Quality Assessment Report:

Graphical Turbulence Guidance, Version 2.5

Prepared By:

Quality Assessment Product Development Team
NOAA/ESRL/GSD

Matthew S. Wandishin[2], Brian P. Pettegrew[2], Melissa A. Petty[3], and Jennifer L. Mahoney[1]

September 8, 2011

Affiliations:

1 – National Oceanic and Atmospheric Administration, Earth System Research Laboratory, Global Systems Division (NOAA/ESRL/GSD)
2- Cooperative Institute for Research in Environmental Sciences (CIRES) and NOAA/ESRL/GSD
3 – Cooperative Institute for Research in the Atmosphere (CIRA) and NOAA/ESRL/GSD

Corresponding Author:
J.L. Mahoney NOAA/ESRL/GSD, 325 Broadway, Boulder, CO 80303;
Jennifer.Mahoney@noaa.gov

Table of Contents

1. Introduction 1
2. Approach 1
3. Data 2
   3.1 Data Period 2
   3.2 Algorithms and Forecasts 3
   3.3 Observations 4
      3.3.1 PIREPs 4
      3.3.2 EDR 4
4. Methodology 5
   4.1 EDR: Event-based Translation 5
   4.2 Incorporation of PIREPs 6
   4.3 Forecast-Observation Matching 6
      4.3.1 Gridded Forecast Neighborhood Approach 7
      4.3.2 Associating Observations to AIRMETs 8
   4.4 Defining Yes/No Events 8
   4.5 Verification Statistics 9
   4.6 Event Length Analysis 10
   4.7 Supplemental Evaluation 11
5. Results 12
   5.1 Overall Results for GTG2.5 Forecasts 12
      5.1.1 Skill as Compared to PIREPs vs. EDR 12
      5.1.2 Skill Stratified by Region 13
      5.1.3 Skill Stratified by Altitude 14
      5.1.4 Performance Characteristics Summary 15
   5.2 GTG2.5 and GTG2 16
      5.2.1 Discrimination of Events 16
      5.2.2 Performance According to Lead time 18

5.2.3 Overall Skill 21
5.2.4 Skill Stratified by Altitude 22
5.2.5 Event Length Analysis 24
5.2.6 Conditional Probabilities 25
5.2.7 Severity Discrimination 27
5.2.8 GTG2.5 and GTG2 Summary 29
5.3 GTG2.5 and (G-)AIRMET 29
5.3.1 Performance Comparison of GTG2.5 to (G-)AIRMET 30
5.3.2 GTG2.5 Supplemental Analysis 31
6. Concluding Statements 34
References 35
Appendix: Climatological Analysis of Observation Data Sets 37
I) Eddy Dissipation Rate (EDR) 37
II) PIREPs 41

List of Tables

Table 2.1: Vertical Stratifications used in GTG2.5 analysis..... 2

Table 4.1: Determination Observation and forecast intensity thresholds..... 9

Table 4.2: Dichotomous summary statistics used in this report. 10

List of Figures

Figure 2.1: Map of the regional stratification used in this report. 2

Figure 4.1: Schematic illustrating the conversion from a report-based to an event-based approach. The top part of the diagram shows a segment of a flight path through a patch of turbulence. Below that is the corresponding time series of EDR reports from the aircraft (color coded by intensity). 6

Figure 4.2: Example gridded forecast neighborhood (shaded) around a given observation (red). 7

Figure 4.3: A sample flight shown in profile (gray line). Non-null turbulence reports are color-coded by intensity (blue=0.15, red = 0.25). Contoured areas denote regions for which a GTG forecast predicts light (green) and MOG (yellow) turbulence. Black lines and numbers along the top and bottom of the plot show the length of observed and forecast turbulence events, respectively. 11

Figure 5.1: Performance measures for GTG2.5 when verified by PIREPs (squares) and EDR (triangles), stratified by severity category. 13

Figure 5.2: As in **Figure 5.1**, but for MOG turbulence forecasts only and stratified by region. 14

Figure 5.3: As in **Figure 5.2**, but stratified by layer. 15

Figure 5.4: Relative operating characteristic (ROC) curves for GTG2.5 (solid) and GTG2 (dashed) when verified by PIREPs (blue) and EDR (brown) for light-or-greater (upper left), MOG (upper right), and severe (lower left) turbulence events. Letters along the curves represent the performance for the light (L), moderate (M), and severe (S) thresholds. Numbers in the key give the area under each ROC curve (AUC). 17

Figure 5.5: Percent volume of MOG turbulence forecasts as a function of lead time for GTG2.5 (solid) and GTG2 (dashed). 19

Figure 5.6: Forecasts of light (green) and MOG (yellow) turbulence from GTG2.5 (bottom row) and GTG2 (top row) for 0-h (left), 6-h (middle), and 12-h (right) lead times. All forecasts are valid at 0300 UTC on 23 January 2011. Red circle highlights area of interest. 20

Figure 5.7: Performance measures for MOG turbulence forecasts as a function of lead time for GTG2.5 (solid) and GTG2 (dashed) when verified by PIREPs (blue) and EDR (brown). 21

Figure 5.8: Performance measures for MOG turbulence forecasts when verified by PIREPs (squares) and EDR (triangles), stratified by algorithm. 22

Figure 5.9: As in **Figure 5.2**, but for GTG2.5 (filled) and GTG2 (hollow) forecasts. Note that EDR forecasts are not used below 20 kft, so there are no EDR scores for the Low region. 23

Figure 5.10: Cumulative distribution functions of observed (black) and forecast (GTG2.5 blue, GTG2 brown) event lengths. Dashed lines and numbers show the 75th percentile event lengths. See text for description of how the event lengths are calculated. 25

Figure 5.11: Conditional probabilities of forecast categories (Null – black, Light – blue, Moderate – brown, Severe – green) matched with each category of observation intensity for

GTG2.5 (left) and GTG2 (right) when verified by PIREPs (top) and EDR (bottom). Numbers below each panel display the marginal probabilities for each intensity category for observations (p(o)) and forecasts (p(f)).....27

Figure 5.12: As in **Figure 5.4**, but for pairwise comparisons of each observed intensity category. Labels above each panel display to two categories being compared (nl – Null and Light, nm – Nul and Moderate, lm – Light and Moderate, ns – Null and Severe, ls – Light and Severe, ms – Moderate and Severe). 28

Figure 5.13: Performance measures for forecasts of MOG turbulence events for GTG2.5 (blue), AIRMETs (brown), and G-AIRMETs (yellow), stratified by GTG2.5 forecast threshold. (AIRMET and G-AIRMET forecasts are identical for each stratification and are repeated for ease of comparison.) 30

Figure 5.14: PODy as a function of the %Volume of the forecasts for GTG2.5 (lines), AIRMETs (stars), and G-AIRMETs (triangles) when verified by PIREPs (black) and EDR (red). Letters mark the performance for the Light (L), Moderate (M), and Severe (S) GTG2.5 thresholds...31

Figure 5.15: Performance of GTG2.5 forecasts (PODy, filled square; PODn, open square) inside (blue) and outside (brown) of AIRMETs, stratified by AIRMET type (AIRMET or G-AIRMET) and observation type (PIREP or EDR).....33

Figure 5.16: Schematic of the supplemental forecast. Large box represents the forecast domain. Small box represents an AIRMET forecast polygon. Gray filled ovals represent GTG forecasts of MOG turbulence. Red 'x's represent observations of MOG turbulence, while blue 'o's represent observations of Null turbulence. All proportions are set to approximate the scores in **Figure 5.15**.....34

Figure A.1: Spatial distribution of EDR reports for the period assessment period (1 December 2010 – 31 March 2011).....38

Figure A.2: Distribution of EDR reports by 1000 ft altitude layer for the assessment period. 39

Figure A.3: Distribution of EDR reports by region. See **Figure 2.1** for map of regions.....39

Figure A.4: Distribution of EDR by intensity category. See **Table 4.1** for category definitions. 40

Figure A.5: Distribution of EDR reports by time of day.....40

Figure A.6: As in **Figure A.1**, but for PIREPs.....42

Figure A.7: As in **Figure A.2**, but for PIREPs.....42

Figure A.8: As in **Figure A.3**, but for PIREPs.....43

Figure A.9: As in **Figure A.4**, but for PIREPs.....43

Figure A.10: As in **Figure A.5**, but for PIREPs.....44

Executive Summary

This report summarizes a quality assessment of Graphical Turbulence Guidance version 2.5 (GTG2.5). The current version of the GTG forecast product is based on the Rapid Update Cycle (RUC) numerical weather prediction system. The RUC is scheduled to be replaced by the Rapid Refresh (RR) prediction system in late 2011, necessitating an interim update in the GTG product.

The Quality Assessment Product Development Team (QA PDT) evaluated the performance of GTG2.5 with a focus on three main themes: 1) general characteristics of the GTG2.5 forecasts, 2) quality of the GTG2.5 forecasts compared to the current GTG product (version 2), and 3) the quality of GTG2.5 relative to Airman's Meteorological Advisories (AIRMETs) and Graphical AIRMETs (G-AIRMETs). This last theme itself consists of two parts: a direct comparison of the performance of GTG2.5 against that of AIRMETs and G-AIRMETs, and an evaluation of GTG2.5 as a supplement to the AIRMET and G-AIRMET forecasts.

The characteristics of the GTG algorithms as they are disseminated to users via ADDS were preserved in the evaluation. Each algorithm was kept at its disseminated grid resolution, and the forecast thresholds corresponding to light and moderate turbulence in the display of the GTG products were used as the thresholds in the assessment.

The GTG2.5 algorithm was analyzed using output generated from 1 December 2010 through 31 March 2011 over the CONUS. Verification was performed with both pilot reports (PIREPS) of turbulence and measurements of turbulence from instrument packages on the tails of select commercial aircraft (in situ Eddy Dissipation Rate measurements).

The primary findings, grouped by the three themes, are:

GTG2.5 Overall Performance

- Using current forecast thresholds, GTG2.5 verifies better against PIREPs than against EDR observations.

- The percent volume of moderate-or-greater GTG2.5 forecasts at cruising altitudes (35,000-40,000 ft) is less than half the size of the forecasts at lower levels, dramatically increasing the number of missed events (the PODy for this layer is nearly half that for lower altitudes).

GTG2.5 compared to GTG2

- Using current forecast thresholds, GTG2.5 is less skillful than GTG2.0 with more misses, but it covers only one-half to one-fourth of the airspace.

- Relative to GTG2, GTG2.5 decreases the number of moderate forecasts and increases the number of light forecasts. While this brings the distribution of intensities closer to that of the observations (both EDR and PIREP), it also increases the misclassification of moderate turbulence events.

GTG2.5 in relation to AIRMETs and G-AIRMETs

- When compared directly to AIRMETs and G-AIRMETs, the GTG2.5 algorithm is more skillful and covers much less of the airspace.

- When evaluated as a supplemental product, the GTG2.5 algorithm is able to narrow the focus of the AIRMETs and G-AIRMETs by capturing most of the moderate-or-greater turbulence events with a much smaller volume, thus reducing the number of false alarms. Outside the AIRMETs and G-AIRMETs, GTG2.5 adds only a small number of false alarms while capturing nearly half of the moderate-or-greater events missed by the AIRMETs and G-AIRMETs.

All of the above results are sensitive to the choice of forecast thresholds. Adjusting the thresholds downward increases the volume of the forecasts and the PODy while reducing the PODn. Currently, the GTG2.5 threshold for moderate turbulence is set at a level that yields a lower PODy than the current operational GTG product.

1. Introduction

This report summarizes a formal quality assessment in support of the transition of the Graphical Turbulence Guidance version 2.5 (GTG2.5) algorithm to National Weather Service (NWS) operations. The current operational algorithm, GTG2, is based on the Rapid Update Cycle (RUC; Benjamin 2004) prediction system. The upcoming replacement of the RUC with the Rapid Refresh (RR; Benjamin 2006) prediction system necessitates an update to the turbulence algorithm.

The report is organized in the following manner. Section 2 provides an overview of the approach taken for the evaluation. The forecast and observed data are described in Section 3, followed by a presentation of the verification methodology in Section 4. The results are described in Section 5. Finally, conclusions are provided in Section 6.

2. Approach

The evaluation consists of three primary assessment areas:

- A general overview of the GTG2.5 forecasts
- A direct comparison of the GTG2.5 and GTG2 forecasts
- The performance of GTG2.5 both in direct comparison to and as a supplement to AIRMETs and G-AIRMETs.

The algorithm is evaluated both for its ability to predict moderate-or-greater (MOG) turbulence and for its ability to distinguish among the different intensity categories. In addition to the entire continental United States (CONUS), the forecasts are analyzed across three regions (**Figure 2.1**) and for three separate vertical layers (**Table 2.1**). The assessment covers the four-month time period from 1 December 2010 through 31 March 2011.

Whereas the GTG2 forecasts are based on the 13-km RUC model, the final display grid is degraded to 20 km. GTG2.5, however, is intended to be displayed at 13-km resolution. Rather than remap one of the forecasts to the grid of the other, each algorithm is evaluated at its native display resolution, i.e., according to how each is seen by the user. The use of a neighborhood technique, described in Section 4, will help mitigate potential representativeness errors inherent in working with different resolutions. **Table 2.1** shows the vertical stratifications used in this evaluations.

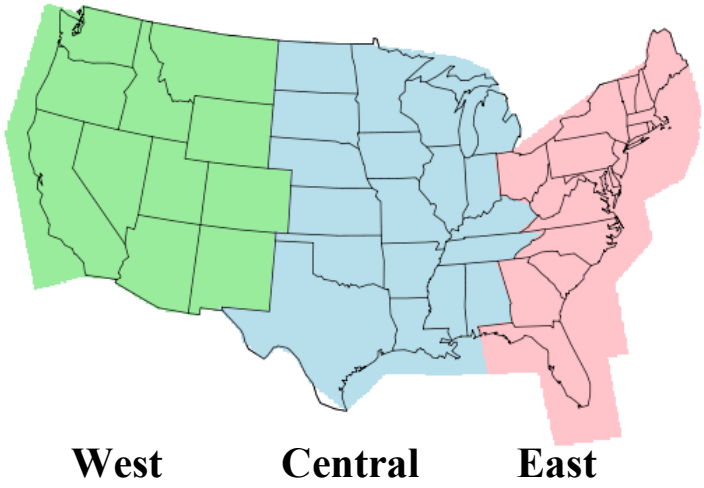

Figure 2.1: Map of the regional stratification used in this report.

Table 2.1: Vertical Stratifications used in GTG2.5 analysis.

Altitude Layers (in kft)	Description
10 – 20	Low
20 – 30	Middle
30 – 45	High

3. Data

3.1 Data Period

Data were collected for analysis from 1 December 2011 through 31 March 2011. Data between 0600 and 1200 UTC were excluded because of the small number of flights from which observations are available.

3.2 Algorithms and Forecasts

In addition to GTG 2.5, the product being assessed, the current operational turbulence forecast products available to users were considered for comparison. The following subsections describe the forecast products that were included in this evaluation.

GTG2: This algorithm uses RUC model output to derive a suite of turbulence diagnostics, which are then combined as a weighted average to produce a forecast of turbulence potential, scaled to yield values between 0 and 1. These forecasts of clear air turbulence (CAT) extend from 10,000 to 45,000 ft. Forecasts are issued every hour with hourly lead times out to 3 h and 3-hourly lead times between 3 and 12 h. For more details on this algorithm see Sharman et al. (2006).

GTG2.5: The adaptation of the GTG algorithm from the RUC model to the RR involved many changes. Perhaps the most significant is a change in the vertical coordinate from a sigma-theta hybrid coordinate, which provides increased vertical resolution near the jet stream—a major source of CAT—to a pure sigma coordinate, which is more uniform with height. These changes forced alterations to the suite of turbulence diagnostics. As a result of the integration of additional eddy dissipation rate (EDR) aircraft measurements, the algorithm now provides an explicit forecast of EDR. The forecast values, therefore, now cover a range from only 0 to 0.8. In addition, the forecasts consist of hourly leads out to 12 h.

AIRMETs: AIRMET turbulence forecasts are a text-based product issued by the AWC four times per day (0300, 0900, 1500 and 2100 UTC) for moderate-or-greater (MOG) turbulence when there is at least a 50% or more probability of occurrence, excluding convectively induced turbulence. These polygon forecasts are valid for a 6-h time period from the issuance and are generated from the production of GAIRMETs (see below). Further, AIRMETs must meet a minimal requirement for areas of at least 3000 mi^2. For this assessment only routine issuances and corrections are used.

