (1200 UTC) is included for the 12-h forecasts, which also resulted in a relatively small total number of forecasts associated with this lead time. The % Volume result may be particularly impacted by the fact that only one issue/valid time is considered for the 12-h forecasts, due to possible diurnal variations in both the PIREPs and the occurrence of CAT.

One other item of interest from Tables 6 and 8-10 is the variation in threshold values required to achieve the same PODy as the AIRMETs. In Brown et al. (1999), it was noted that the required threshold values seemed to change with lead time. In contrast, the results presented here for TURB2000 indicate that the threshold values are relatively stable with regard to variations in lead time.

Table 8: As in Table 6, but for 6-h forecasts.

Algorithm	Threshold	PODy (All)	PODy (MOG)	PODn	TSS	ROC Curve Area	Average % Vol	VE
AIRMETs	--	0.65	0.70	0.62	0.32	--	24.5	2.9
Brown-1	8×10^{-5}	0.65	0.68	0.52	0.21	0.66	34.5	2.0
CCAT	4×10^{-9}	0.68	0.70	0.41	0.11	0.56	37.1	1.9
DTF3	0.50	0.67	0.69	0.58	0.27	0.69	36.0	1.9
DTF5	0.12	0.67	0.69	0.52	0.21	0.66	37.3	1.8
Dutton	20.00	0.63	0.66	0.54	0.20	0.65	32.2	2.0
Ellrod-2	3×10^{-7}	0.68	0.71	0.57	0.28	0.70	30.6	2.3
Endlich	0.175	0.68	0.70	0.44	0.13	0.61	41.3	1.7
ITFA	0.10	0.66	0.69	0.54	0.23	0.67	32.6	2.1
Richardson	7.0	0.72	0.73	0.52	0.25	0.68	41.8	1.7
SCATR	0.0004	0.71	0.70	0.25	-0.04	0.45	45.6	1.5
Shear	0.005	0.70	0.71	0.45	0.16	0.63	40.7	1.7

Table 9: As in Table 6, but for 9-h forecasts.

Algorithm	Threshold	PODy (All)	PODy (MOG)	PODn	TSS	ROC Curve Area	Average % Vol	VE
AIRMETs	--	0.65	0.70	0.62	0.32	--	24.5	2.9
Brown-1	7.5×10^{-5}	0.68	0.71	0.52	0.22	0.68	38.9	1.8
CCAT	4×10^{-9}	0.67	0.70	0.43	0.13	0.57	36.0	1.9
DTF3	0.40	0.69	0.70	0.50	0.20	0.65	40.5	1.7
DTF5	0.10	0.71	0.72	0.44	0.15	0.63	43.6	1.7
Dutton	18.0	0.66	0.68	0.49	0.17	0.63	36.5	1.9
Ellrod-2	2.5×10^{-7}	0.72	0.74	0.52	0.26	0.69	35.2	2.1
Endlich	0.15	0.71	0.71	0.39	0.10	0.58	45.4	1.6
ITFA	0.10	0.65	0.68	0.52	0.21	0.66	31.0	2.2
Richardson	7.0	0.67	0.67	0.51	0.18	0.64	39.6	1.7
SCATR	0.0004	0.73	0.72	0.23	-0.05	0.46	46.0	1.9
Shear	0.005	0.65	0.66	0.44	0.10	0.60	37.6	1.8

Table 10: As in Table 6, but for 12-h forecasts.

Algorithm	Threshold	PODy (All)	PODy (MOG)	PODn	TSS	ROC Curve Area	Average % Vol	VE
AIRMETs	--	0.65	0.70	0.62	0.32	--	24.5	2.9
Brown-1	8×10^{-5}	0.66	0.68	0.52	0.20	0.67	32.5	2.1
CCAT	4×10^{-9}	0.66	0.67	0.40	0.07	0.53	34.6	1.9
DTF3	0.45	0.66	0.69	0.55	0.25	0.69	33.4	2.1
DTF5	0.12	0.64	0.68	0.54	0.22	0.67	31.6	2.2
Dutton	20.0	0.63	0.68	0.55	0.20	0.64	27.5	2.5
Ellrod-2	3×10^{-7}	0.66	0.70	0.61	0.32	0.71	25.5	2.7
Endlich	0.15	0.65	0.68	0.43	0.10	0.58	41.9	1.6
ITFA	0.10	0.67	0.71	0.52	0.23	0.67	29.5	2.4
Richardson	7.0	0.68	0.70	0.52	0.22	0.67	37.2	1.9
SCATR	0.0004	0.71	0.70	0.21	-0.09	0.43	46.6	1.5
Shear	0.005	0.68	0.71	0.44	0.14	0.63	36.0	2.0

Variations of the statistics with lead time are considered directly for three algorithms (DTF3, Ellrod-2, and ITFA) in Figs. 31-33. These figures show the curves relating PODy to 1-PODn and % Volume for these algorithms, with separate curves on each plot for the four lead times. The curves in Fig. 31 indicate that the results for DTF3 are quite similar for all lead times except the 9-h lead time, for which the ROC and PODy vs. % Volume curves are somewhat lower than the curves for other lead times. The 6-h curve for PODy vs. % Volume also is slightly lower than the 3- and 12-h curves.

For Ellrod-2 (Fig. 32), the ROC curves for all four lead times are practically coincident. Moreover, the PODy vs. % Volume curve for the 12-h lead time is somewhat above the curves for the other lead times, except for the 3-h lead time. The curves in Fig. 32b confirm the results in Table 10, which suggest that the Ellrod-2 12-h forecasts appear to be as skillful as the Ellrod-2 3-h forecasts. However, as mentioned earlier, this result may be at least partially an artifact related to the limited issue/valid times available to verify the 12-h forecasts. In the case of ITFA (Fig. 33), the curves for all lead times are approximately coincident.

