Graphical Turbulence Guidance 2 (GTG2): Quality Assessment Report

FAA Aviation Weather Research Program
Quality Assessment Product Development Team[1]

Agnes Takacs[2], **Lacey Holland**[2], **Mike Chapman**[2], **Barbara Brown**[2] , **Jennifer Mahoney**[3], **and Chris Fischer**[3,4]

31 July 2004

---

[1] Contacts: B.G. Brown (bgb@ucar.edu) or J.L. Mahoney (Jennifer.Mahoney@noaa.gov)

[2] Research Applications Program, National Center for Atmospheric Research, Boulder CO 80307-3000

[3] NOAA Research – Forecast Systems Laboratory, Boulder, Colorado

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

Contents

Graphical Turbulence Guidance 2 (GTG2): Quality Assessment Report

July 2004

Aviation Weather Research Program
Quality Assessment Product Development Team

Summary

This report summarizes the quality of middle- and upper-level turbulence forecasts produced by the second generation of the Graphical Turbulence Guidance (GTG2) forecasting system. GTG2 is an enhanced version of the operational Graphical Turbulence Guidance (GTG) algorithm that runs operationally at the NOAA National Center for Environmental Prediction Aviation Weather Center (AWC), and was originally developed as the Integrated Turbulence Forecast Algorithm (ITFA). GTG2 was developed by the Turbulence Product Development Team (TPDT) of the Federal Aviation Administration's Aviation Weather Research Program (FAA/AWRP). The main enhancement to the algorithm is that it provides forecasts at mid-levels (10-20,000 ft) in addition to the upper levels (above 20,000 ft) considered by GTG. GTG2 is currently being considered for transition to experimental status through the Aviation Weather Technology Transfer (AWTT) process.

The performance of GTG2 forecasts was evaluated over one winter (January through April 2004) by the AWRP Quality Assessment PDT (QAPDT). Ongoing real-time and long-term evaluations are available on the Real-Time Verification System (RTVS; http://www-ad.fsl.noaa.gov/fvb/rtvs/turb/index.html), developed by the National Oceanic and Atmospheric Administration's Forecast Systems Laboratory (NOAA/FSL). Most of the results presented in this report are based on the RTVS analyses. Additional analyses of the results were undertaken by the Verification Group at the Research Applications Program at the National Center for Atmospheric Research (NCAR/RAP). Both the real-time and post-analysis evaluations provided meteorological/statistical verification of the turbulence forecasts. Performance of the GTG2 forecasts is compared to the performance of several other turbulence forecasts, including ITFA and GTG (the current operational version of the algorithm).

The forecasts were verified using Yes and No turbulence observations from pilot reports (PIREPs) indicating either moderate-or-greater (MOG) turbulence severity or no turbulence. GTG2 and the other turbulence 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 forecast coverage. In addition, forecasts based on Airmens' Meteorological Advisories (AIRMETs), the operational forecasts issued by the AWC, were evaluated to provide a standard of comparison. More than 1,200 individual GTG2 forecasts were considered for both mid- and upper levels in this evaluation. The number of Yes (No) PIREPs considered in the evaluation for

upper-level forecasts ranged from 2,100 to 7,086 (975 to 2,975) depending on the forecast lead time. For mid-level forecasts, the number of Yes (No) PIREPs ranged from 415 to 1,408 (230 to 842).

Results of the evaluation indicate that GTG2 is skillful at discriminating between Yes and No turbulence conditions at both upper and middle levels and that it is significantly more skillful than GTG, ITFA and the Ellrod-1 Index. GTG2 also provides relatively efficient forecasts, covering comparatively small volumes for a given turbulence detection rate. Using a threshold of 0.25, GTG2 correctly classifies 89% of the Yes PIREPs in mid-level regions and 69% in upper-level regions; and 69% of the No PIREPs in mid-level regions and 84% in upper-level regions, while covering approximately 34% of the airspace volume in the mid-levels and 27% in the upper levels over the CONUS. The forecast performance is relatively insensitive to lead time, especially at mid-levels, and is consistent through the atmosphere (10,000 ft and higher). Skill and efficiency measures vary somewhat from day-to-day, but less than for some other types of turbulence forecasts. Regional analyses indicate that the best performance is in the West region for mid-level forecasts and in the East and Central regions for upper-level forecasts.

In summary, this evaluation of GTG2 demonstrates that it is a skillful turbulence forecasting algorithm that has more capability to discriminate between Yes and No turbulence PIREPs than previous generations of the algorithm, with relatively efficient forecasts. The algorithm is skillful at both upper and mid-levels. The quality of GTG2 forecasts is relatively insensitive to variations in the PIREPs used for the analyses and does not degrade with altitude.

1. Introduction

This report summarizes basic results of an evaluation of the forecasting capability of an enhanced version of the Graphical Turbulence Guidance (GTG) algorithm to be denoted "GTG2". This algorithm is under consideration for transition to experimental status through the Aviation Weather Technology Transfer (AWTT) process. GTG2, which was developed by the Turbulence Product Development Team (TPDT) of the Federal Aviation Administration's Aviation Weather Research Program (FAA/AWRP) is designed to predict clear-air turbulence (CAT) at altitudes above 10,000 ft over the continental U.S. (CONUS). This assessment of GTG2 was performed by the AWRP's Quality Assessment Product Development Team (QAPDT) in specific algorithm intercomparison studies. These studies were conducted using the Real-Time Verification System (RTVS) developed by the National Oceanic and Atmospheric Administration's Forecast Systems Laboratory (NOAA/FSL) (Mahoney et al. 1997, 2002) and in a post-analysis by the Verification Group at the National Center for Atmospheric Research, Research Applications Program (NCAR/RAP). The analyses in this report focus on turbulence forecasts for winter (January through April) 2004.

The report is organized as follows. The study approach is presented in Section 2. Section 3 briefly describes the algorithms and forecasts 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 and results of the study are presented in Section 6. Finally, Section 7 includes the conclusions and discussion.

2. Approach

A subset of the algorithms that were included in the winter 2004 RTVS and post-analysis evaluations of turbulence forecasts are considered in this report. 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 NCEP. Model forecasts issued at 1200, 1500, 1800, and 2100 UTC, with lead times of 3, 6, 9, and 12 hours and valid times between 1500 and 0000 UTC, were included in the study. In addition, the turbulence Airmen's Meteorological Advisories (AIRMETs), which are one of the operational turbulence forecasts issued by the AWC, were included for comparison purposes (i.e., this report is not intended as an evaluation of turbulence AIRMETs). Due to the emphasis placed on forecasting mid- and upper-level turbulence, the evaluation focused on the layers in the atmosphere between 10-20,000 ft, and above 20,000 ft. In addition to the entire CONUS, forecasting performance across three large and 15 small geographic sub-regions was also considered. Forecasts issued during the period 1 January through 30 April 2004 were included in the analyses.

The verification approach applied in the winter 2004 evaluation is identical to the approach taken in previous studies. In particular, the algorithm forecasts and AIRMETs were verified using Yes and No PIREPs of turbulence. 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 were 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. In addition, the amount of airspace impacted by the forecasts was considered. Although all PIREPs were included in the analyses, all of the analyses reported here were based on Yes PIREPs reporting MOG turbulence severity as well as PIREPs explicitly reporting No-turbulence conditions.

In evaluating an algorithm or forecast, it is important to compare the quality of the forecasts to the quality of one or more standards of reference. Thus, the quality of the GTG2 forecasts is compared to the quality of several other automated forecasting algorithms (e.g., Ellrod-1, ITFA; see Section 3), as well as to the quality of the operational forecasts (i.e., AIRMETs). However, it is important to emphasize that the algorithm forecasts and the AIRMETs are very different types of forecasts, with different objectives. GTG2 forecasts generally are understood to be valid at a particular time. The AIRMETs, on the other hand, are valid over a 6-h period and are designed to capture turbulence conditions as they move through the AIRMET area over the period. Due to the differences between these forecasts, it is difficult to clearly compare their performance. However, in order to understand the quality of GTG2, it is necessary for comparisons between various forecasts to be made, and for GTG2 forecasts to be compared to the operational standard, especially since both types of information will be available to users. The comparisons are made in such a way as to be as fair as possible to both the AIRMETs and GTG2, as described in Section 4, while still obtaining the information needed. Nevertheless, users of these statistics should keep these assumptions in mind when evaluating the strengths and weaknesses of each type of forecast.

3. Algorithms and forecasts

The algorithms and forecasts that are considered in most of the analyses presented in this report are briefly described in this section. Further information about the algorithms and their development can be found in the references that are provided and in Sharman et al. (2002b, and 2004); information about the algorithms included on RTVS is available through a link from the RTVS web site and in Sharman et al. (2002b). Operational forecasts of turbulence are also described.

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).

ITFA : The ITFA forecasting technique uses fuzzy logic to integrate available turbulence observations (in the form of PIREPs) 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, 2000a, 2002a,b). The suite of algorithms that is included is described in Sharman et al. (2002b). This algorithm was developed by the TPDT of the AWRP. The version of the algorithm considered in this study is an upgrade of the algorithm that was transferred to the AWC, and includes forecasts for both upper and middle altitude layers.

GTG: This algorithm was originally developed as ITFA (Sharman et al. 2002b), and now runs operationally at the AWC. GTG only provides turbulence forecasts for upper levels (20,000 ft and above).

GTG2: GTG2 expands the capabilities of GTG by providing turbulence predictions at both mid-levels (10-20,000 ft msl) and upper levels (≥20,000 ft). In addition, new turbulence diagnostics were included in the suite of diagnostic turbulence algorithms that are utilized in GTG2. Within GTG2, the mid- and upper-level forecasts are computed separately, and the results merged at the 20,000-ft boundary. This merging is necessary since it was found that (a) the best sets of turbulence diagnostics (in terms of discriminating between Yes and No turbulence observations) differs between mid- and upper levels; (b) the optimum threshold values also differ; and (c) the number of available PIREPs is substantially smaller at mid-levels than at upper levels, so different PIREP time windows must be used in the two altitude regimes.

Given a set of turbulence diagnostics, the method for combining them to derive the optimum turbulence forecasts was unchanged from that used in GTG (formerly ITFA). The combination process is described in Sharman et al. (2002b). The algorithm is also described in Sharman and Cornman (1998), Sharman et al. (1999), Sharman et al. (2000b), Sharman et al. (2002a), Sharman et al. (2004), and Tebaldi et al. (2002). Examples of a GTG2 and a GTG forecast are presented in Fig. 1. In Fig. 1(a), GTG2 values are shown for a particular layer (15,000 ft msl), and in Fig. 1(b) the composite for a GTG forecast for 20,000 ft and above is shown.. These figures represent how GTG2 would appear on the Aviation Digital Data Service (ADDS) web-site if it becomes operational.

AIRMETs: AIRMETs are the operational forecasts of turbulence conditions. These forecasts are produced by AWC forecasters every six hours and are valid for up to six hours (NWS 1991). AIRMETs may be amended as needed between the standard issue times. The forecasts are in a textual form that can be decoded into latitude and longitude vertices, with tops and bottoms of the turbulence regions defined in terms of altitude. Unfortunately, the more descriptive elements of the AIRMETs cannot be decoded and thus are not considered. For comparison with the forecasts from GTG2 and other algorithms, the AIRMETs are evaluated over the same time window as the model-based algorithms.

4. Data

The data that were used in the evaluation include model output and PIREPs. Although lightning data were used in some previous evaluations to eliminate the effects of PIREPs related to convection (Brown et al. 2000a), it was determined in that study that this stratification had little impact on the results. Thus, lightning data are not considered in this study.

Model output was obtained from the RUC-2 model, which is run operationally at NOAA's NCEP, Environmental Modeling Center (Benjamin et al. 1998). The model vertical

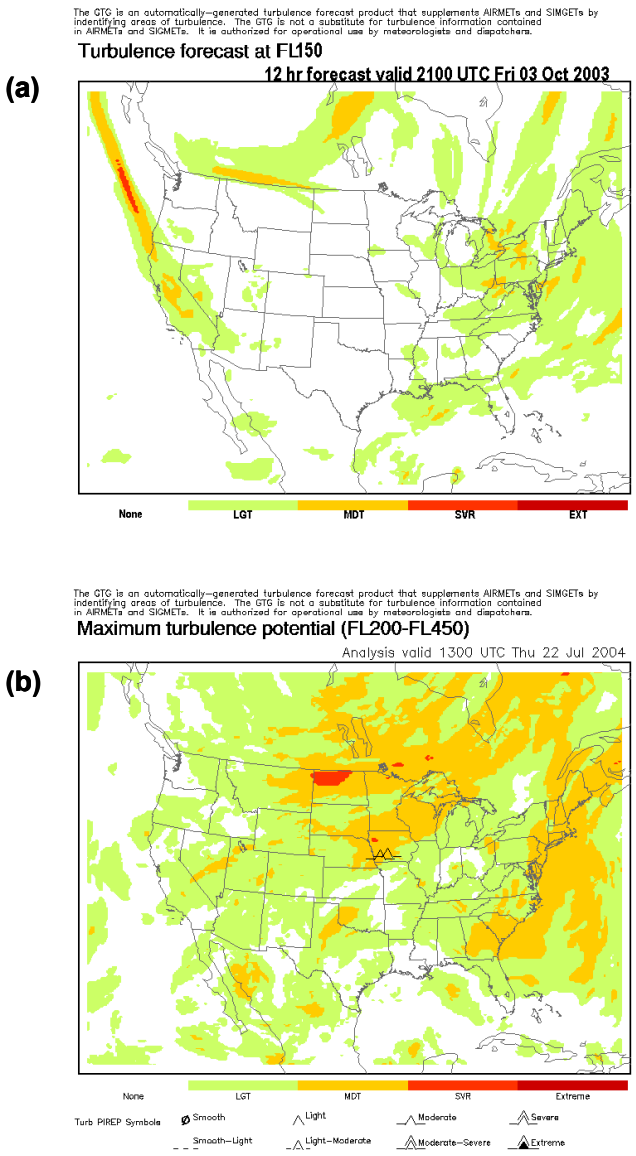


Figure 1. Examples of output from Graphical Turbulence Guidance forecasts: (a) GTG2 for FL150 as a mid-level forecast, (b) GTG as a composite for FL200 and above. These figures represent how GTG2 would appear on the Aviation Digital Data Service (ADDS) web-site if it becomes operational.

coordinate system is based on a hybrid isentropic-sigma vertical coordinate, and the horizontal grid spacing is approximately 20 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 described in Section 2 were used in this study. The verification analyses were limited to a subset of the RUC-2 domain covered by the AIRMETs, which is shown in Fig. 2.

Figure 2. Outline around continental U.S. denotes the total domain of the AIRMETs. All analyses were limited to this domain.

The turbulence algorithms were applied to the model output files to create turbulence forecast files. This part of the process was undertaken by the TPDT. As part of this process, the turbulence forecasts were interpolated to flight levels (i.e., every 1,000 ft) rather than the raw model levels. The AIRMETs were decoded to extract the relevant location, altitude range, and other information.

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 MOG category. Information about turbulence type (e.g., "Chop," "CAT") frequently is missing, and was ignored in this analysis.

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 GTG2 and other forecasts.

5.1 Matching methods

As in previous evaluations (e.g., Brown et al. 2000a,b,c; Mahoney et al. 2001b, Brown et al. 2002), each PIREP was connected to the forecasts at the nearest eight forecast grid points (four surrounding grid points; two levels vertically). Specifically, the RTVS uses bi-linear

interpolation to compute the appropriate forecast value, whereas the post-analysis system matches the PIREP to the most extreme forecast value among the four surrounding gridpoints. 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 results for winter 2004 are the same as the methods used in previous studies and are consistent with the approach described by Brown et al. (1997, 2002). More details on the general concepts underlying verification of turbulence forecasts can be found in Brown and Mahoney (1998). These methods are briefly described 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 for Ellrod-1, GTG, GTG2, and ITFA are listed in Table 1; thresholds for other algorithms included on RTVS can be found on the RTVS web pages. Thus, the basic verification approach makes use of 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 1: Threshold values used to convert algorithm forecasts to Yes/No forecasts.

Algorithm	*Thresholds*
Ellrod-1	10^{-8}, $30x10^{-8}$, $40x10^{-8}$, $50x10^{-8}$, $70x10^{-8}$, $200x10^{-8}$
GTG	0.060, 0.125, 0.150, 0.250, 0.375, 0.500, 0.625
GTG2	0.060, 0.125, 0.150, 0.250, 0.375, 0.500, 0.625
ITFA	0.060, 0.080, 0.150, 0.200, 0.300, 0.375, 0.400

Table 2: Contingency table for evaluation of dichotomous (Yes/No) forecasts. Elements in the cells are the counts of forecast-observation pairs.