G-AIRMETs: G-AIRMETs are issued concurrently with the AIRMETs and serve as the basis for the AIRMET forecasts. G-AIRMETs are defined by the AWC as a decision tool based on weather "snapshots" at short intervals of time (Aviation Weather Service, 2010). These short intervals are lead times every 3 h out to 12 h. As a result, the text AIRMET is now a fully consistent product that is generated from the 0, 3, and 6-h forecasts of the G-AIRMET. G-AIRMETs are intended to provide a finer scale forecast both spatially and temporally than the text-based AIRMETs and according to the Federal Aviation Administration's (FAA) Aeronautical Information Manual (FAA AIM; 2011), the use of AIRMETs and G-AIRMETs in concert will improve the clarity of impacted regions and further enhance situational awareness. Consistent with the AIRMET requirements, G-AIRMETs must also meet a minimal size requirement of at least 3000 mi^2.

3.3 Observations

Two sets of turbulence observations are used in this report: NWS Pilot Reports (PIREPs) and *in situ* measurements of eddy dissipation rate (EDR). See the Appendix for a climatological analysis of each reporting platform for the period of the assessment.

3.3.1 PIREPs

PIREPs are voice-recorded pilot reports of aviation hazards, including turbulence. PIREPs are reported by pilots at their discretion and based on their experience with turbulent events. As a result, PIREPs suffer from the following deficiencies:

1. They are subjective. Different-sized aircraft can respond differently to the same turbulent patch. Pilots may have different concepts of what constitutes "moderate" turbulence.

2. They are inconsistently reported. Not every encounter with turbulence generates a report, particularly for lighter turbulence. More importantly, there are no guidelines for when a "Null" report should be issued. Consequently, non-events are under-represented in the PIREP data set.

3. They are imprecise. Reports are categorical, using vague terminology (e.g., "light chop with occasional moderate"). Also, reports are issued after the fact leading to temporal and spatial errors.

3.3.2 EDR

EDR observations are taken in-situ from various United Airlines 737 and 757 aircraft. On-board equipment measures and reports vertical accelerations in the aircraft while in flight. These measurements are converted into an EDR value and then reported back to a database where they undergo quality control processes. The EDR observing system reports a maximum and median value every minute. Due to equipment sensitivity during ascent/descent stages of flight, EDR observations beneath 20,000 ft are not utilized (Cornman et al. 2004).

EDR observations are based on tail measurements of vertical accelerations from select United Airlines 737 and 757 aircraft[1]. As a result:

1. They are objective. EDR observations are based on actual properties of the plane's encounter with turbulence, but transformed into an aircraft-independent measure.

[1] Currently, EDR observations are also available from Delta Airlines. However, the Delta Airlines data employs a different reporting methodology and the data were made available too late to allow the two reporting methods to be reconciled. This report, then, uses the data from the United aircraft only.

2. They are consistently reported. Measurements are automatically sent to ground receiving stations every minute, regardless of severity.

3. They are precise. Because of the consistent reporting, the exact time and location of each encounter with a turbulent patch (including the duration of the encounter) is recorded.

However, there are shortcomings with the EDR data set, as well. Due to equipment sensitivity during the ascent and descent stages of flight, EDR observations beneath 20,000 ft are not utilized (Cornman et al. 2004). This, combined with the small number of planes reporting EDR, leads to a more limited spatial coverage.

4. Methodology

This section describes the overall methodology used for verification, such as the approaches taken to incorporate the two observation sets, the way in which the forecast is paired with observations, and the definition of yes and no events for the forecast and observation sets. It also describes the approach taken in the evaluation of GTG 2.5 as a supplement to the operational (G)-AIRMET.

4.1 EDR: Event-based Translation

Because EDR is reported every minute, multiple reports may contain information about the same turbulence encounter, leading to the event being oversampled. This oversampling can be mitigated by switching from individual observations of turbulence to an event-based approach. For this assessment, turbulence events have been defined as the set of all turbulence observations separated by four minutes[2] or less. The intensity of an EDR event containing more than a single observation is defined as the maximum intensity of the set of observations defining the event.

Figure 4.1 provides an example of the event-based approach. The six reports of non-null turbulence are transformed into a single event of six-minute duration. The magnitude of the event is taken to be the maximum intensity of the six constituent reports, namely the 0.25 value from the 6th report (at 12:20).

Because the vast majority of the atmosphere does not contain turbulence—entire flights can be turbulence free—a different approach must be taken to translate null reports into events. Null events are defined as contiguous 15-minute segments of null turbulence reports, with the location of the event defined as the midpoint of the 15-minute segment.

[2] The FAA AIM (2011) defines intermittent turbulence as turbulence occurring during at least one-third of a given time period. A pair of turbulence reports separate by a four-minute gap would satisfy this one-third requirement.

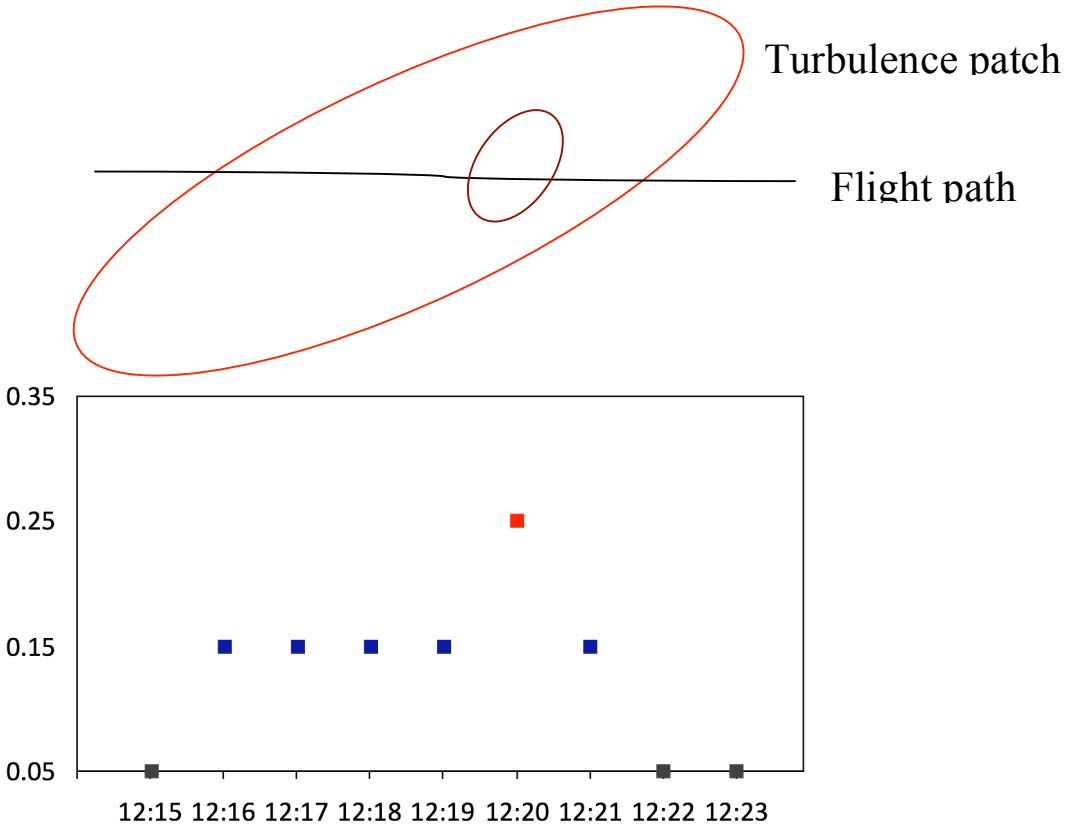

Figure 4.1: Schematic illustrating the conversion from a report-based to an event-based approach. The top part of the diagram shows a segment of a flight path through a patch of turbulence. Below that is the corresponding time series of EDR reports from the aircraft (color coded by intensity).

4.2 Incorporation of PIREPs

PIREPs are already event based—pilots do not issue multiple reports for the same turbulence encounter—and so no translation from report to event is necessary. When a PIREP is reported over a vertical range, the entire range is considered in its association to a forecast, as will be described in subsequent sections.

4.3 Forecast-Observation Matching

The following subsections describe how the various forecast and observation types are associated.

4.3.1 Gridded Forecast Neighborhood Approach

Since the absence of a report of turbulence does not necessarily mean an absence of turbulence, verification of turbulence forecasts must be observation based. That is, verification is based on the set of observations, and the forecasts are then matched to these observations. In this report, gridded forecasts are paired with observations using a neighborhood approach. First, observations are matched vertically to the nearest forecast grid level and then horizontally to the nearest forecast grid box. All of the forecast grid boxes contained within a given horizontal radius of the observation at the matched grid level (**Figure 4.2**), plus one grid level above and below the matched level are included in the neighborhood. Observation times are rounded to the nearest valid time, e.g., events at 1830 UTC and 1929 UTC will both be matched to forecasts valid at 1900 UTC.

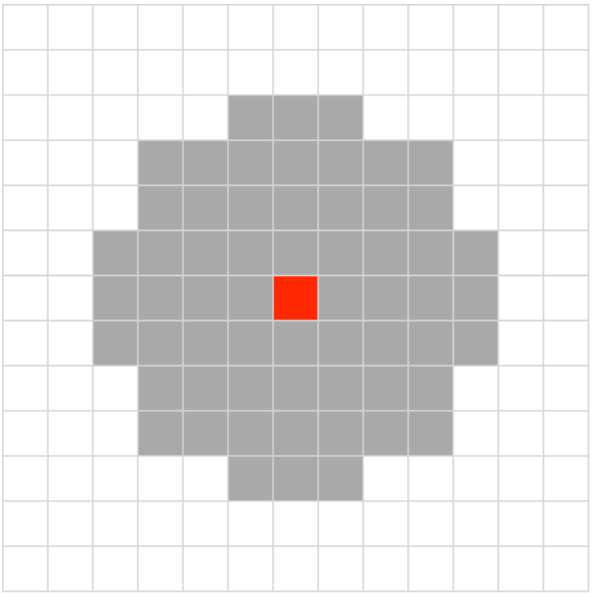

Figure 4.2: Example gridded forecast neighborhood (shaded) around a given observation (red).

4.3.1.1 EDR

Nearly 90% of all observed EDR turbulence events are 50 km in length or shorter. The radius of the forecast neighborhood around an EDR report is taken to be one 90%-event length. However, in order to establish a better alignment between the 13- and 20-km forecast grids the radius is adjusted to 60 km. For EDR events containing more than one report, the final neighborhood is simply the union of the individual neighborhoods around each report within the event.

4.3.1.2 PIREPs

For PIREPs, a larger radius neighborhood is used to account for the spatial uncertainties associated with these reports. Based on an analysis of the spatial relationship between EDR observations and PIREPs (Wandishin et al. 2010), the radius is set to 150 km[3]. Some PIREPs report turbulence over a vertical range. For these reports, the final neighborhood consists of the horizontal neighborhood at each level within the vertical range plus one grid level above the top reported height and one grid level below the minimum reported height.

4.3.1.3 Choice of Representative Forecast Value

For pairing observations with GTG2.5 or GTG2 forecasts, the maximum value within the forecast neighborhood is taken to be the forecast turbulence intensity. Other summary measures (mean, median, mode) were also examined, but only the maximum is presented herein because of its operational significance.

4.3.2 Associating Observations to AIRMETs

For determining whether an observation is inside an AIRMET or G-AIRMET, the following criterion was used. If any part of an observed turbulence event is inside an AIRMET or G-AIRMET, the entire event is considered to be within the advisory volume. Nearly all observed events are either entirely within or entirely without an AIRMET or G-AIRMET and so the results are not sensitive to this threshold. Similar to the case of the GTG algorithms, G-AIRMETs are matched to any observations reported within 30 minutes before or after the forecast valid time. For AIRMETs, the forecast is matched to all observations reported within the six-hour time window over which the AIRMET forecast is valid.

4.4 Defining Yes/No Events

The following criteria are used to define events for the various forecasts and observations:

- GTG forecasts: If the maximum value within the forecast neighborhood meets or exceeds an event category threshold, it is considered a forecast of that event.
- (G-)AIRMETs: Everywhere within the forecast polygon is by definition considered a forecast of MOG turbulence.
- PIREPS: If the PIREP intensity meets or exceeds the event threshold, it is considered an observed event.

[3] The sensitivity of the results to the neighborhood size was examined by using the 60-km EDR neighborhood radius around PIREPs as well. Using a smaller radius has a quantitative impact, but does not alter the relationships of the scores presented in the results.

- EDR: An observed event occurs if the maximum intensity within an EDR event is greater than or equal to the event threshold.

Note that non-events are not limited to explicit nulls, but rather include all categories less than the event category.

Table 4.1 shows the event threshold for the observations and the GTG algorithms. Note that the GTG2.5 thresholds are even lower than what is expected from the smaller range of the forecast values (from 0-1 to 0-0.8). Since GTG2.5 is explicitly a forecast of EDR, the thresholds were chosen to align with the ICAO definitions of Light, Moderate, and Severe turbulence.

Table 4.1: Determination Observation and forecast intensity thresholds.

Intensity	EDR	PIREP	GTG2.5	GTG2
Null	0.05	0	0	0
Light	0.15	1	0.15	0.3
Moderate	0.25	3	0.31	0.475
Severe	0.45	5	0.54	0.8

4.5 Verification Statistics

Table 4.2 lists the dichotomous statistics calculated from the entries in a contingency table. Another dichotomous summary measure, or more precisely, a collection of dichotomous statistics, is the Relative Operating Characteristic (ROC) curve. For a defined event, one can create a series of PODy and PODn pairs by adjusting the forecast threshold. Plotting PODy as a function of 1-PODn produces a ROC curve. The further the curve lies toward the upper left corner of the diagram (PODy=1, PODn=1), the better the forecast system is at discriminating between events and non-events. This ability to discriminate between events and non-events is summarized by the area under the curve (AUC).

In addition to the dichotomous techniques, two categorical measures are employed: conditional probabilities measure the relative occurrence of a forecast category associated with each observed intensity category, and multi-category ROC curves disclose the ability of forecasts to discriminate between difference pairs of observed intensity categories.

Table 4.2: Dichotomous summary statistics used in this report.

Statistic	Formula	Description
PODy	YY / (YY + NY)	Proportion of events detected correctly
PODn	NN / (YN + NN)	Proportion of non-events detected correctly
TSS	PODY + PODn -1	True Skill Statistic

Lastly, the size of the forecast is measured by the percent volume where

$$\% \text{ Volume} = 100 * \text{Volume}_{\text{forecast}} / \text{Volume}_{\text{possible}}.$$

The % Volume measures the percent of the possible volume (the forecast domain) that is covered by the forecast. Unfortunately, it is not possible to measure the % Volume of the observations and so a direct measure of bias is not possible. However, an approximate measure of the bias is available, as described in the next sub-section.

4.6 Event Length Analysis

Consider a flight path represented by aircraft-reported EDR measurements (**Figure 4.3**). Observed turbulence events can be identified as described in Section 4.1. The length of each event can then be calculated from the locations of the individual reports comprising the event. Similarly, one may locate the flight path within a forecast grid and, for each EDR report, replace the observed intensity with the forecast value from the grid box in which the report is located. These individual forecast values can then be converted to forecast events in the same manner as is done for the observations and the corresponding lengths of the forecast events calculated. One may then compare the distribution of observed and forecast event lengths; the difference in these distributions becomes a measure of the bias of the forecast.