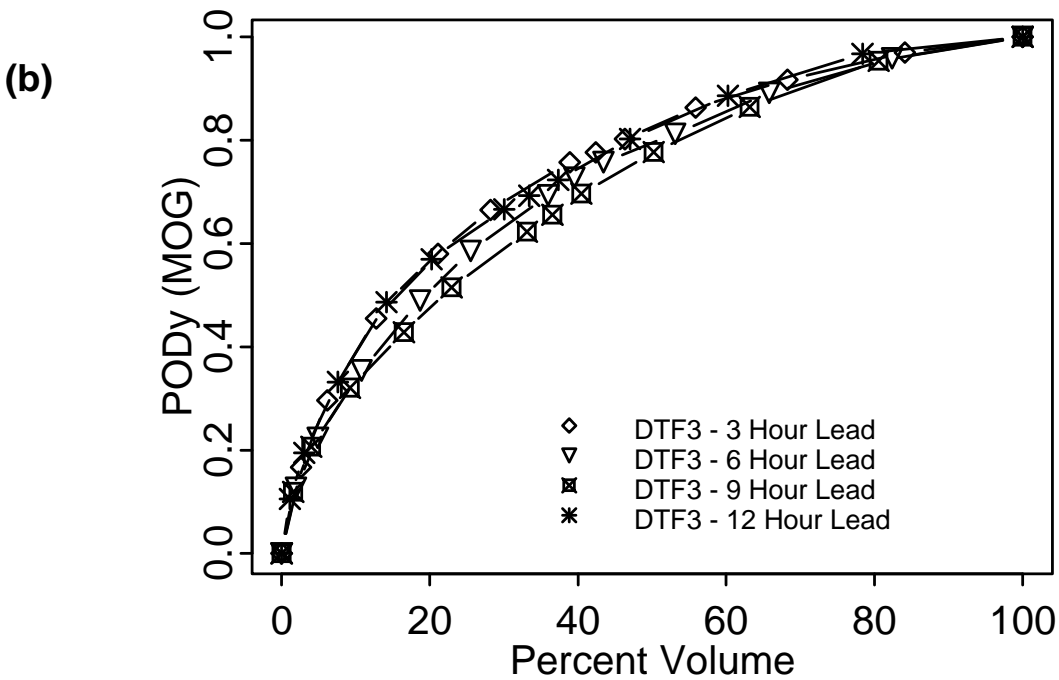
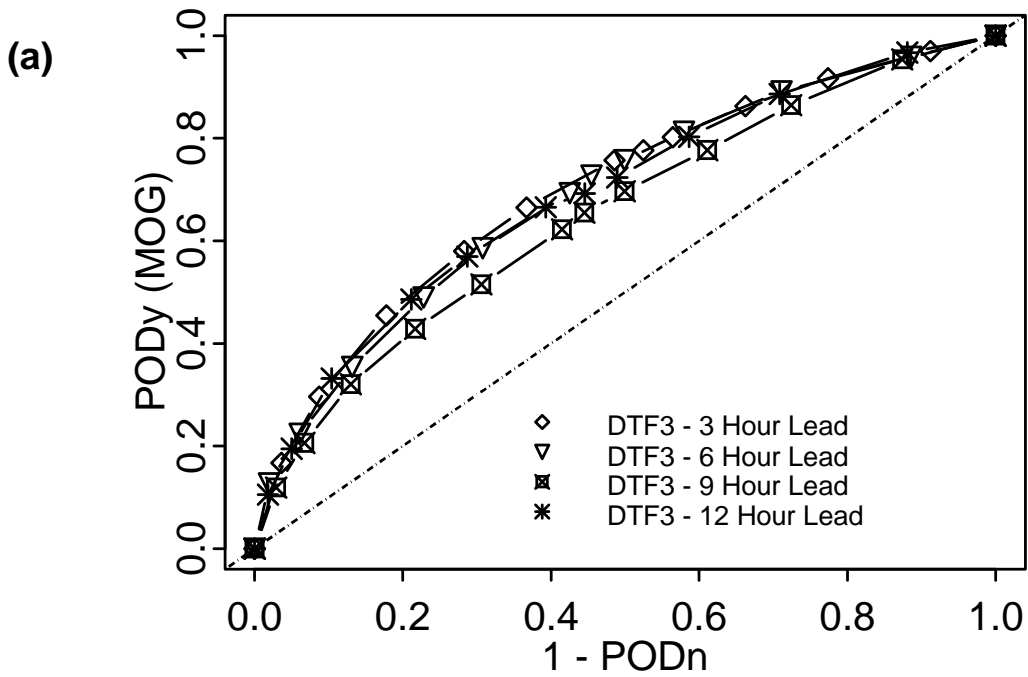


Figure 31. Plots of (a) POD_y vs. $1-POD_n$ and (b) POD_y vs. % Volume for DTF3, showing variations with lead time.

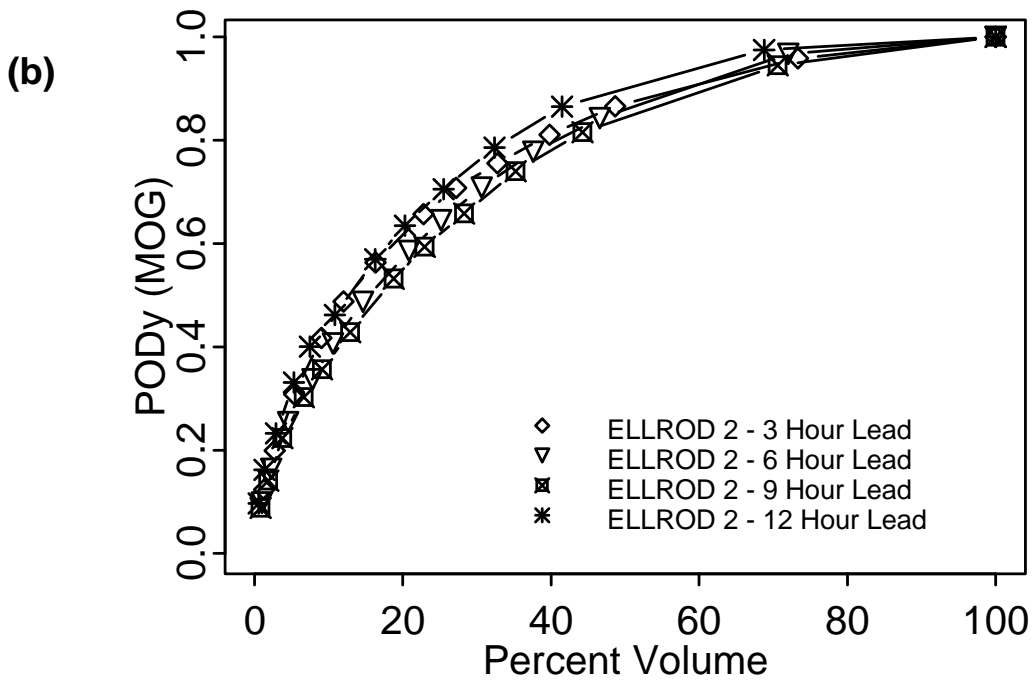
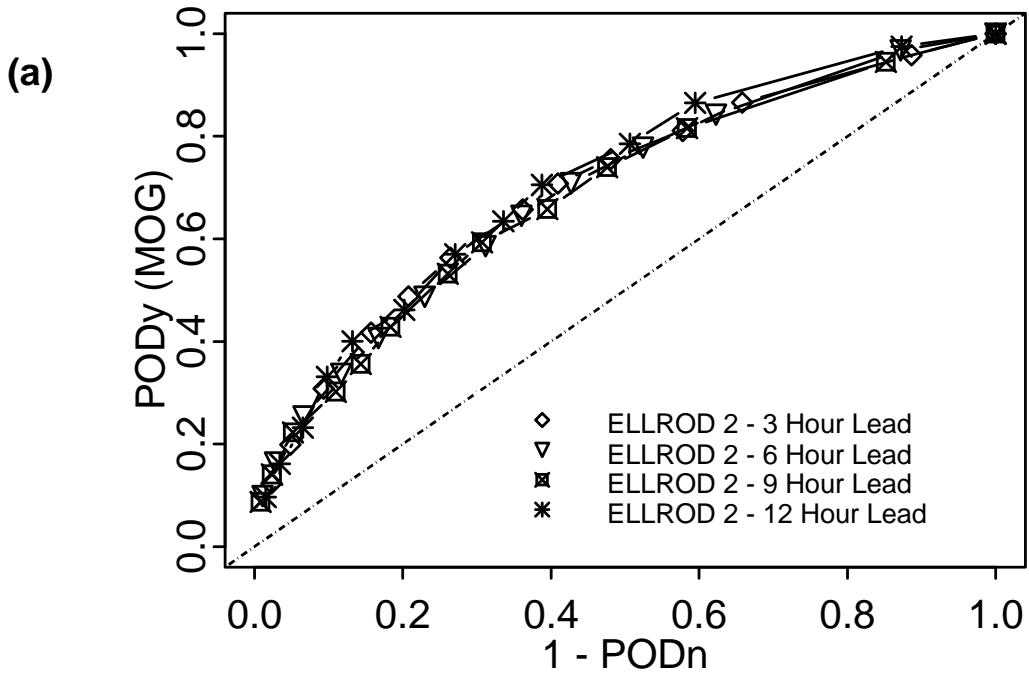


