Quality Assessment Report:

Graphical Turbulence Guidance (GTG) version 2.3

Michael P. Kay*[1,2], Judy K. Henderson[1], Stacey A. Krieger[1,2], Jennifer L. Mahoney[1],
Lacey D. Holland[3], and Barbara G. Brown[3]

[1]NOAA Earth System Research Laboratory, Boulder, CO
[2]Cooperative Institute for Research in Environmental
Sciences/University of Colorado, Boulder, CO
[3]National Center for Atmospheric Research, Boulder, CO

Quality Assessment Product Development Team

30 August 2006

Table of Contents

List of Figures

List of Tables

Quality Assessment Report:

Graphical Turbulence Guidance (GTG) version 2.3

Michael P. Kay*[12], Judy K. Henderson[1], Stacey A. Krieger[12], Jennifer L. Mahoney[1], Lacey D. Holland[3], and Barbara G. Brown[3]

[1]NOAA Earth System Research Laboratory, Boulder, CO

[2]Cooperative Institute for Research in Environmental Sciences/University of Colorado, Boulder, CO

[3]National Center for Atmospheric Research, Boulder, CO

Quality Assessment Product Development Team

SUMMARY

Results of an evaluation of GTG version 2.3 (GTG2.3) are presented in this report. The algorithm was analyzed from 1 November 2005 through 31 January 2006. Additionally, GTG2.3 was compared to several existing operational turbulence forecasts including GTG version 1.0 (GTG). Forecasts were verified with pilot reports of turbulence.

The primary findings are:

The two versions of GTG studied here (GTG2.3E and GTG2.3P) had nearly identical performance. The introduction of eddy dissipation rate information into the algorithm did not have any adverse effects on the results.

GTG2.3E performed well in the mid- and upper levels for forecasts of moderate or greater turbulence.

GTG2.3E showed limited ability to forecast the correct intensity of turbulence. It performs best for the None and Moderate categories.

When compared to Airman's Meteorological Advisories (AIRMETs), GTG2.3E performed well in both mid- and upper levels. GTG2.3E forecast volumes were much

smaller than the volumes associated with the AIRMETs.

GTG2.3E and Significant Meteorological Advisories (SIGMETs) did a poor job forecasting severe turbulence as was indicated by the statistics that were derived from the limited number of severe turbulence reports. GTG2.3E forecast volumes were several orders of magnitude smaller than those produced by SIGMETs. This result may have been due in part to the small numbers of PIREPs reporting severe turbulence.

1. INTRODUCTION

This report summarizes the quality of mid- and upper-level turbulence forecasts produced by the second generation (version 2.3) Graphical Turbulence Guidance (GTG) product (denoted as GTG2.3), which is under consideration for transition from experimental to operational status within the Aviation Weather Technology Transfer (AWTT) process. Takacs et al. (2004) evaluated the quality of the previous GTG product (version 2.0), which was accepted by the AWTT Board as an experimental product in 2004.

The GTG2.3 algorithm combines input from numerous data sources to provide forecasts of clear-air turbulence over the continental United States (CONUS) at altitudes greater than 10,000ft (Sharman et al. 2004). GTG2.3 has been developed by the Federal Aviation Administration Weather Research Program's (AWRP) Turbulence Product Development Team (TPDT). The AWRP's Quality Assessment Team (QA PDT) evaluated GTG2.3 through specific algorithm comparison studies. The studies were conducted using the Real-Time Verification System (RTVS; Mahoney et al. 2002) developed by staff at the National Oceanic and Atmospheric Administration's Earth System Research Laboratory Global Systems Division.

The report is organized in the following manner. Section 2 provides an overview of the approach taken in evaluating GTG2.3. Section 3 describes the algorithms and forecasts that are assessed in this evaluation. The data used are described next in Section 4. Section 5 presents the verification methods that are employed for the evaluation while results are presented in Section 6. Finally, the report concludes with discussion and a summary of results in Section 7.

2. APPROACH

GTG2.3 was evaluated with respect to other operational turbulence forecasts, which included the operational version of GTG, Airman's Meteorological Advisories (AIRMETs) and Significant Meteorological Advisories (SIGMETs). It should be noted that this report is not intended as an evaluation of the turbulence AIRMETs and SIGMETs. The intercomparison is made in such a way as to treat all forecasts as equitably as possible. More explanation is provided in Section 5. Users of these statistics should keep these assumptions in mind when evaluating the strengths and weaknesses of each type of forecast.

Due to the emphasis placed on forecasting mid- and upper-level turbulence, the evaluation focused on the layers of the atmosphere from 10,000 to 20,000ft, and 20,000 to 40,000ft. In addition to the entire CONUS, forecasting performance across three large and fifteen small geographic subregions was also considered. Forecasts issued during the period 1 November 2005 through 31 January 2006 were analyzed. The verification approach applied in this evaluation is identical to the approach taken in previous studies (e.g., Takacs et al. 2004). Additional analyses that focus on the qualitative trend between the forecast and observed turbulence intensity categories are also included in this report.

3. ALGORITHMS AND FORECASTS

This report is focused on the evaluation of GTG2.3 and its transition to National Weather Service (NWS) operations. The turbulence forecasts used for intercomparison with GTG2.3 in this report represent the current operational guidance available to forecasters. The forecasts considered in this report are:

GTG: This algorithm, formerly known as the Integrated Turbulence Forecast Algorithm (ITFA; Sharman et al. 2002), is intended to forecast moderate or greater (MOG) clear-air turbulence at altitudes from 20,000 to 40,000 ft. GTG forecasts are created by dynamically combining, and optimally weighting, a series of turbulence diagnostics using a fuzzy logic system. The Rapid Update Cycle (RUC; Benjamin 1998) model is used to provide the background fields from which the diagnostics are computed. The GTG is produced operationally by the National Weather Service© Aviation Weather Center (NWS/AWC).

GTG2.3 (versions E and P): The GTG2.3 forecast system represents an incremental improvement to the current experimental version of GTG (version 2.0) and expands the capability of the operational version of GTG (denoted as GTG) by providing turbulence predictions at both mid- (10,000 to 20,000 ft) and upper levels (20,000 ft and above). Additional changes include new diagnostics and the use of the 13-km RUC model for the large-scale atmospheric processes. For information on the performance of the experimental version of GTG (GTG2) and support for its transition to experimental status, see Takacs et al. (2004).

Two versions of GTG2.3 are currently produced by the TPDT: GTG2.3P and GTG2.3E. The configurations of the algorithms are identical except for the turbulence observations used in the forecast tuning, and therefore, the internal weighting of the various diagnostics. GTG2.3P is tuned using pilot reports (PIREPs) whereas GTG2.3E utilizes *in situ* eddy dissipation rate (EDR) measurements that are available from numerous commercial aircraft in addition to PIREPs. EDR data are used to augment many of the shortcomings of PIREP data: they provide high frequency, objective, quantitative observations of turbulence that are independent of aircraft size (Cornman et al. 2004). The EDR observations also lead to increased numbers of ªNoº reports of turbulence. No PIREPs of turbulence are much less frequent than Yes turbulence PIREPs despite the fact that much of the atmosphere is turbulence-free. Understanding the variations in forecast quality that are due to the differences between the two versions of GTG2.3 is important, since situations could occur when the EDR data are unavailable in operations. In these situations, the GTG2.3E version of the algorithm will revert to version GTG2.3P.

AIRMETs: AIRMETs are advisories issued for en-route hazardous weather phenomena (NWS; 2003). In this study, only AIRMET forecasts issued for turbulence were considered. Turbulence AIRMETs are issued for moderate or greater turbulence conditions. Forecasts are issued four times per day for periods up to six hours and may be amended as needed. In this study, only non-amended AIRMETs are considered. The temporal aspect of AIRMETs, as it relates to the intercomparison with GTG2.3, is discussed further in Section 6. Attributes used from these forecasts include the areal extent of the forecast and the vertical layer where turbulence is expected. While AIRMETs provide more detailed information that could potentially aid in the analysis, this information is not encoded in a standard way and therefore cannot be decoded and used systematically in verification studies.

SIGMETs: SIGMETs are in-flight advisories that warn of internationally specified weather phenomena of an intensity and/or extent that concerns pilots and operators of all aircraft (NWS, 2003). SIGMETs can be issued at any time and are valid for up to four hours. In the conterminous United States, SIGMETs have been separated into two types: convective (i.e., thunderstorm-related) and nonconvective. In this study, only nonconvective turbulence SIGMETs are considered. Hereafter, the term SIGMET will be used to represent nonconvective SIGMETs.

4. DATA

Data were collected for analysis from 1 November 2005 - 31 January 2006. A subset of all possible GTG2.3E and GTG2.3P issuance and lead times were used in this study (Table 1). The study focuses on the valid time period between 1500 and 0000 UTC in order to maximize the number of pilot reports available for verification. GTG2.3 forecasts with issuance times of 1800 UTC were not used in this study since they were used by the TPDT to alter weighting parameters within GTG2.3. GTG2.3 algorithms were applied to data from the 13-km RUC model output obtained from the NWS.

Table 1: The set of issue times and lead times for GTG2.3 and operational GTG used in this report.

Issue Time (UTC)	Lead Time (h)
1200	3, 6, 9, 12
1500	3, 6, 9
2100	3

PIREPs of turbulence were used as the observational dataset for this evaluation. PIREPs are subjective, non-systematic reports of aircraft encounters with weather hazards such as turbulence and icing. They can also be issued by pilots when a hazard is expected but none is observed (these are referred to as null PIREPs in this study). PIREPs represent the best available operational source of turbulence observations available today. The attributes of PIREPs that were considered include the report location (latitude, longitude, and altitude) and the intensity of the turbulence encountered. Because a PIREP may be issued for a hazard over a vertical range instead of a single level, they are broken down to create a series of one or more reports for each PIREP. For instance, a PIREP for turbulence between 34,000ft and 37,000ft would be split into a series of four PIREPs having the same latitude, longitude, and time but with altitudes 34,000,35,000, 36,000 and 37,000 ft, respectively. These reports are then used for verification. Throughoutthis report the term PIREP will refer to these post-processed reports instead of the original PIREPs unless otherwise noted. For more information on PIREPs and their characteristics, see Schwartz (1996). No attempt to stratify the PIREPs by origin (i.e., mountain wave, convection, and clear-air) was pursued for this evaluation.

5. METHODS

This section describes the verification methodology and statistics employed to assess GTG2.3. The methodology is similar to past evaluations of turbulence by the QAPDT. More detail and background can be found in Brown and Mahoney (1998). Verification results were obtained from the Real-Time Verification System (RTVS) (Mahoney et al. 2002).