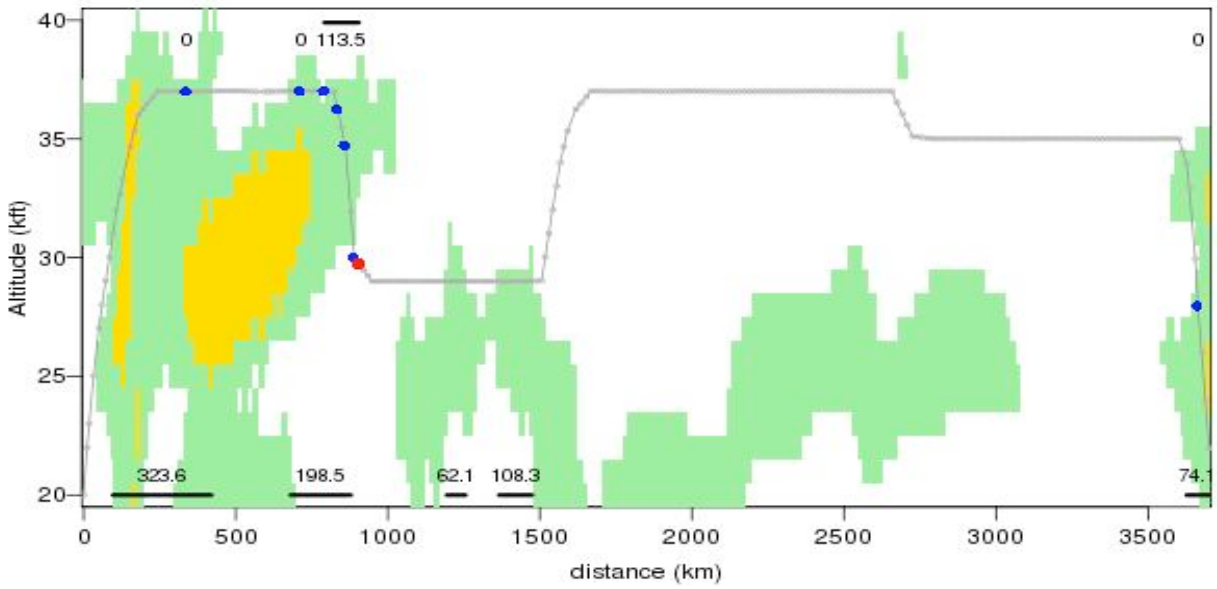

Figure 4.3: A sample flight shown in profile (gray line). Non-null turbulence reports are color-coded by intensity (blue=0.15, red = 0.25). Contoured areas denote regions for which a GTG forecast predicts light (green) and MOG (yellow) turbulence. Black lines and numbers along the top and bottom of the plot show the length of observed and forecast turbulence events, respectively.

4.7 Supplemental Evaluation

In addition to evaluating GTG 2.5's performance as a standalone forecast, its performance was assessed as a supplement to the operational (G)-AIRMET product. This section describes the supplemental relationship defined to frame the evaluation.

AIRMETs and G-AIRMETs already partition the airspace into regions in which widespread MOG turbulence is expected and regions in which MOG turbulence is not expected. If the forecasts are at all skillful (and they are), then observations within the (G-)AIRMET will be very different than the observations outside the (G-)AIRMET. Specifically, the region within the AIRMET will contain a higher proportion of events and the region outside the AIRMET will contain a higher proportion of non-events. Therefore, evaluating the success of GTG2.5 as a supplement to the (G-)AIRMET forecasts entails focusing on the two regions separately:

- Inside the (G-)AIRMET, where turbulence is expected, GTG2.5 should open the airspace (i.e., reduce the forecast volume and, relatedly, the number of false alarms.
- Outside the (G-)AIRMET, where turbulence is not expected, GTG2.5 should reduce the exposure to unforecasted events (i.e., capture events missed by the (G-)AIRMET) without unduly restricting the airspace.

5. Results

5.1 Overall Results for GTG2.5 Forecasts

The following section is a summary of the general performance characteristics of GTG2.5 as verified by PIREPs and EDR.

5.1.1 Skill as Compared to PIREPs vs. EDR

GTG2.5 is more skillful for MOG events when verified by PIREP observations than when verified by EDR events (**Figure 5.1**), whereas for LOG events there is no difference in skill. As severity increases, the number of events is reduced (the event becomes more rare) and it becomes more difficult for GTG2.5 to capture the stronger turbulence events; this is indicated by the decrease in PODy and increase in PODn for both observation sets. Similarly, since turbulence events are more commonly observed with PIREPs than with EDR (cf. **Figure A.4** and **Figure A.9)**, the forecasts have a higher PODy (and lower PODn) when verified by PIREPs. These results are sensitive to the choice of forecast thresholds. Reducing the GTG2.5 threshold for MOD turbulence makes the forecast event more common, leading to an increase in the PODy and a decrease in the PODn. When applied to GTG2.5 verified against EDR, this change in threshold can bring the skill closer to that when verified by PIREPs.

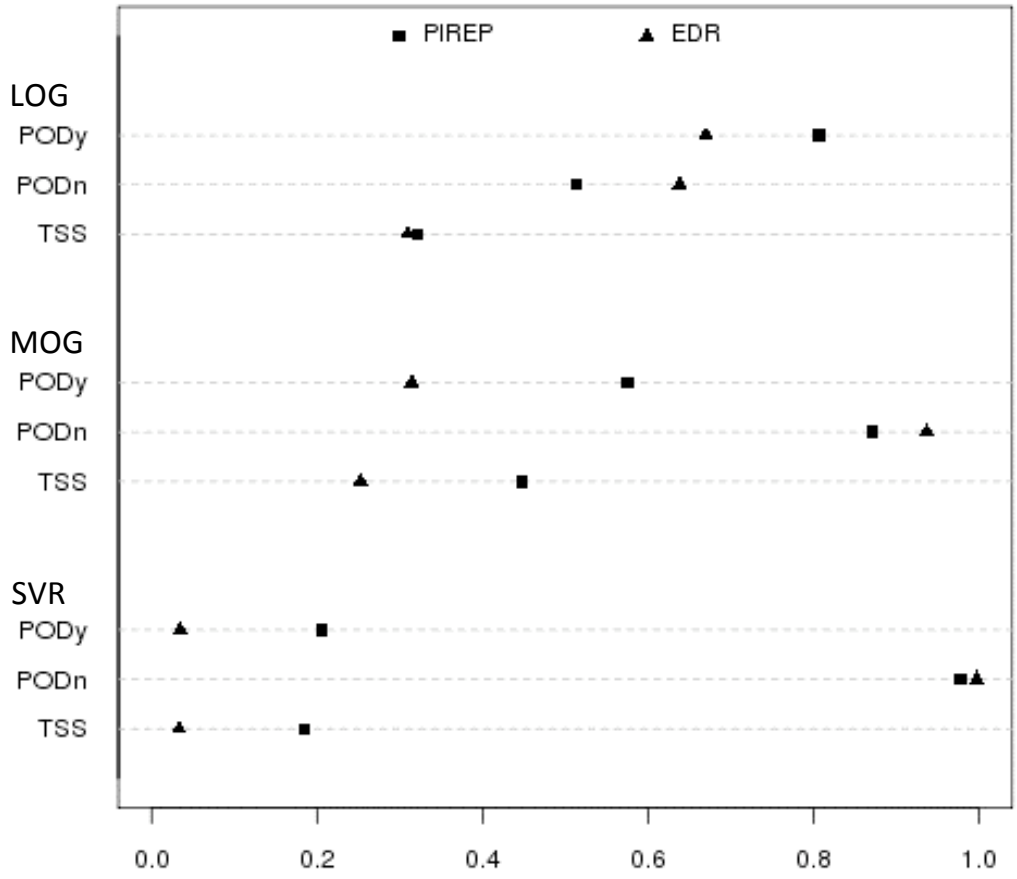

Figure 5.1: Performance measures for GTG2.5 when verified by PIREPs (squares) and EDR (triangles), stratified by severity category.

5.1.2 Skill Stratified by Region

Stratifying performance by regions reveals somewhat better skill in the East (**Figure 5.2**). Forecasts in the East are able to capture more of the turbulence events (larger PODy) than in the Central region, while correctly avoiding more non-events (larger PODn) than in the West. This difference demonstrates that GTG2.5 forecasts in the East are not superior to both the other regions in all respects, but only when considered in total.

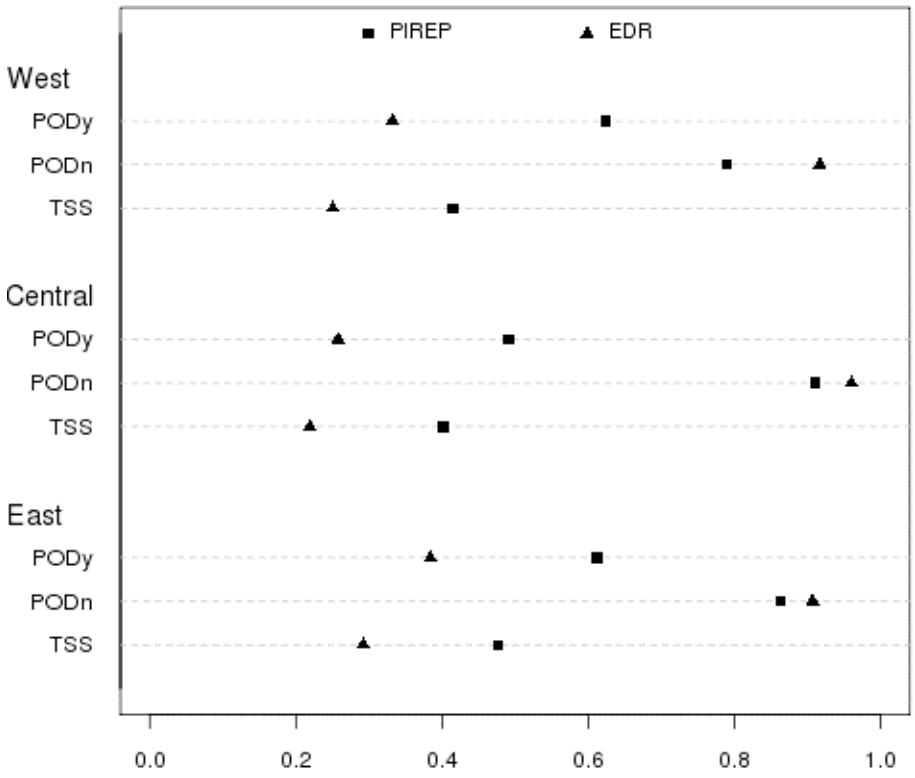

Figure 5.2: As in **Figure 5.1**, but for MOG turbulence forecasts only and stratified by region.

5.1.3 Skill Stratified by Altitude

The ability of GTG2.5 to capture MOG turbulence events is reduced with altitude (decreasing PODy; **Figure 5.3**), while areas with less-than-MOD turbulence are better identified (increasing PODn). This pattern is consistent with the variation of the % Volume of the forecasts with height: as the altitude increases the % Volume decreases (from 2.9% in the Low layer to 2.6% in the Middle layer to 1.0% in the High layer). In fact, the % Volume in the 30-45 kft layer is less than half that for the 10-20 kft layer. This decrease in forecast volume leads to a reduction in the PODy and an increase in the PODn—the more narrowly focused forecast volume will contain both fewer events and fewer non-events. Despite this large decrease in volume and subsequent near halving of the PODy for the High layer, the skill of the GTG2.5 forecasts is very similar for the three layers (when verified by PIREPs).

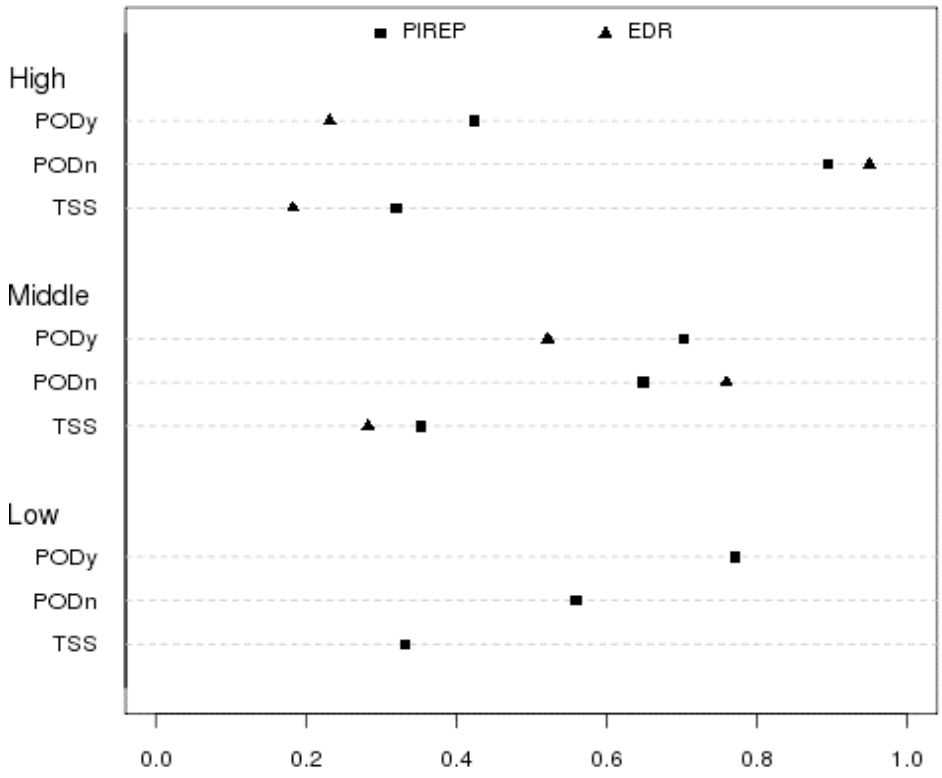

Figure 5.3: As in **Figure 5.2**, but stratified by layer.

5.1.4 Performance Characteristics Summary

In summary, GTG2.5 forecasts are more skillful when verified by PIREPs than when verified by EDR, particularly for MOG turbulence. When verified by PIREPs, the forecasts have a much higher PODy and a lower PODn. This difference in performance is due, in part, to the different intensity distributions of the two observing platforms. In addition, the results are sensitive to the choice of forecast thresholds. Lowering the MOD forecast threshold will increase the PODy (and lower the PODn). When this is applied to GTG 2.5 when verified by EDR, it can raise the skill of the forecasts closer to that when verified by PIREPs. Another finding is that the forecasts are sensitive to altitude; specifically, forecasts in the layer containing cruising altitudes are less than half the size of the forecasts in lower levels. As a result, the number of turbulence events missed by the forecasts increases substantially (i.e., PODy values are nearly cut in half, as well).

5.2 GTG2.5 and GTG2

This section discusses the skill of GTG2.5 as compared to that of GTG2, the current operational product. In comparing the performance of these two products, it is important to remember that the two algorithms differ in several ways, including the resolution of the forecast grids. Both forecast products are based on 13-km model output, but the GTG2 forecasts are degraded to 20 km resolution for the ADDS display, while GTG2.5 will be displayed on its native grid. The increase in grid resolution is expected to impact verification scores, but the neighborhood method employed herein should reduce this impact.

5.2.1 Discrimination of Events

Relative Operating Characteristic (ROC) curves plot PODy as a function of (1-PODn) and provide a measure of a forecast system's ability to discriminate between events and non-events. ROC curves are typically summarized by the area beneath the curve (AUC), which is related to the degree of separation between the distribution of forecast scores associated with events and the distribution of forecast scores associated with non-events. The further the ROC curve extends toward the upper-left corner, and thus the higher the area under the curve, the better the forecast system is at distinguishing events from non-events. Moving along the curve from the upper-right corner to the lower-left corner, the points on the curve represent increasing larger forecast thresholds.

Figure 5.4 provides ROC curves according to observed event severity. As the severity of the observed event increases, the ROC areas increase. (It is not uncommon in meteorology for rarer and stronger phenomena to be easier to distinguish from non-events.) For all severities, however, the ROC curves for the GTG2 forecasts (dotted) are slightly higher than those for the GTG2.5 forecasts (solid), whether they are being verified against PIREPs (blue) or EDR (brown).