Figure 32. As in Fig. 31 for Ellrod-2.

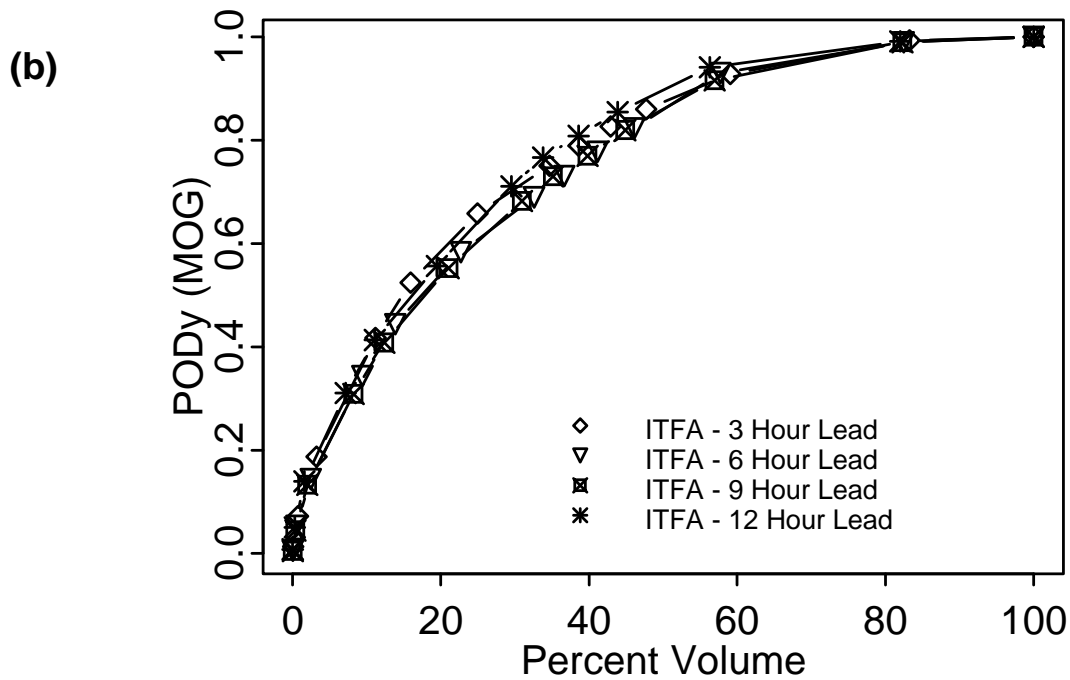
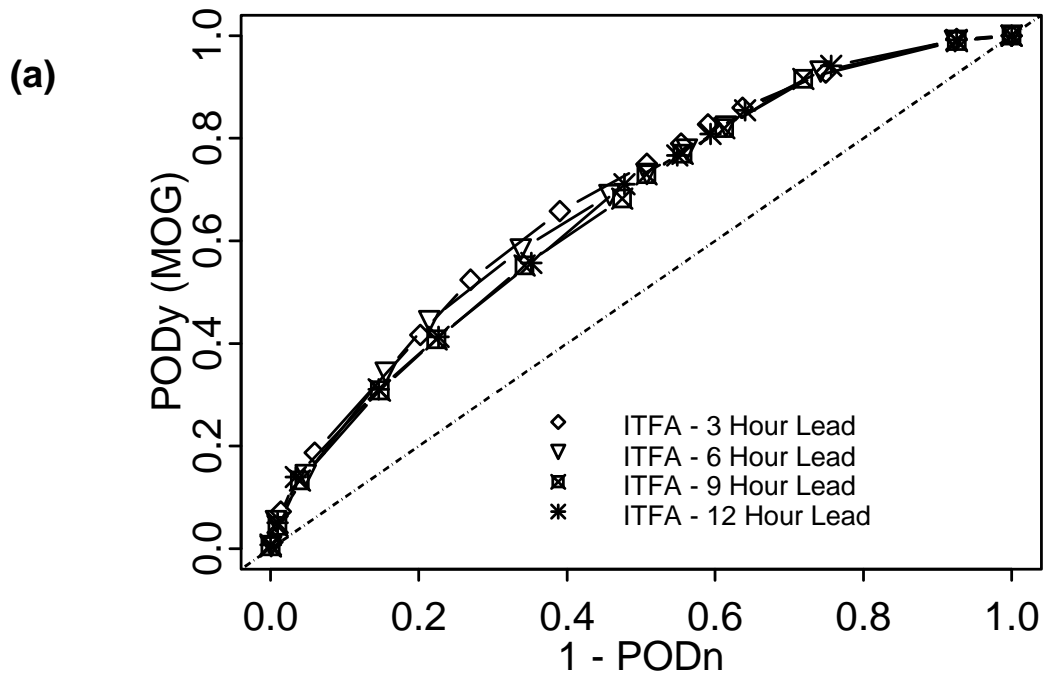


Figure 33. As in Fig. 31 for ITFA.

7.2.3 Comparisons among valid times

To consider possible effects of valid time on the skill of the forecasts, Figs. 34-36 present the verification curves by valid time for 3-h forecasts by three of the algorithms. In particular, the curves in these figures show verification results for 3-h DTF3, Ellrod-2, and ITFA forecasts, respectively, valid at 1500, 1800, 2100, and 0000 UTC. In general, the curves seem to vary only a little with valid time. For DTF3 (Fig. 34), the ROC curves suggest a slight improvement for forecasts valid at 1500 UTC, and in terms of PODy vs. % Volume, the forecasts valid at 0000 UTC appear to have slightly worse performance than the other forecasts. Little variation in performance with valid time is apparent for the Ellrod-2 forecasts (Fig. 35). Greater variation in performance with valid time is apparent for ITFA (Fig. 36). In particular, forecasts valid at 0000 UTC have the worst performance, both in terms of the ROC curves and with regard to PODy vs. % Volume. ITFA forecasts valid at 1500 UTC also had slightly poorer performance than those forecasts that were valid at either 1800 or 2100 UTC.