5.1 Creation of forecast/observation pairs

In order to assess the accuracy of GTG2.3, the forecast values must be matched in space and time to the PIREP observations. Because GTG2.3 is a gridded product, and PIREPs are point observations, the product is only assessed at observation locations. The gridded forecast values are bilinearly interpolated to the PIREP positions (latitude, longitude, and altitude) using the four surrounding grid points representing the bounding volume for a PIREP observation. In order to allow for timing differences between reports and forecast valid times, a temporal window of 60 min. is used to collect PIREPs for each forecast valid time. This window also increases the set of observations available for verification.

Table 2: Contingency table for a dichotomous forecast situation.

		Observed	
		Yes	No
Forecast	Yes	YY	YN
	No	NY	NN

Once the forecast/observation pairs have been generated, verification is performed in one of two fundamental ways. The first approach is to treat the forecast dichotomously (i.e., Yes/No) by thresholding the forecast values to derive a 2x2 contingency table (Table 2). The choice of forecast thresholds for both GTG and GTG2.3 and their associated turbulence categories (as used in PIREPs) are discussed later in this section. For all analyses, unless otherwise noted, PIREPs representing moderate or greater intensities are treated as ªYesº observations of turbulence. ªNoº observations are represented by PIREPs with reported intensities less than moderate. Moderate intensity is used as the threshold for Yes and No events because moderate-or-greater (MOG) turbulence represents a greater hazard to aviation than do lesser intensities. Is the most often observed level of turbulence intensity (aside from no turbulence).

Because of the limitations of PIREP data, which do not sample the airspace systematically, not all scalar dichotomous summary statistics can be computed (Brown and Young 2000). The three statistics that will be the focus of this report are the probability of detecting an event (PODy), the probability of detection of a non-event (PODn), and the True Skill Statistic (TSS), which can be represented as PODy + PODn ± 1. TSS is a measure of a forecast©s ability to distinguish between Yes (turbulence) and No (no turbulence) events. Due to the nature of the non-systematic observations, PODy and PODn must not be considered true probabilities but instead as proportions of the observed set of Yes and No PIREPs that are correctly categorized by the forecasts. An additional summary statistic that is used within the report is Percent Volume (or % Volume). This statistic measures the percent of the total air space volume where turbulence is forecast. Possible values range from 0 to 100 %. The value is not itself a measure of accuracy but should instead be used in conjunction with other scores such as PODy to gain greater understanding of the forecast quality as a function of areal forecast coverage. A description of these statistics is given in Table 3.

A sample GTG2.3E forecast obtained from RTVS is presented in Fig. 1 to illustrate the verification mechanics described above[1]. The top panel of the display shows the 6-h lead time forecast issued at 1500 UTC on 01 December 2005. A threshold of 0.375 was used to create a dichotomous forecast situation. The spatial distribution of PIREPs is illustrated by the numbers on the map. The lower panels show the vertical profile of the GTG2.3E forecast at each of the PIREP locations. Within each column, multiple PIREP values may be noted depending on the depth of the layer reported within the original PIREP. Recall that PIREPs are broken into multiple reports each having a vertical depth of 1000 ft. The set of PIREPs and the corresponding forecast values at the PIREP locations make up the cells of the 2x2 contingency table (Table 2) from which the relevant statistics are computed (Table 3).

1 GTG2.3 forecasts on the Experimental Aviation Digital Data Service are available from the following URL: http://www.weather.aero/turbulence/

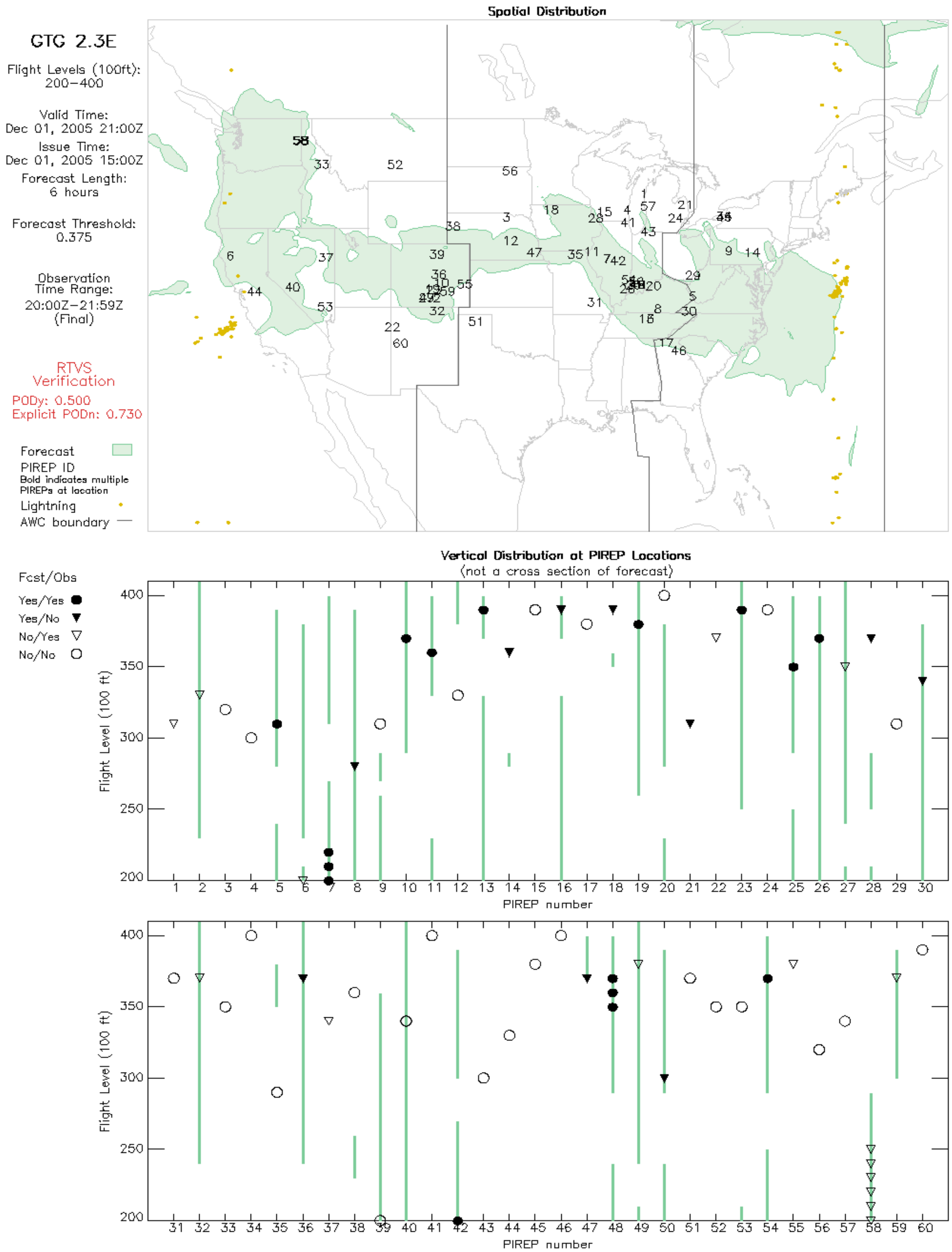

Fig. 1: RTVS display of GTG version 2.3E 6-h forecast on 01 Dec. 2005. Top panel shows plan view of forecast using a threshold of 0.375, observation locations, and lightning data. Bottom panels show vertical distribution of forecasts and observations at all observation locations.

Table 3: Dichotomous summary statistics used in this report. Terms in definitions column are linked to the contingency table presented in Table 2.

Statistic	Description	Definition
PODy	Proportion of events detected correctly	YY/(YY+NY)
PODn	Proportion of non-events detected correctly	NN/(NN+YN)
TSS	True Skill Statistic	PODy+PODn-1
% Volume	Percent of the possible volume covered by the forecast	$100*\text{Volume}_{forecast}/\text{Volume}_{possible}$

The second analysis approach involves the use of signal detection theory and, more specifically, the Relative Operating Characteristic (ROC) diagram (Mason 1982). Rather than choosing a single decision threshold (such as 0.25) from the forecast values to compute the dichotomous statistics, a set of thresholds is chosen and for each threshold the dichotomous statistics PODy and PODn are computed. Each of these pairs of points is then plotted on a diagram called a ROC diagram where the x-axis is 1-PODn and the y-axis is PODy. The line connecting these points is the Relative Operating Characteristic curve. If a forecast shows no ability to distinguish between Yes and No events, the PODy and PODn values will be identical and values will lie on the diagonal of the diagram for all decision thresholds. Forecasts for which PODy exceeds 1-PODn have skill in separating the events from nonevents; for these forecasts the points will lie above the diagonal on the diagram. Perfect forecasts will have points near the upper left-hand corner of the diagram where correct forecasts are maximized and false alarms are minimized. The area under the ROC curve, commonly referred to as the AUC, is used as a summary measure of performance. Possible values for the AUC range from 0 to 1, with values of 0.5 indicating no skill. For the ROC computations, additional thresholds were used to create sufficient data points to resolve the curves. Slight differences in thresholds were necessary for GTG and GTG2.3 owing to differences in thresholds used to define the categories None, Light, Moderate, and Severe. The categories are discussed later in this section. The sets of thresholds used for each algorithm are presented in Table 4.

Table 4: Dichotomous thresholds applied to GTG and GTG2.3 for POD and PODn computations used to resolve ROC curves.

Algorithm	Thresholds
GTG	0.06, 0.125, 0.15, 0.20, 0.25, 0.312, 0.375, 0.437, 0.50, 0.562, 0.625, 0.75
GTG2.3	0.06, 0.125, 0.15, 0.20, 0.25, 0.312, 0.375, 0.475, 0.50, 0.562, 0.625, 0.75, 0.80

The focus for the overall evaluation of GTG2.3 is the accuracy of its predictions of moderate or greater (MOG) intensity turbulence. This aspect of GTG2.3 performance is well captured by the techniques described above. However, when GTG2.3 becomes an operational forecast product and its output is displayed to end users through the operational Aviation Digital Data Service (ADDS), forecasts of specific turbulence intensity categories of None, Light, Moderate, and Severe will be provided. Therefore, it is imperative that the ability of GTG2.3 to predict the correct category of turbulence intensity also be evaluated. An analysis will be performed that focuses on the qualitative trend between the forecast and observed categories. Table 5 shows the mapping between the categorical labels and the associated forecast thresholds that define the lower bound of the range of values tied to each label. For instance, GTG2.3 forecasts of moderate intensity turbulence are associated with forecast values greater than or equal to 0.475 and less than 0.8.

Table 5: Mapping of forecast and observed values to categories used in this report. Values for GTG and GTG2.3 represent lower bounds of the ranges of data associated with each category.