Forecast	*Observation*		*Total*
	Yes	*No*	
Yes	YY	YN	YY+YN
No	NY	NN	NY+NN
Total	YY+NY	YN+NN	YY+YN+NY+NN

Table 3 lists the verification statistics used in this evaluation. As shown in this table, PODy and PODn are the primary verification statistics used for the evaluation of GTG2 and the other turbulence algorithms. Together, PODy and PODn measure the ability of the forecasts to discriminate between (or correctly categorize) 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**.**

The relationship between PODy and 1-PODn for different algorithm thresholds is the basis for the verification approach known as "Signal Detection Theory" (SDT). For a given algorithm, this relationship can be represented 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.

It should be noted that Table 3 does not include the False Alarm Ratio (FAR), a statistic that is commonly computed from the 2x2 table. Due to the non-systematic nature of PIREPs, it is not appropriate to compute FAR using these observations. This conclusion, which also applies to statistics such as the Critical Success Index and Bias, is documented analytically and by example in Brown and Young (2000). In addition, due to characteristics of PIREPs and their limited numbers, other verification statistics (e.g., PODy and PODn) should not be interpreted in an absolute sense, but can be used for comparisons among algorithms and forecasts. Moreover, PODy and PODn should not be interpreted as probabilities, but rather as proportions of PIREPs that are correctly forecast.

Table 3: Verification statistics used in this study.

Statistic	*Definition*	*Description*	*Interpretation*	*Range*
PODy	YY/(YY+NY)	Probability of Detection of Yes observations	Proportion of Yes observations that were correctly forecasted	0-1 Best: 1 Worst: 0
PODn	NN/(YN+NN)	Probability of Detection of No observations	Proportion of No observations that were correctly forecasted	0-1 Best: 1 Worst: 0
TSS	PODy + PODn – 1	True Skill Statistic; Hanssen-Kuipers discrimination	Level of discrimination between Yes and No observations	-1 to 1 Best: 1 No skill: 0
Curve Area	Area under the curve relating PODy and 1-PODn	Area under the curve relating PODy and 1-PODn (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	Percent of the total air space volume that is impacted by the forecast	Percent of the total air space volume that is impacted by the forecast	0-100 Smaller is better
Volume Efficiency (VE)	(PODy x 100) / % Volume	PODy (x 100) per unit % Volume	PODy relative to airspace coverage	0-infinity Larger is better

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 volume[5] 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.

[5] The total possible area (limiting coverage to the area of the continental United States that can be included in AIRMETs) is 9.5 million km^2. For the two altitude ranges (10-20,000 ft and above 20,000 ft), the total possible volume values are about 29 million km^3, and 57 million km^3, respectively

Use of these statistics is considered in somewhat greater detail in Brown et al. (2000a). In general, however, the argument presented in the previous paragraph can be extended to all of the statistics in Table 3; none of the statistics should be considered in isolation – all should be examined in combination with the others to obtain a complete picture of forecast quality.

Emphasis will be placed on PODy, PODn, and % Volume as measures of forecast performance. 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 is often 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 QAPDT has developed an alternative confidence interval method based on re-sampling statistics, which is 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

All of the evaluations are limited to PIREPs and algorithm output in the altitude ranges 10-20,000 ft (mid-levels) and above 20,000 ft (upper levels). 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. In addition, for most analyses, only PIREPs available through regular FAA/NOAA sources are included; additional reports received from United and Northwest Airlines are included in selected comparisons. In almost all cases the results are stratified by lead time.

Stratification by regions was also included in this study. Figure 3 shows the three large regions utilized by the RTVS. These regions are based on the regional boundaries used by forecasters at the AWC. Verification statistics for the turbulence forecasts are computed separately for the West, Central, and East regions, for both the mid- and upper levels. In addition, 15 small regions (Figure 4) were defined based on climatological characteristics. Mid- and upper-level verification statistics for GTG2 and ITFA are considered for these smaller regions.

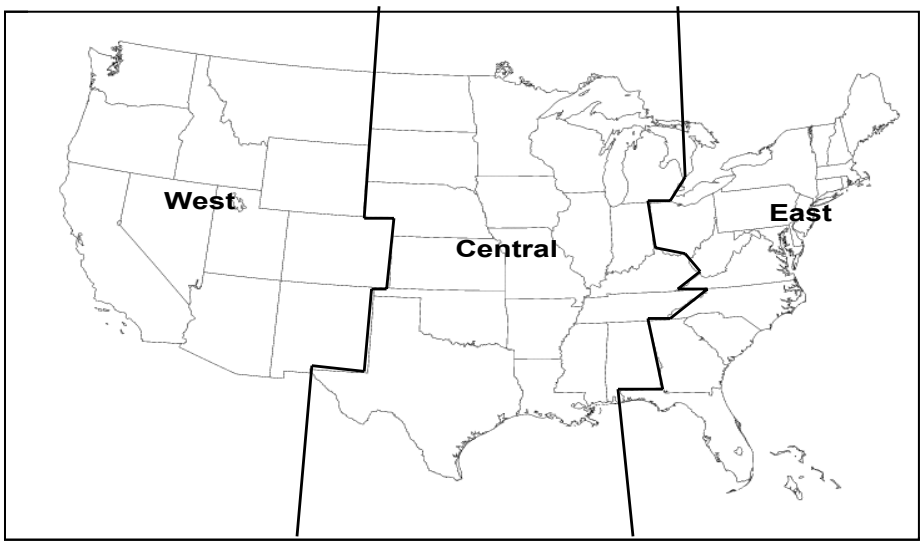


Figure 3: Map of large regions used by AWC forecasters.

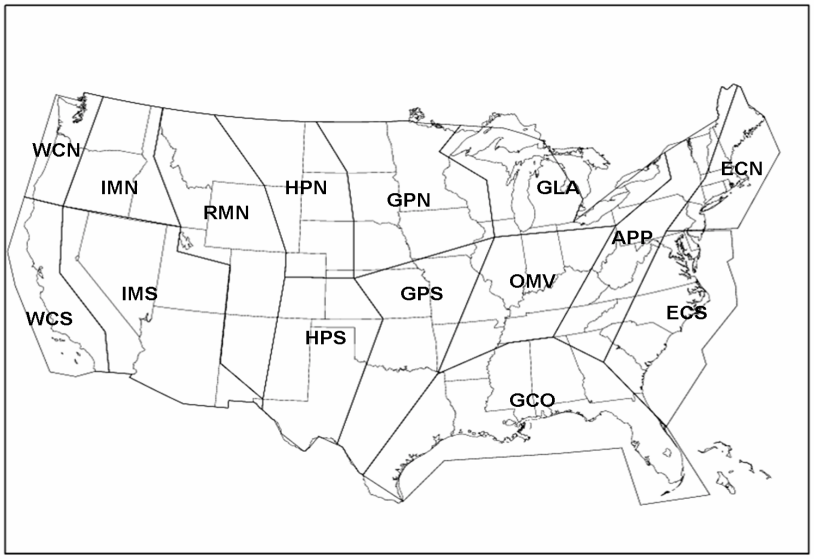


Figure 4. Map of small regions used for regional evaluations of turbulence forecast performance.

6. Results

This section summarizes results from the 2004 evaluation of GTG2. Results for other algorithms are also presented and compared to the GTG2 verification statistics. Except for the small-region results, all of the statistics were retrieved from the RTVS. The period examined includes 1 January – 30 April 2004.

6.1 Overall results

Overall turbulence forecast verification results for mid-level and upper-level forecasts, stratified by lead time, for the GTG2, Ellrod-1, ITFA, GTG (upper levels only) and AIRMET (6-h only) forecasts, are presented in ROC curves in Figs. 5 and 6. In these diagrams, the individual points on the algorithm curves represent particular thresholds used to create Yes/No forecasts. More skillful forecasts are represented by curves that are located closer to the upper left corner of the diagram. These figures show that for each lead time GTG2 is better than the other turbulence forecasts at classifying (or discriminating between) Yes and No turbulence observations, at both mid- and upper levels. In particular, the ROC curve for GTG2 is located much closer to the upper left corner than the curves for the other algorithms.

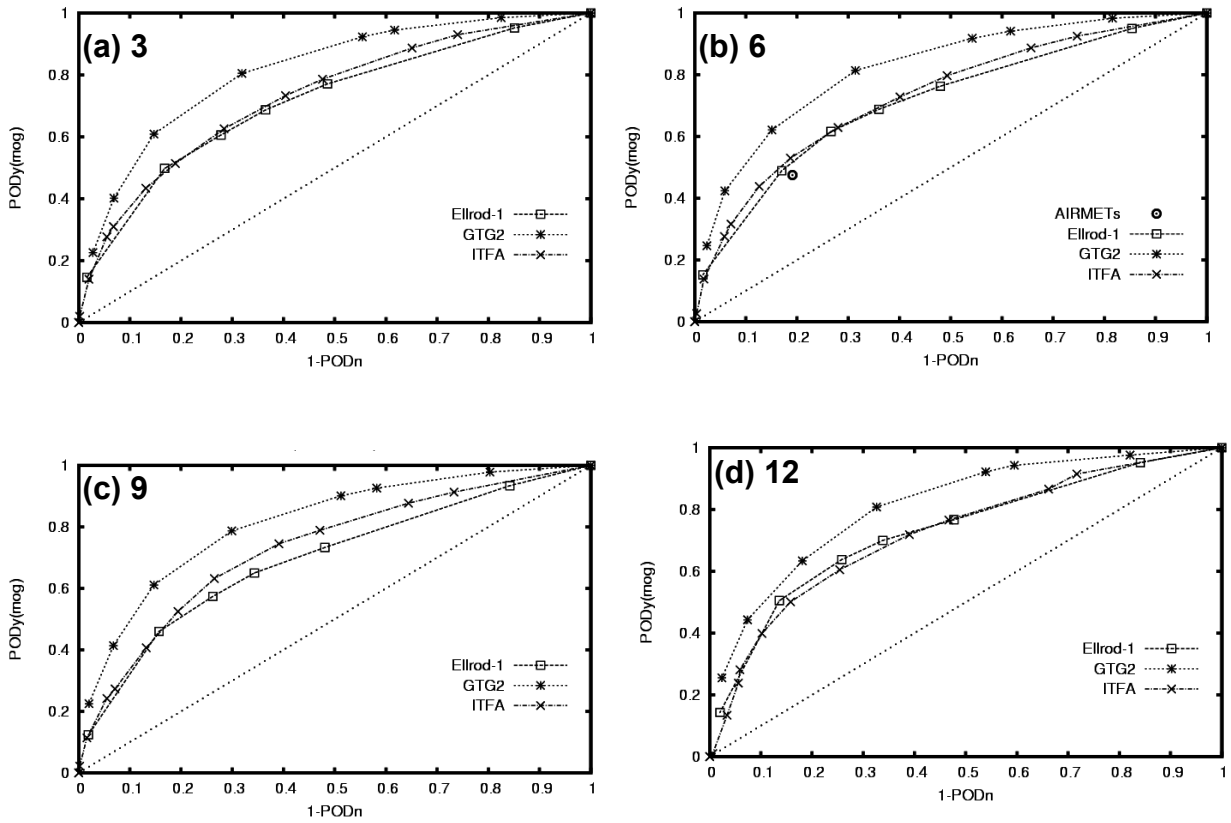


Figure 5. ROC diagrams for Ellrod-1, GTG2, and ITFA for mid-levels for lead times: (a) 3 h, (b) 6 h, (c) 9 h, and (d) 12 h.

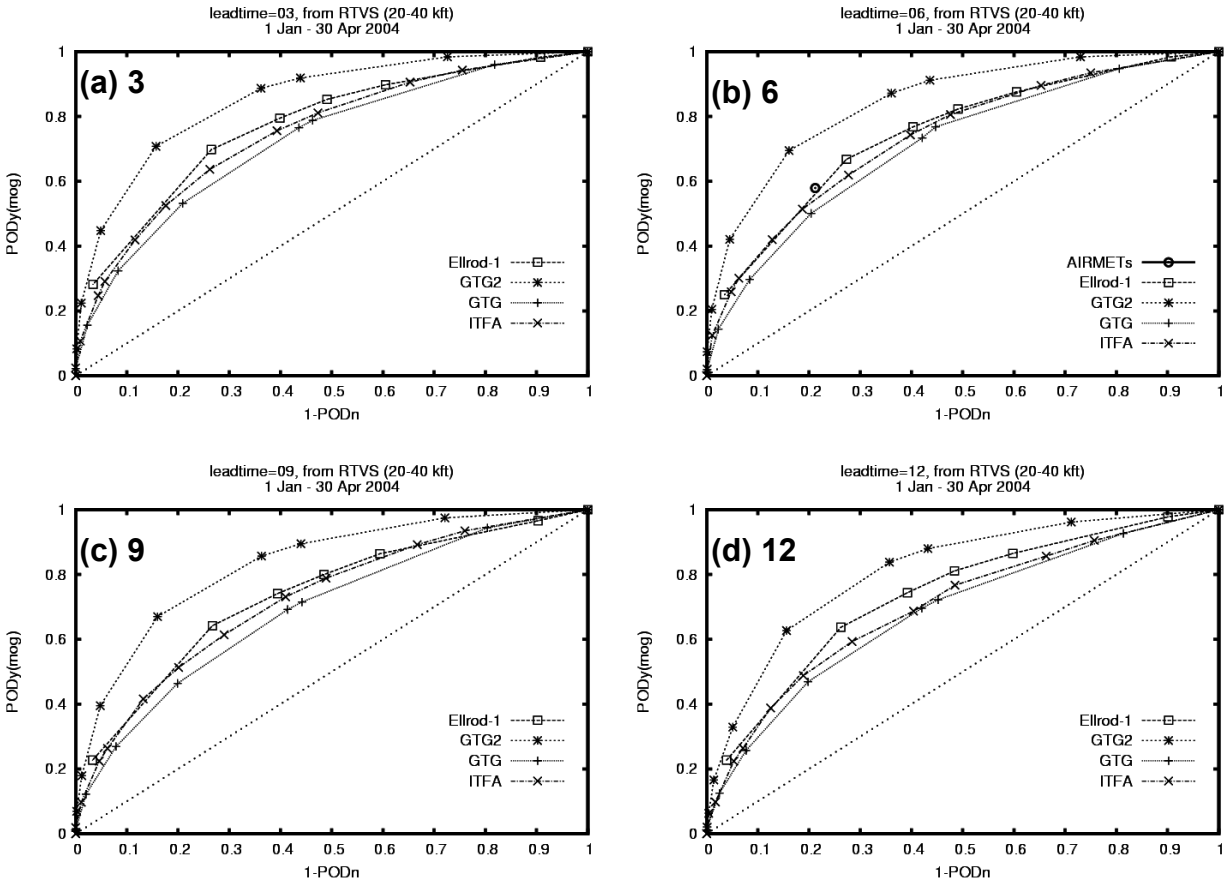


Figure 6. As in Fig. 5, for upper levels.

Confidence intervals provide an objective way of comparing the forecasting performance of the algorithms. Because Ellrod-1 largely performs better than the other algorithms except GTG2, it is particularly useful to compare GTG2 and Ellrod-1 performance by examining their ROC confidence intervals. Figure 7 shows 95% confidence intervals for both GTG2 and Ellrod-1 for 6-h forecasts. Because the confidence bands do not overlap (except *very* slightly for large thresholds at mid-levels), it is clear that GTG2 has significantly greater skill than Ellrod-1, and hence, the other algorithms. At upper levels, the differences between the confidence interval for GTG2 and Ellrod-1 is larger than at mid-levels, indicating even larger differences between the two algorithms in favor of GTG2.

Figures 8 and 9 show plots of PODy(MOG) vs. % Volume for each algorithm for mid- and upper levels, stratified by lead time. As in the ROC plots, curves located closer to the upper left corner of the diagram are more skillful. In particular, for the algorithms with better performance, incremental improvements in PODy are associated with smaller increases in % Volume.