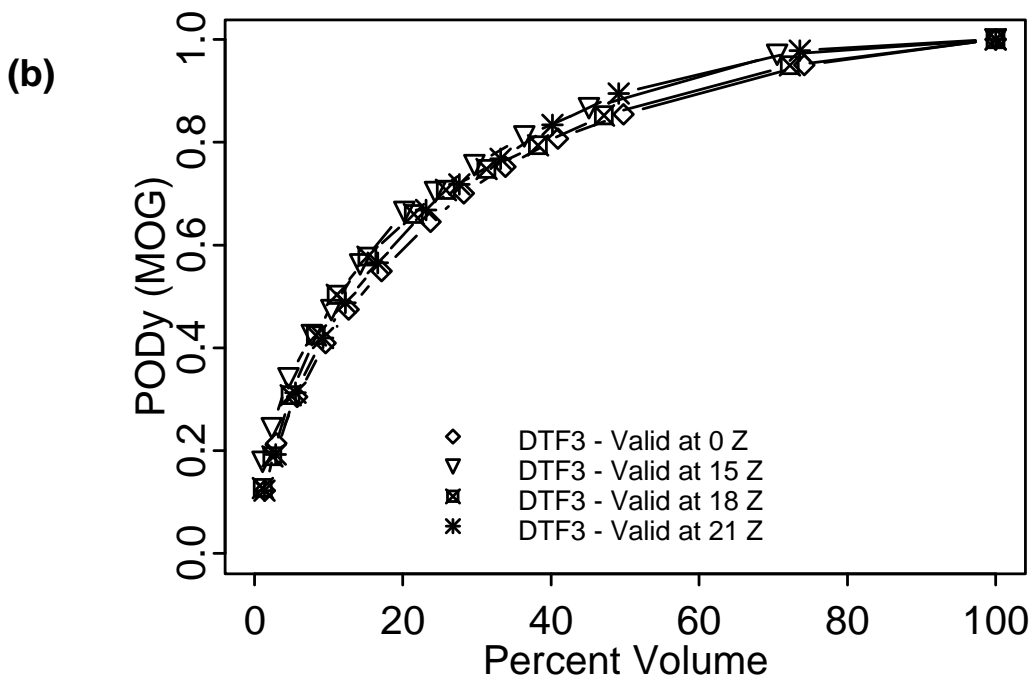
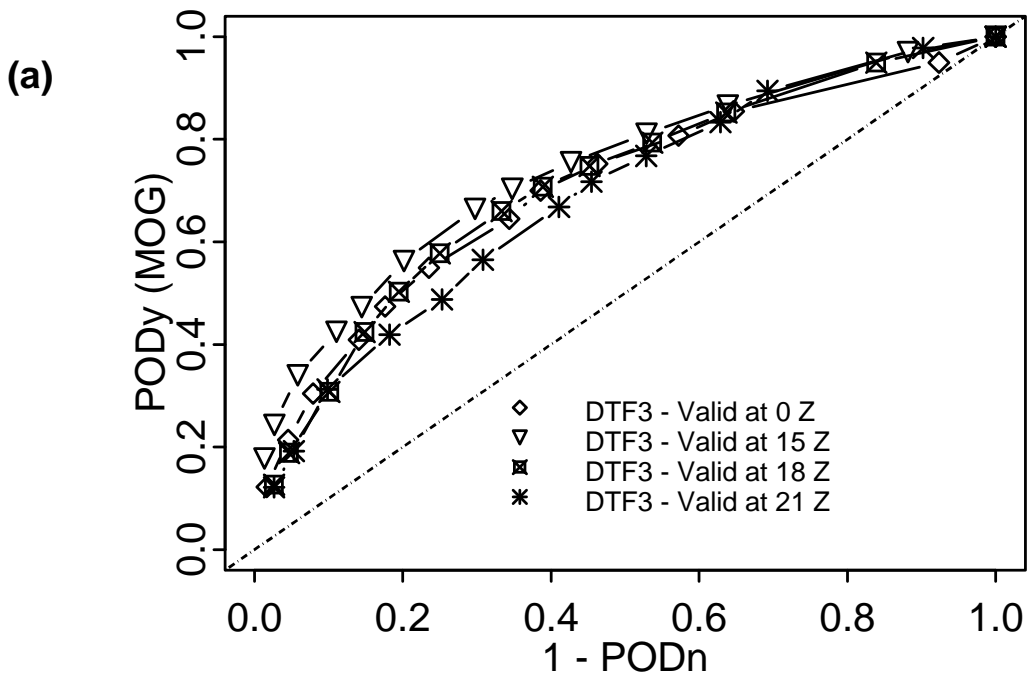


Figure 34. Plots of (a) POD_y vs. $1-POD_n$ and (b) POD_y vs. % Volume for DTF3, showing variations with valid time.