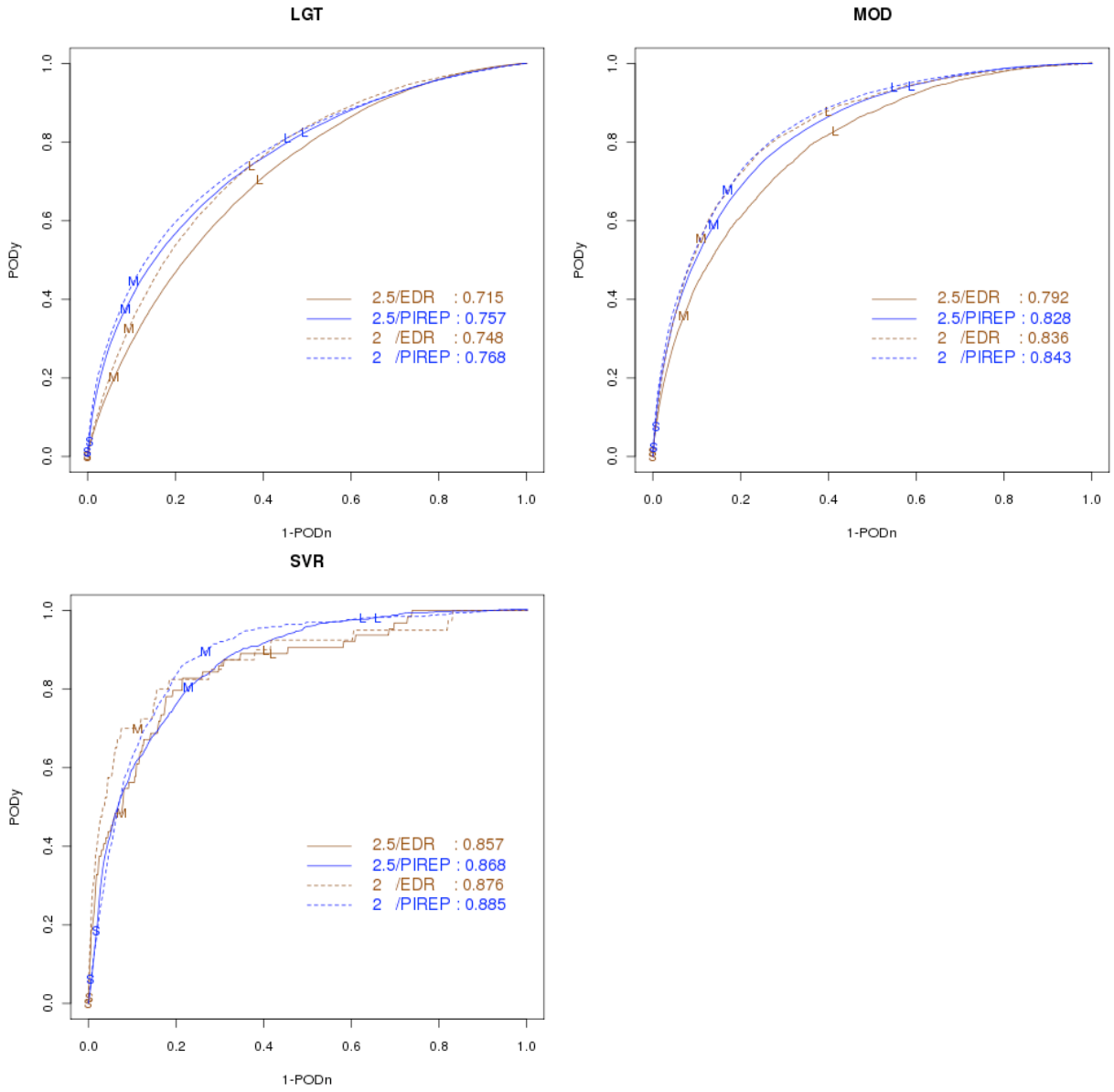

Figure 5.4: Relative operating characteristic (ROC) curves for GTG2.5 (solid) and GTG2 (dashed) when verified by PIREPs (blue) and EDR (brown) for light-or-greater (upper left), MOG (upper right), and severe (lower left) turbulence events. Letters along the curves represent the performance for the light (L), moderate (M), and severe (S) thresholds. Numbers in the key give the area under each ROC curve (AUC).

Another consistent pattern is that the forecast thresholds for the GTG2.5 forecasts are further along the curves toward the lower-right corner, particularly the MOD threshold. This signifies that MOG events are less frequently forecast in GTG2.5 than in GTG2, implying that the GTG2.5 forecasts will occupy less of the airspace than the GTG2 forecasts.

5.2.2 Performance According to Lead time

5.2.2.1 % Volume Analysis

The Volume analysis supports observations from the preceding section by demonstrating the larger volumes of the GTG2 vs GTG2.5. From **Figure 5.5** it can be seen that the volume of the GTG2.5 analysis is nearly half that of the GTG2 analysis. Additionally, the GTG2 forecasts grow in extent after the 1-h lead time such that the forecasts at the 12-h lead time are more than twice the size of the analyses (i.e., the 0-h lead). By contrast, the GTG2.5 forecasts have completely eliminated the growth as a function of lead time. **Figure 5.6** illustrates these results for a single set of forecasts all sharing the same valid time. The red circles highlight a common area of MOG forecasts. The GTG2 MOG forecast (mustard color; top panels) fills nearly half the space at hour 0, and by 12 h only the northeast portion of the highlighted domain remains at less-than-MOG. The GTG2.5 forecast (bottom panels) grows somewhat between 0 and 6 h before shrinking back at 12h. Throughout, the GTG2.5 forecast fills much less of the highlighted domain than the GTG2 forecasts.

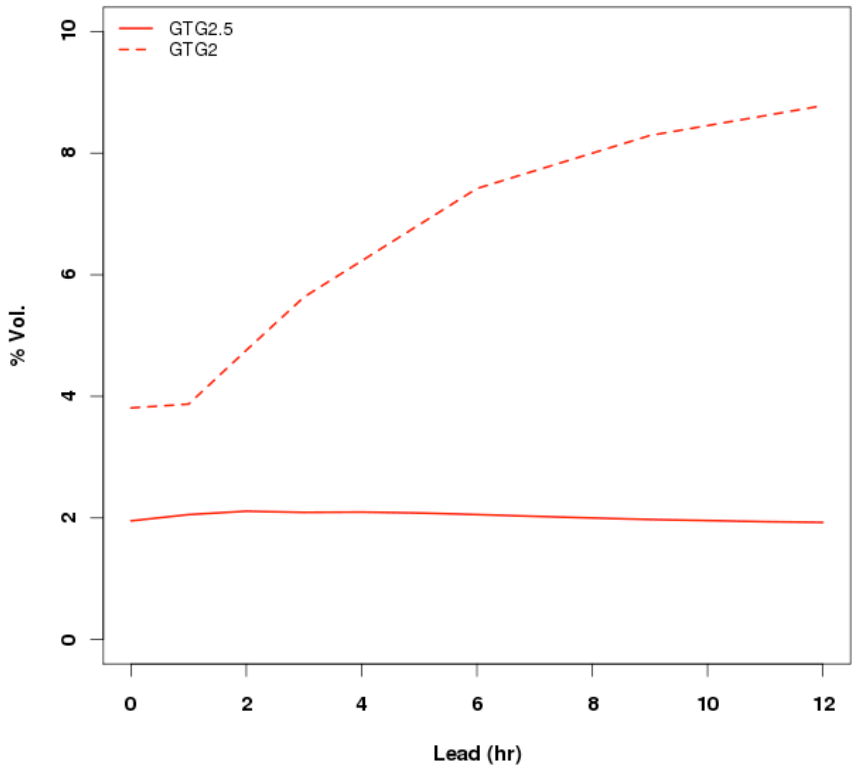

Figure 5.5: Percent volume of MOG turbulence forecasts as a function of lead time for GTG2.5 (solid) and GTG2 (dashed).

.

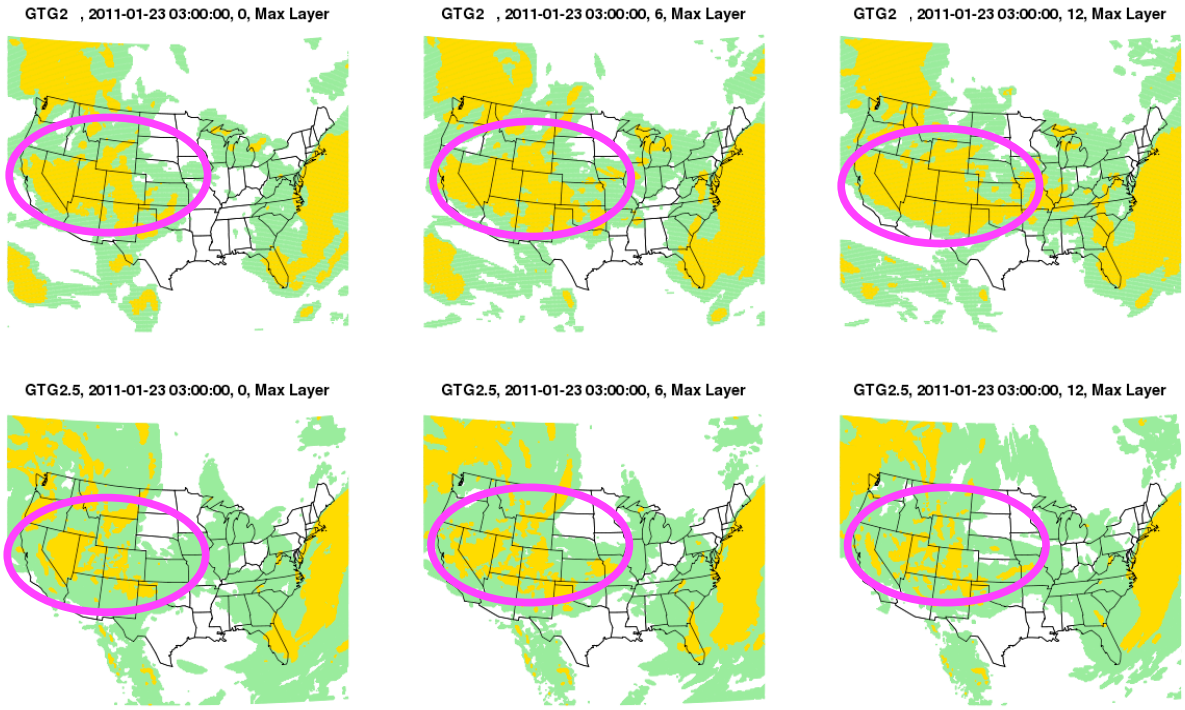

Figure 5.6: Forecasts of light (green) and MOG (yellow) turbulence from GTG2.5 (bottom row) and GTG2 (top row) for 0-h (left), 6-h (middle), and 12-h (right) lead times. All forecasts are valid at 0300 UTC on 23 January 2011. Red circle highlights area of interest.

5.2.2.2 Skill Analysis

As expected, the larger forecast volumes will capture both more events and more non-events, translating into higher PODy and lower PODn values, (**Figure 5.7;** all scores are for MOG turbulence events). Not only do the GTG2.0 forecasts have a higher PODy than GTG2.5, but that gap increases with increasing lead time as the growing GTG2 forecasts lead to even higher PODy values. The trend in PODn values is the opposite of that for PODy: the GTG2 PODn scores are lower than those for GTG2.5 at the analysis time and decrease further with increasing lead time. In fact, the trends in the two scores appear to balance out almost completely, as the TSS varies little with lead time and the gap between GTG2 and GTG2.5 skill is much smaller than it is for PODy and PODn. As seen with the ROC scores, the gap in performance between the two algorithms is smaller when verified by PIREPs than when verified by EDR events.

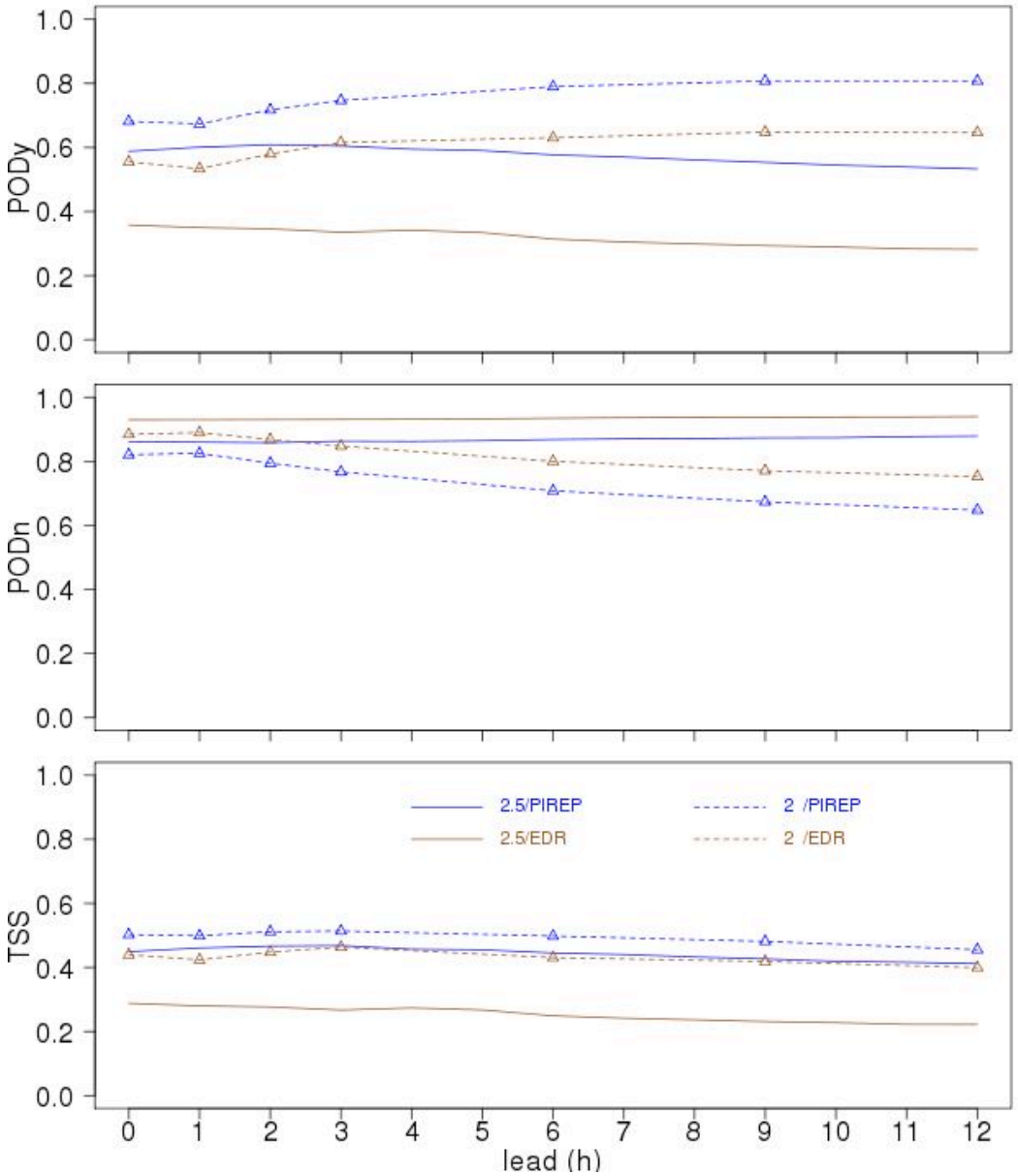

Figure 5.7: Performance measures for MOG turbulence forecasts as a function of lead time for GTG2.5 (solid) and GTG2 (dashed) when verified by PIREPs (blue) and EDR (brown).

5.2.3 Overall Skill

Patterns discussed previously also hold when viewing the scores for all stratifications (all issues, leads, regions, and vertical levels) grouped together (**Figure 5.8**): GTG2 has substantially larger PODy and somewhat lower PODn values, leading to TSS values that are a little better when verified by PIREPs and almost twice as high when verified by EDR. Once more it is important to note that these results are sensitive to the choice of forecast thresholds. By

lowering the GTG2.5 forecast thresholds, thus making the forecasts of a MOG event more prevalent, one could increase the PODy at the cost of an increase in the PODn. In other words, the number of missed forecasts could be reduced at the cost of an increase in false alarms. Furthermore, considering the large reduction in MOG forecast coverage for GTG2.5 compared to GTG2, the forecast thresholds could be reduced substantially while still maintaining smaller forecast volumes. However, the thresholds listed in Section 4.4 are the values slated for the ADDS display and so only results for those thresholds are considered in this evaluation.

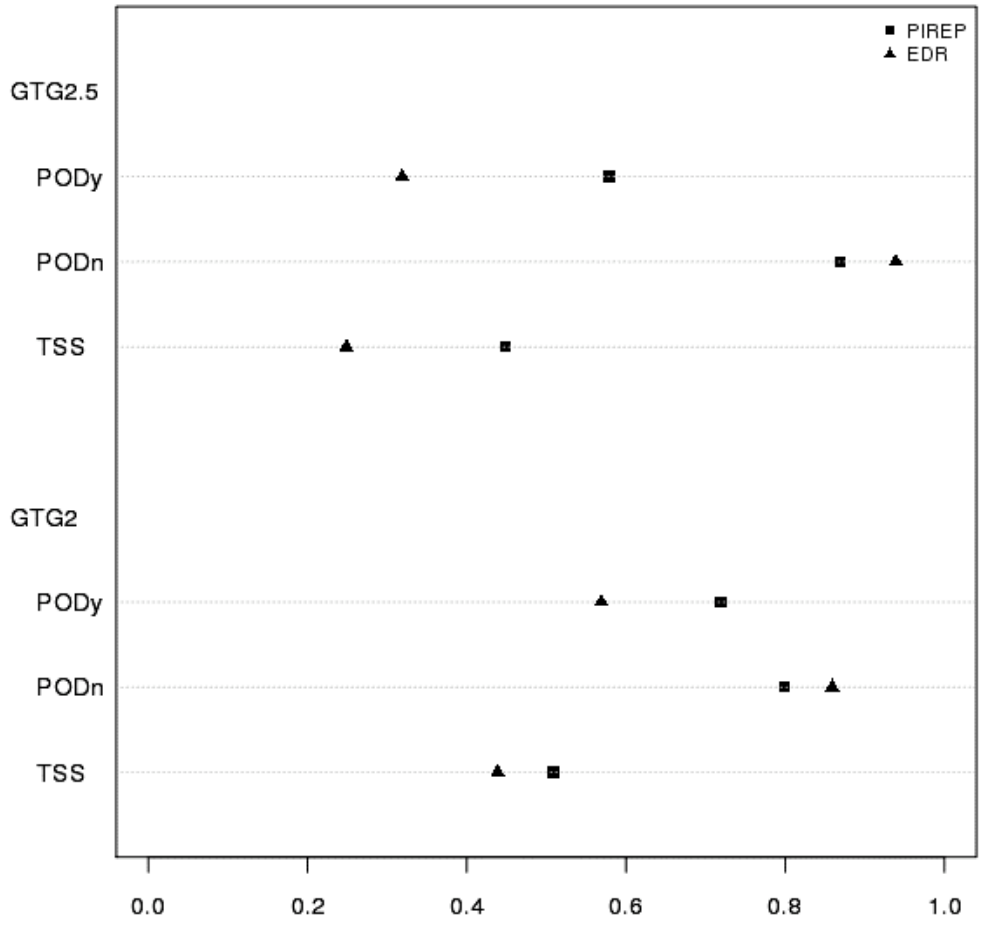

Figure 5.8: Performance measures for MOG turbulence forecasts when verified by PIREPs (squares) and EDR (triangles), stratified by algorithm.