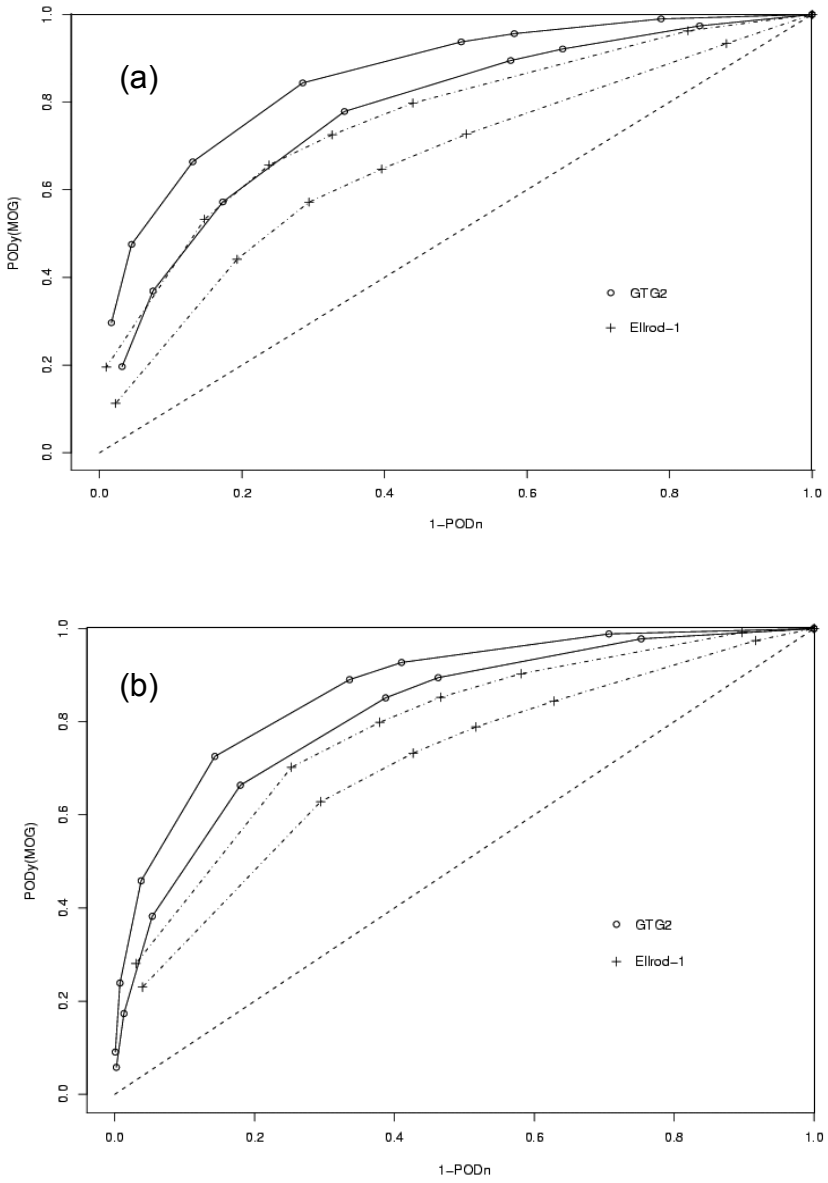


Figure 7. ROC 95% confidence intervals for GTG2 and Ellrod-1 for 6-h forecasts for (a) mid- and (b) upper levels.

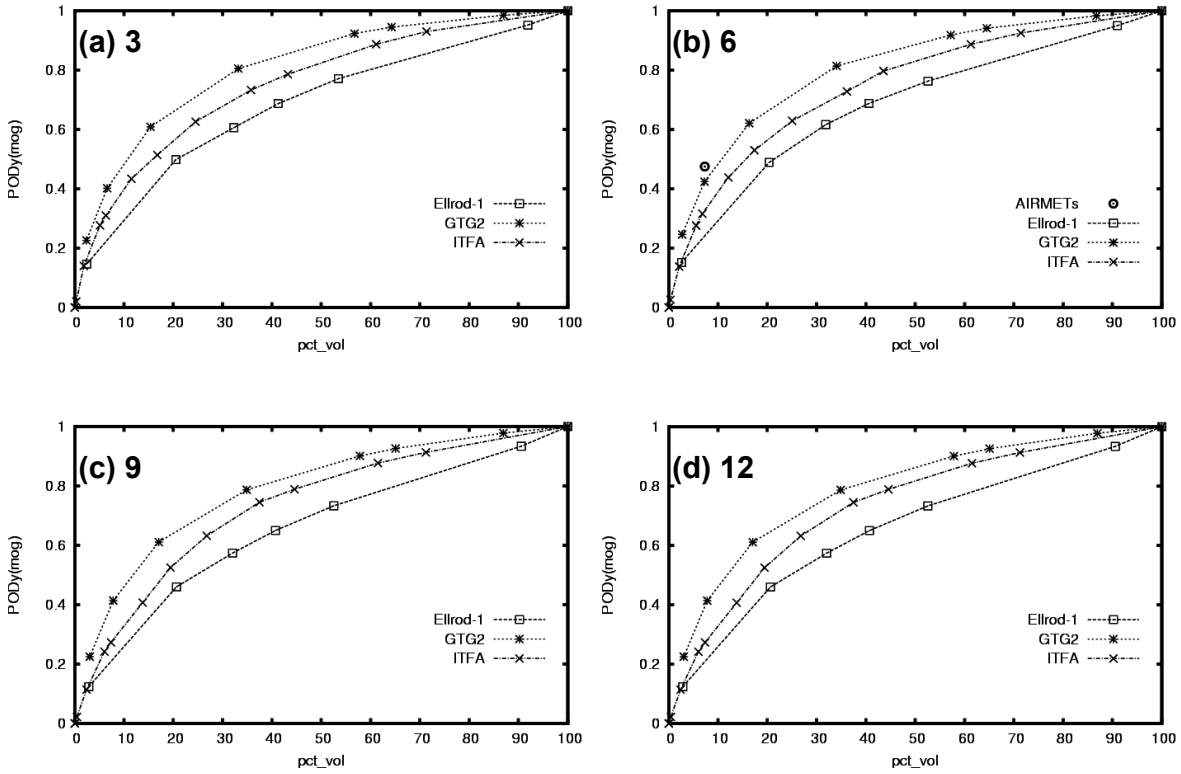


Figure 8. PODy(MOG) vs. % Volume plots for Ellrod-1, GTG2, and ITFA for mid-levels, for lead times: (a) 3 h, (b) 6 h, (c) 9 h, and (d) 12 h.

Figure 8 shows that GTG2 has better performance in this regard than the other forecasts, except the AIRMETs (Fig. 8b). For upper levels (Fig. 9), the PODy vs. % Volume results are similar for all of the algorithms, with GTG2 only slightly more successful than the others in some cases. As was the case for mid-levels, the GTG2 and AIRMET results for upper levels are quite similar, although the AIRMETs have slightly better statistics.

Tabular verification statistics for 6-h forecasts are shown in Table 4. The thresholds applied to these algorithms were selected because their corresponding PODy(MOG) value most closely matches the value attained by the AIRMETs. In some cases more than one threshold was used for an algorithm, usually to allow the PODy(MOG) values to span the AIRMET PODy(MOG) value. In addition, the threshold of 0.375 is also shown for GTG2, ITFA, and GTG, as this threshold is used by the algorithms as an indication of moderate turbulence severity. Compared to the other algorithms, for the 0.250 threshold, the GTG2 has the largest TSS for both mid- and upper levels. The best VE values are achieved by the AIRMETs for mid-levels and GTG2 with a threshold of 0.375 for upper levels.

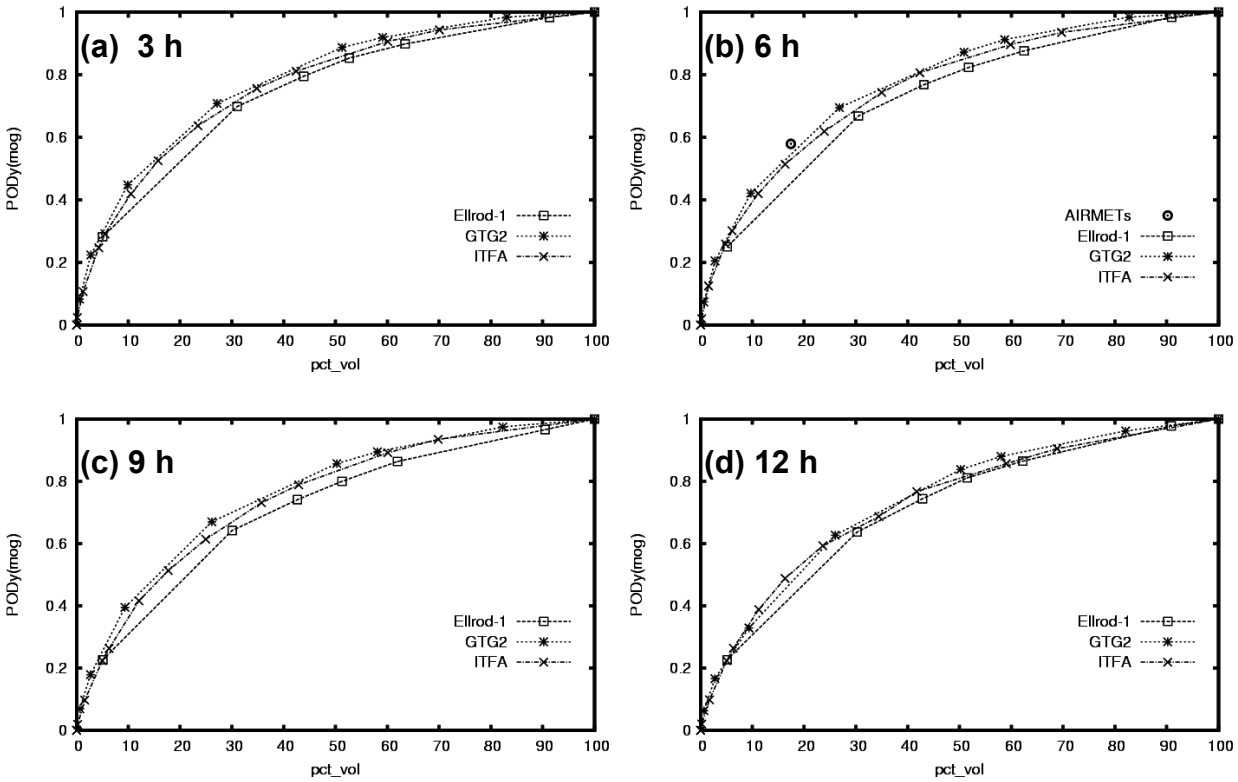


Figure 9. As in Fig. 8, for upper levels.

Table 4: Overall verification statistics for 6-h AIRMETs, GTG2, Ellrod-1, ITFA and GTG forecasts for selected thresholds, for both mid- and upper-level forecasts.

Algorithm	*Threshold*	*PODy (all)*	*PODy (MOG)*	*PODn*	*TSS*	*Ave. % Volume*	*VE*
Mid-levels							
AIRMETs	N/A	0.457	0.475	0.809	0.266	7.31	**6.25**
GTG2	0.250	0.782	0.805	0.686	**0.468**	33.96	2.30
	0.375	0.585	0.615	0.849	0.434	16.18	3.62
Ellrod-1	7e-7	0.458	0.488	0.837	0.295	20.53	2.23
ITFA	0.200	0.611	0.626	0.723	0.334	25.24	2.24
	0.375	0.291	0.316	0.929	0.220	6.83	4.26
Upper levels							
AIRMETs	N/A	0.565	0.579	0.788	0.353	17.43	3.24
GTG2	0.250	0.662	0.695	0.839	**0.502**	26.80	2.47
	0.375	0.390	0.419	0.952	0.342	9.67	**4.03**
GTG	0.250	0.483	0.501	0.795	0.278	15.69	3.08
	0.375	0.287	0.297	0.916	0.203	7.27	3.95
Ellrod-1	7e-7	0.650	0.668	0.727	0.377	30.47	2.13
ITFA	0.200	0.601	0.619	0.723	0.340	23.89	2.51
	0.375	0.287	0.301	0.937	0.224	6.09	4.71

The areas under the ROC curves (i.e., curves shown in Figs. 5 and 6) are shown in Table 5 for each algorithm, for all lead times and for mid-level and upper-level forecasts. As shown in Table 5, GTG2 has a larger ROC area for both mid- and upper-level forecasts, indicating better overall skill when compared to the skill for the other algorithms.

Table 5: ROC areas for Ellrod-1, GTG2, GTG, and ITFA

	Algorithm			
Altitude layer	***Ellrod-1***	***GTG2***	***GTG***	***ITFA***
Mid-level	0.71	0.82	N/A	0.73
Upper level	0.71	0.84	0.71	0.74

6.2 GTG2 comparisons among lead times

Figures 10 and 11 show the ROC and % Volume plots for GTG2 only, stratified by lead time, for mid- and upper-level forecasts, respectively. For the mid-level forecasts, the performance is nearly constant across lead times, with only a slight hint of degradation for the 9-h lead-time in the PODy vs. % Volume plot. In contrast, a gradual degradation in GTG2 performance with lead time is evident for the upper level forecasts, as shown in Fig. 11.

The ROC areas for GTG2 for each lead time and altitude layer are shown in Table 6. From this table, it is evident that at mid-levels, the skill does not vary with lead time to any significant extent. However, for the upper layer, the skill gradually decreases with increasing lead time. These differences in performance between the mid- and upper levels are described further in the next section.

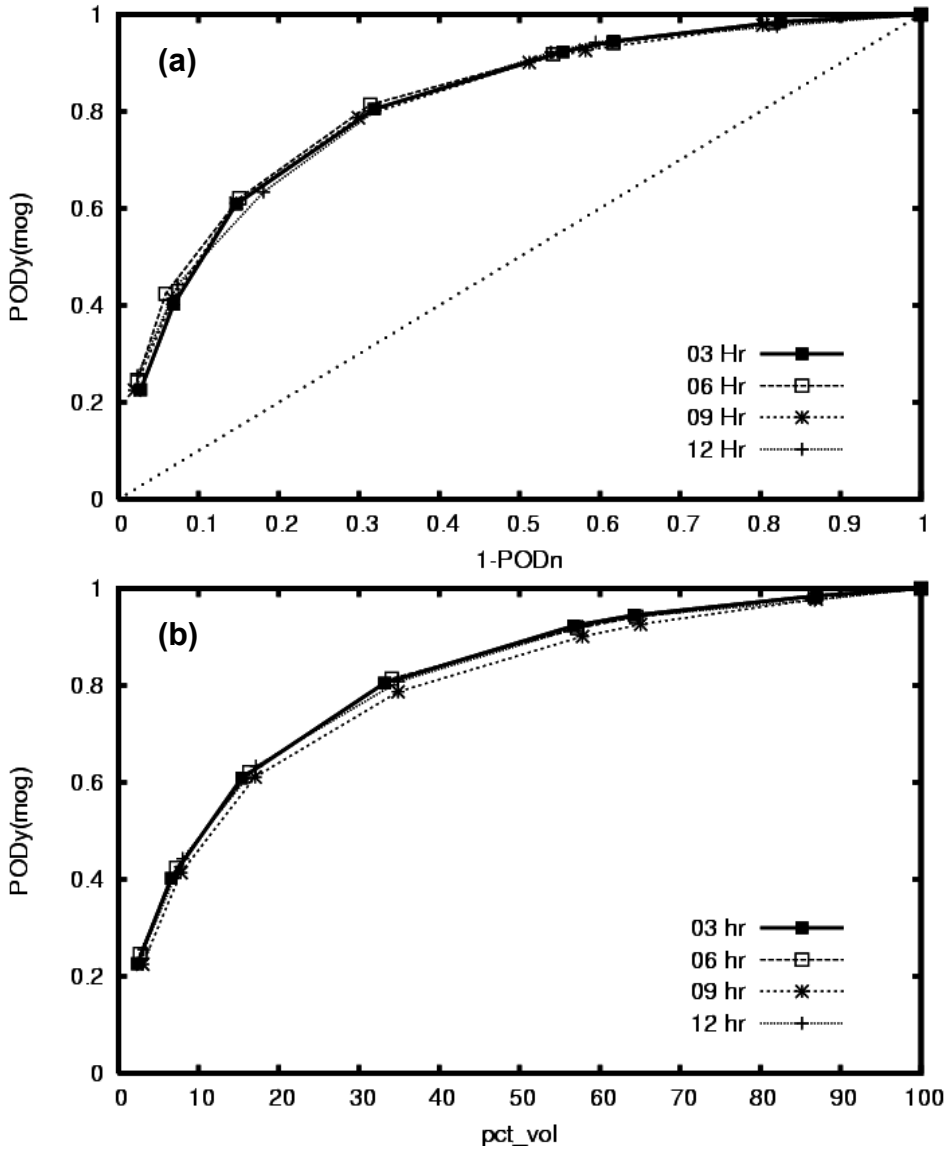


Figure 10. GTG2 verification statistics stratified by lead time for mid-level forecasts (a) ROC diagrams and (b) PODy(MOG) vs. % Volume.