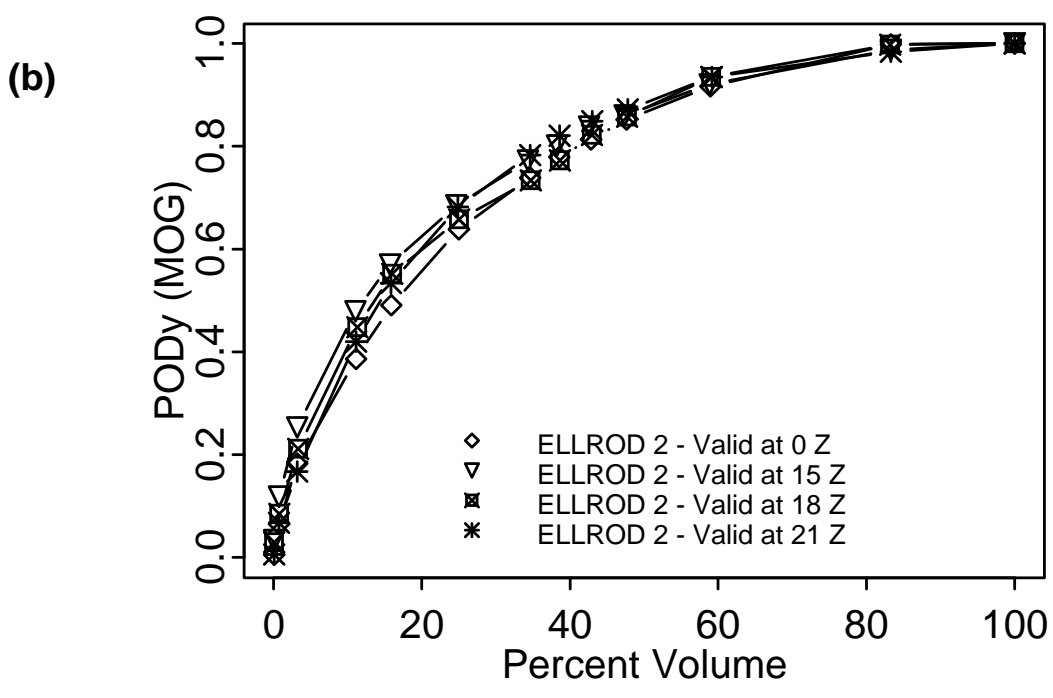
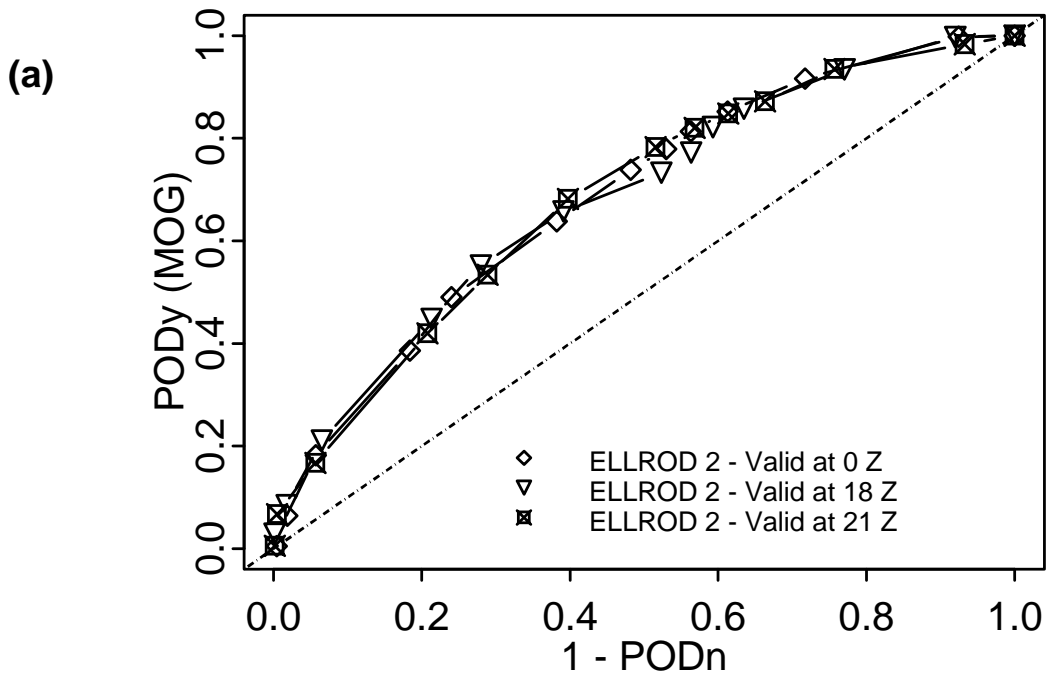


Figure 35. As in Fig. 34 for Ellrod-2.

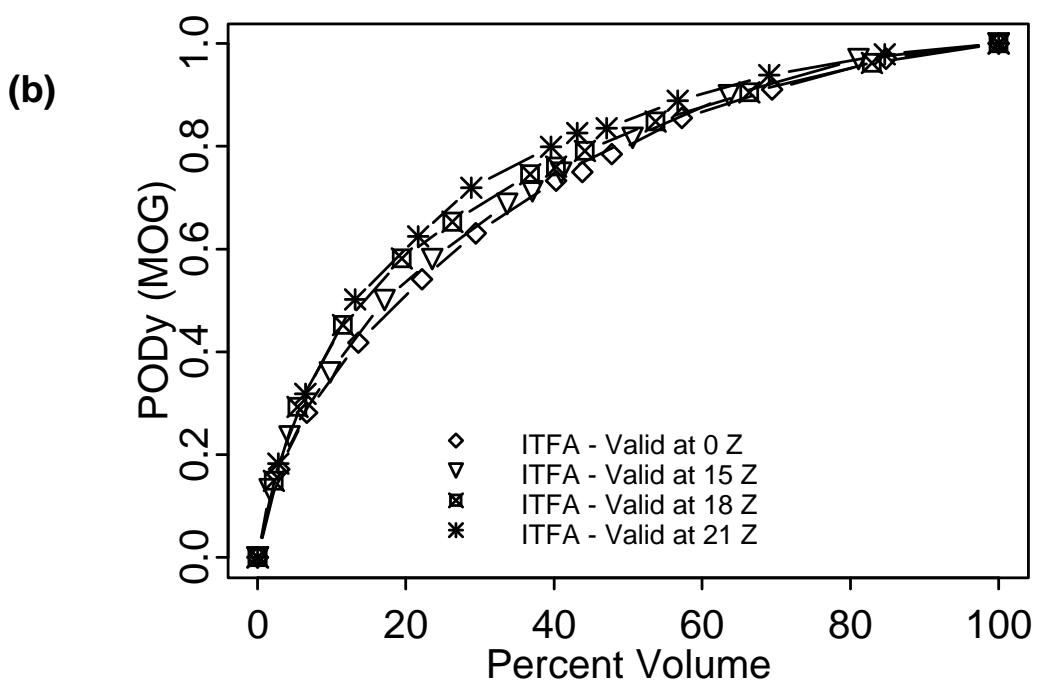
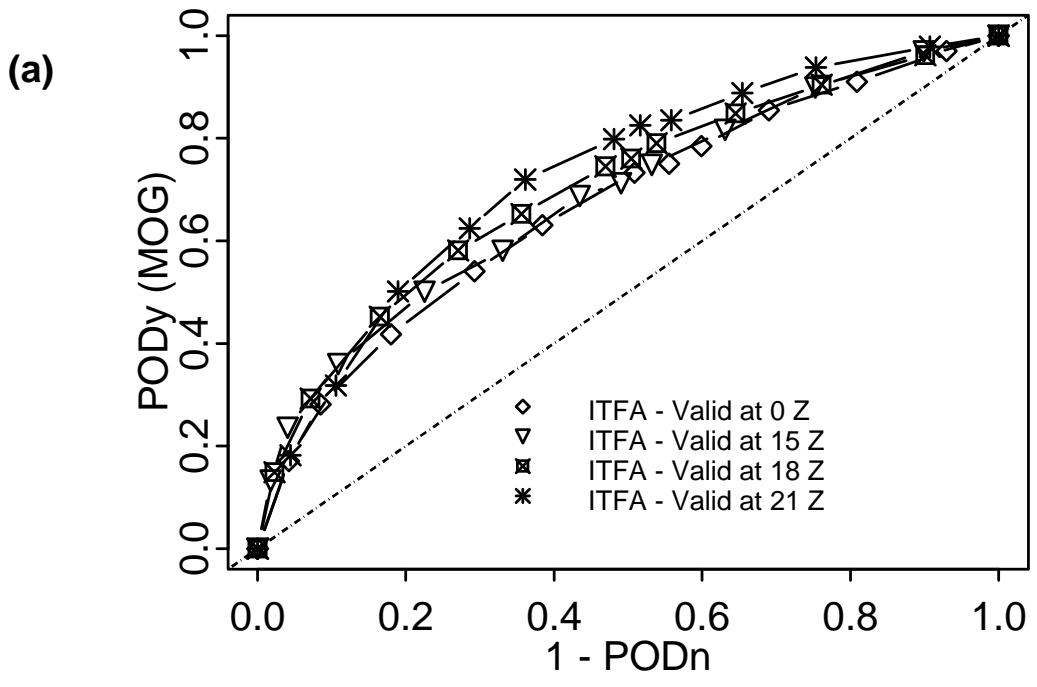


Figure 36. As in Fig. 34 for ITFA.