Category	GTG	GTG2.3	PIREP Intensities
None	0.0	0.0	Smooth/None, Smooth to occasional light
Light	0.125	0.3	Light to occasional moderate
Moderate	0.375	0.475	Moderate, Moderate to occasional severe
Severe	0.625	0.8	Severe, Severe to occasional extreme, Extreme

5.2 Stratifications

Data were analyzed over the CONUS and nearby oceanic regions according to the east, west, and central forecast regions used by the AWC (Fig. 2). The national domain is simply the aggregate of the three regions. Additionally, verification results for 15 smaller regions that are based upon differing climatological attributes (Table 6 and Fig. 3) are considered in Section 6 with results presented for each region. For mid- and upper-level results, forecast/observation counts were aggregated vertically through the 10,000 to 20,000 ft and 20,000 to 40,000 ft layers, respectively. For evaluation of the forecast performance across vertical profiles, data were aggregated into 5,000 ft layers.

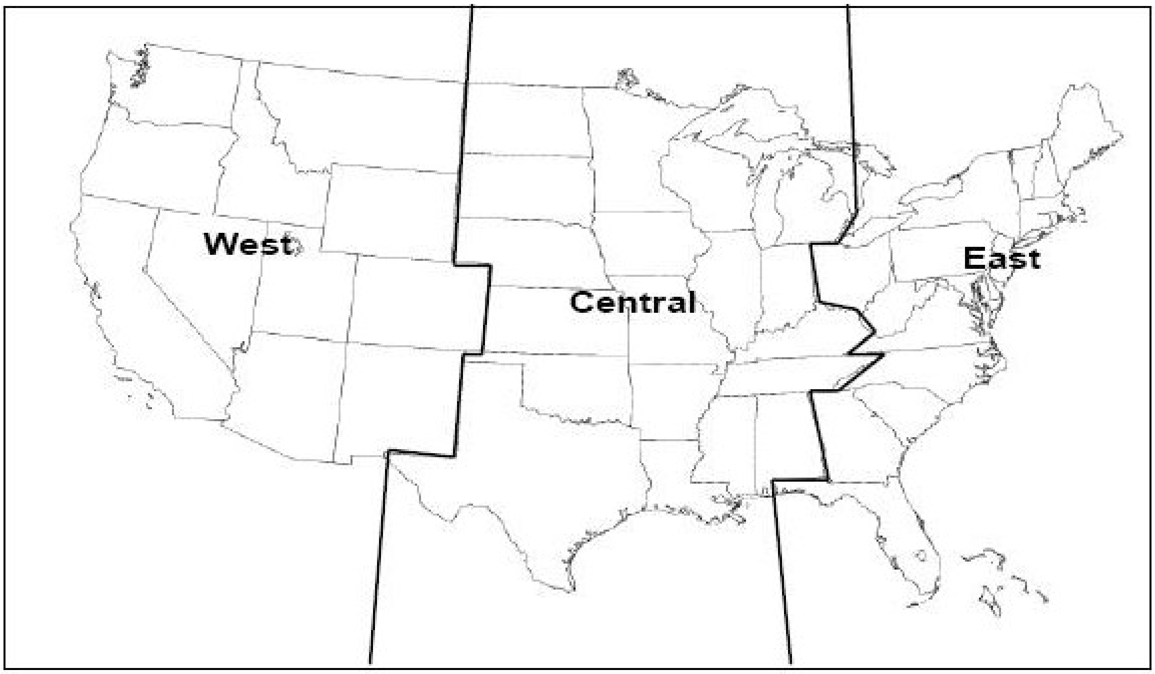

Fig. 2: AWC forecast regions.

Table 6: List of the climatologically-defined regions within the Continental U.S. used in this study and their abbreviations.

Abbreviation	Region
WCN	West Coast North
WCS	West Coast South
IMN	Intermountain North
IMS	Intermountain South
RMN	Rocky Mountain
HPN	High Plains North
HPS	High Plains South
GPN	Great Plains North

Abbreviation	Region
GPS	Great Plains South
GLA	Great Lakes
OMV	Ohio and Mississippi Valley
GCO	Gulf Coast
APP	Appalachians
ECN	East Coast North
ECS	East Coast South

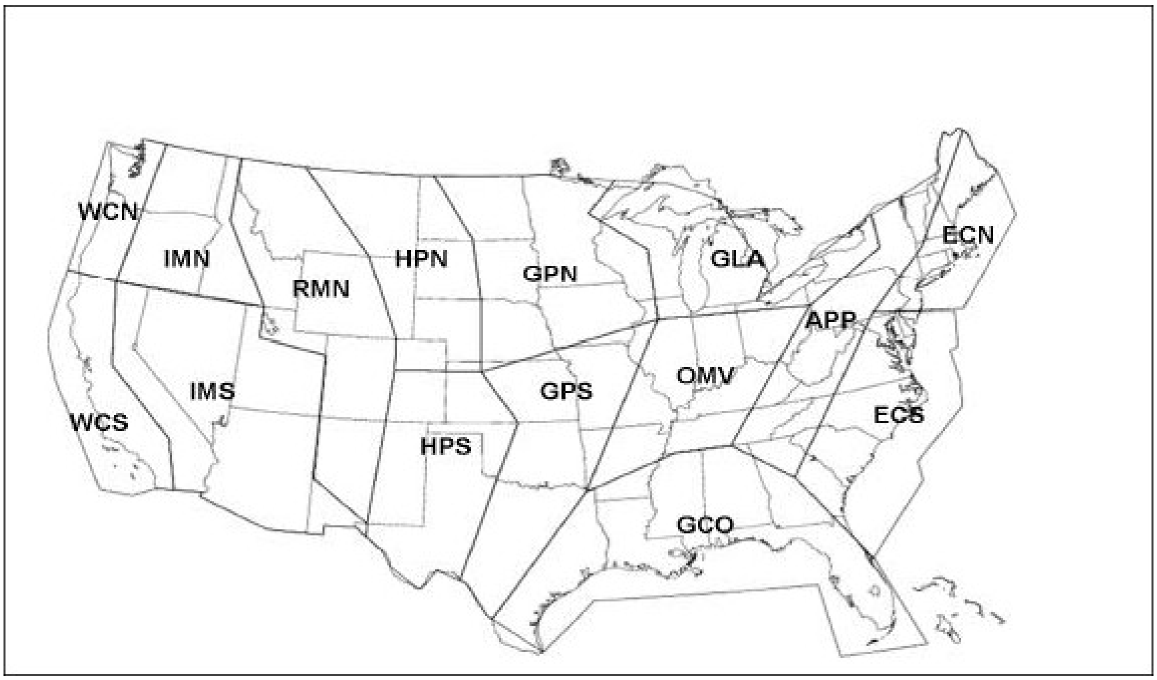

Fig. 3: Climatological regions used for subcontinental characterization of GTG performance.

6. RESULTS

6.1 Comparison of GTG2.3P with GTG2.3E

The purpose of this section is to compare the performance of the two versions of GTG2.3 that are being considered for transition to operations, GTG2.3E and GTG2.3P. The two algorithms are identical, but use differing observations at initialization: GTG2.3P uses PIREPs alone whereas GTG2.3E incorporates EDR measurements in addition to PIREPs. Recall the importance of this evaluation. If EDR observations are unavailable in operations, GTG2.3E will revert to GTG2.3P. This intercomparison should illuminate any differences that may arise owing to the incorporation of EDR data into GTG2.3E.

Overall performance, depicted through ROC curves, is shown in Fig. 4 for the 3-, 6-, 9-, and 12-h lead times for the upper levels (20,000 to 40,000 ft) and in Fig. 5 for midlevels (10,000 to 20,000 ft). Both forecasts show convex curves indicating significant skill at discriminating between Yes and No observations of turbulence throughout the airspace. Minor differences appear between the MOG PODy and 1-MOG PODn values for the two forms of the algorithm, particularly at the lower thresholds, but these differences do not appear to be significant. Areas under the ROC curves are identical for each forecast system at each lead time, with values of 0.87, 0.85, 0.84, and 0.84 for the 3-, 6-, 9-, and 12-h lead times, respectively. The two algorithms appear identical in the overall results for both upper- and midlevels.

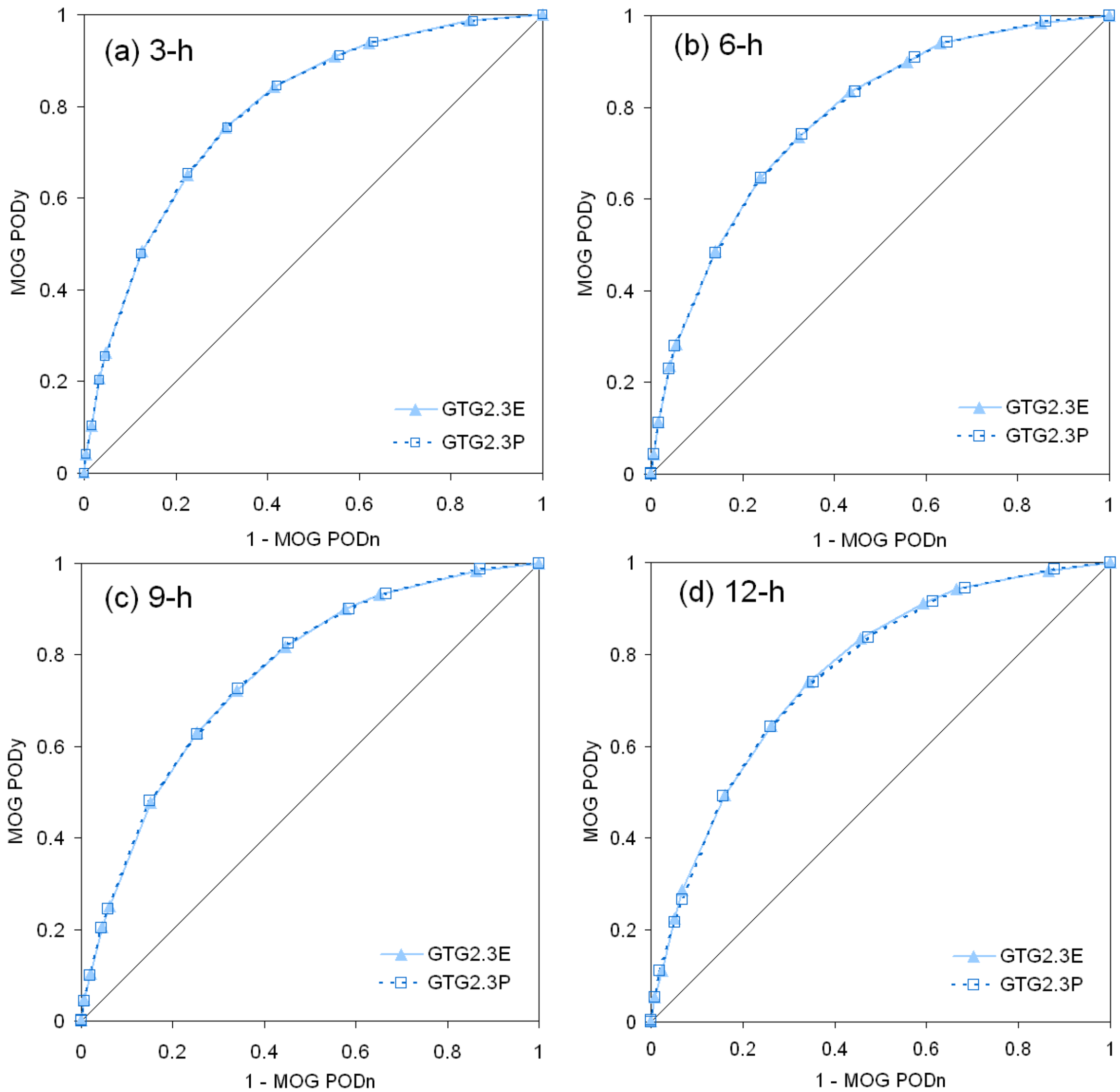

Fig. 4: ROC diagrams for upper levels (20,000 to 40,000 ft) for the (a) 3-h, (b) 6-h, (c) 9-h, and (d) 12-h forecasts for GTG2.3E and GTG2.3P.