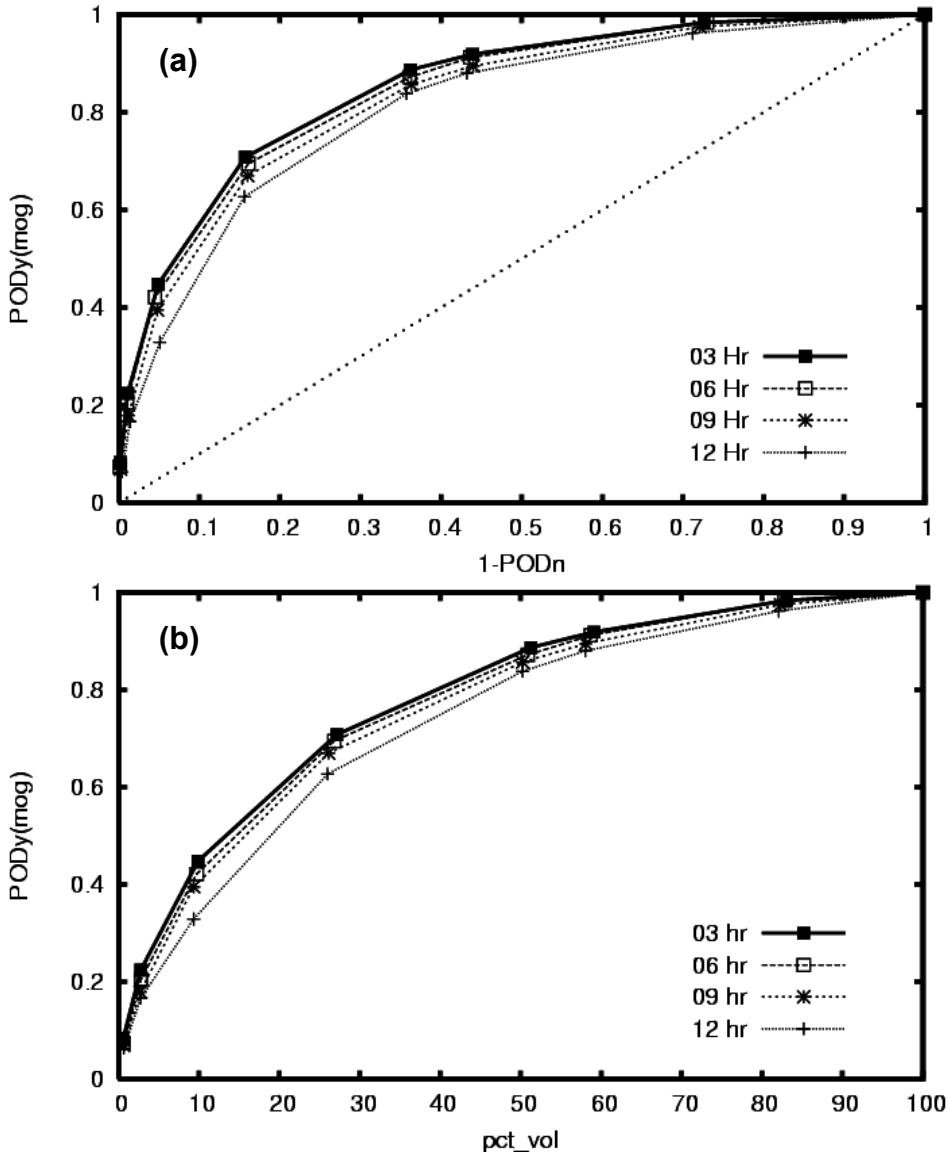


Figure 11. As in Fig. 10, for upper levels.

Table 6: ROC areas for 3-, 6-, 9-, and 12-h lead times for mid- and upper-level GTG2 forecasts.

	Lead time			
Layer	***3 h***	***6 h***	***9 h***	***12 h***
Mid-levels	0.81	0.82	0.81	0.81
Upper levels	0.85	0.85	0.83	0.81

6.3 Comparison by altitude

Figure 12 compares GTG2 performance at mid- and upper levels for all issue and lead times combined. As indicated by the ROC plots (Fig. 12a), GTG2 is better at discriminating between Yes and No observations of turbulence at upper levels than at mid-levels. Conversely, as shown by the PODy vs. % Volume plot, GTG2 provided more efficient volumetric forecasts at mid-levels than at upper levels.

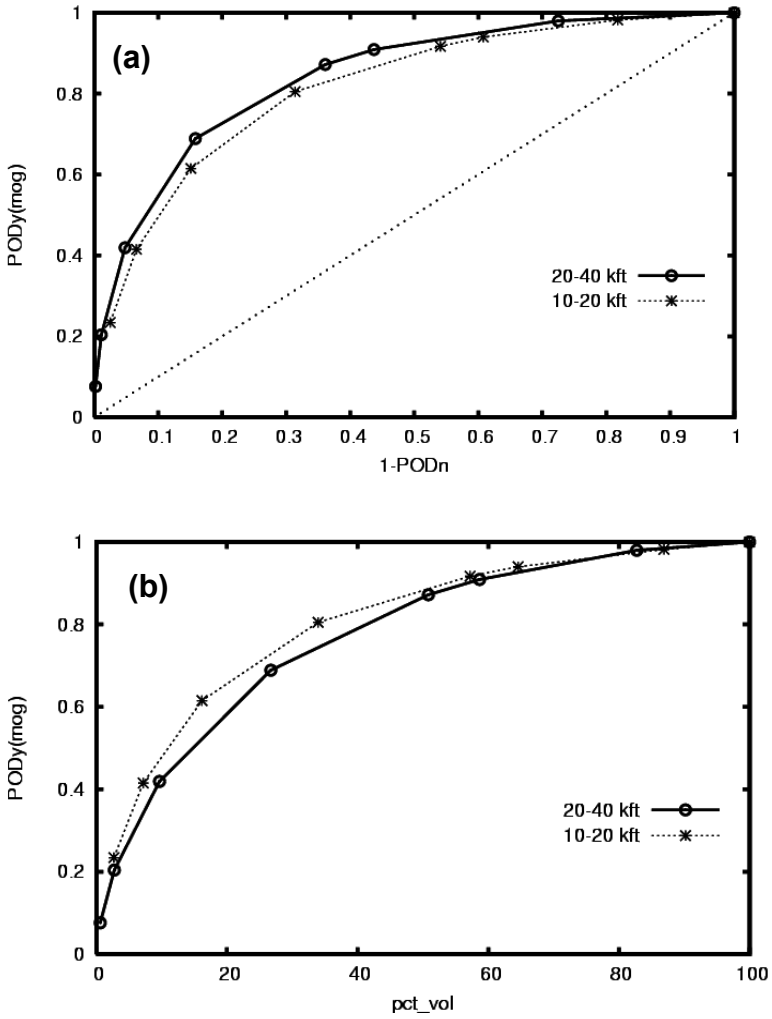


Figure 12: GTG2 verification statistics as a function of altitude range (mid- and upper levels) for all issue and lead times combined: (a) ROC diagrams and (b) PODy(MOG) vs. % Volume diagrams.

Height series plots of PODy(MOG) and PODn for 6-h GTG2, ITFA, and GTG forecasts, as well as AIRMETs are shown in Fig. 13. These plots show how the verification statistics vary with altitude. For the plots in Fig. 13, a threshold of 0.2 (0.25) was applied to the ITFA (GTG and GTG2) forecasts to create the Yes/No forecasts. Figure 14 shows the same plots with a threshold of 0.375 applied to ITFA, GTG2, and GTG. These plots indicate that all of the algorithms as well as the AIRMETs perform fairly consistently at all altitudes. The PODy(MOG) for GTG2 has a tendency to decrease slightly at the highest altitudes, whereas the PODy(MOG) for the AIRMETs is best at the higher altitudes. Results based on the two different sets of thresholds (Figs. 13 and 14) are consistent with one another.

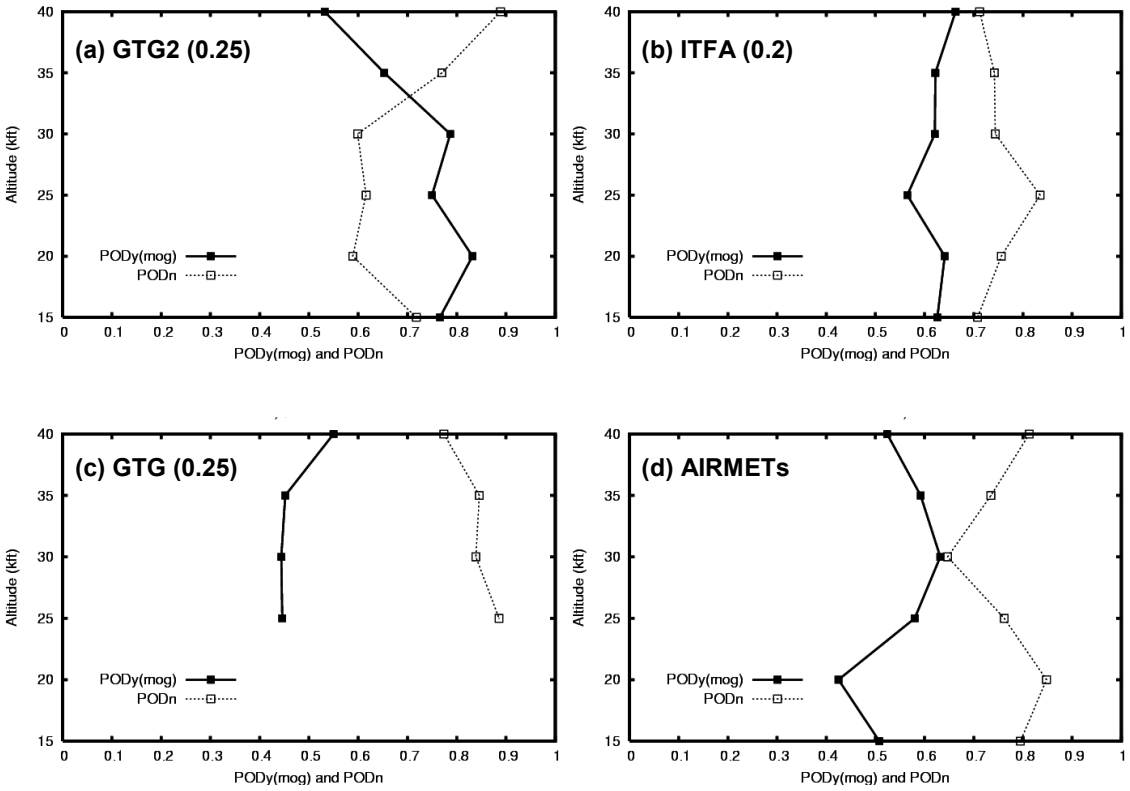


Figure 13: Variations in PODy(MOG) and PODn with altitude, for 6-h forecasts based on (a) GTG2 with a threshold of 0.25, (b) ITFA with a threshold of 0.20, (c) GTG with a threshold of 0.250, and (d) AIRMETs. The altitudes listed represent the top of each layer.

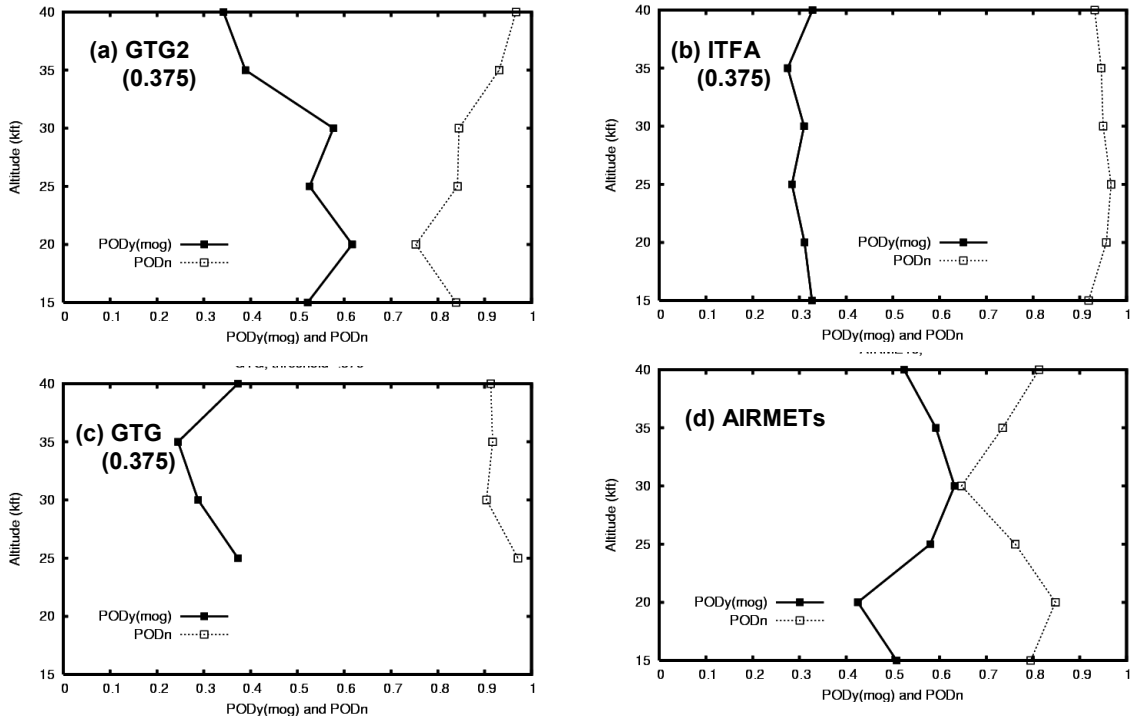


Figure 14. As in Fig. 13, with a threshold of 0.375.

Height series plots of TSS are shown in Fig. 15 for all the algorithms and the AIRMETs. The thresholds applied in Fig. 15a are 0.20 for ITFA and 0.25 for GTG and GTG2; a threshold of 0.375 was applied to all algorithms in Fig. 15b. As shown in Fig. 15a, GTG2 is more skillful than the other algorithms in terms of TSS, at nearly every level except 25,000 ft. At 25,000 ft ITFA has a slightly larger but comparable TSS value. For the second set of thresholds (Fig. 15b), GTG2 is more skillful from the mid-levels up to 30,000 ft, but above 30,000 ft the AIRMETs display somewhat better forecast skill. While the TSS value for the AIRMETs remains fairly constant, especially for the 0.375 threshold value, the TSS values for GTG2 gradually decrease with increasing altitude.