5.2.4 Skill Stratified by Altitude

Focusing on the results when verifying by PIREPs, the difference in PODy values between GTG2 and GTG2.5 decreases with decreasing altitude layer (**Figure 5.9**; squares). At the same time, the difference in PODn values does not change between the High and Middle layer, but

decreases between the Middle and Low layers. Consequently, the gap in skill between the two algorithms decreases in the lower layers. In fact, the skill of the two algorithms is identical in the Middle layer. Remember that while the volume of the GTG2.5 forecast more than doubles from the High to the Middle layer, the volume of the Low layer is larger still, so this increase alone cannot explain the improvement in the GTG2.5 forecast relative to GTG2. For results when verified by EDR, the difference in skill between the two algorithms also decreases between the High and Middle layers. However, in contrast to the PIREP scores, the decrease comes not from a relative improvement in the PODy, but from a near doubling of GTG2.5's superior PODn values. (Recall that no EDR observations below 20 kft are used for this report; therefore there are no EDR scores for the Low layer.)

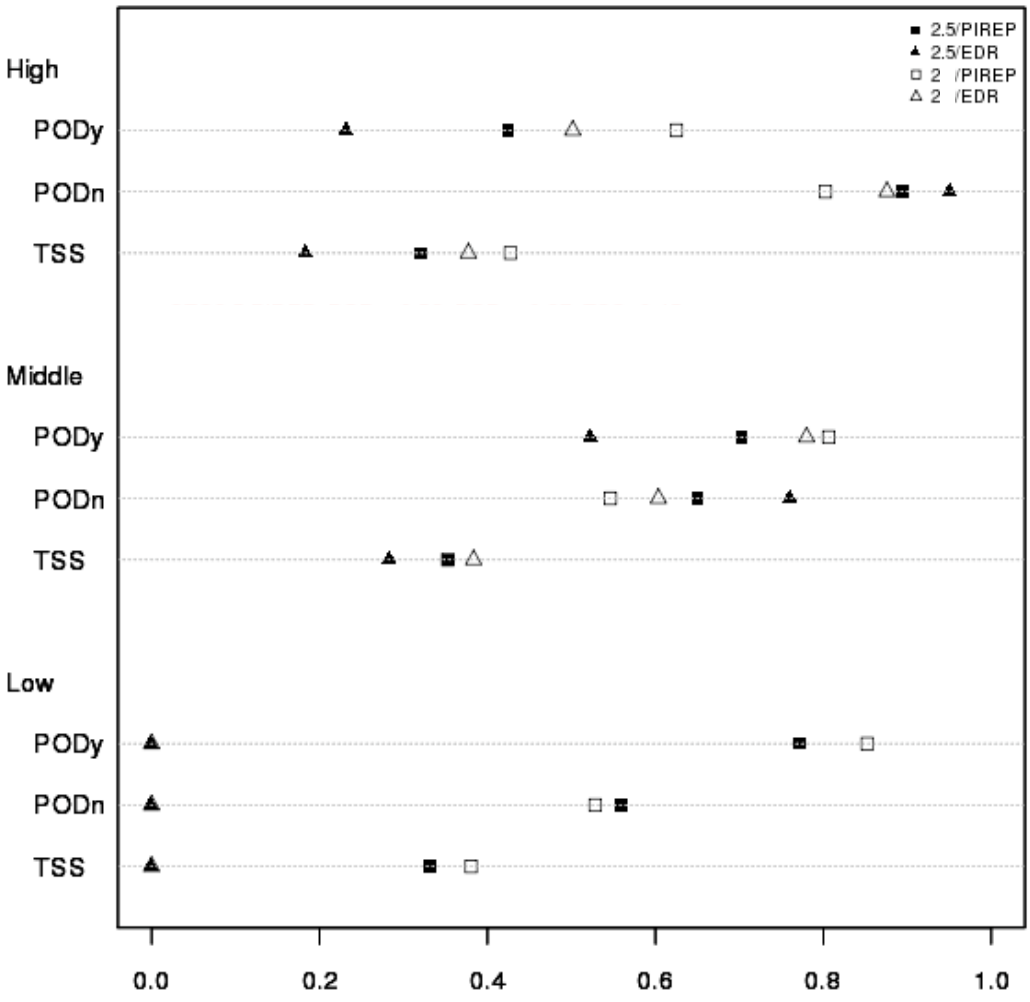

Figure 5.9: As in **Figure 5.2**, but for GTG2.5 (filled) and GTG2 (hollow) forecasts. Note that EDR forecasts are not used below 20 kft, so there are no EDR scores for the Low region.

5.2.5 Event Length Analysis

Recall from Section 4.6 that forecast event lengths can be estimated by taking actual flight paths and replacing the observed EDR values with forecast values from the same time and location. The same method of converting individual EDR reports into EDR events can in turn be applied to forecast values. The distribution of the lengths of the forecast and observed EDR turbulence events (i.e., patches of non-null turbulence) can then be compared as a measure of forecast bias (**Figure 5.10**). Both GTG2.5 and GTG2 produce turbulence events that are considerably larger than those observed, but GTG2.5 makes progress in shrinking the size of its turbulence patches. For example, the 75th percentile is found at 17km, meaning that 75% of all observed turbulence events are shorter than 17 km (nearly 60% of all observed turbulence events consist of a only a single report). By contrast, for GTG2.5 the 75th percentile is 183 km, and for GTG2 the 75th percentile is found at 229 km. So, whereas GTG2.5 forecast turbulence objects are still much larger than those observed, they are a marked improvement over GTG2. It is important to note that it is likely not desirable for the forecast distribution to exactly match the observed distribution. With more than half of all events containing just a single report, a forecast display with the same distribution would consist of a large number of single pixel events making it difficult for forecast users to interpret the display.

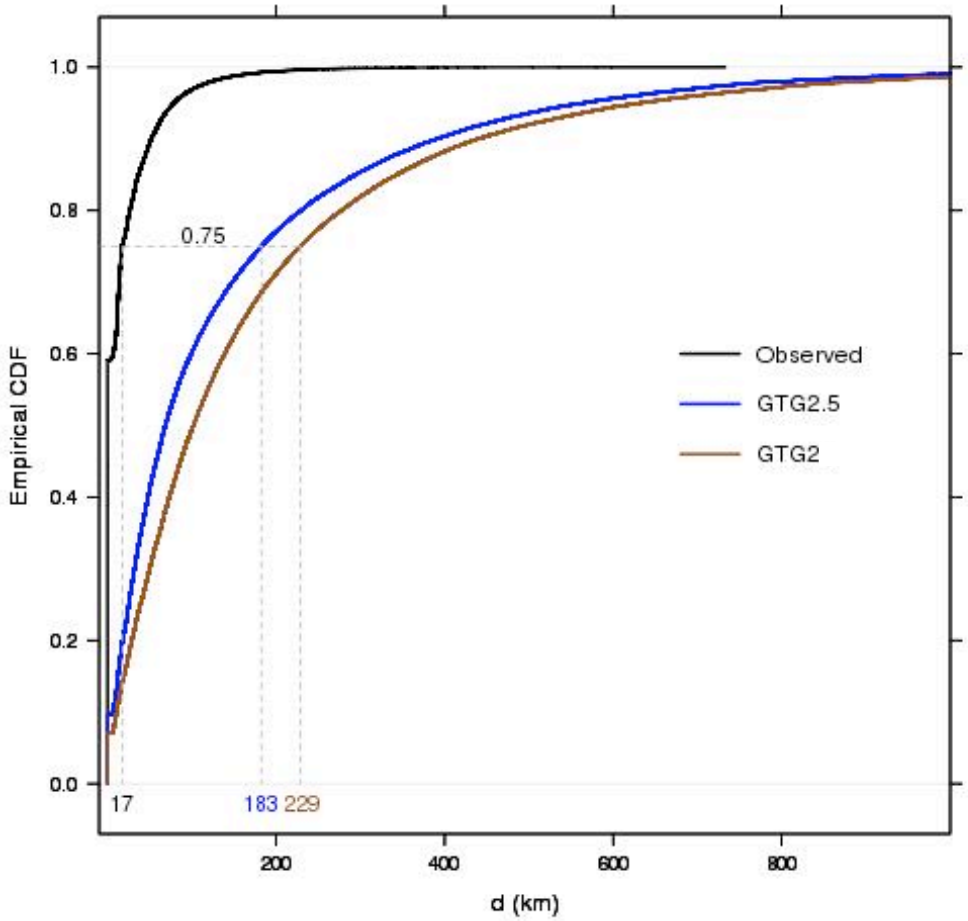

Figure 5.10: Cumulative distribution functions of observed (black) and forecast (GTG2.5 blue, GTG2 brown) event lengths. Dashed lines and numbers show the 75th percentile event lengths. See text for description of how the event lengths are calculated.

5.2.6 Conditional Probabilities

It is instructive to examine not only how often a forecast is able to match the intensity of the observation with which it is paired (i.e. PODy), but also to understand the relative frequency at which each forecast intensity is matched to a particular observed intensity, i.e., the conditional probability. In other words, for a given observed intensity, what is the distribution of the intensities from the matching forecasts? This information is given greater context with the inclusion of the marginal probabilities: how often each category occurs overall in the forecasts and observations, respectively.

Comparing the change in marginal probabilities for the GTG2 and GTG2.5 forecasts (**Figure 5.11**), one sees a substantial shift in forecast intensities from GTG2 to GTG2.5, from the MOD category to the LGT category for PIREPs (top panel) and from the MOD to the LGT and NULL categories for EDR. (It is important to keep in mind that the forecast marginal probabilities are

based on the maximum forecast value within the neighborhood of the observations and so do not reflect the distribution of forecast intensities as a whole.) This shift of intensities away from the MOD category brings the proportion of MOD forecasts more in line with that observed (in the case of PIREPs, exactly so). However, it also leads to a stronger overforecast of LGT turbulence, while the underforecast of NULL events remains. The large overall quantity of forecast turbulent events results in NULL events being matched with non-NULL forecasts.

The impact of this shift in the marginal probabilities from GTG2 to GTG2.5 is an increase in the conditional probabilities of LGT forecasts for each observed intensity category and a corresponding decrease in the conditional probability of MOD forecasts. As a result, there is an increase in the misclassification of MOD events in GTG2.5: only 30% of MOD PIREP events are paired with less-than-MOD GTG2 forecast, but almost 50% of MOD PIREP events are paired with less-than-MOD GTG2.5 forecasts. Similarly, the NULL, MOD, and SVR observed categories are each more likely to be paired with a LGT forecast in GTG2.5 than in GTG2. For EDR, the result of the shift in the forecast distribution is that MOD EDR events are now more likely to be paired with LGT forecasts than MOD forecasts.

The conditional and marginal probabilities reveal the challenge inherent in forecasting turbulence. Adjusting the forecast thresholds can change the marginal probabilities, but this will not necessarily lead to better pairings of forecasts and observations.

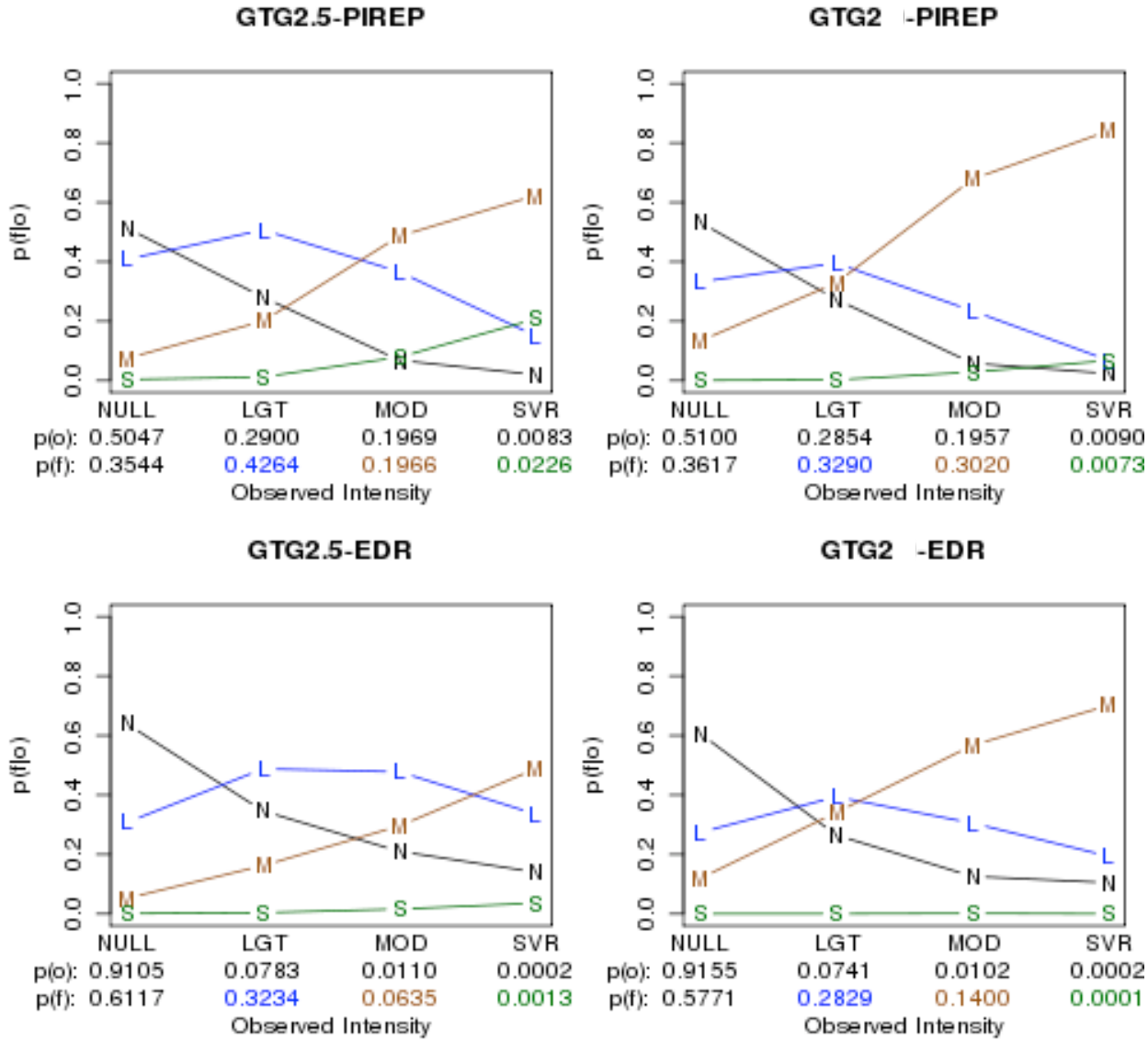

Figure 5.11: Conditional probabilities of forecast categories (Null – black, Light – blue, Moderate – brown, Severe – green) matched with each category of observation intensity for GTG2.5 (left) and GTG2 (right) when verified by PIREPs (top) and EDR (bottom). Numbers below each panel display the marginal probabilities for each intensity category for observations (p(o)) and forecasts (p(f)).