Height series of both MOG PODy and MOG PODn, using a threshold of 0.475 and shown in Fig. 6, illustrate good agreement between the two algorithms from 10,000 to 40,000 ft. Minor differences are apparent in the layer from 20,000 ft to 35,000 ft. GTG2.3E has slightly larger MOG PODy values while having slightly smaller MOG PODn values than GTG2.3P.

The results presented above suggest that the introduction of EDR measurements into GTG2.3 does not decrease the skill of the algorithm, nor do they appear to significantly enhance the algorithm. This may be due to the fact that EDR measurements are still not particularly widespread. Moreover, the *in situ* EDR observations were not used in the verification analyses; inclusion of these observations likely would have some impacts on the verification results. Given the known benefits of EDR data at detecting and

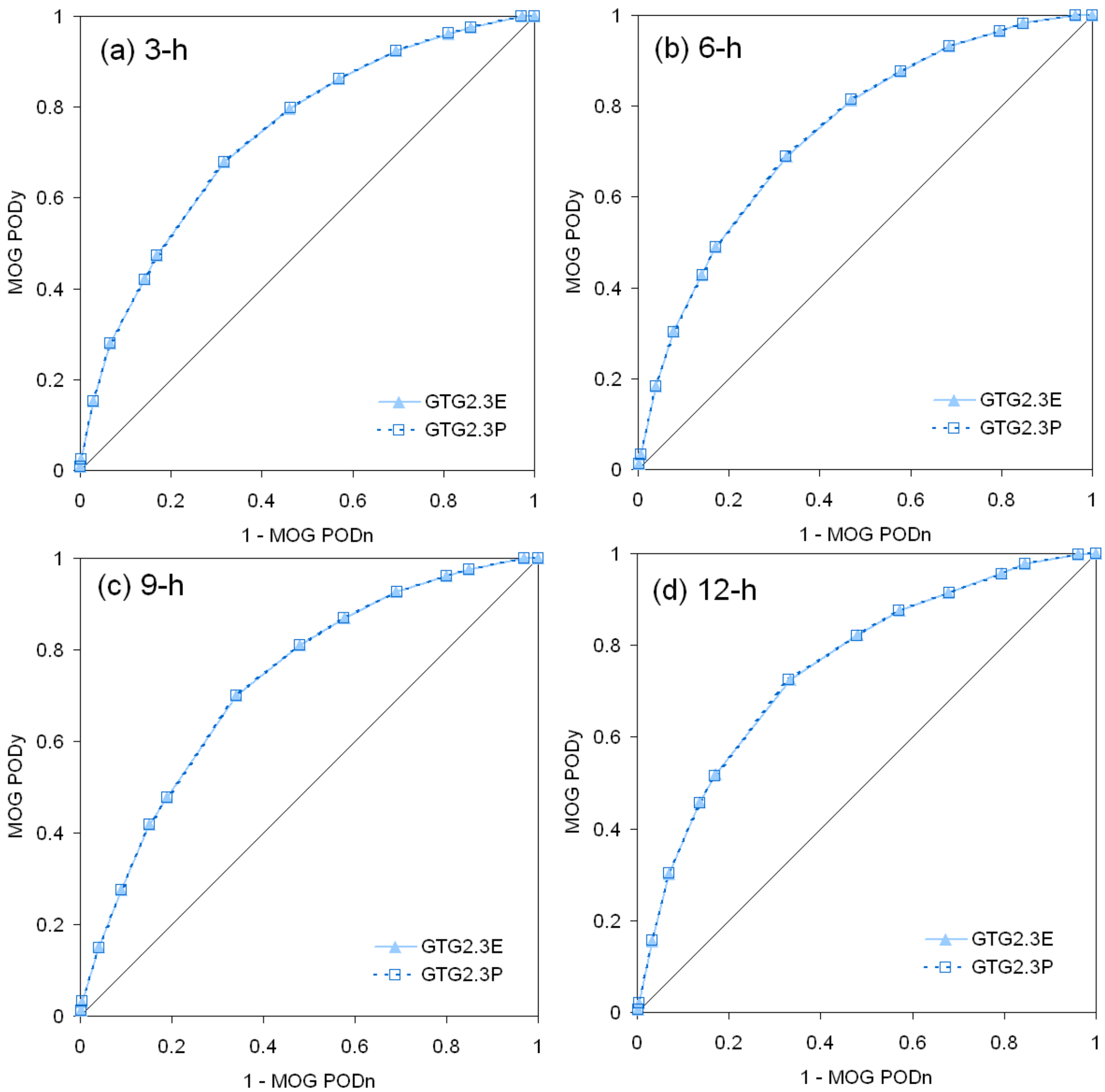

Fig. 5: As in Fig. 4 except for midlevels (10,000 to 20,000 ft).

quantifying turbulence in the free atmosphere, the rest of this report will focus solely on the quality of GTG2.3E.

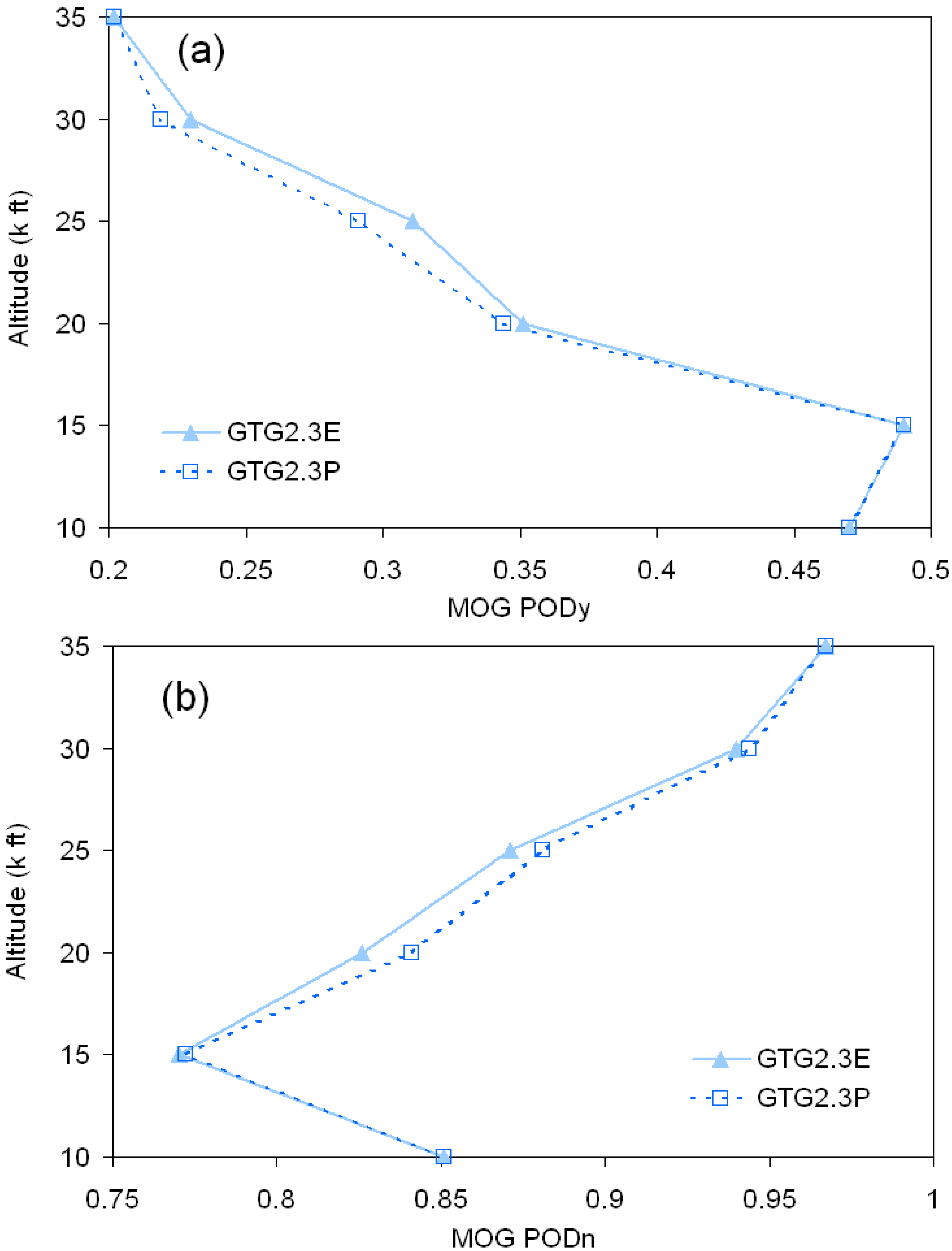

Fig. 6: Height series of (a) MOG PODy and (b) MOG PODn for GTG2.3E and GTG2.3P. Data are plotted at the bottom of each 5,000 ft vertical layer. Forecast threshold is 0.475.

6.2 GTG2.3E performance

The performance of GTG2.3E is presented in this section. Results are summarized by forecast lead time, height, and region.

GTG2.3E performance at upper levels (20,000 to 40,000 ft) is very similar for all lead times (Fig. 7). The 3-h forecasts perform best while the 9-h and 12-h forecasts perform slightly worse than the 6-h forecasts. Areas under the curves for the 3-, 6-, 9-, and 12-h forecasts are 0.789, 0.776, 0.757, and 0.761, respectively, indicating positive skill at all lead times. An identical pattern of performance is seen when one considers how probability of detection varies along with the volume of the airspace where turbulence is forecast (Fig. 8). For lower thresholds, such as the interval between 0.125 through 0.25, the forecasts at all lead times achieve roughly comparable MOG PODy values along with decreased volumes of impacted airspace.