7.2.4 Comparisons with *TURB98-99* results

Figures 37-39 compare the *TURB2000* overall verification statistics for 3-h DTF3, Ellrod-2, and ITFA forecasts, to the comparable statistics obtained in *TURB98-99*. The curves in these plots lie almost completely on top of each other. Thus, these plots suggest that the *TURB98-99* results are consistent with those obtained in the previous evaluation.

However, as shown in Figs. 37-39, the basic AIRMET statistics for *TURB2000* were quite different from the AIRMET statistics obtained in *TURB98-99*. For example, PODy(MOG) was 0.64 in *TURB98-99*, compared to 0.70 for *TURB2000*. Correspondingly, the PODn value in *TURB2000* (0.62) was smaller than the PODn value obtained in *TURB98-99* (0.70). Because both PODy and PODn varied between the two exercises, the value of TSS was approximately constant across the winters (0.34 in *TURB98-99*; 0.32 in *TURB2000*). The average % Volume for the AIRMETs was slightly larger for *TURB2000* than for *TURB98-99* (24.5 vs. 22.9), which is consistent with the larger PODy values obtained in *TURB2000*.

Table 11 shows the variations in the ROC curve areas between the two exercises. Only algorithms and lead times that were included in both *TURB98-99* and *TURB2000* are presented in Table 11.

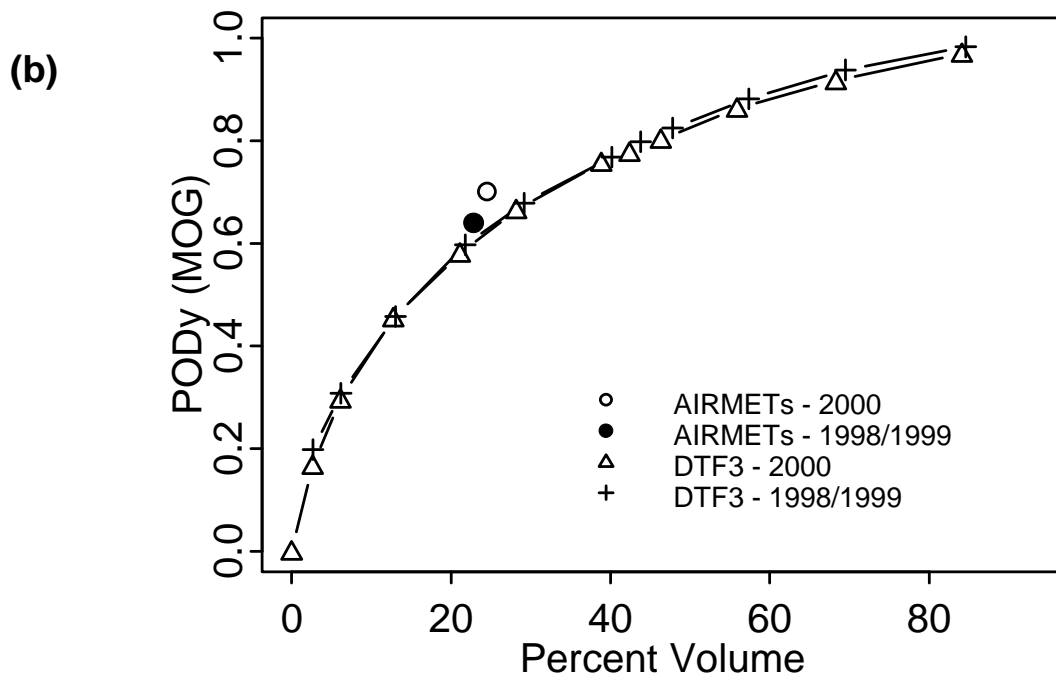
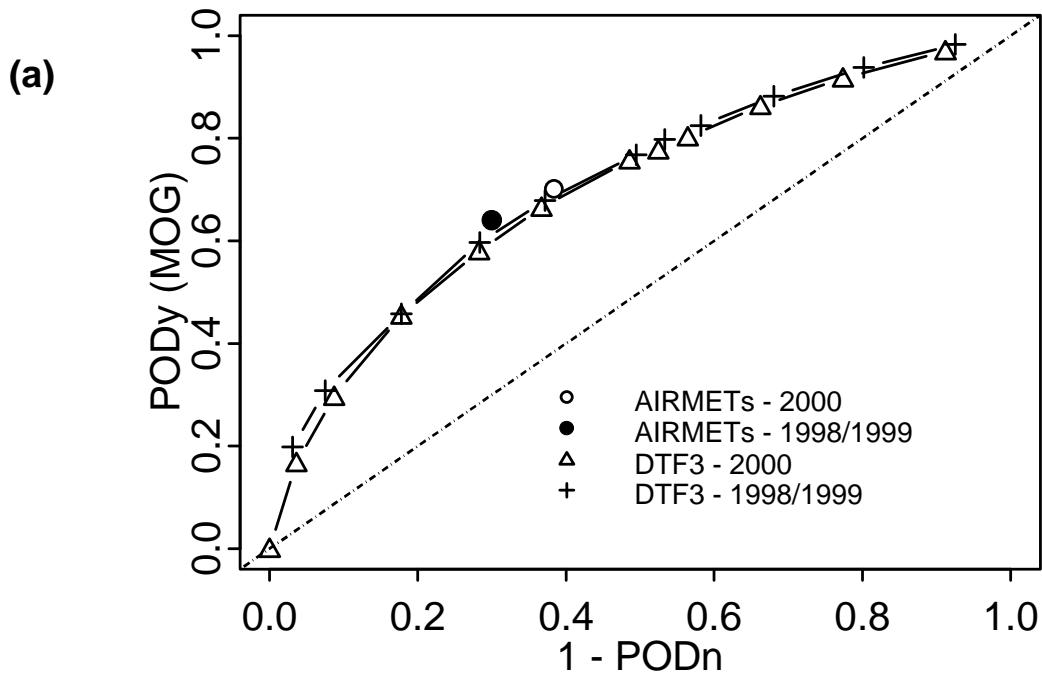


Figure 37. Plots of (a) PODY vs. 1-PODn and (b) PODY vs. % Volume for DTF3, showing variations with year.