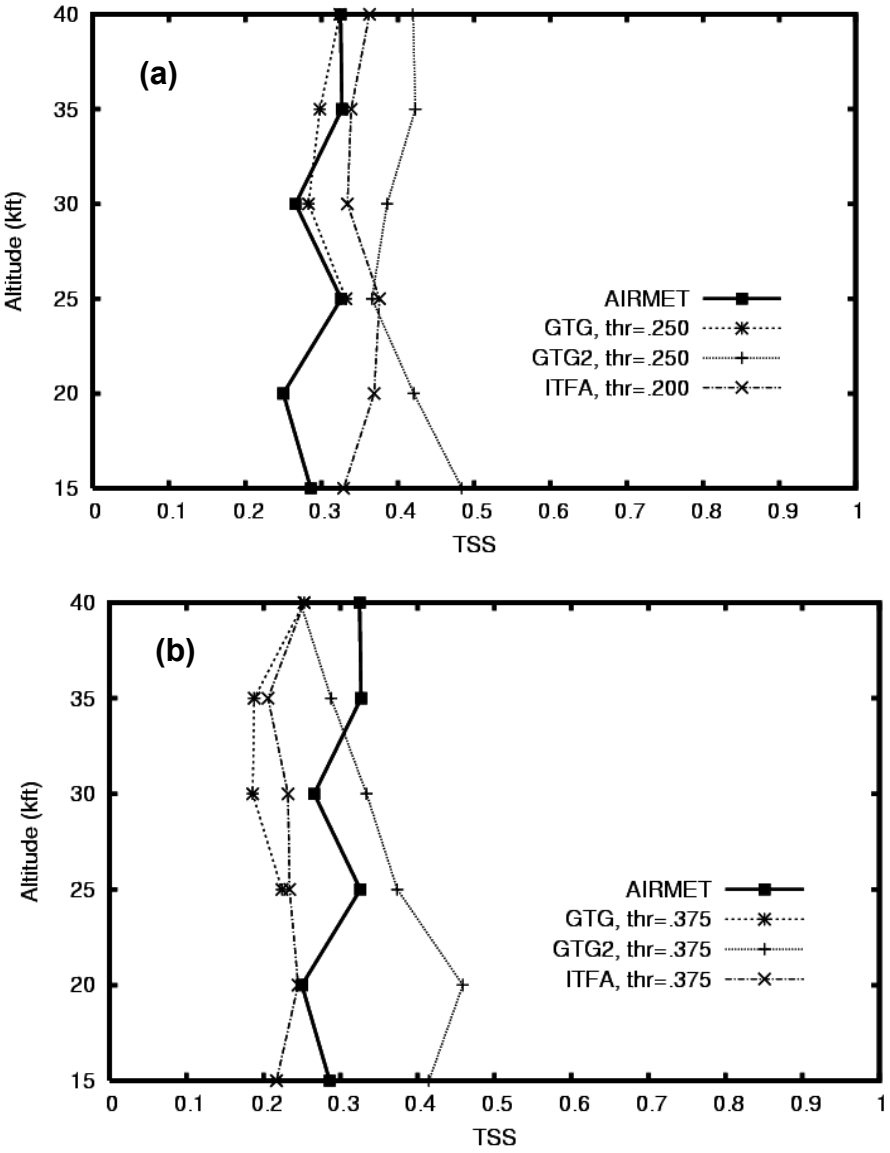


Figure 15: Variations in TSS for individual algorithms by altitude, for 6-h forecasts: (a) AIRMETs, GTG (threshold of 0.25), GTG2 (threshold of 0.25), and ITFA (threshold of 0.20); and (b) AIRMETs, GTG, GTG2, and ITFA (all with threshold of 0.375).

6.4 Comparisons by regions

This section presents comparisons of turbulence forecast performance for individual regions, based on the large and small regions defined in Section 5.

6.4.1 Large regions

Figures 16 and 17 show ROC plots for the three large regions (West, Central, and East) for the 6-h GTG2, Ellrod-1, ITFA, GTG (upper levels only), and AIRMET forecasts, for the mid- and upper levels, respectively. For the West region at mid-levels (Fig. 16a) the results indicate that the GTG2 has considerably more skill at discriminating between Yes and No observations than Ellrod-1, ITFA, and the AIRMETs. For the Central and East regions (Figs. 16b, and c) at the mid-levels, GTG2 also achieves greater skill than the other forecasts, but the differences are smaller, especially for the East region. For upper levels (Fig. 17) the results are similar to those shown in Fig. 16. However, for this altitude range, GTG2 has substantially greater skill in all three regions than the forecasts based on the AIRMETs and other algorithms.

Figures 18 and 19 show plots of PODy(MOG) vs. % Volume for the three large regions for the 6-h forecasts by the various algorithms and the AIRMETs, for mid- and upper levels, respectively. The results for the West region at mid-levels (Fig. 18a) indicate the best skill is attained by GTG2 and the AIRMETs. For the Central region at mid-levels, for high thresholds GTG2 and ITFA are both somewhat more skillful than Ellrod-1 and the AIRMETs. For lower thresholds, GTG2 forecast skill is somewhat greater than the forecast skill for ITFA. AIRMET skill for this region and altitude layer is quite low, as indicated by the PODy (MOG) and % Volume values (Fig. 18b). For the East region (Fig. 18c), the results for GTG2, ITFA, and Ellrod-1 are very similar to those in the Central region. The statistics for the AIRMETs over the East region at mid-levels are much better than those for the Central region.

At upper levels over the West region (Fig. 19a), all forecasts exhibit similar skill, with the exception of Ellrod-1, which has slightly less skill than the others. In the Central region (Fig. 19b), the results show an increase in forecast skill for GTG2 and the AIRMETs in comparison to the other algorithms. In the East region (Fig. 19c), the results show that the AIRMETs have the best skill as indicated by the GTG2 and ITFA curves, which fall just below the AIRMET point, and are slightly above the curves for GTG and Ellrod-1.

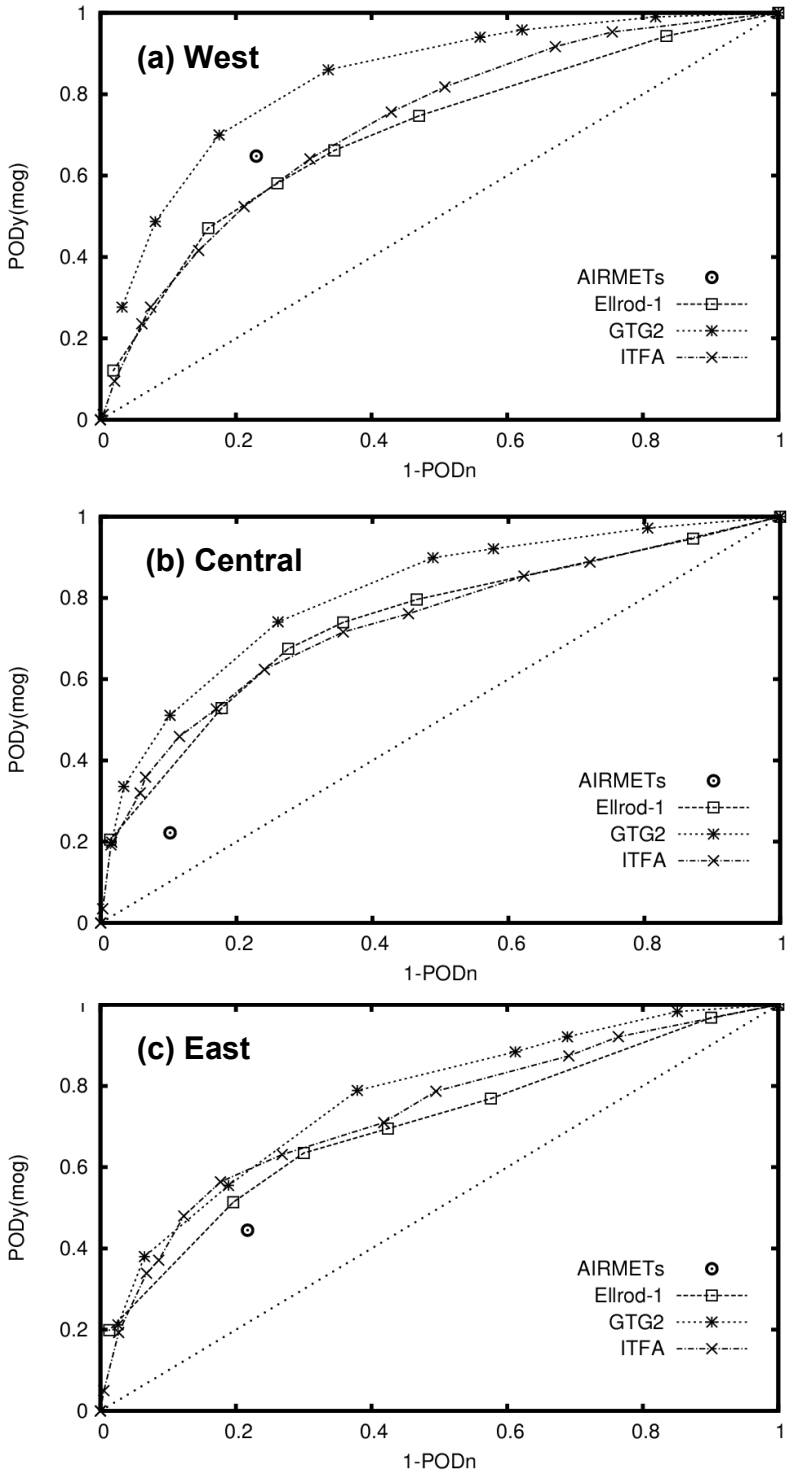


Figure 16. ROC diagrams for large regions for 6-h forecasts at mid-levels: (a) West; (b) Central; (c) East.

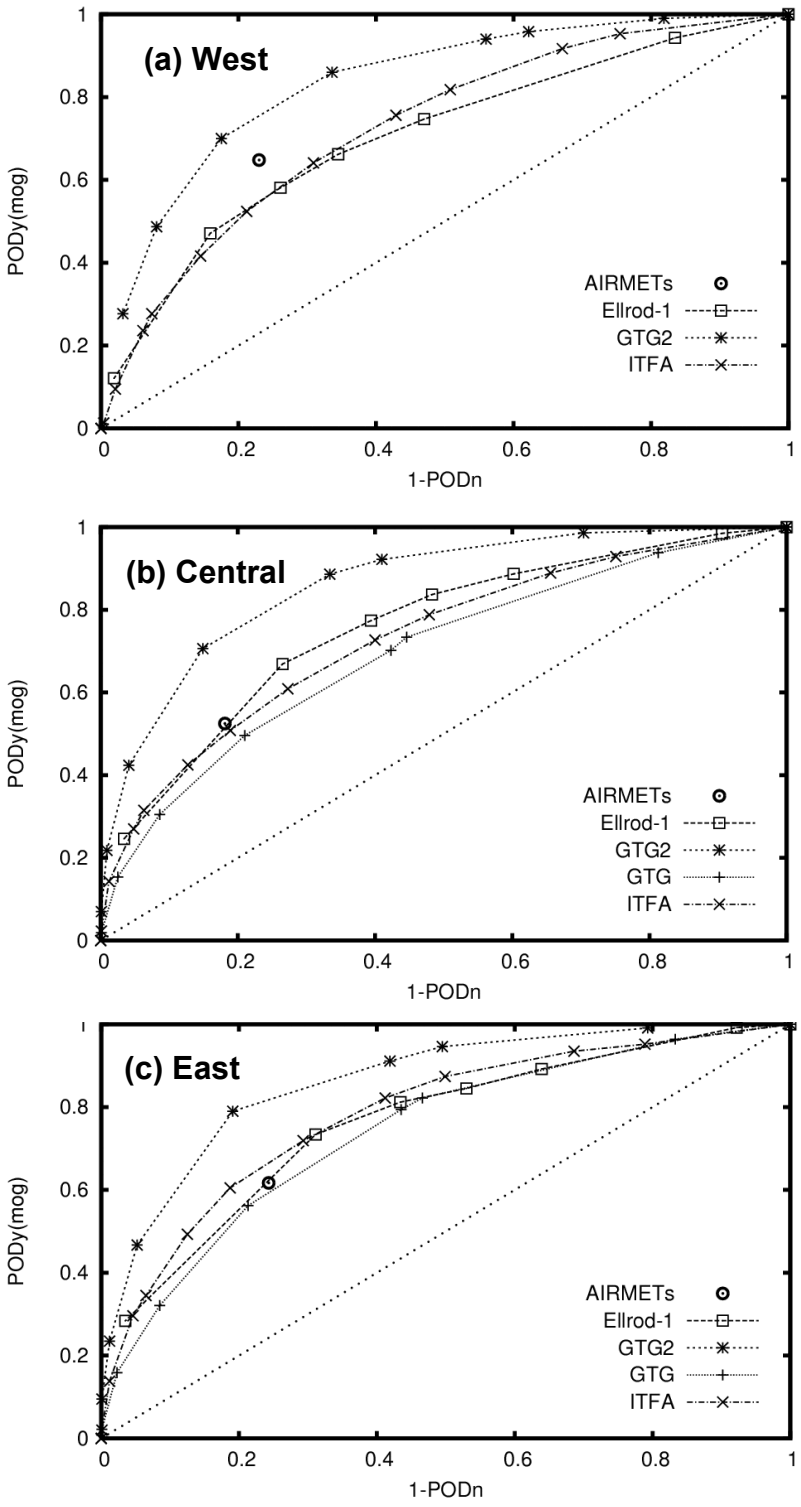


Figure 17. As in Fig. 16 for upper levels.

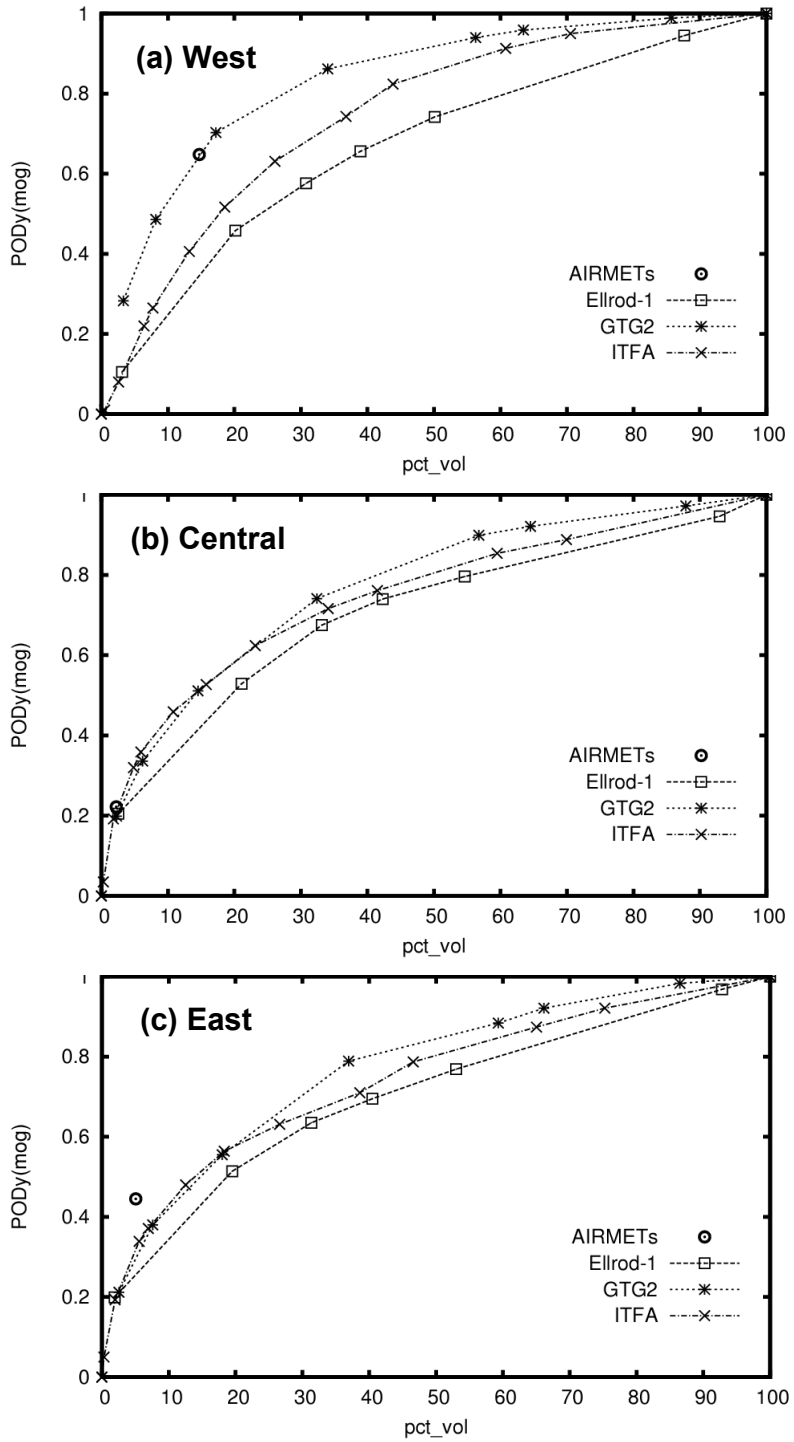


Figure 18. Relationship between PODy(MOG) and %Volume for large regions, for 6-h forecasts at mid-levels: (a) West; (b) Central; (c) East.