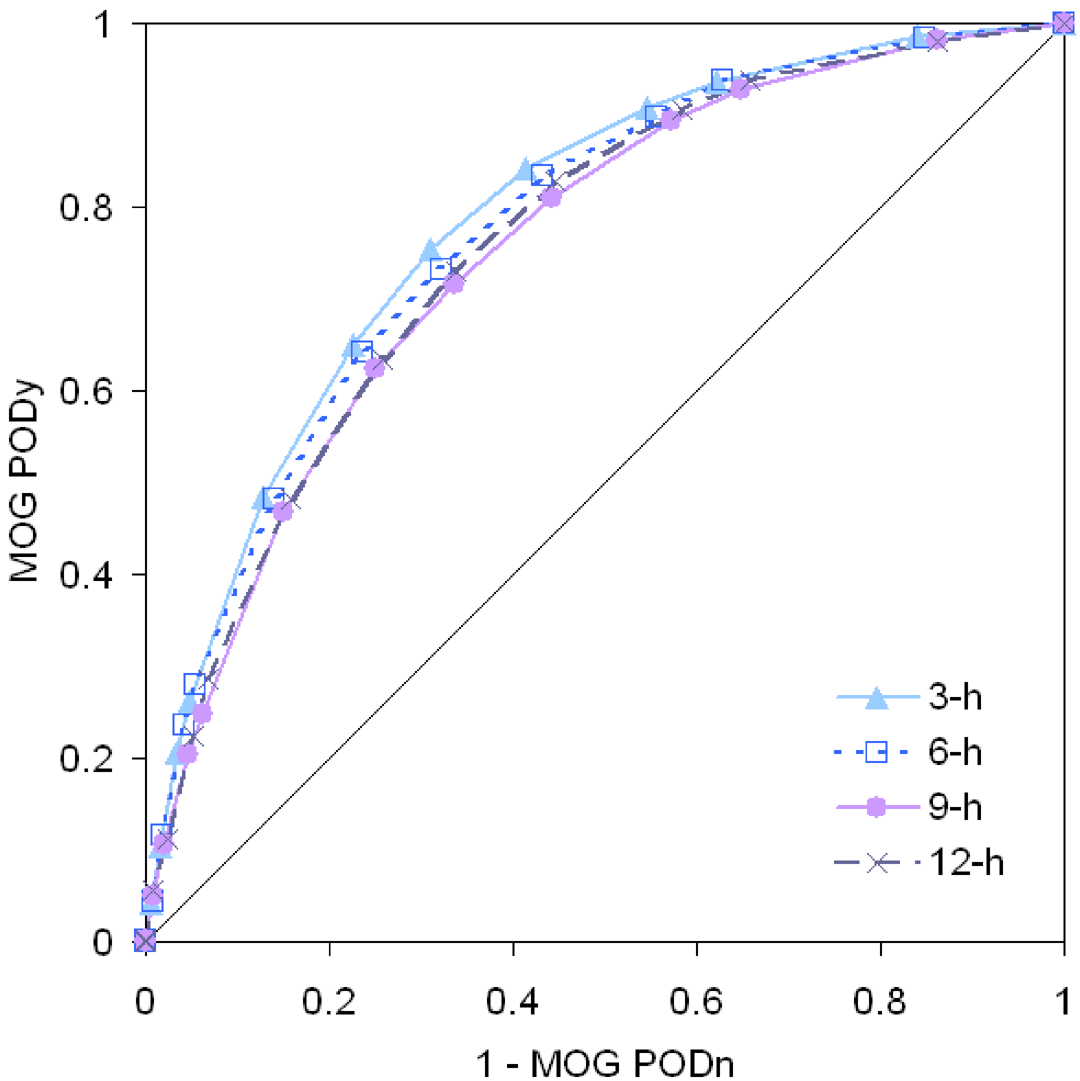

Fig. 7: ROC diagram for GTG2.3E for the 3-, 6-, 9-, and 12-h lead times for upper levels (20,000 to 40,000 ft).

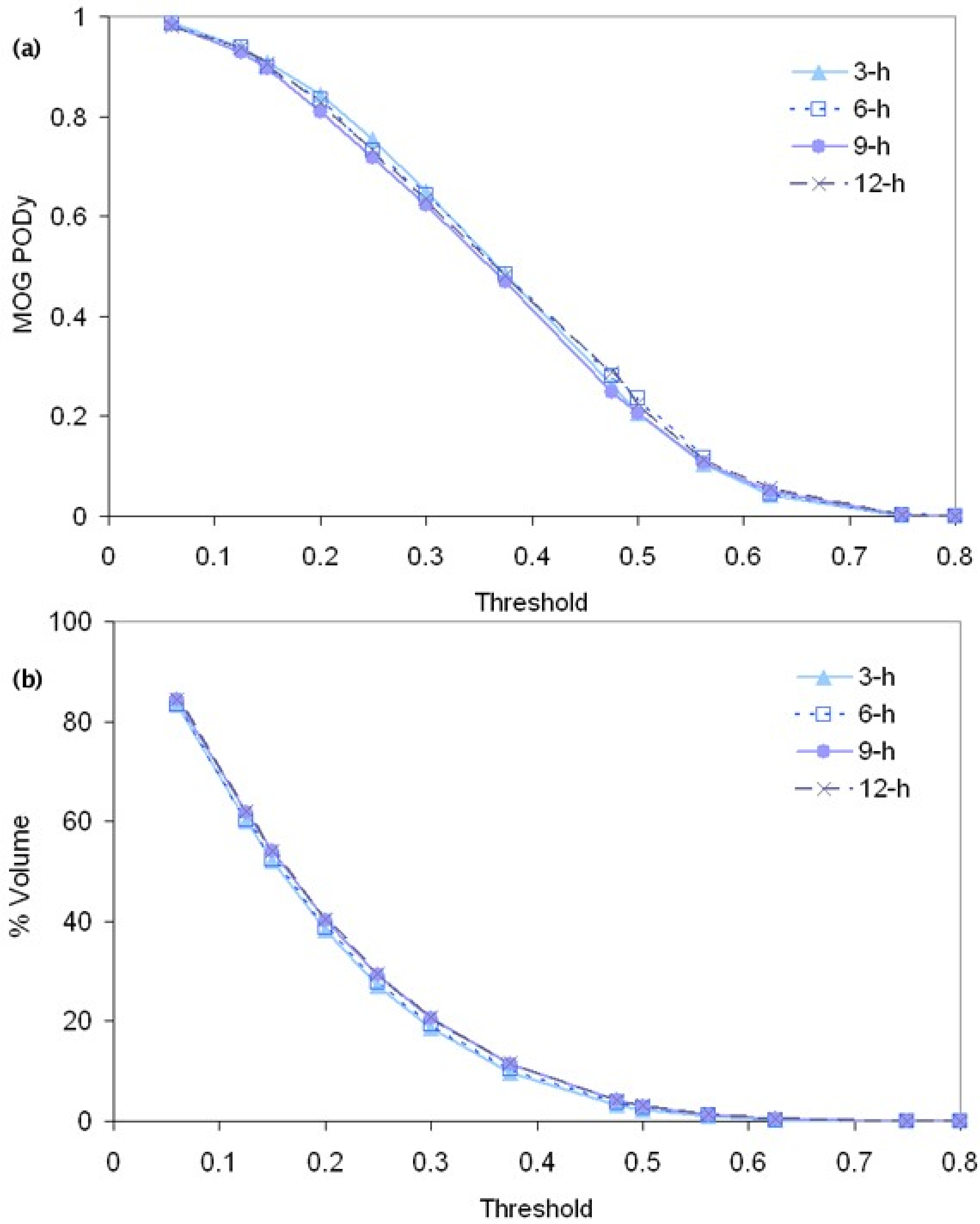

Fig. 8: Threshold plots of (a) MOG PODy and (b) % Volume for GTG2.3E at upper levels for the 3-, 6-, 9- and 12-h lead times.

The ROC diagram for midlevel (10,000 to 20,000 ft) forecasts from GTG2.3E is shown in Fig. 9. These results vary somewhat from those shown for upper levels (Fig. 7). The

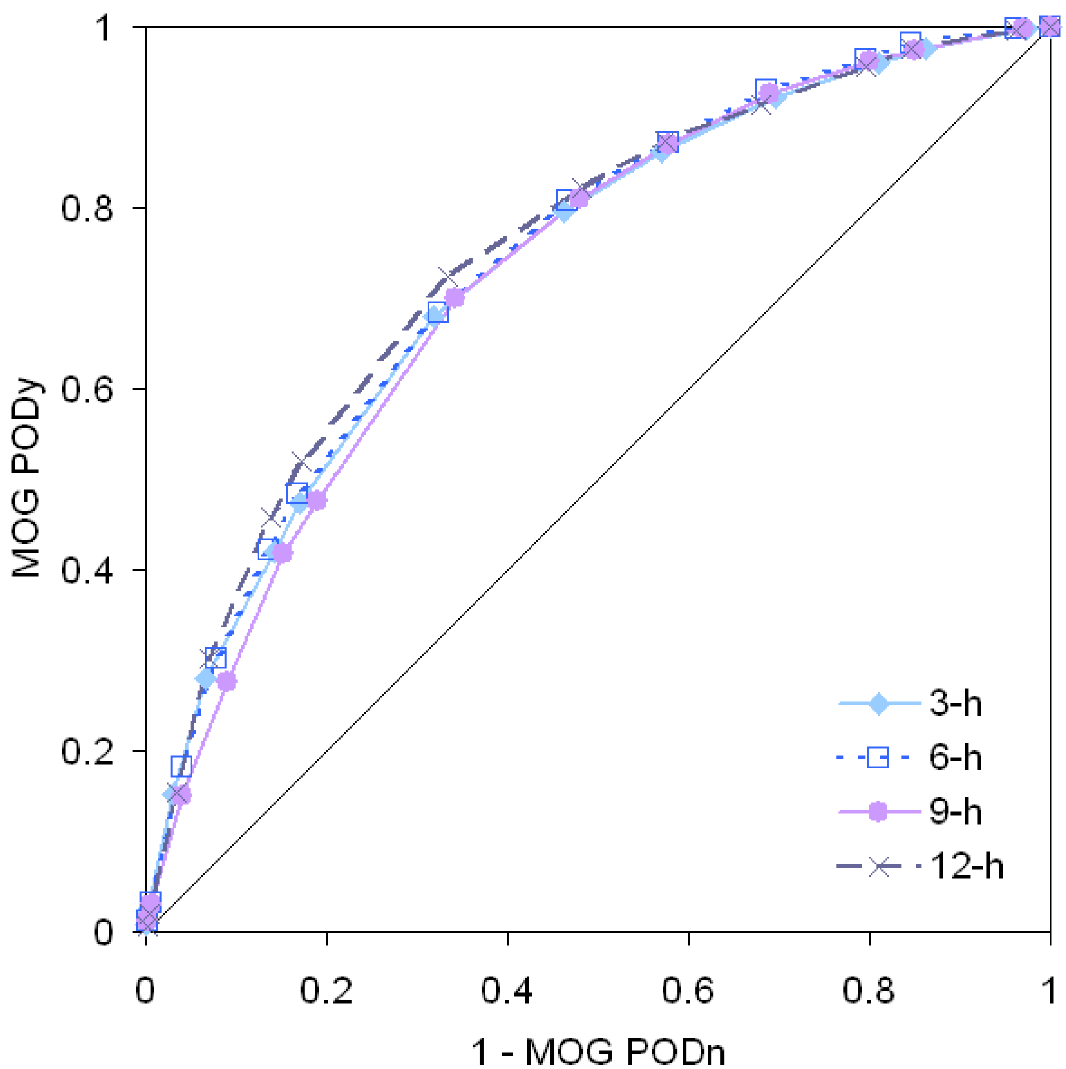

Fig. 9: As in Fig. 7 except for midlevels (10,000 to 20,000 ft).