5.2.7 Severity Discrimination

As mentioned previously, ROC curves measure the separation, or distance, between a pair of distributions. For the turbulence forecasts there are four such distributions according to severity category: the distribution of forecast values associated with NULL, LGT, MOD, and SVR events, respectively. (The conditional probabilities above are discretized representation of these distributions.) There are six possible pairings of these four distributions leading to the six ROC curves shown in **Figure 5.12**. Each curve shows the ability of the GTG algorithms to

distinguish between the two types of observed events. For example, the topmost panel shows how well the algorithms can distinguish between NULL and LGT events. The rightmost panel shows the ability to distinguish between MOD and SVR events. The curves along the main diagonal are for pairings of adjacent categories (NULL-LGT, LGT-MOD, MOD-SVR), pairings that are more difficult to separate. Therefore the ROC areas are expected to be lower for these curves.

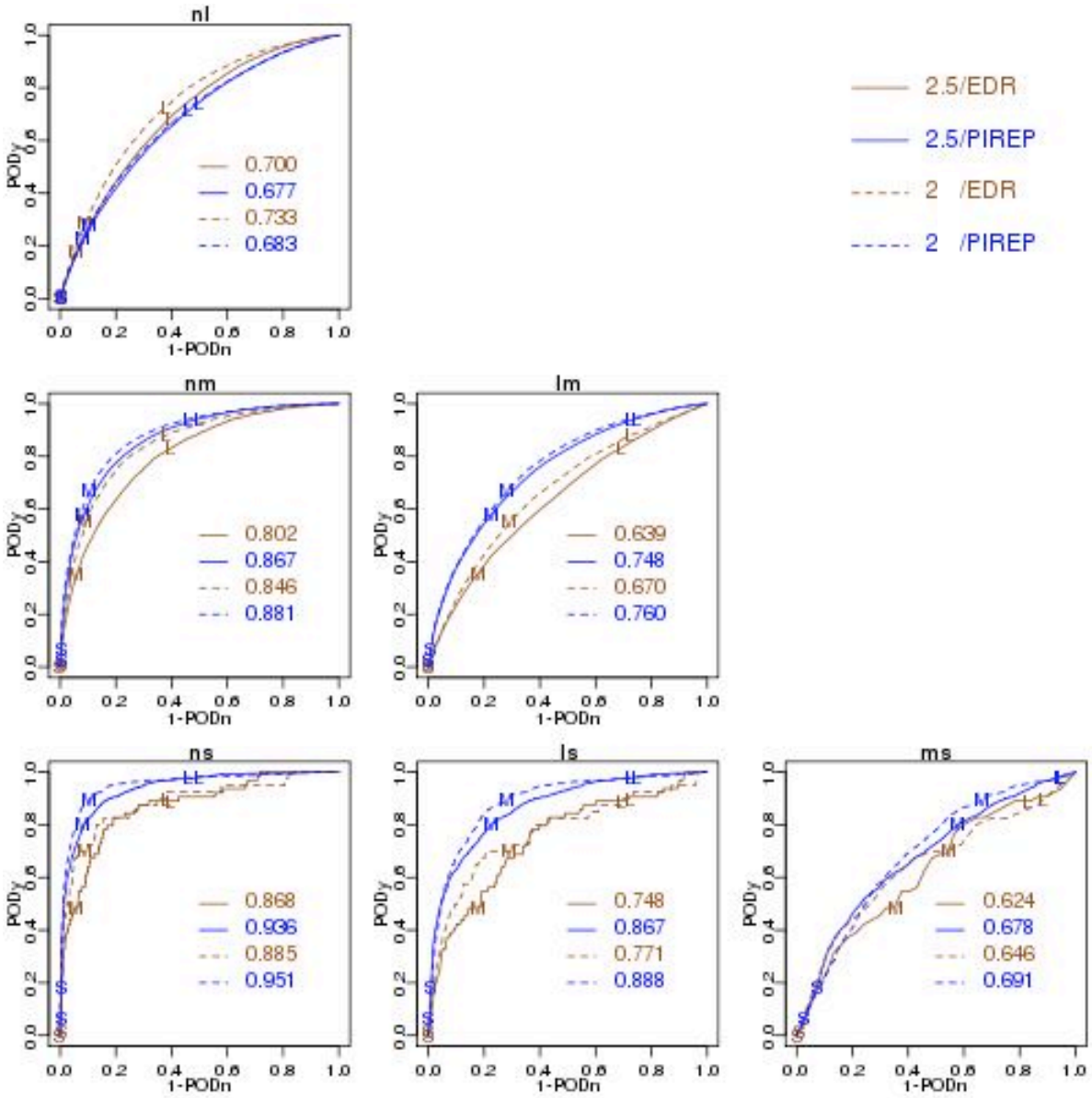

Figure 5.12: As in **Figure 5.4**, but for pairwise comparisons of each observed intensity category. Labels above each panel display to two categories being compared (nl – Null and Light, nm – Nul and Moderate, lm – Light and Moderate, ns – Null and Severe, ls – Light and Severe, ms – Moderate and Severe).

Specifically for PIREP results, among the pairings of adjacent categories, the algorithms are best able to distinguish between LGT and MOD observed events, as demonstrated by the AUC. For the pairings of distributions two categories apart the algorithms are equally capable of distinguishing between NULL and MOD as between LGT and SVR.

For EDR events, the results are considerably different. The algorithms are much better at distinguishing between NULL and LGT events than between LGT and MOD or MOD and SVR events. Similarly, both algorithms are much better at distinguishing between NULL and MOD than between LGT and SVR events.

5.2.8 GTG2.5 and GTG2 Summary

In summary, as verified by PIREPs, for nearly all verification scores, GTG2 outperforms GTG2.5 on forecasts of MOG turbulence. Similarly, GTG2 outperforms GTG2.5 above 30 kft, but the difference in algorithm performance is nearly absent below 30 kft.

The GTG2.5 analysis volumes are nearly half that of GTG2.0 and remain steady with lead time, whereas GTG2 volumes nearly double between the analysis and 12-h lead time (from ~4% to ~8%). Put differently, GTG2.5 trades an increase in misses in exchange for fewer false alarms. While both algorithms produce turbulence events of much greater length than those observed, the events in the GTG2.5 forecasts are substantially smaller than the events in the GTG2 forecasts.

When broadening the focus to examine each of the event categories, the effect of the smaller forecast volumes in GTG2.5 relative to GTG2 is to alter the marginal probabilities of the forecasts such that they more closely resemble the marginal probabilities of the observations. However, this shift in the distribution of forecast intensities does not result in conditional probabilities. That is, while the overall distribution of the forecasts more closely resembles that of the observations, the one-to-one matching of forecast and observed intensities is not as good with GTG2.5 as it is with GTG2.

As noted in the previous section, all of these results are sensitive to the choice of forecast threshold. Lowering the forecast threshold can increase the skill of GTG2.5 but, at the same time, may eliminate the improvement in the (smaller) size of the turbulence events in GTG2.5 relative to GTG2.

5.3 GTG2.5 and (G-)AIRMET

The comparison of the GTG2.5 and (G-)AIRMET forecasts consists of two parts: a direct comparison (i.e., is GTG2.5 better than (G-)AIRMET?) and an examination of GTG2.5's performance as a supplement to the (G-)AIRMET forecasts. Since the (G-)AIRMET polygons are, by definition, forecasts of MOG turbulence only observed MOG events will be considered.

5.3.1 Performance Comparison of GTG2.5 to (G-)AIRMET

In contrast to the charts in the previous sections, the three stratifications in **Figure 5.13** represent variations in the forecast threshold only. For example, the top section shows the performance when using the GTG2.5 LGT threshold to forecast MOG events. The (G-)AIRMET values are necessarily identical for each stratification and are repeated only for ease of comparison.

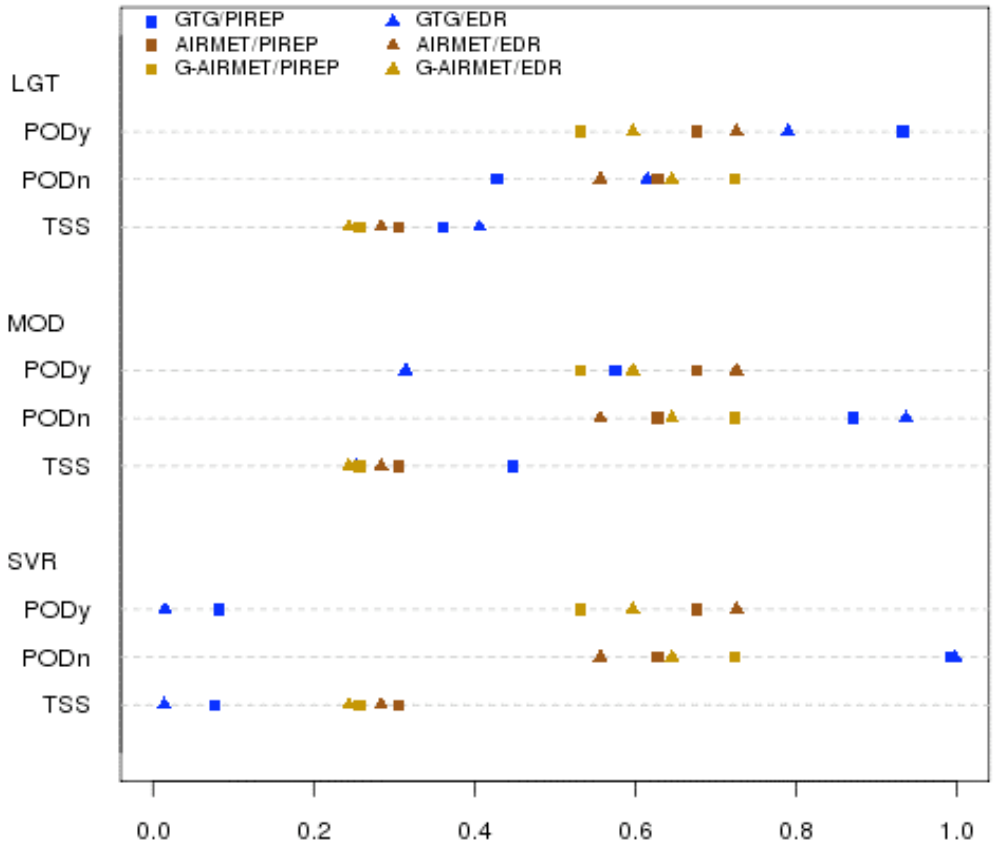

Figure 5.13: Performance measures for forecasts of MOG turbulence events for GTG2.5 (blue), AIRMETs (brown), and G-AIRMETs (yellow), stratified by GTG2.5 forecast threshold. (AIRMET and G-AIRMET forecasts are identical for each stratification and are repeated for ease of comparison.)

Using the either the LGT or MOD threshold, GTG2.5 forecasts are more skillful than the (G-)AIRMETs, but for different reasons. For the LGT threshold, the superior skill comes from substantially higher PODy but somewhat lower PODn (for PIREP events). The higher PODy is obtained despite a slightly smaller forecast volume for the GTG2.5 LGT forecasts than for the (G-)AIRMETs (**Figure 5.14**). For the MOD threshold, the relationship is reversed: GTG2.5 is now

somewhat more likely to miss events (lower PODy, though the PODy remains slightly higher for GTG2.5 than for G-AIRMETs), but is much less likely to false alarm (higher PODn). In addition, the volume of the GTG2.5 MOD forecasts is smaller than the volume of the (G-)AIRMETs by nearly a factor of ten. Choosing a GTG2.5 threshold value somewhere between the LGT and MOD values, one could achieve higher PODy and still retain a much smaller forecast volume.

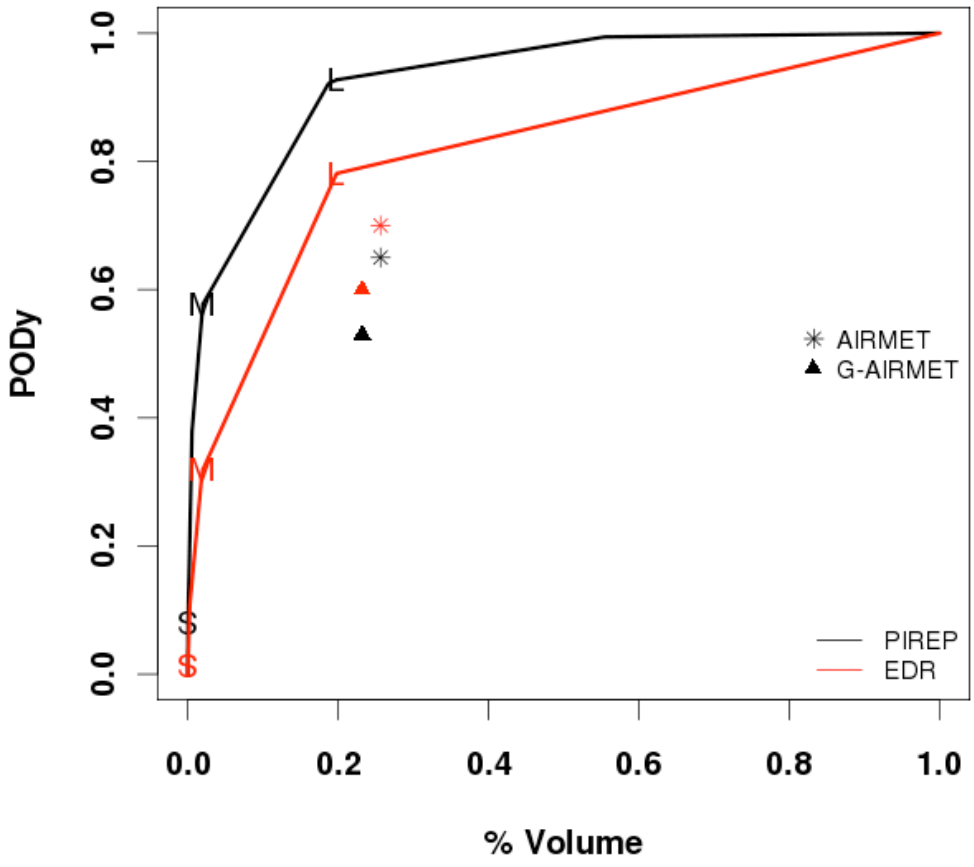

Figure 5.14: PODy as a function of the %Volume of the forecasts for GTG2.5 (lines), AIRMETs (stars), and G-AIRMETs (triangles) when verified by PIREPs (black) and EDR (red). Letters mark the performance for the Light (L), Moderate (M), and Severe (S) GTG2.5 thresholds.

5.3.2 GTG2.5 Supplemental Analysis

To determine the performance of GTG2.5, the GTG2.5 forecasts were evaluated separately for two scenarios: 1) GTG located inside an (G-)AIRMET forecast, and 2) GTG located outside of an (G-)AIRMET forecast. As described in Section 4.7, the definition of a successful supplemental forecast differs for each scenario, so performance for each is considered separately.

Inside an AIRMET, where turbulence is already expected, a successful supplemental forecast will reduce the forecast volume as much as possible, thus opening up the airspace, while still capturing most of the observed turbulence. Put in terms of the scores, when inside the AIRMET, the supplemental forecast will have a high PODn, while maintaining a high PODy.

Outside the AIRMET, where large areas of turbulence are not expected, the goal of the supplemental forecast is reversed. The successful forecast will capture as many of the events missed by the AIRMET as possible without unduly increasing the forecast volume, thereby further restricting the airspace. Again, in terms of the scores, the supplemental forecast, outside the AIRMET, will have a reasonable PODy while maintaining a high PODn.

Because the composition of the observations inside and outside the AIRMETs is so different, the PODy and PODn values inside the AIRMET are not compared with the scores outside the AIRMETs. A comparison would provide little or no benefit: if the AIRMETs are successful at distinguishing between areas of substantial turbulence and areas without, then even random forecast should have a PODy higher inside the AIRMET and a higher PODn outside the AIRMET.

Inside the (G-)AIRMETs, GTG2.5 achieves a PODn of around 0.8 while maintaining a PODy of around 0.7 (**Figure 5.15**). **Figure 5.16** is a schematic that pictorially represents the overall results. The GTG2.5 forecast substantially opens up the airspace (the GTG2.5 forecast contains only 6% of the total AIRMET volume) but misses about 30% of all MOG turbulence events. Outside, the (G)-AIRMETs, GTG2.5 is able to capture about 40% of the turbulence events while adding less than 5% to the false alarms. The GTG2.5 forecasts contain only about 1% of the airspace outside of an (G-)AIRMET. In other words, GTG2.5 has reduced the number of misses with only a minor increase in restricted airspace.