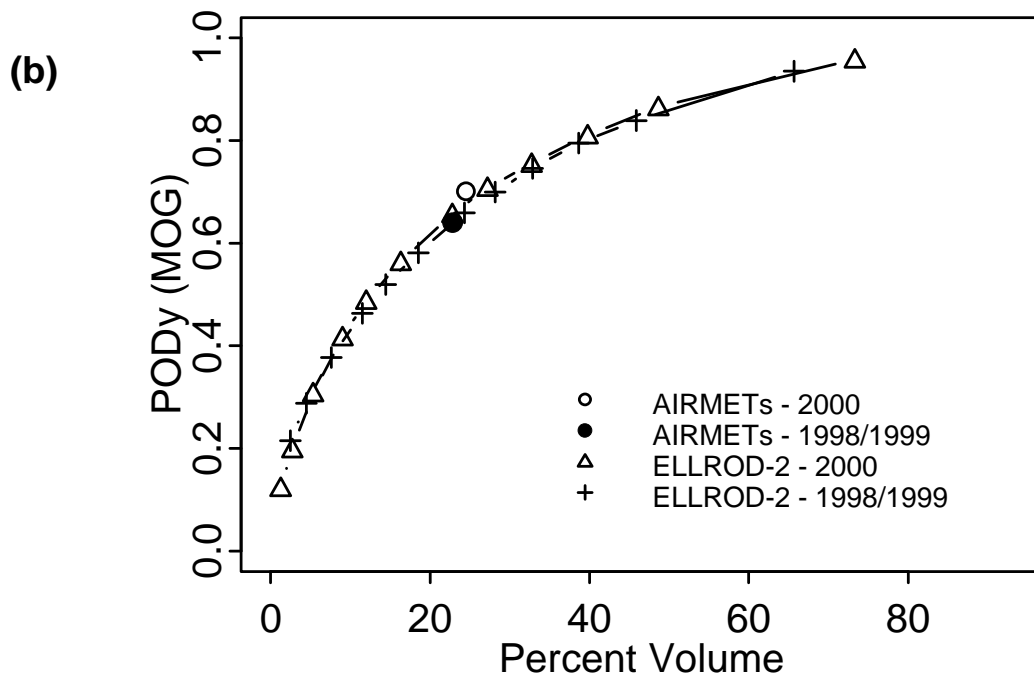
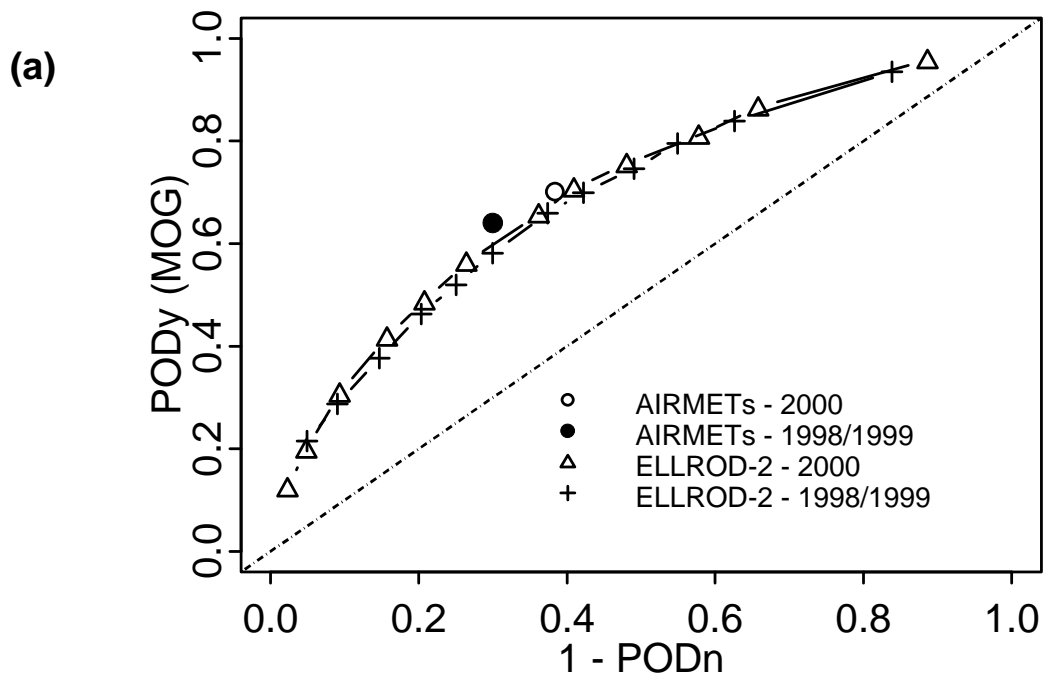


Figure 38. As in Fig. 37, for Ellrod-2.

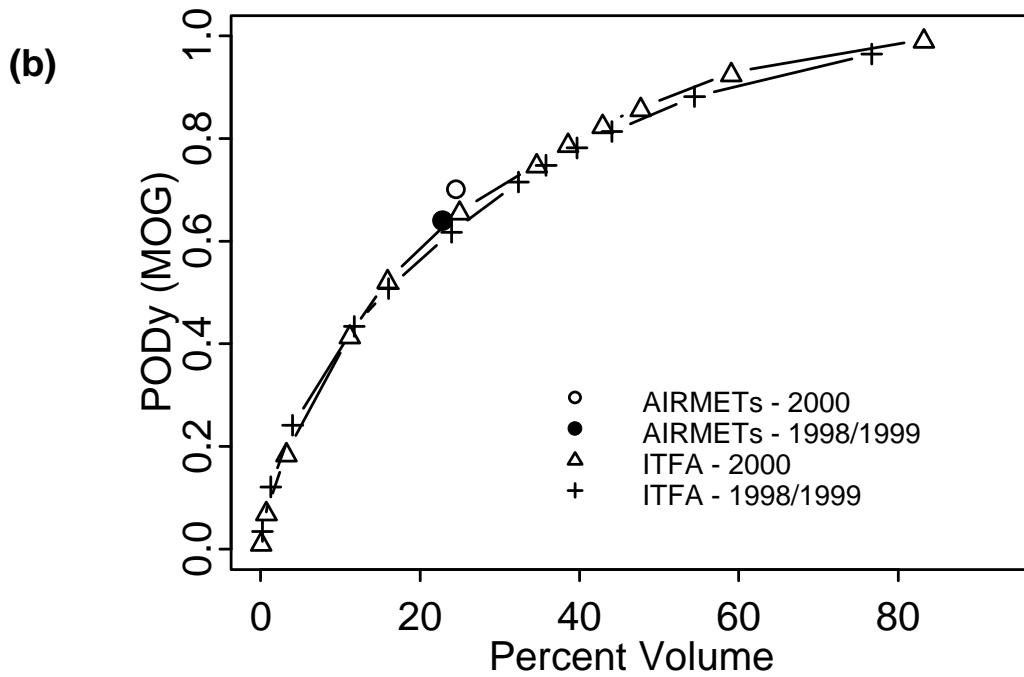
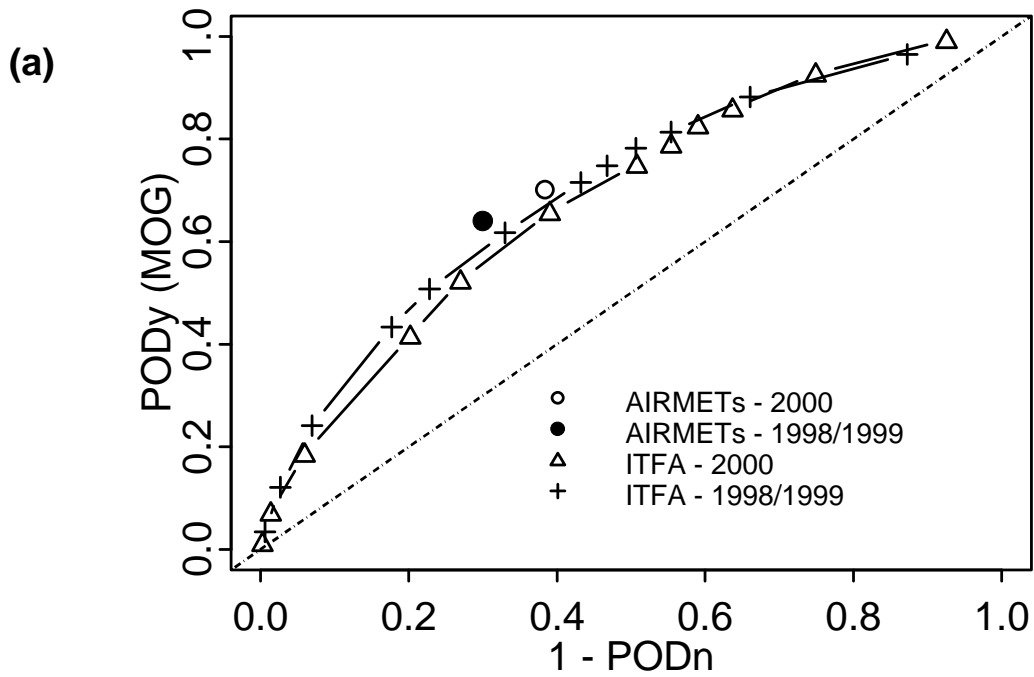