AUCs are 0.738, 0.744, 0.730, and 0.752 for the 3-, 6-, 9-, and 12-h forecasts, respectively. These values are slightly smaller than those found for the upper level forecasts. This difference may be an indication that the midlevel forecasts from GTG2.3E are not quite as mature as those from the upper levels where the algorithm was initially developed. In addition, fewer PIREPs and EDR observations are typically found at midlevels, which may have an impact on the algorithm performance since these observations are one of the important inputs to GTG2.3E. The 12-h forecasts are the most skillful forecasts at midlevels according to the AUC calculations. MOG PODy and % Volume statistics as a function of threshold (not shown) are similar to the results found for upper levels (Fig. 8).

The performance of GTG2.3E at different vertical levels is somewhat variable (Fig.

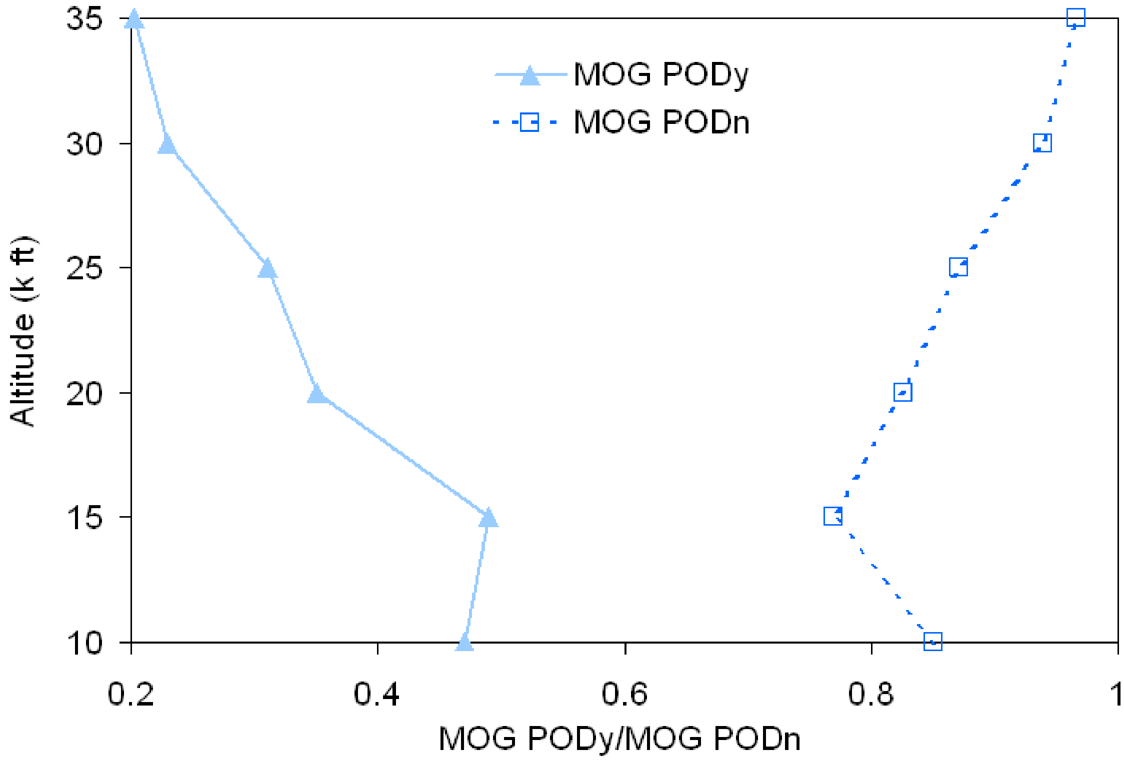

Fig. 10: MOG PODy and MOG PODn height series for GTG2.3E. Forecast threshold is 0.475.

10). MOG PODy values are maximized in the 15,000to 20,000-ft layer. Approximately 51% of all MOG PIREPs are captured correctly in that layer with slightly more than 20% correct in the 35,000 to 40,000-ft layer (Fig. 10). MOG PODn values change in an opposite sense as PODy with the lowest value occurring for the 15,000-20,000 ft layer and a maximum of 97% of PIREPs with intensities less than moderate having correct forecasts in the 35,000to 40,000-ft layer. The vertical profile of MOG PODn is also less variable than the MOG PODy profile.

Vertical profiles of MOG PODy and MOG PODn as a function of lead time are shown in Figs. 11 and 12, respectively. Few differences are seen in the MOG PODy values among the 3-, 6-, and the 9-h forecasts from GTG2.3E. However, the 12-h forecasts show a statistically significant departure from the other lead times for the 25,000 to 30,000-ft layer. MOG PODn values cluster tightly with several minor differences between the profiles. The only statistically significant differences are the 3- and 9-h values for the 15,000 to 20,000-f layer and the 3- and 12-h values for the 20,000 to 25,000-ft layer. No statistically significant differences exist for any other MOG PODn values given the number of PIREPs available during the evaluation period.

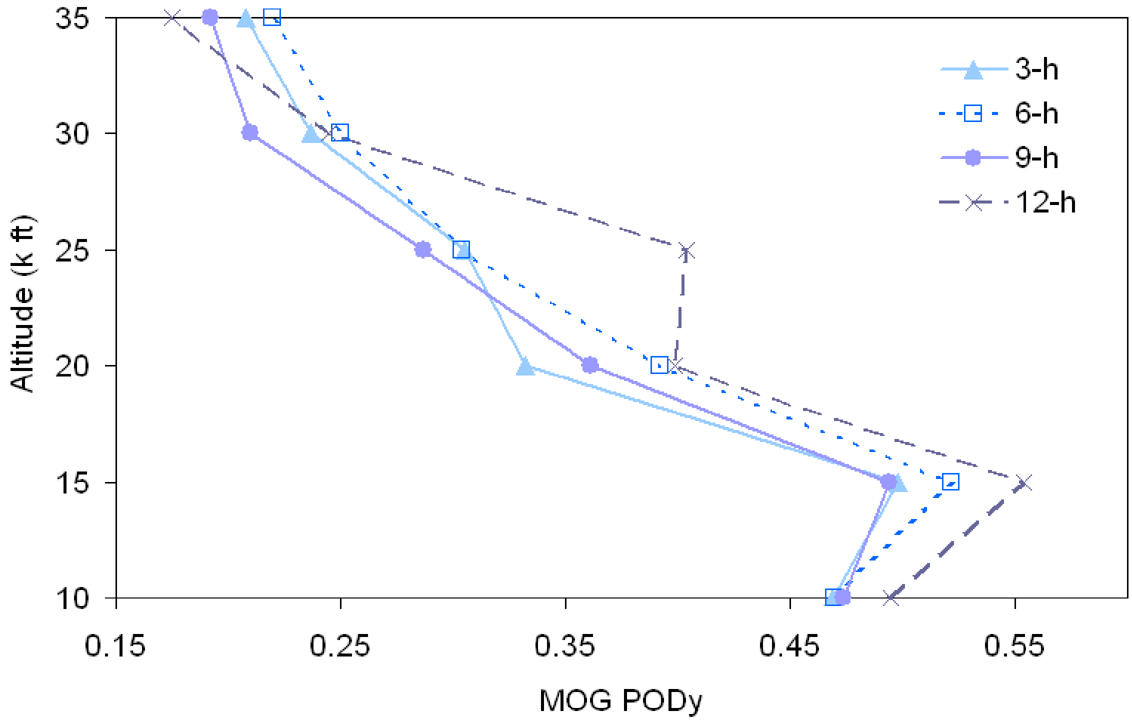

Fig. 11: MOG PODy height series for GTG2.3E for 3-, 6-, 9-, and 12-h lead times. Forecast threshold is 0.475.

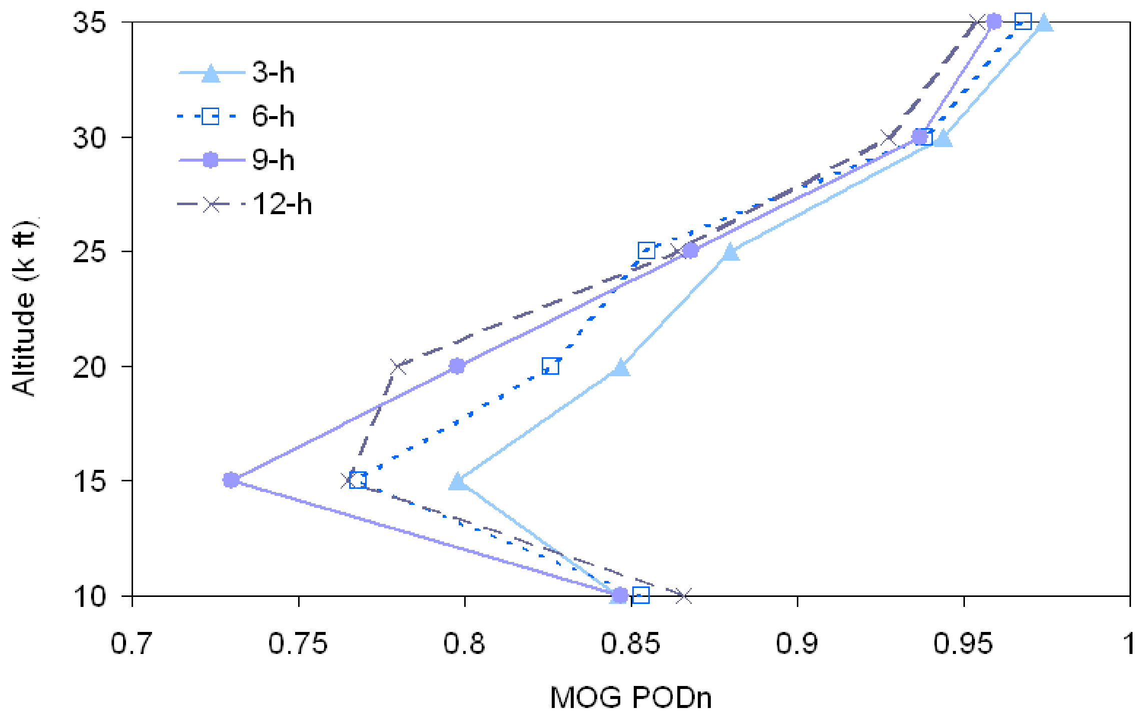

Fig. 12: As in Fig. 11 except for MOG PODn.