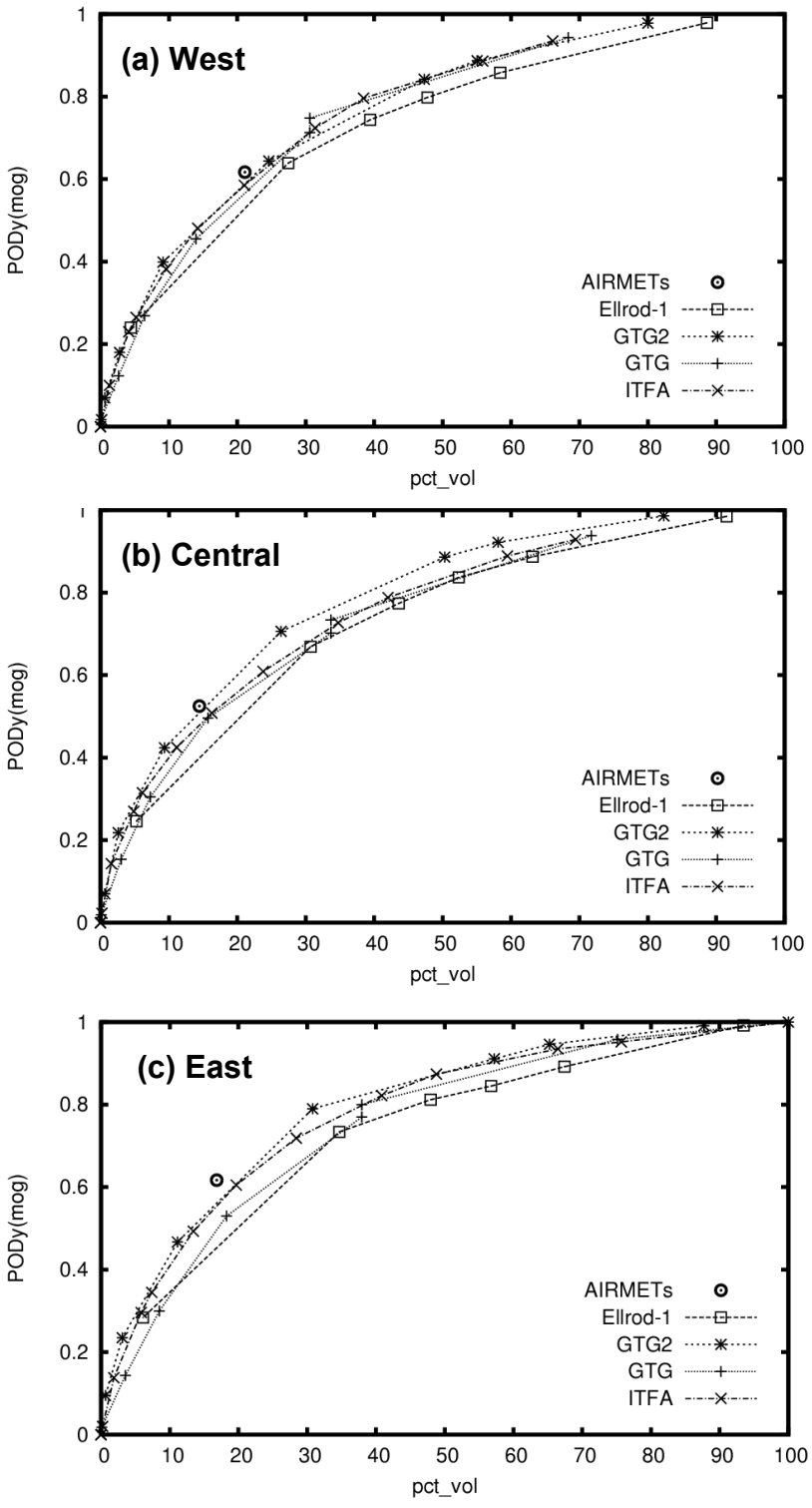


Figure 19. As in Fig. 18 for upper levels.

Figure 20 shows ROC diagrams for GTG2 over the three regions, for both mid- and upper level forecasts. The results for mid-levels (Fig. 20a) show that GTG2 has greater skill in the West than in the other regions at this altitude layer, with the least skill associated with the East region. For upper levels (Fig. 20b), the best skill for GTG2 is attained over the East and Central regions.

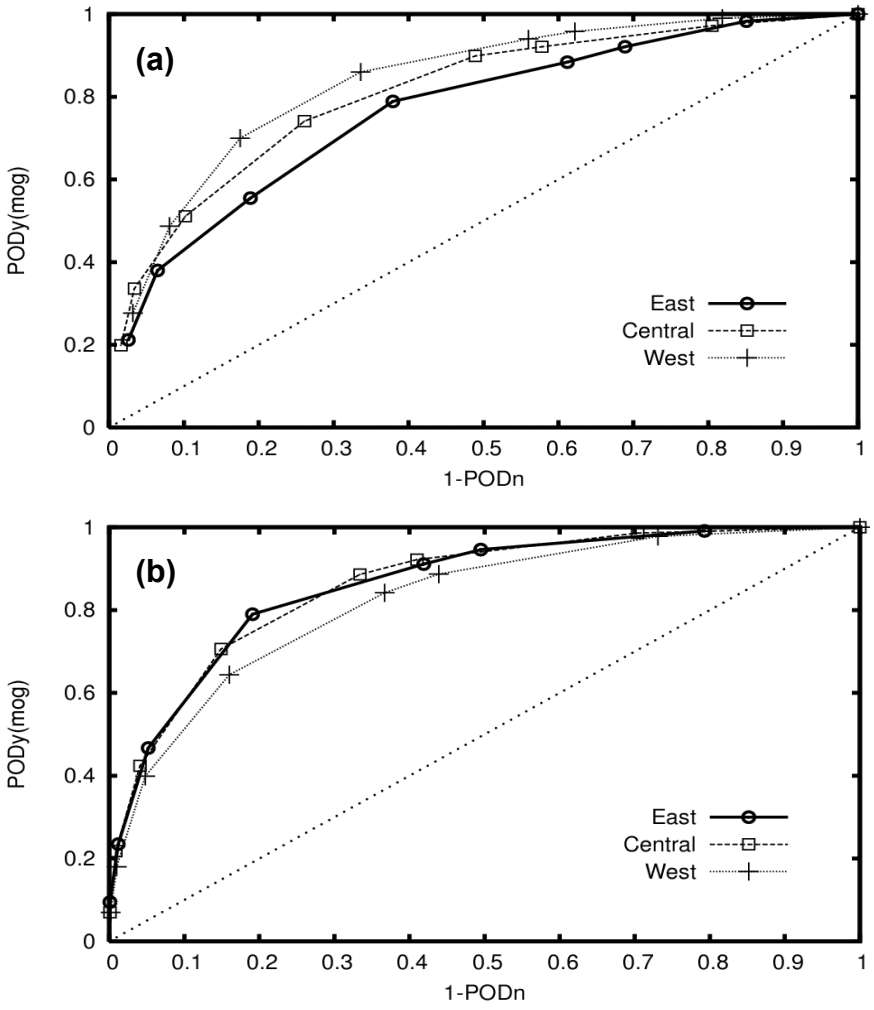


Figure 20. Relationship between PODy(MOG) and 1-PODn for GTG2: (a) Mid-levels (10-20,000 ft.); (b) Upper levels.

6.4.2 Small regions

Verification results for the 15 climatologically-defined regions are presented in this section. As shown in Fig. 4, the 15 regions are the following: West Coast North (WCN), West Coast South (WCS), Inter-mountain North (IMN), Inter-mountain South (IMS), Rocky Mountains (RMN), High Plains North (HPN), High Plains South (HPS), Great Plains North (GPN), Great Plains South (GPS), Great Lakes (GLA), Ohio-Mississippi Valley (OMV), Gulf Coast (GCO), Appalachians (APP), East Coast North (ECN), and East Coast South (ECS).

Mid-level GTG2 and ITFA forecasts. Tables 8 and 9 present verification statistics for the 15 small regions for mid-level 6-h GTG2 and ITFA forecasts with a threshold of 0.30. The statistics for these algorithms are also diagrammed in Fig. 21. Over all of the regions combined, the PODy(MOG) for GTG2 (0.774) (Table 8) is notably larger than the value for ITFA (0.565) and the overall PODn for GTG2 (0.679) is somewhat smaller than the value for ITFA (0.750). It is difficult to compare these values since the larger PODy values are generally also associated with larger % Volume values and smaller PODn values. For this reason, TSS can provide a clearer comparison of the algorithms' skill. The overall TSS values (for the CONUS; Fig. 21) are 0.315 for ITFA and 0.453 for GTG2, thus indicating better skill for GTG2 than for ITFA over the 15 regions combined. The TSS results (Table 9) show larger skill scores for the GTG2 in comparison with ITFA in all of the regions except for GCO, GPN, HPN, and OMV. However, the TSS values for the GPN region are large for both algorithms, and the difference is subtle. For GCO, HPN, and OMV, a smaller PODn value is the cause of the associated smaller TSS value. The largest differences appear in the APP, IMN, and WCN regions. In these regions, the PODy(MOG) is substantially larger for GTG2, leading to much different TSS values for the two algorithms: the TSS values for ITFA (GTG2) are 0.022 (0.565) for APP; 0.064 (0.528) for IMN; and 0.195 (0.391) for WCN. It must also be noted that the number of reports for both the "MOG" and "NO" categories is sparse in several of the regions, which can lead to variability in the results.

In contrast to the TSS results in Table 9, the VE results indicate that the ITFA forecasts were at least somewhat more efficient than GTG2 in the amount of volume coverage relative to the value of PODy(MOG). Exceptions are for the APP and IMN regions, where the VE values for GTG2 are somewhat larger than those for ITFA. In part, this result is related to the differences in PODy(MOG) shown in Table 8; as noted in Section 5.2, because VE is a ratio it can be easy to obtain a large VE value when the PODy(MOG) value is relatively small.

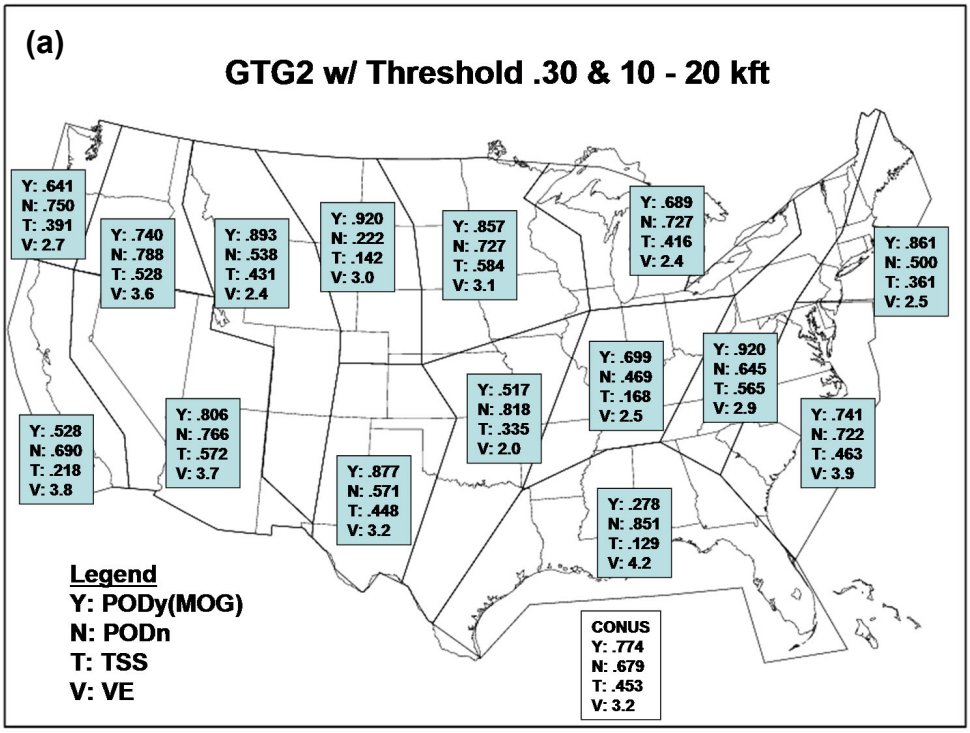
Upper-level GTG2 and ITFA forecasts. Tables 10 and 11 show results for the 15 smaller regions for 6-h upper-level (defined for these analyses as 20-46,000 ft) GTG2 and ITFA forecasts based on a threshold of 0.30. The results are also diagrammed in Fig. 22. Over all of the regions, GTG2 has a larger PODn in 7 of the 15 regions and larger PODy(MOG) and %Volume values in all of the 15 regions (Table 10); however, the increase in % Volume is not as drastic as was noted for the mid-levels. The differences in these statistics for the two algorithms may be partially indicative of differences in the calibration of ITFA and GTG2 (i.e., a threshold of 0.3 essentially may not mean the same thing for ITFA and GTG2 due to changes in the formulation of the algorithm). In any case, the larger PODn and PODy(MOG) values for GTG2 are reflected by the larger TSS values for all of the regions (Table 11). The regions with the biggest

Table 8: Regional results for 6-h mid-level ITFA and GTG2 forecasts.

Regions	*No. of MOG PIREPs*	*PODy(MOG)*		*No. of No PIREPs*	*PODn*		*% Volume*	
		ITFA	*GTG2*		*ITFA*	*GTG2*	*ITFA*	*GTG2*
APP	25	0.280	**0.920**	31	**0.742**	0.645	**12.8**	31.4
ECN	36	0.472	**0.861**	6	**0.667**	0.500	**15.7**	34.0
ECS	27	0.667	**0.741**	18	**0.778**	0.722	**8.4**	18.9
GCO	18	0.278	0.278	47	**0.915**	0.851	**3.4**	6.7
GLA	90	0.456	**0.689**	11	**0.909**	0.727	**10.7**	28.8
GPN	42	0.690	**0.857**	11	**0.909**	0.727	**11.3**	27.7
GPS	60	0.367	**0.517**	33	0.788	**0.818**	**12.3**	25.8
HPN	75	0.760	**0.920**	9	**0.556**	0.222	**16.0**	30.5
HPS	73	0.849	**0.877**	56	**0.661**	0.571	**16.4**	27.4
IMN	50	0.220	**0.740**	45	**0.844**	0.788	**9.6**	20.6
IMS	93	0.559	**0.806**	141	0.745	**0.766**	**14.2**	22.1
OMV	123	0.545	**0.699**	32	**0.781**	0.469	**11.6**	27.5
RMN	244	0.732	**0.893**	106	**0.604**	0.538	**22.3**	37.2
WCN	64	0.266	**0.641**	28	**0.929**	0.750	**7.0**	23.4
WCS	36	0.361	**0.528**	87	**0.759**	0.690	**6.1**	14.0

Table 9: Regional TSS and Volume Efficiency values for 6-h mid-level ITFA and GTG2 forecasts (10-20,000 ft).

Regions	*TSS*		*VE*	
	ITFA	*GTG2*	*ITFA*	*GTG2*
APP	0.022	**0.565**	2.2	**2.9**
ECN	0.139	**0.361**	**3.0**	2.5
ECS	0.445	**0.463**	**7.9**	3.9
GCO	**0.193**	0.129	**8.2**	4.2
GLA	0.365	**0.416**	**4.3**	2.4
GPN	**0.599**	0.584	**6.1**	3.1
GPS	0.155	**0.335**	**3.0**	2.0
HPN	**0.316**	0.142	**4.8**	3.0
HPS	0.510	**0.448**	**5.2**	3.2
IMN	0.064	**0.528**	2.3	**3.7**
IMS	0.304	**0.572**	**4.0**	3.6
OMV	**0.326**	0.168	**4.7**	2.5
RMN	0.336	**0.431**	**3.3**	2.4
WCN	0.195	**0.391**	**3.8**	2.7
WCS	0.120	**0.218**	**5.9**	3.8



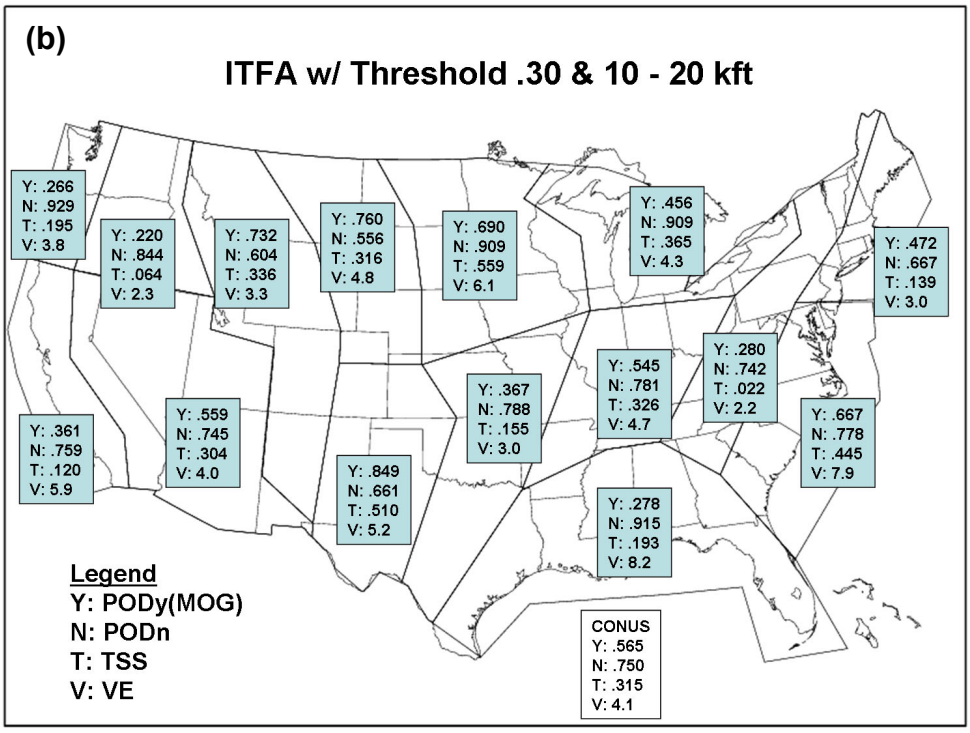


Figure 21. Regional results for 6-h forecasts for mid-level forecasts: (a) GTG2; (b) ITFA.