Figure 39. As in Fig. 37 for ITFA.

Table 11: ROC curve areas for TURB98-99 and TURB2000. [NEED TO ADD THE TURB2000 NUMBERS TO THIS TABLE]

Algorithm	Lead time					
	3 hr		6 hr		9 hr	
	TURB98-99	TURB2000	TURB98-99	TURB2000	TURB98-99	TURB2000
Brown-1	0.67	0.66	0.67	0.66	0.69	0.68
CCAT	0.59	0.58	0.64	0.56	0.62	0.57
DTF3	0.71	0.70	0.70	0.69	0.70	0.65
DTF5	0.67	0.66	0.67	0.66	0.67	0.63
Dutton	0.65	0.64	0.65	0.65	0.64	0.63
Ellrod-2	0.69	0.70	0.69	0.70	0.69	0.69
Endlich	0.63	0.60	0.62	0.61	0.60	0.58
ITFA	0.70	0.68	0.69	0.67	0.68	0.66
Richardson	0.70	0.68	0.69	0.68	0.68	0.64
SCATR	0.46	0.44	0.50	0.45	0.48	0.46
Shear	0.67	0.63	0.68	0.63	0.66	0.60

8. Summary, conclusions and discussion

This report has summarized the results of the TURB2000 exercise, in which the forecasting performance of a number of turbulence algorithms was tested. This exercise was the second such intercomparison that has taken place over the last two winters. The results obtained in TURB2000 are consistent with those obtained in the previous exercise, TURB98-99. In particular, both exercises identified several algorithms that appear to have somewhat better skill at forecasting CAT than some of the other algorithms. Other algorithms appear to have relatively little – or even negative – skill in forecasting CAT. The algorithms that appear to perform the best are DTF3, Ellrod-2, and ITFA.

TURB2000 again consisted of two components – a real-time component and a post-analysis. The real-time component facilitated a forecaster assessment of the algorithms at AWC (Mahoney and Brown 2000), and provided ongoing information to the algorithm developers and verification team. Through the real-time component, verification statistics were generated in near-real-time by RTVS and were provided to anyone interested through statistical displays on the Web. This process provided near real-time feedback (i) to model developers so that thresholds and techniques in the models could be identified and adjusted; (ii) to forecasters so that information on algorithm quality could be used during the forecasting process; and (iii) to those evaluating the algorithms so that information could be easily shared and compared. New displays on the RTVS facilitated interpretation and comparison of the algorithms' performance. The post-analysis allowed a closer examination of the verification results, as reported here. Results from both components were consistent. Evaluation of the Random algorithm by both verification systems provided reassurance that the systems are functioning correctly.

Overall, the results indicate that each of the algorithms, except for SCATR, has some skill in forecasting turbulence, with DTF3, the Ellrod-2 index and ITFA showing the best performance overall. Since AWC forecasters often rely on output from the Richardson Number and Shear algorithms for guidance to the location, duration, and movement of turbulence, it is worth noting that these algorithms generally did not perform as well as Ellrod-2, ITFA, DTF3, Dutton, or Brown. However, Shear and Richardson Number did, in most instances, perform better than Endlich and CCAT. In general, the AIRMETs had greater skill overall than the algorithms. Of course, these human-generated forecasts benefit from forecaster experience and the use of the model-based turbulence algorithms as guidance. The time series plots indicated a large variability in algorithm performance from week to week and between the various forecast issue and lead times. The real-time and post-analysis results both indicated that the 6-h forecasts often performed as well as or better than those the 3-h forecasts. This result suggests that using this forecast lead time as guidance is reasonable.

Numerous further analyses of the TURB2000 data will be undertaken. A few additional algorithms will be added to the evaluation, as will new/improved versions of ITFA (as noted in Section 3, the version of ITFA included in TURB2000 is an early version of the algorithm). In addition, data from TURB98-99 will be combined with the TURB2000 data to provide a long-term evaluation of the algorithms.

A third intercomparison of the algorithms will be undertaken during winter 2000-2001. Because a new 20-km version of the RUC model will be available in early 2001, it would be desirable – if possible – to evaluate the algorithms on both versions of the model simultaneously. This exercise would allow evaluation of the impacts of model resolution on the performance of the CAT algorithms. In order to accommodate forecasts based on both models, it likely would be necessary to reduce the number of algorithms considered, perhaps by eliminating those algorithms with poorest performance in both TURB98-99 and TURB2000.

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 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), 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., G. Thompson, R.T. Buintjes, 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., 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.
- 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., 1964: The mesoscale structure of some regions of clear-air turbulence. *Journal of Applied Meteorology*, **3**.
- 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. Report to the FAA. Available from J.L. Mahoney, FSL, 325 Broadway, Boulder, CO 80303.
- 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.
- Orville, R.E., 1991: Lightning ground flash density in the contiguous United States – 1989. *Monthly Weather Review*, **119**, 573-577.
- 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.
- Vogel, G.N. and C.R. Sampson, 1996: Clear air turbulence indices derived from U.S. Navy numerical model data: a verification study. Naval Research Laboratory, Monterey, NRL/MR/7543-96-7223, 30 pp.
- Wahl, G.M., R.L. Minton, P.A. Mandics, 1997: NIMUS: An advanced meteorological data acquisition, processing, and distribution system. *Preprints, Thirteen International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Long Beach, CA. Amer. Meteor. Soc., Boston, MA.
- Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.