Regional analysis is important as GTG2.3 is meant to be used as guidance to support forecasting efforts, such as those performed at the AWC where regional forecasting is required. Additionally, the varied topography of the CONUS is such that terrain-induced turbulence may cause regional differences in the performance of GTG2.3E to become apparent. This knowledge is beneficial to forecasters as well as forecast users.

At upper levels, GTG2.3E performs the worst in the western U.S. while performing best in the central U.S. (Fig. 13). AUC values for the West, Central, and East regions are 0.724, 0.810, and 0.792, respectively. It is possible that mountain wave activity is influencing the results in the West region. Sharman et al. (2006) note that GTG attempts only to forecast clear-air turbulence. Even though some of the diagnostics used to create the GTG product may capture mountain wave conditions, the current algorithm is not expected to forecast them well. Additionally, the PIREP dataset is not filtered to remove any PIREPs that may be due to mountain waves.

For the midlevels, overall performance is degraded somewhat from that found for upper levels in the East and Central regions (Fig. 14). In the West region, performance is slightly better. AUC values are 0.763, 0.729, and 0.699 for the West, Central, and East regions, respectively. The West region has noticeably larger MOG PODy values compared to the East and Central regions. The larger MOG PODy appears to be directly related to the fact that MOG turbulence was forecast over a much larger volume in the West (Fig. 15) than in the other regions.

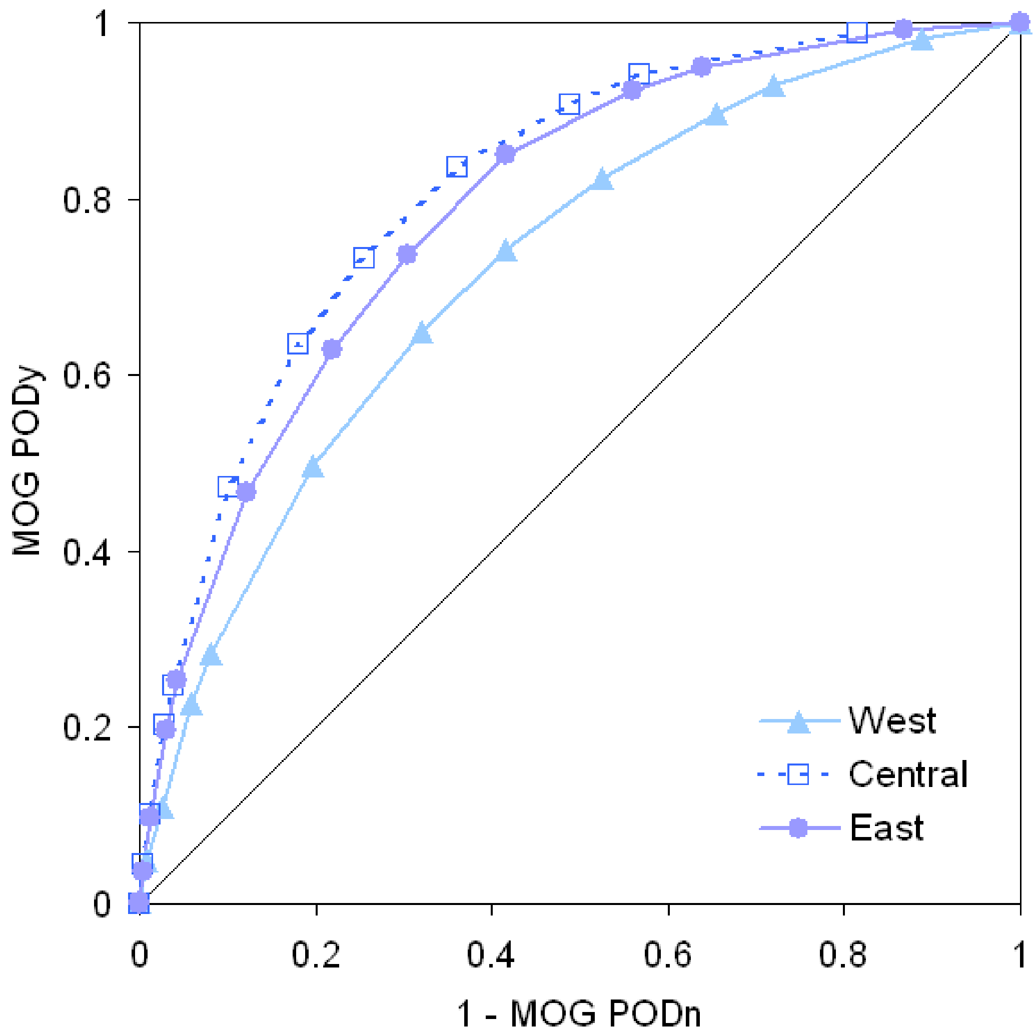

Fig. 13: ROC diagram for GTG2.3E stratified by AWC forecast region for upper levels (20,000 to 40,000 ft).

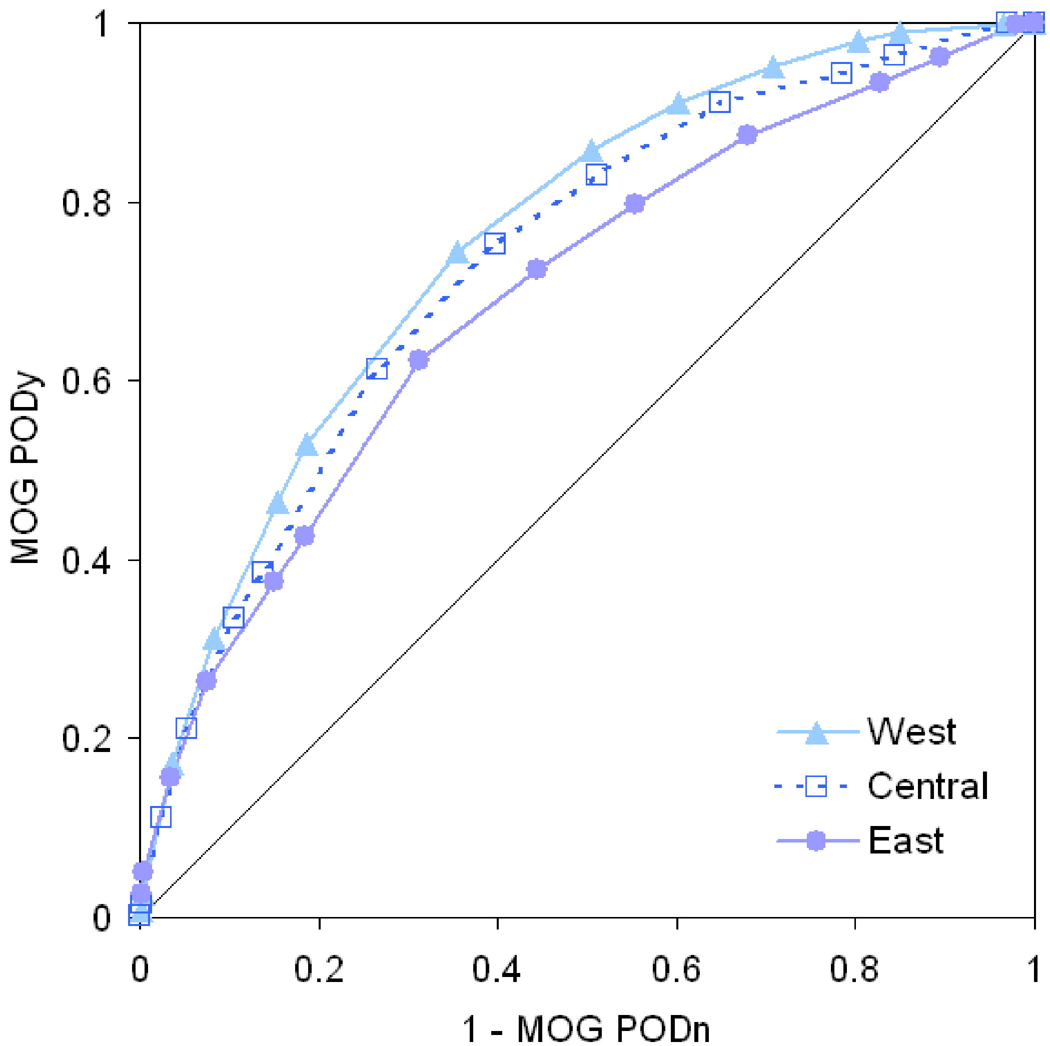

Fig. 14: As in Fig. 13 except for midlevels (10,000 to 20,000 ft).

GTG2.3E performance in each of the climatological regions (Fig. 3; Table 6) are discussed next. For this analysis, all issue times and lead times are combined to increase sample sizes within each region. Results are presented for moderate or greater turbulence only (GTG2.3E threshold of 0.475).