Table 10: Regional results for 6-h upper-level ITFA and GTG2 forecasts.

Regions	*No. of MOG PIREPs*	*PODy(MOG)*		*No. of No PIREPs*	*PODn*		*% Volume*	
		ITFA	*GTG2*		*ITFA*	*GTG2*	*ITFA*	*GTG2*
APP	319	0.564	**0.677**	175	**0.857**	0.811	**16.2**	20.8
ECN	50	0.580	**0.680**	23	0.870	**0.913**	**15.0**	18.3
ECS	167	0.473	**0.802**	117	**0.846**	0.803	**12.9**	22.5
GCO	420	0.324	**0.581**	248	**0.879**	0.827	**6.9**	17.0
GLA	380	0.513	**0.634**	139	0.827	**0.899**	**11.0**	13.4
GPN	280	0.625	**0.654**	166	0.849	**0.867**	**8.3**	12.2
GPS	432	0.454	**0.606**	235	0.740	**0.796**	**8.5**	17.1
HPN	143	0.497	**0.594**	140	0.786	**0.900**	**8.3**	13.9
HPS	290	0.493	**0.545**	151	0.781	**0.841**	**9.8**	17.5
IMN	476	0.340	**0.592**	108	**0.889**	0.806	**7.4**	13.4
IMS	444	0.426	**0.493**	247	**0.846**	0.834	**8.9**	14.3
OMV	541	0.497	**0.671**	257	0.829	**0.868**	**12.4**	17.7
RMN	510	0.482	**0.667**	282	**0.855**	0.770	**9.2**	18.7
WCN	318	0.267	**0.484**	31	0.839	0.839	**7.2**	11.4
WCS	150	0.307	**0.533**	78	**0.833**	0.782	**14.5**	11.4

Table 11: Regional TSS and Volume Efficiency values for 6-h upper-level ITFA and GTG2 forecasts.

Regions	*TSS*		*VE*	
	ITFA	*GTG2*	*ITFA*	*GTG2*
APP	0.421	**0.488**	**3.5**	3.3
ECN	0.450	**0.593**	**3.9**	3.7
ECS	0.319	**0.605**	**3.7**	3.6
GCO	0.203	**0.408**	**4.7**	3.4
GLA	0.340	**0.533**	4.7	4.7
GPN	0.474	**0.521**	**7.5**	5.4
GPS	0.194	**0.402**	**5.3**	3.5
HPN	0.283	**0.494**	**6.0**	4.3
HPS	0.274	**0.386**	**5.0**	3.1
IMN	0.229	**0.398**	**4.6**	4.4
IMS	0.272	**0.327**	**4.8**	3.4
OMV	0.326	**0.539**	**4.0**	3.8
RMN	0.337	**0.437**	**5.2**	3.6
WCN	0.106	**0.323**	3.7	**4.2**
WCS	0.140	**0.315**	2.1	**4.7**

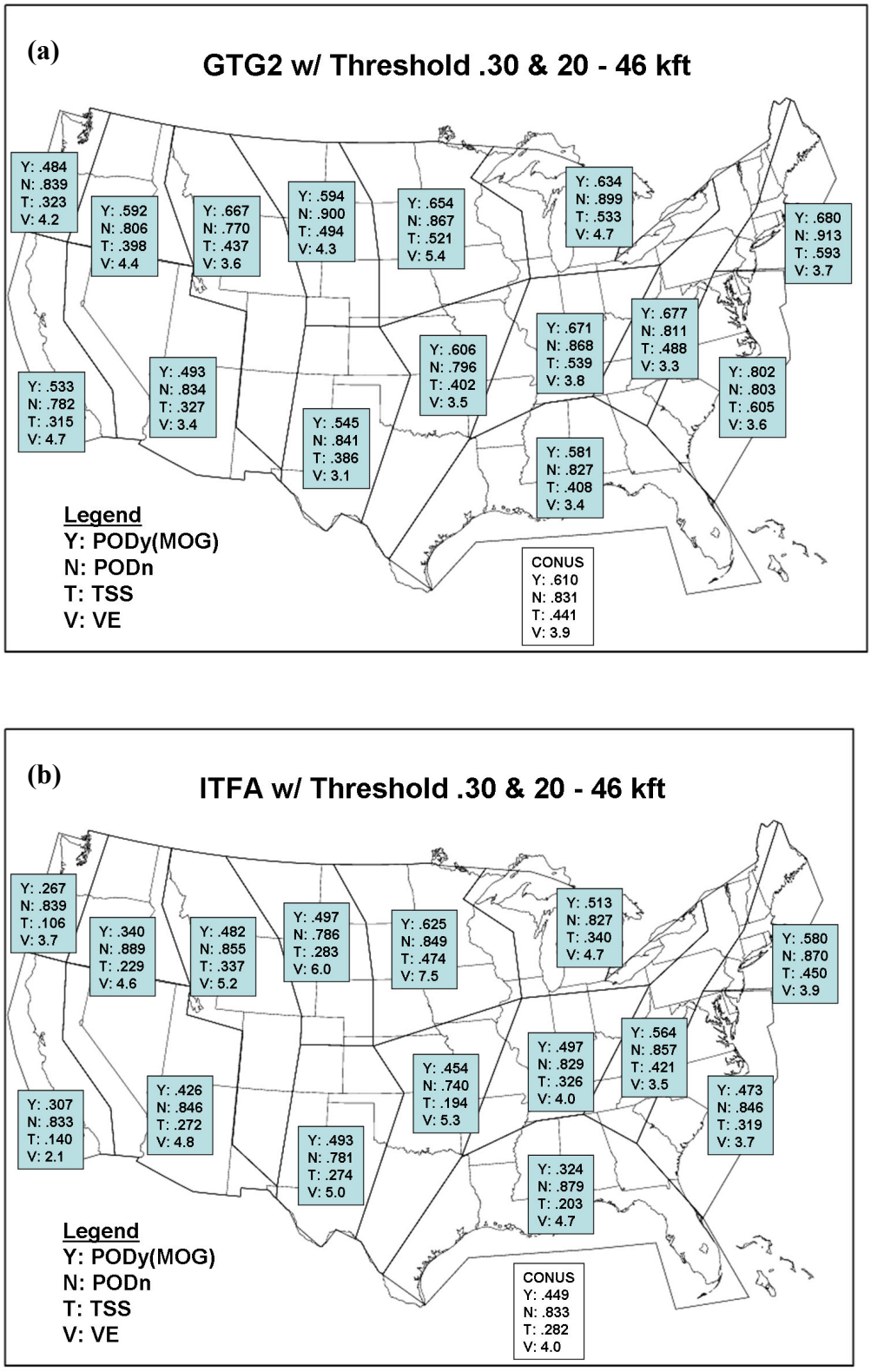


Figure 22 As in Fig. 21 for upper levels.

differences in the TSS values are scattered across the CONUS in the following regions: ECS, GCO, GLA, GPS, HPN, OMV, and WCN. The TSS values for these regions are at least 0.10 larger for GTG2 compared to ITFA. The overall TSS value for the layer, over the CONUS, is substantially larger for GTG2 (0.441) than for ITFA (0.282), and is a larger difference than was found for mid-levels.

Mid- and upper-level GTG2 forecasts for the small regions. Figure 21a shows a map of the GTG2 verification results for the 15 regions for the mid-levels. For this layer, the algorithm performs best in the APP, IMN, and IMS regions. All of these regions had a TSS value greater than 0.50. GTG2 performs least skillfully in the GCO, the HPN, and the OMV regions. Each of these regions had a TSS value less than about 0.20. The poor performance in GCO could be attributed to the nature of the turbulence reports in these regions, where there might be a larger number of convective turbulence reports; these reports would most likely lower the PODy(MOG) values.

Figure 22a shows a map of the GTG2 verification results for the 15 regions for the upper levels (20-46,000 ft). For this layer, the TSS values in the Western regions (WCN, WCS, IMN, and IMS) are not as large (TSS<0.4) as the values for the other regions in the Central and Eastern CONUS. The number of "No" observations is relatively small in some of these regions, and this could explain at least some of the reduction in skill. However, the PODy(MOG) values in these regions are also smaller than the values in the other regions, even with an adequate number of MOG reports. These results are consistent with the large-region results in Fig. 20b.

6.5 Day-to-day variations

It is important to consider day-to-day variations in the skill of the forecast algorithms because such variations may impact the usefulness of the forecasts. Figure 23 shows time series plots of TSS for the mid-level forecasts. The plots in this figure compare TSS values for 6-h GTG2 forecasts to (a) the corresponding AIRMET statistics, and (b) the corresponding ITFA statistics. In general, GTG2 seems to perform somewhat better than the AIRMETs in terms of daily variations of TSS. There are some noticeable drops in the AIRMET TSS values, where the GTG2 performs much better. Another example is the period from 6 March – 17 April when the GTG2 performance showed a marked improvement over the AIRMETs with the exception of brief decreases in skill on 8 and 12 March. In Fig. 23b it appears that GTG2 also has more skill, in terms of TSS values, than ITFA. Figure 24 shows the same basic results for upper levels, when comparing the GTG2 and ITFA daily statistics. However, the AIRMETs and GTG2 perform more similarly to each other at the upper levels (Fig. 24a), with a short period of increased TSS values for GTG2 relative to the AIRMET values in early to mid-April.

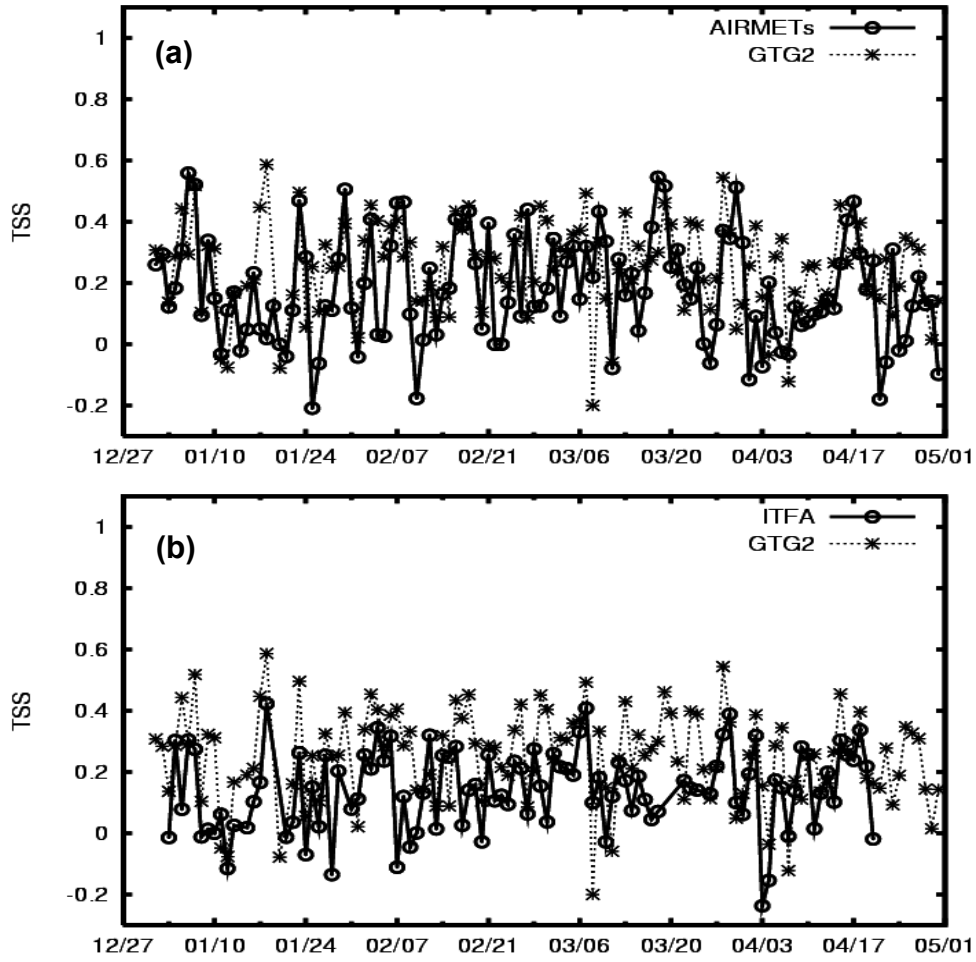


Figure 23. Time series of TSS for 6-h forecasts for mid-levels: (a) GTG2 vs. AIRMETs; (b) GTG2 vs. ITFA.

Day-to-day variations in the verification statistics can also be represented by box plots, which show various quantiles of the distributions. The central box includes the 0.25^{th}, 0.50^{th}, and 0.75^{th} quantiles, and the top and bottom lines ("whiskers") extend to cover the range of the data between the 0.05^{th} and 0.95^{th} quantiles (i.e., percentiles). Figures 25 and 26 show box plots of the distributions of TSS and VE for 6-h forecasts at mid- and upper levels, respectively. Thresholds used to define the Yes/No forecasts are 0.20 for ITFA and 0.25 for GTG (upper levels only) and GTG2. For mid-levels, some interesting differences among the statistics for the various forecasts are apparent in the shapes of the boxes. For example, the TSS box plot for ITFA has a fairly large spread, indicating a large variation in TSS values for ITFA at this altitude. The placement of the box in the TSS boxplot for GTG2 indicates that it performed better than the AIRMETs and ITFA at upper levels (as measured by this particular statistic). The VE values for both GTG2 and ITFA vary less than the VE values for the AIRMETs at mid-levels. The AIRMETs have higher median VE, but also exhibit a larger range of values.

At upper levels (Fig. 26), the TSS distributions have similar spread. However, the TSS values for GTG2 have larger median and quartile values than the TSS values for the other algorithms. In general, the GTG2 VE distribution for upper levels is somewhat below the corresponding distributions for the AIRMETs and GTG, but is similar to the ITFA distribution. The VE distributions for GTG2 and ITFA are also somewhat less variable than the VE distributions for the AIRMETs and GTG.