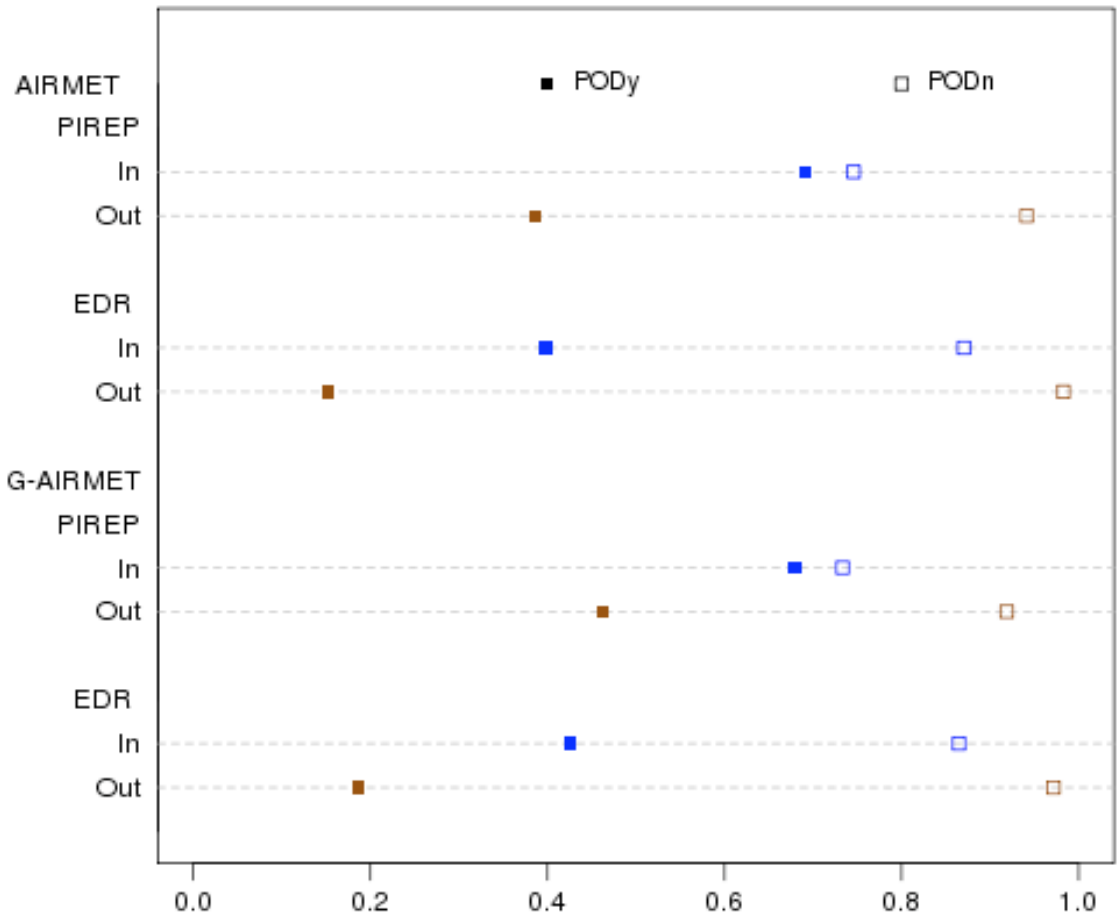

Figure 5.15: Performance of GTG2.5 forecasts (PODy, filled square; PODn, open square) inside (blue) and outside (brown) of AIRMETs, stratified by AIRMET type (AIRMET or G-AIRMET) and observation type (PIREP or EDR).

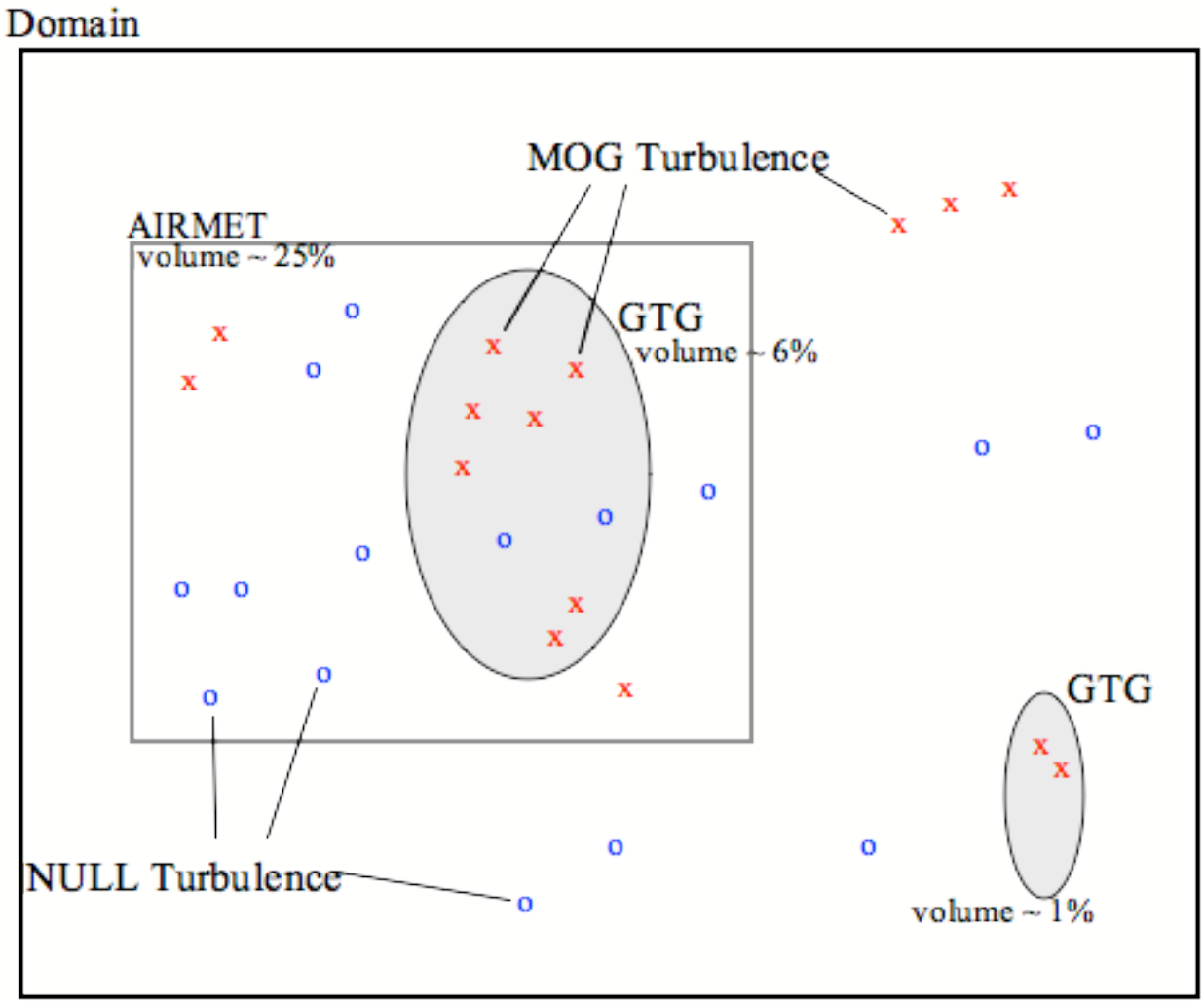

Figure 5.16: Schematic of the supplemental forecast. Large box represents the forecast domain. Small box represents an AIRMET forecast polygon. Gray filled ovals represent GTG forecasts of MOG turbulence. Red 'x's represent observations of MOG turbulence, while blue 'o's represent observations of Null turbulence. All proportions are set to approximate the scores in **Figure 5.15**.

6. Concluding Statements

GTG2.5 is considered an interim product, occasioned by the forthcoming change in the numerical weather prediction model on which the GTG algorithm is based from the RUC model to the Rapid Refresh model. However, as a result of the changes in the underlying model (e.g., the vertical coordinate, physics packages, etc.), along with the algorithm's change from predicting turbulence potential to predicting eddy dissipation rate, GTG2.5 is a significantly different algorithm than the current operational GTG product.

Using the forecast thresholds as currently set for the ADDS display, GTG2.5 forecasts turbulence less frequently than GTG2—the volume of the GTG2.5 forecasts is less than one-half the volume of the GTG2 forecasts. As a result, GTG2.5 captures fewer events, i.e., there are more forecast misses, particularly in the layer of the atmosphere containing aircraft cruising altitudes. GTG2.5 volumes are particularly small in this layer.

Compared to AIRMETs and G-AIRMETs, GTG2.5 captures roughly the same proportion of moderate-or-greater turbulence events, despite a forecast volume many times smaller than the AIRMET and G-AIRMET volumes. Viewed as a supplemental forecast, GTG2.5 is able to substantially open up the airspace within an AIRMET while still capturing 70% of MOG events contained within the AIRMET. Outside the AIRMET, GTG2.5 is somewhat less successful, but still adds value. The algorithm avoids any notable increase in false alarms, but reduces the number of missed turbulence events by less than half.

All of these results are sensitive to the choice of forecast thresholds. The GTG2.5 threshold for moderate turbulence is set to yield fewer forecasts of moderate-or-greater turbulence than that produced by GTG2. In effect, using the current forecast thresholds involves an implicit choice to have fewer false alarms at the cost of an increase in missed events. Judicious selection of the forecast threshold permits the user to optimize this tradeoff between increasing the airspace available to aircraft and increasing their risk of encountering unexpected strong turbulence.

References

Aviation Weather Center, 2010: Product Description Document: Graphical Airman's Meteorological Advisory (G-AIRMET). 3pp. http://www.aviationweather.gov/products/gairmet/docs/G_AIRMETPDD_2010.pdf .

Benjamin, S.G., D. Devenyi, S.S. Weygandt, K.J. Brundage, J.M. Brown, G.A. Grell, D. Kim, B.E. Schwartz, T.G. Smirnova, T.L. Smith, and G.S. Manikin, 2004: An hourly assimilation/forecast cycle: The RUC**.** *Mon. Wea. Rev.,* **132,** 495-518.

Benjamin, S.G., D. Devenyi, R. Smirnova, S. Weygandt, J.M. Brown, S. Peckham, K. Brundage, R.L. Smith, G.Grell, and T. Schlatter, 2006: From the 13-km RUC to the Rapid Refresh. *12th Conf. on Aviation, Range, and Aerospace Meteorology,* Atlanta, GA, Amer. Meteor. Soc., CD-ROM. 9.1.

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. *Wea. Forecasting*, **12**, 890-914.

Cornman, L.B., G. Merymaris, and M. Limber, 2004: An update on the FAA Aviation Weather Research Program's *in situ* turbulence measurement and report system. Preprints. *11th Conf. on Aviation, Range, and Aerospace Meteorology,* Hyannis, MA, Amer. Meteor. Soc. CD-ROM, P4.3.

Federal Aviation Administration, 2007: *Aviation Weather Services Handbook* (section 3).

Federal Aviation Administration, 2011: *Aeronautical Information Manual* (section 7-1-7). http://www.faa.gov/air_traffic/publications/ATpubs/AIM.

Kane, T.L., B.G. Brown, and R. Bruintjes, 1998: Characteristics of pilots reports of icing. Preprints, *14th Conf. on Probability and Statistics*, Phoenix, p.90-95.

Schwartz, B., 1996: The quantitative use of PIREPs in developing aviation weather guidance products. *Wea. Forecasting,* **11,** 372-384.

Sharman, C. Tebaldi, G. Wiener, and J. Wolff, 2006: An integrated approach to mid- and upper-level turbulence forecasting. *Wea. Forecasting,* **21,** 268-287.

Wandishin, M.S., B.P. Pettegrew, M.A. Petty, and J.L. Mahoney, 2010: EDR for use in turbulence verification. Prepared for the FAA Aviation Weather Research Program.

Appendix: Climatological Analysis of Observation Data Sets

I) Eddy Dissipation Rate (EDR)

While PIREPs, in the past, have been the observation of choice in turbulence verification, the Eddy Dissipation Rate (EDR), a turbulence measure taken directly from the reporting aircraft in-situ, is also assimilated into the GTG2.5 algorithm and will play a prominent role in the assessment of turbulence forecasts. The benefit of this observation is the removal of subjectivity from the recording of turbulence events. The reporting of turbulence from commercial aircraft uses existing sensors, computational capabilities, and the ACARS communication network to derive and transmit a quantitative measure of vertical accelerations every minute that are calibrated for the type of aircraft.

There are over 200 United Airlines planes flying with this system on board. According to Cornman et al. (2004), automatic EDR observations are intended to augment (and someday possibly replace) PIREPs. For this report, EDR measurements are only analyzed from flights operated by United Airlines (UAL) B737 and B757 aircraft. Each aircraft's EDR data contains both median and peak EDR values. Cornman et al. (2004) state that retrieved data pass through two quality control filters, one on board the aircraft, and one applied by ground systems. EDR data are binned every 0.10 m2/3 s-1 starting at 0.05 m2/3 s-1. Observations from Delta Airlines (DAL) are currently operational and in the data stream, but due to a differing reporting system and limited study time, they were not used.

While this form of sampling is objective, there are characteristics of the reporting system that must be accounted for when using these observations as the basis for verification. The equipment measuring EDR in commercial aircraft are calibrated for the size and weight of the aircraft they are on, which, in the UAL fleet are 737 and 757's. This provides a measure of consistency and objectivity in the intensity and location reporting of the observed turbulence. However, there are large reporting gaps due to flight routes taken by the available airlines. UAL covers a large area of the center of the U.S. while the newly operational DAL EDR observations cover portions of the gap in the southeastern U.S. left by UAL flights (not shown). Other shortcomings include the oversampling, especially by systems on the UAL flights. Observations from the UAL dataset are reported every minute and binned by intensity level groups. Given that turbulence is a rare event, as seen in the climatology, over 98% of reports are NULL reports.

Since all EDR observations used in this evaluation are from commercial UAL aircraft, the predominant layer in which observations occur is between 35,000 and 40,000 ft. Due to equipment sensitivity during ascent/descent stages of flight, EDR observations beneath 20,000 ft are not utilized. According to Cornman et al. (2004), due to more frequent reporting during rapid ascent/descent (i.e., at altitudes below 20,000 ft), it is not unusual to see unexpected high values of turbulence reported. Many of these are flagged as bad reports due to the report time being so short that an accurate peak/median value of EDR cannot be obtained. Regionally, the UAL aircraft tend to cover a large majority of the central coast-to-coast regions of the U.S.

leaving gaps in the northwest, southwest, and southeast U.S. With time and new airlines providing observations, these coverage gaps will slowly diminish. Areas such as the north central U.S. will likely always have coverage gaps in EDR observations due to lack of commercial aircraft traffic generally found in that region. In general, EDR observations favor sampling in the western and central U.S. This may be in large part due to the location of UAL hubs and general air traffic. This pattern follows when looking explicitly at the occurrence of MOG observations. However, when looking at MOG observations as a percentage of the total observation set, just as with PIREPs, the western and eastern U.S. show a greater density of MOG turbulence. Temporally, the greatest density of EDR observations, similarly to PIREPs, is between 1200 UTC and 0500 UTC, which follows the general flow of air traffic density.

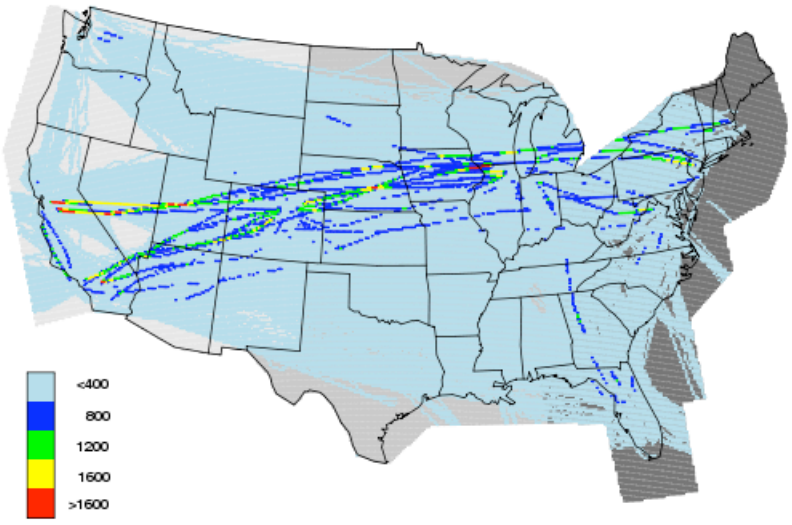

Figure A.1: Spatial distribution of EDR reports for the period assessment period (1 December 2010 – 31 March 2011).