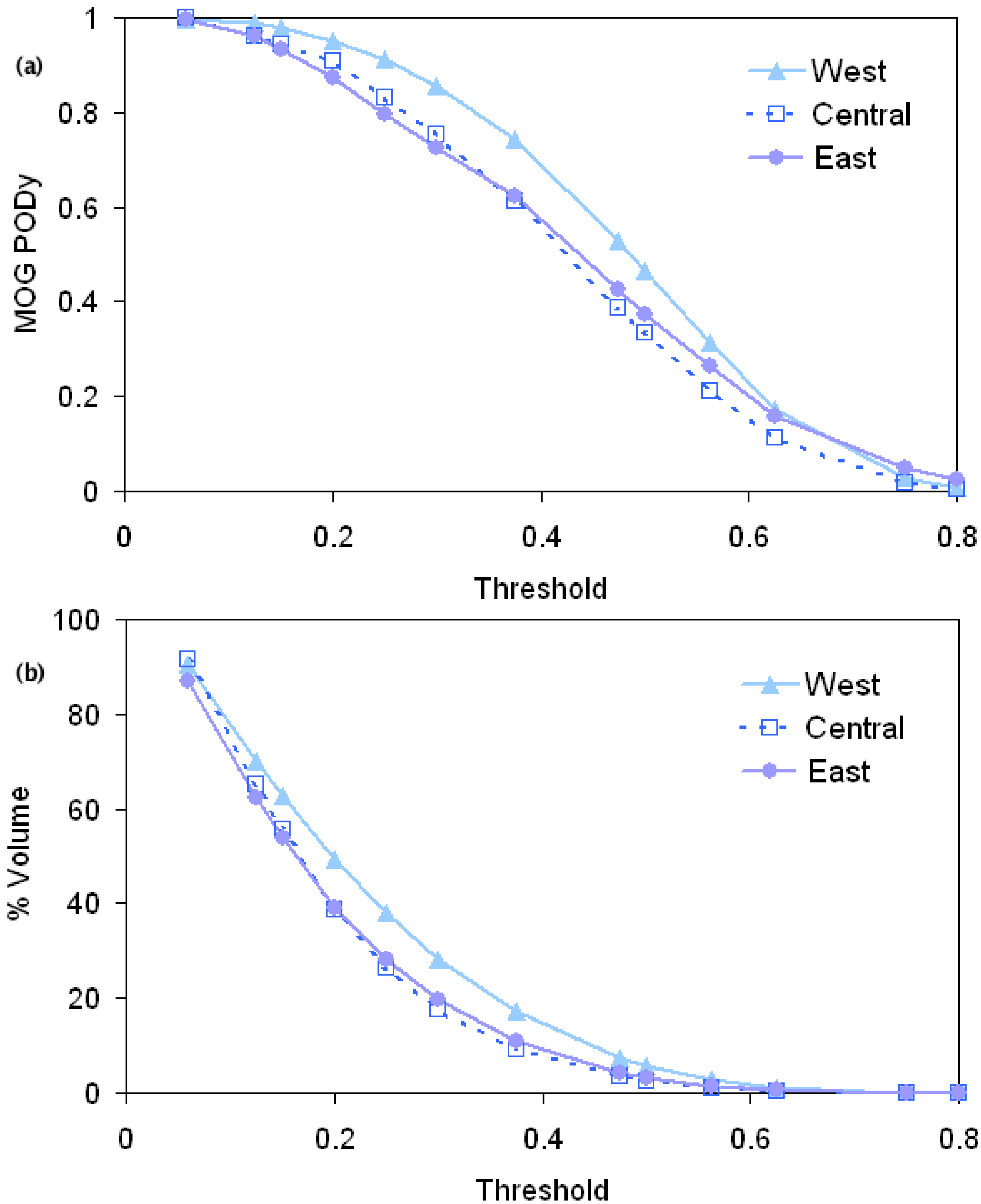

Fig. 15: Threshold plots of (a) MOG PODy and (b) % Volume for GTG2.3E at midlevels for the stratified by AWC forecast region.

For upper levels, GTG2.3E has the largest MOG PODy in the Rocky Mountains North (RMN) region (MOG PODy of 0.405; Fig. 16a). The RMN region is also associated with the largest percent volume of any region (5.8%; nearly double the largest value of any other region). The MOG TSS values are less than 0.300 for most regions (Fig. 16d).

The statistics for the WCN, WCS, GCO, and APP regions indicate that GTG2.3E has relatively little skill in these regions compared to other regions. The algorithm has the greatest skill in distinguishing between MOG PIREPs and PIREPs of lesser intensities in the GLA and RMN regions (MOG TSS values of 0.327 and 0.306, respectively).

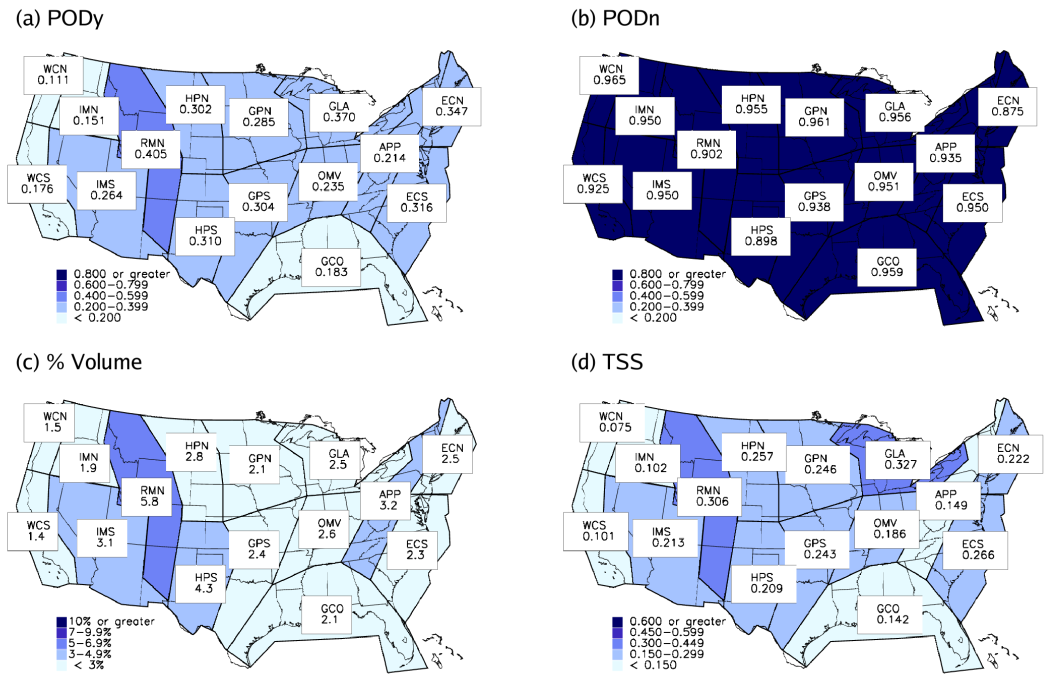

Fig. 16: Overall (a) PODy, (b) PODn, (c) % Volume, and (d) TSS values for GTG2.3E at upper levels for the climatological regions. Forecast threshold is 0.475.

For the midlevels, the performance in the climatological regions is much more variable than at upper levels (Fig. 17). While some regions, such as HPN, have a large PODy value for MOG turbulence (MOG PODy of 0.873), other regions like GPN have an extremely small MOG PODy value (0.130). Furthermore, the MOG PODy value in the GPN region is quite different from the values for adjacent regions. The largest TSS value for GTG2.3E occurs in the HPS region (Fig. 17d); the algorithm also has relatively good skill in the RMN region. The MOG TSS values for most regions are greater than 0.200, but the TSS score for the GPN region is negative, indicating negative skill for this region. The negative skill is due to the poor MOG PODy value here (0.130). GTG2.3E performs well over the mountains and high plains, while having smaller TSS values over the coastal areas (except for the ECS region).

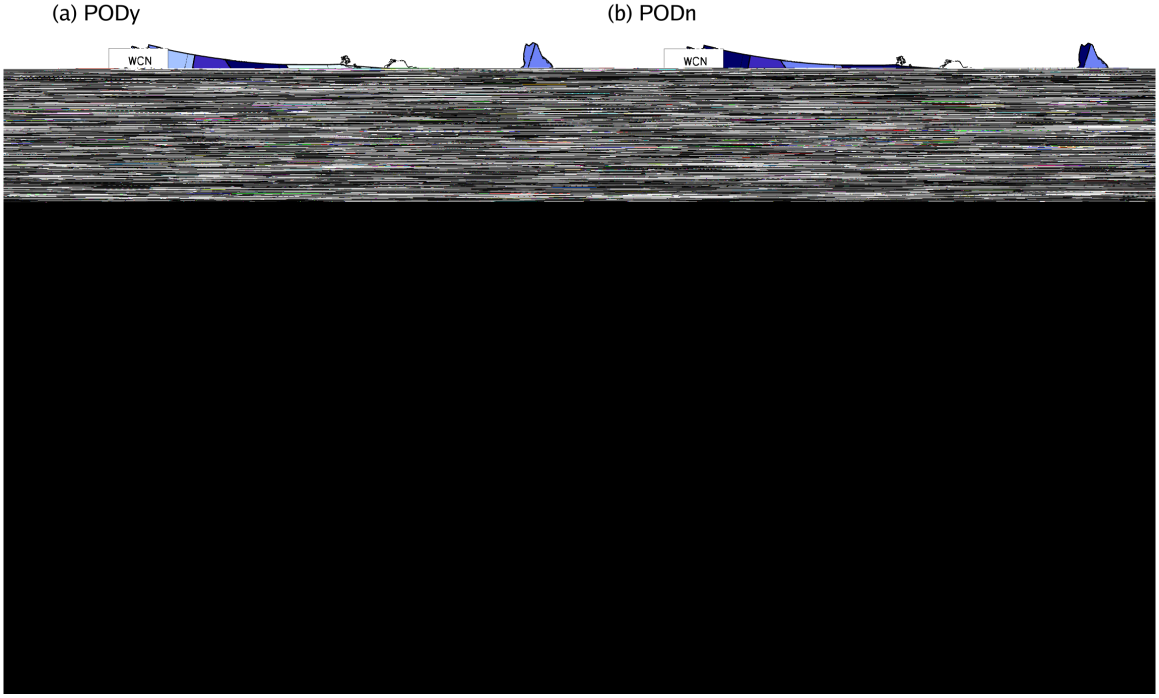

Fig. 17: As in Fig. 16 except for midlevels.

GTG2.3E is displayed to users through ADDS using four categories: None, Light, Moderate and Severe. The forecast and observation values used to define the categories were listed previously in Table 5. Because the forecasts will be provided in this way to users and decision makers, it is important to provide an assessment of its ability to discern the correct relationship with PIREP observations. The analysis will be based upon the joint distribution of forecasts and observations. This distribution provides all nontime-dependent information about the forecasts and observations (Murphy and Winkler, 1987). From the joint distribution one can derive two additional distributions that provide important information about the forecast performance. For this report, only results for the distribution of forecast intensities given the observation intensities, denoted p(f|x), will be presented. The conditioning variable provides the frame of reference within which the results are interpreted.

Ideally, the forecasts and observations match identically. Realistically, they do not match perfectly and the conditional distributions allow for greater understanding about the behavior of the forecasts and observations. It is important to reiterate that in the results that follow, the data are only valid wherever there was a PIREP. Since the true distribution of turbulence in the atmosphere is unknown, all forecasts points cannot be considered.

The joint distribution for upper levels is shown in Table 7. The observations are not uniform and are instead dominated by reports of None or Moderate intensity. Forecasts