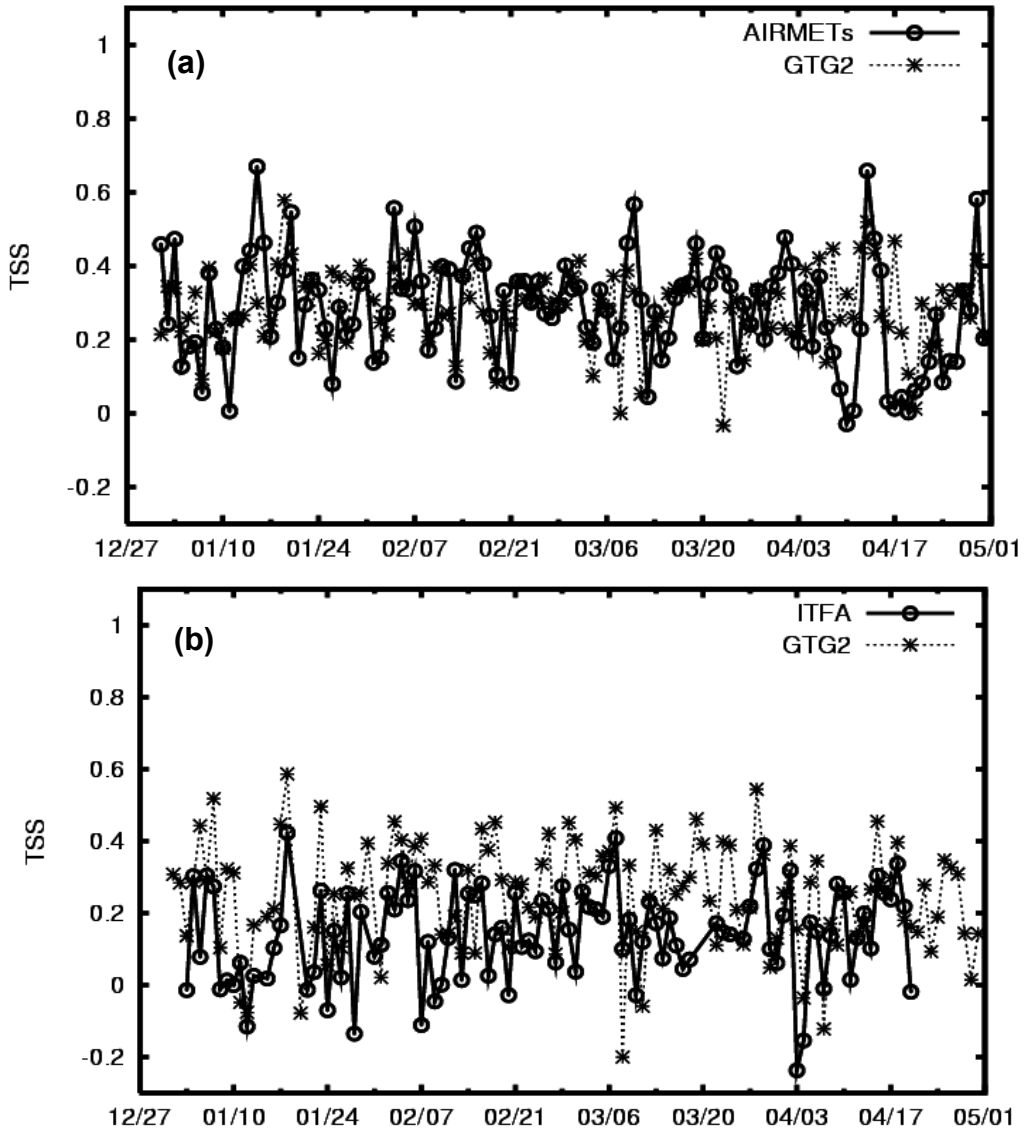


Figure 24. As in Fig. 23 for upper-level forecasts.

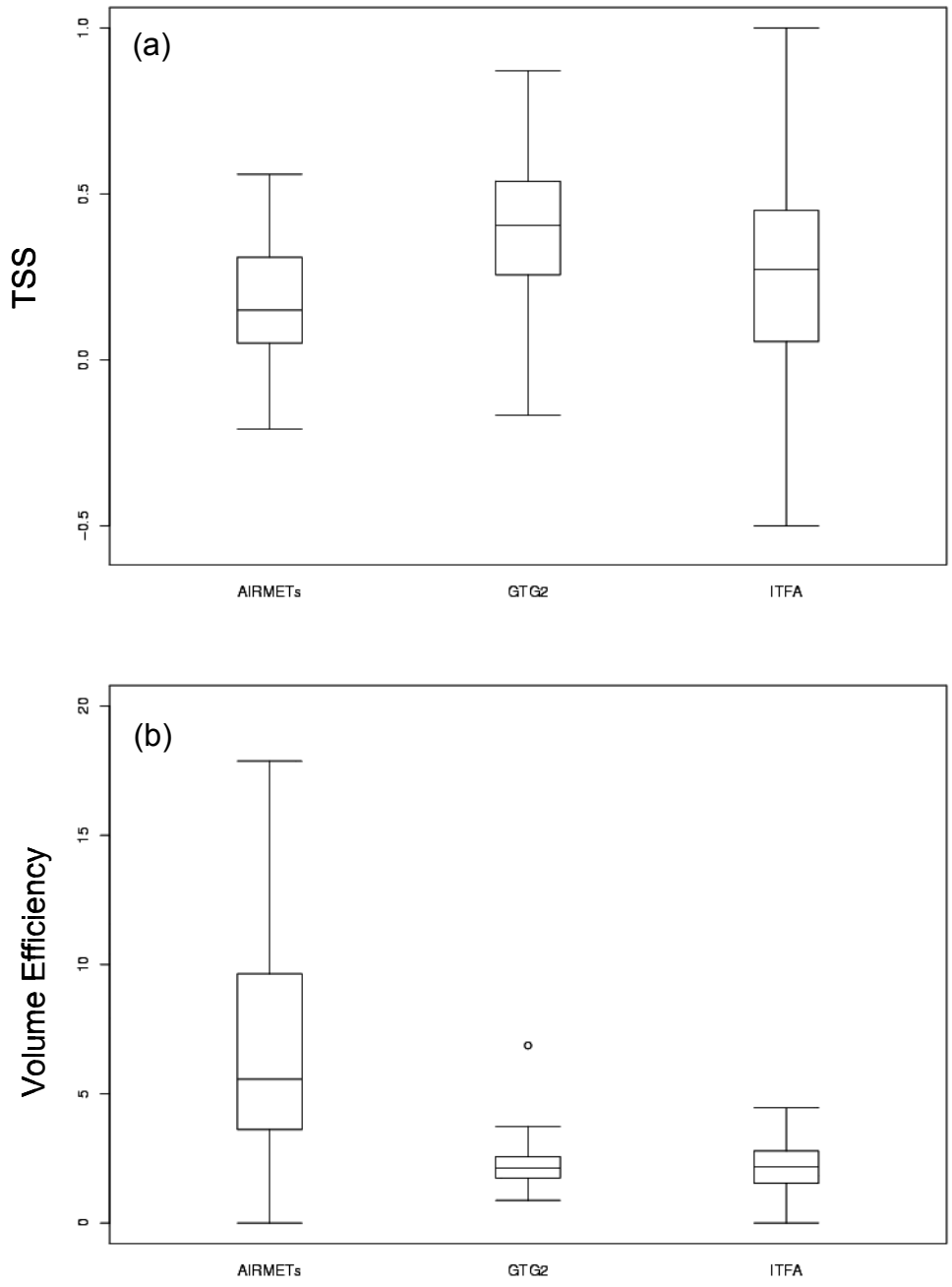


Figure 25. Box plots for the mid-level verification statistics showing the distributions of (a) TSS and (b) VE. GTG2 threshold is 0.25; ITFA threshold is 0.20. See text for explanation of boxplots.

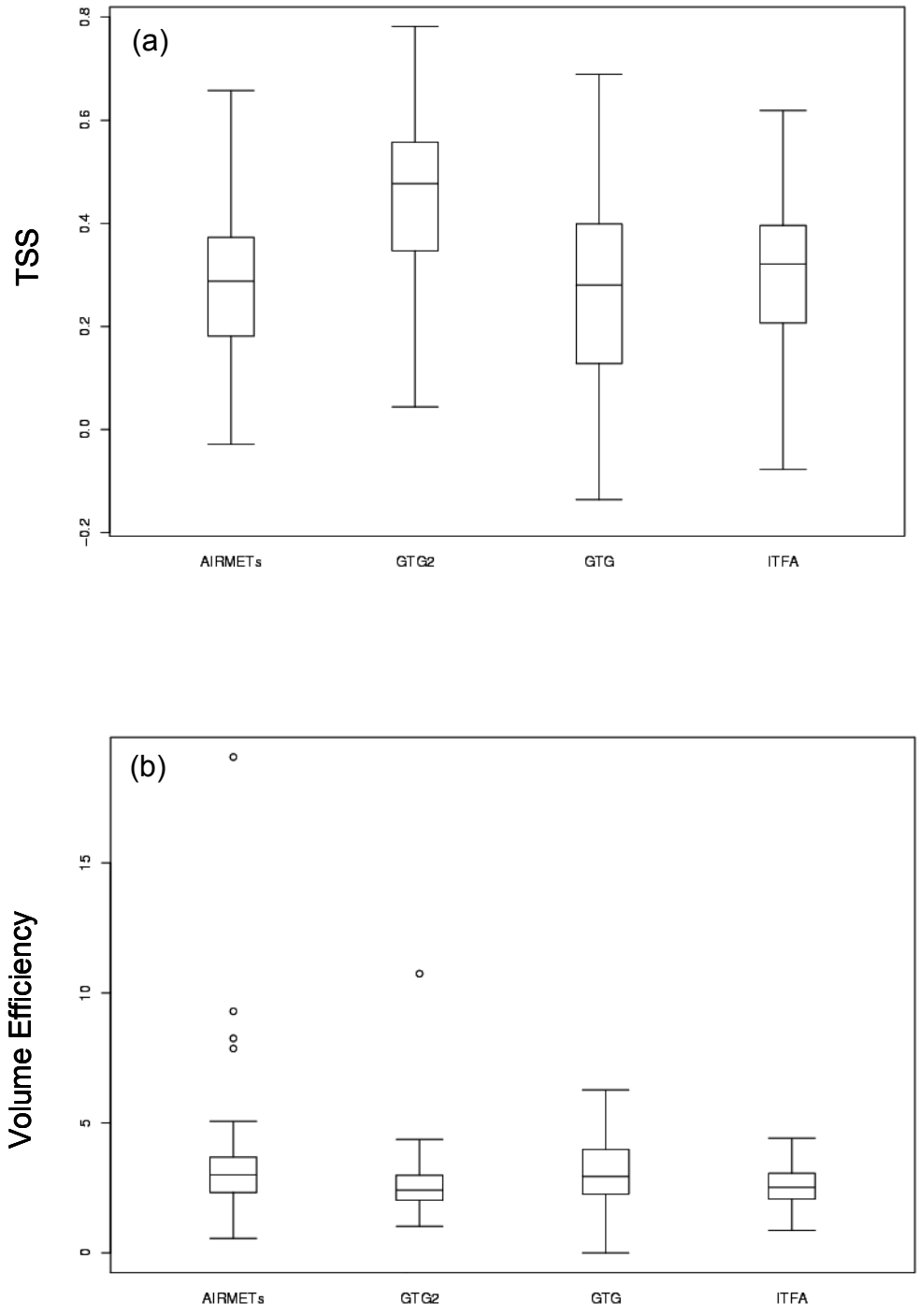


Figure 26. As in Fig. 25 for upper levels and with the addition of GTG (threshold = 0.25).

7. Conclusions and discussion

This report has summarized an evaluation of the mid- and upper-level turbulence forecasts produced by GTG2. The study was based on an intercomparison of GTG2 performance with the performance of other turbulence algorithms, including the operational version of GTG, during winter (January through April) 2004. The results of the study indicate the following conclusions:

- GTG2 forecasts are skillful, as measured by their ability to discriminate between Yes and No PIREPs of turbulence. This result is true for forecasts at both mid- (10-20,000 ft) and upper (above 20,000 ft) level forecasts.

- Comparisons of the performance of GTG2 forecasts to the performance of other algorithms indicate that GTG2's ability to correctly classify Yes and No PIREP observations is significantly better than the capability of other algorithms, including the current operational version of GTG and the previous generation of the algorithm (ITFA).

- GTG2 is quite efficient at limiting the amount of additional airspace covered by a Yes forecast as the value of PODy is increased. At mid-levels GTG's capability in this respect is much better than the capability of other turbulence algorithms (e.g., GTG, Ellrod-1).

- GTG2 forecasts perform consistently at all altitudes at 10,000 ft and above, with only small decreases in performance at the highest altitudes.

- Regional analyses of GTG2 performance indicate that GTG2 performance at mid-levels is best in the Western region and at upper levels GTG2 performance is better in the Central and Eastern regions.

The results described in this report are a small fraction of the verification results that are available. For example, a wide variety of verification information for GTG2, other algorithms, and the AIRMETs is available at the RTVS web site (http://www-ad.fsl.noaa.gov/fvb/rtvs/turb/index.html).

This study demonstrates the strength of the GTG approach in being able to evolve as new approaches are developed for forecasting turbulence and for combining indices. Adjustments to the algorithm, and the addition of mid-level forecasts have significantly improved the performance of the algorithm.

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 the independent verification effort over the last several years. We also thank Jamie Wolff 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. In addition, we would like to express our appreciation to other members of the NCAR and FSL verification teams who provided excellent support for this effort.

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*, Phoenix, AZ, American Meteorological Society (Boston), 249-252.

Brown, B.G., G. Thompson, R.T. Bruintjes, R. Bullock, and T. Kane, 1997: Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *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., 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, B.G., J.L. Mahoney, R. Bullock, J. Henderson, and T.L. Fowler, 2000a: Turbulence Algorithm Intercomparison: 1998-99 Initial Results. NOAA Technical Memorandum OAR FSL-25, 64 pp.

Brown, B.G., J.L. Mahoney, R. Bullock, T.L. Fowler, J. Hart, J. Henderson, and A. Loughe, 2000b: Turbulence Algorithm Intercomparison: Winter 2000 Results. NOAA Technical Memorandum OAR FSL-26, 62 pp.

Brown, B.G., J.L. Mahoney, J. Henderson, T.L. Kane, R. Bullock, and J.E. Hart, 2000c: 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. Bullock, M.B. Chapman, C. Fischer, T.L. Fowler, J.E. Hart, and J.K.Henderson, 2002: Integrated turbulence forecasting algorithm (ITFA): Quality assessment report. Report to the FAA. Available from B.G. Brown, NCAR, P.O. Box 3000, Boulder, CO 80307, 10pp.

Ellrod, G.P. and D.I. Knapp, 1992: An objective clear-air turbulence forecasting technique: verification and operational use. *Weather and 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.

Mahoney, J.L., J.K. Henderson, and P.A. Miller, 1997: A description of the Forecast Systems Laboratory's Real-Time Verification System (RTVS). *Preprints, 7th Conference on Aviation, Range, and Aerospace Meteorology*, Long Beach, CA, American Meteorological Society (Boston), J26-J31.

Mahoney, J.L., B.G. Brown, R. Bullock, C. Fischer, J. Henderson, and B. Sigren, 2001b: Turbulence Algorithm Intercomparison: Winter 2001 Results. Report to the FAA. Available from J.L. Mahoney, FSL. 325 Broadway, Boulder, CO 80303.

Mahoney, J.L, J. K. Henderson, B.G. Brown, J.E. Hart, A. Loughe, C. Fischer, and B. Sigren, 2002: The Real-Time Verification System (RTVS) and its application to aviation weather forecasts. *Preprints, 10th Conference on Aviation, Range, and Aerospace Meteorology*, 13-16 May, Portland, OR, American Meteorological Society (Boston), 323-326.

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

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

Sharman, R., L. Cornman, 1998: An integrated approach to clear-air turbulence prediction. **AIAA 98-0382**. *AIAA 36th Aerospace Sciences Meeting and Exhibit, 10-13 January 1998*, AIAA, Reno, Nevada.

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 (Boston), 68-71.

Sharman, R, B. Brown, and S. Dettling, 2000a: 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.

Sharman, R., G. Wiener, and B. Brown, 2000b: Description and verification of the NCAR Integrated Turbulence Forecasting Algorithm (ITFA). **AIAA 00-0493**. *AIAA 38th Aerospace Sciences Meeting and Exhibit, 10-13 January 2000*, AIAA, Reno, NV.

Sharman, R., C. Tebaldi, J. Wolff, and G. Wiener, 2002a: Results from the NCAR Integrated Turbulence Forecasting Algorithm (ITFA) for predicting upper-level clear-air turbulence. *Preprints, 10th Conference on Aviation, Range, and Aerospace Meteorology*, Portland, OR, 13-16 May, American Meteorological Society (Boston), 351-354.

Sharman, R., J. Wolff, G. Wiener, and C. Tebaldi, 2002b: Technical Description Document for the Integrated Turbulence Forecasting Algorithm (ITFA). Report, submitted to the Federal Aviation Administration Aviation Weather Research Program (FAA/AWRP), available from R. Sharman (sharman@rap.ucar.edu).

Sharman, R., J. Wolff, G. Wiener, and C. Tebaldi, 2004: Technical Description Document for the Graphical Turbulence Guidance Product 2 (GTG2). Report, submitted to the Federal Aviation Administration Aviation Weather Research Program (FAA/AWRP); available from R. Sharman (sharman@ucar.edu).

Tebaldi, C., D. Nychka, B.G. Brown, and R. Sharman, 2002: Flexible discriminant techniques for forecasting clear-air turbulence. *Environmetrics*, in press.

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