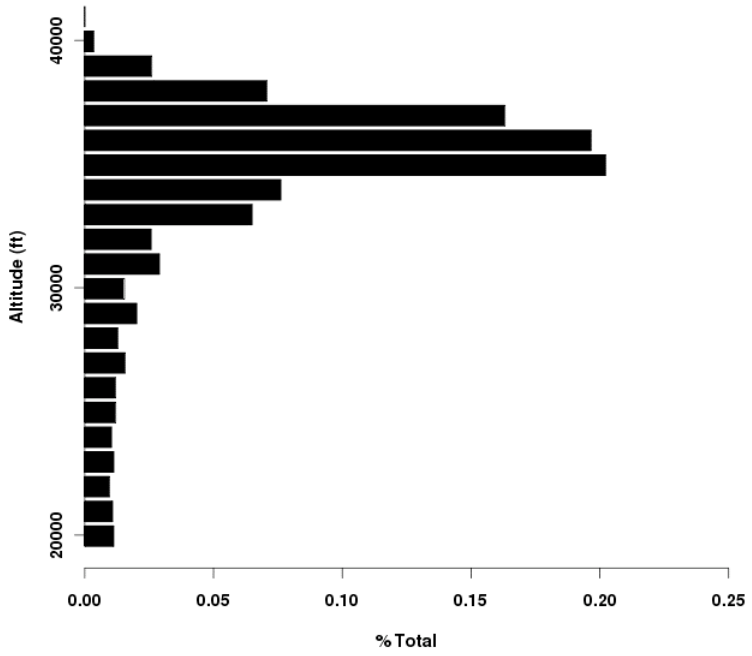

Figure A.2: Distribution of EDR reports by 1000 ft altitude layer for the assessment period.

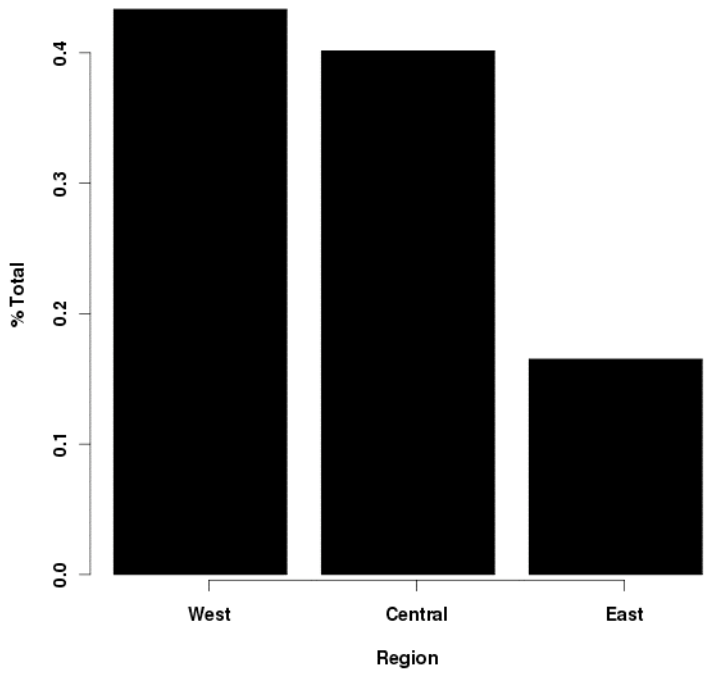

Figure A.3: Distribution of EDR reports by region. See **Figure 2.1** for map of regions.

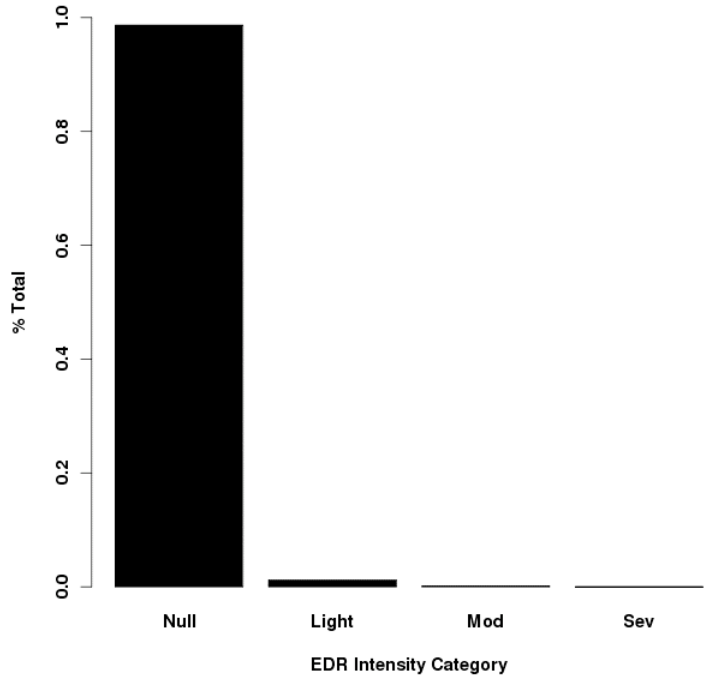

Figure A.4: Distribution of EDR by intensity category. See **Table 4.1** for category definitions.

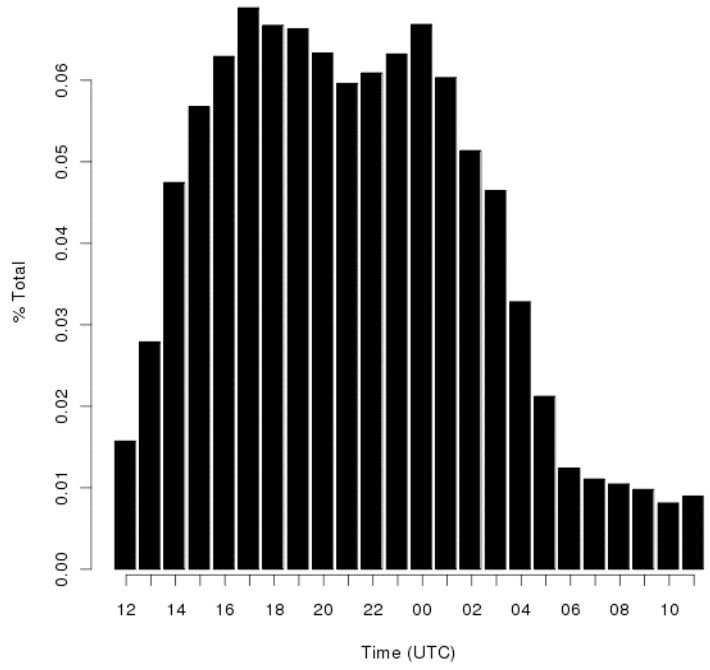

Figure A.5: Distribution of EDR reports by time of day.

II) PIREPs

National Weather Service Pilot Reports (PIREPs) are a report of aviation hazards disseminated by pilots during the flight to describe in-flight weather phenomena such as turbulence, icing, sky conditions, or even clear-air and smooth flight conditions. These observations report, specifically for this instance, the hazard, the intensity of the hazard, the latitude and longitude location of the hazard occurrence and the altitude above sea level of the hazard (FAA 2007). For years PIREPs were really the only source of routine, upper-air turbulence reports. With the advent of measures taken directly from aircraft aloft, PIREPs are no longer the single most reliable source of such observations. Advantages to PIREPs are that they are event based, meaning that a report is triggered when something does or does not occur. However, there are numerous limitations to the use of PIREPs for verification of turbulence. Most notably is the erratic nature of PIREP issuances and reporting in space and time, and the subjectivity of pilots submitting the reports (Kane et al. 1998 and Schwartz 1996). It has also been shown that PIREPs have a bias towards positive event issuance (Brown et al. 1997), which significantly reduces the ability to use dichotomous statistics to measure the skill in forecasting this phenomenon. Further errors in pilot reporting are associated with inconsistent use of acronyms and spellings for different event types.

A brief climatology of PIREP observations performed on data during the evaluation period of December 2010 through March 2011 shows the greatest density PIREPs reported between 1200 and 0400 UTC. Furthermore, PIREPs have two altitude ranges of prominence, one representing traffic in the general aviation sector (0-10,000 ft) and the primary one between 35,000 and 40,000 ft where heavy commercial air traffic exists. PIREPs are randomly sampled throughout the CONUS. Regionally, however, the sampling favors most states in the central U.S. with a significant amount existing in the western portions as well. No connection between this sampling and the actual occurrence of turbulence is made. When looking at just MOG observations, this distribution tilts back towards the western U.S. The distribution of MOG observations as a percentage of all observations is greater in the western and eastern U.S. Thus, while the central U.S. has a greater overall density of PIREP observations, there is a greater density of moderate-or-greater PIREP observations over the western and eastern U.S. Also, a greater occurrence of MOG observations tends to occur over the western U.S., which could be attributed to the rugged terrain.

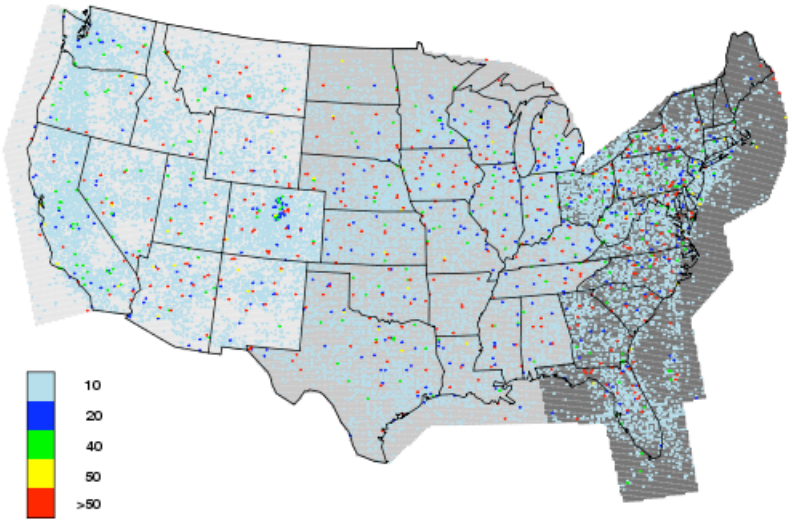

Figure A.6: As in **Figure A.1**, but for PIREPs.

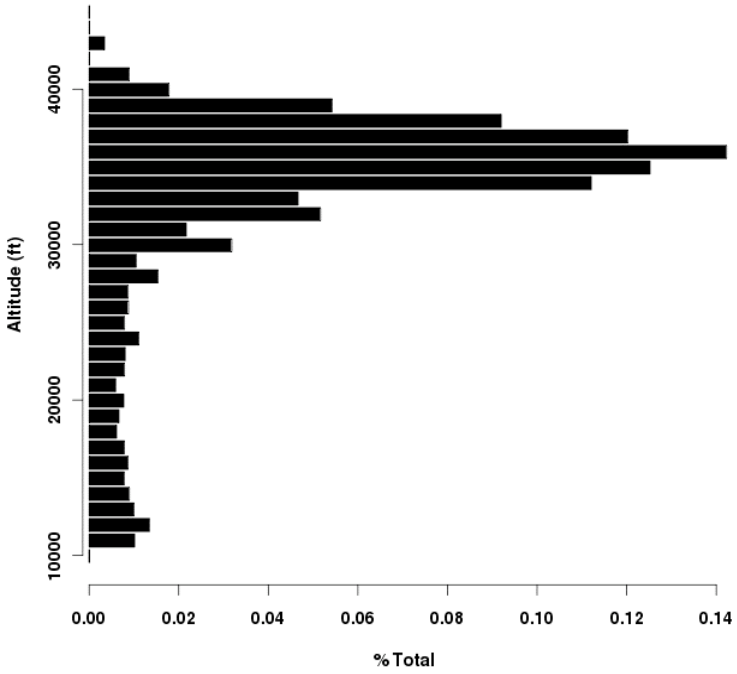

Figure A.7: As in **Figure A.2**, but for PIREPs.

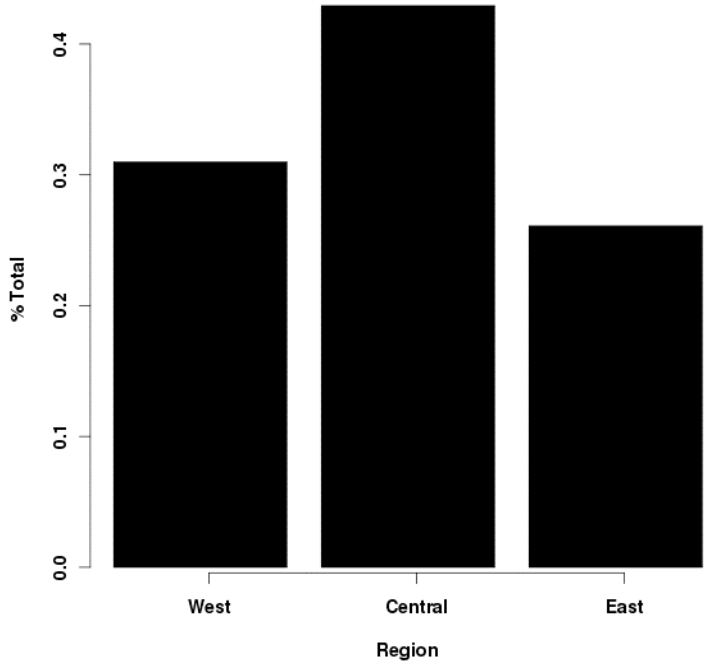

Figure A.8: As in **Figure A.3**, but for PIREPs.

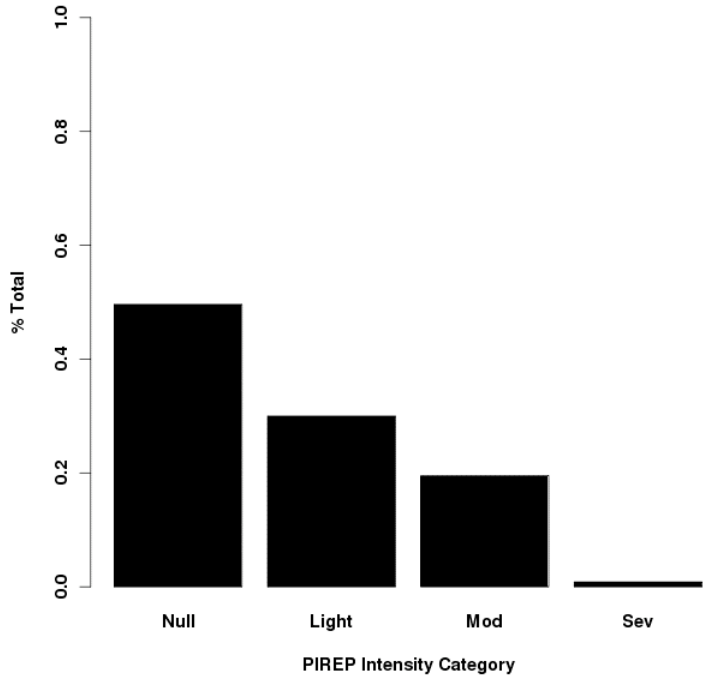

Figure A.9: As in **Figure A.4**, but for PIREPs.

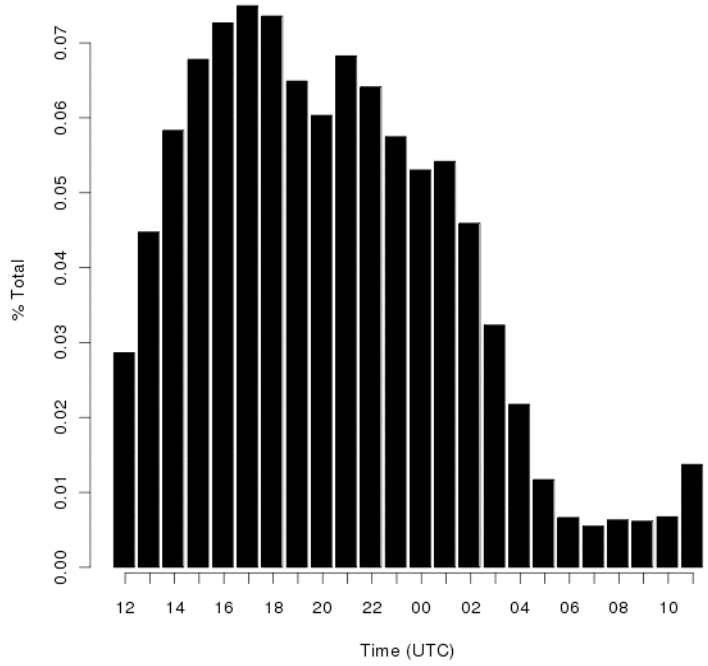

Figure A.10: As in **Figure A.5**, but for PIREPs.