NOAA Technical Memorandum OAR GSD-53

https://doi.org/10.7289/V5/TM-OAR-GSD-53

Assessment of the Graphical Turbulence Guidance, Version 3 (GTG3)

September 2014

Matthew S. Wandishin
Laura Paulik
Joan Hart
Brian Etherton
Melissa A. Petty

Earth System Research Laboratory
Global System Division
Boulder, Colorado
September 2014

noaa NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION / Office of Oceanic and Atmospheric Research

https://doi.org/10.7289/V5/TM-OAR-GSD-53

NOAA Technical Memorandum OAR GSD-53

Assessment of the Graphical Turbulence Guidance Version 3 (GTG3)

Matthew S. Wandishin[2]
Laura Paulik[2]
Joan Hart[2]
Brian Etherton[1]
Melissa A. Petty[3]

[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

Acknowledgements

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA. The authors would like to thank the Turbulence Product Development Team for providing the forecast data that was needed for the evaluation.

UNITED STATES DEPARTMENT OF COMMERCE

Wilbur Ross
Secretary

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Tim Gallaudet, Ph.D., USN Ret.
Under Secretary for Oceans
And Atmosphere/NOAA Administrator

Office of Oceanic and Atmospheric Research

Craig N. McLean
Assistant Administrator

EXECUTIVE SUMMARY

The QA PDT was tasked to assess the quality of the Graphical Turbulence Guidance, version 3, (GTG3) algorithm developed by the National Center for Atmospheric Research (NCAR). This product is to replace the current GTG2.5 algorithm currently being used for operational aviation turbulence decisions. Changes between GTG2.5 and GTG3 include: 1) an extension of the forecast domain down to 100-ft altitude (from 10,000 ft), 2) an increase in forecast leads from 12 to 18 hours, 3) the addition of an explicit mountain-wave (MW) turbulence component, and 4) an upgrade to the conversion of the raw algorithm output to the Eddy Dissipation Rate (EDR).

The assessment has five main areas of investigation and incorporates output from the operational GTG2.5 algorithms, the GTG3, and the National Weather Service (NWS)-produced Graphical Airmen's Meteorological Advisories (G-AIRMETs), as well as, PIREPs and EDR values derived from *in situ* measurements. The forecasts were analyzed using output generated from 1 January – 31 March 2013 and 1 July – 30 September 2013 over the CONUS.

Primary findings include:

- GTG3 distributions are noticeably different than the distributions for GTG2.5—the GTG3 distribution is more constrained (i.e., lower variance, weaker tails) and the peak of the distribution is shifted from near-zero values to around 0.1.
- GTG3 is consistently better at discriminating events from non-events than GTG2.5, at all observed thresholds.
- When the forecast threshold is constrained to match the observed threshold (i.e., no calibration), GTG3 is more skillful than GTG2.5 for only a small range of thresholds; however, this range can be expanded with proper calibration.
- With calibration, GTG3 outperforms GTG2.5 for events with an EDR greater than about 0.14, while GTG2.5 outperforms GTG3 for events with an EDR less than about 0.14.
- GTG3 captures more Moderate-or-greater (MOG) events than G-AIRMETs for the same forecast volume, or by choosing a different forecast threshold, GTG3 captures the same number of events as G-AIRMETs while using only one-third of the volume.
- Performance of GTG3 in the Near-surface layer (below 10,000 ft) is not as skillful as other layers, but GTG3 outperforms G-AIRMET in this layer.
- Mountain-wave component:
    - Very effective (99%) at capturing Light-or-greater intensity explicit MW PIREPs, but with a higher number of false alarms (60%).
    - Captures 70% of Moderate-or-greater MW PIREPs, with very few false alarms (6%).
    - Using all reports (MW and others), forecasts using the clear-air (CAT) algorithm with the MW component) are equally skillful as forecasts using only the CAT algorithm.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The QA PDT was tasked to assess the quality of the Graphical Turbulence Guidance, version 3, (GTG3) algorithm developed by the National Center for Atmospheric Research (NCAR). This product is to replace the current GTG2.5 algorithm currently being used for operational aviation turbulence decisions. Changes between GTG2.5 and GTG3 include: 1.) an extension of the forecast domain down to 100-ft altitude (from 10,000 ft), 2.) an increase in forecast leads from 12 to 18 hours, 3.) the addition of an explicit mountain-wave (MW) turbulence component, and 4.) an upgrade to the conversion of the raw algorithm output to Eddy Dissipation Rate (EDR).

The assessment incorporates output from the operational GTG2.5 algorithm, GTG3, and the National Weather Service (NWS)-produced Graphical Airmen's Meteorological Advisories (G-AIRMETS) as well as PIREPs and EDR values derived from *in situ* measurements, to establish a performance baseline, and has five main areas of investigation:

1. Forecast and observation distributions
2. Overall performance and accuracy of the GTG3
3. Comparison of GTG3 with GTG2.5
4. Comparison of GTG3 with the G-AIRMET forecasts
5. Performance of the mountain-wave component of GTG3

The results and conclusions obtained from the QA PDT assessment will be provided to a Technical Review Panel as input to the decision on whether the GTG3 algorithm is ready for transition to operations at the NWS.

2 DATA

This section describes the forecast and observation data that were included in the assessment, along with the principal stratifications that were used. The time period for this study consists of a winter period, 1 January – 31 March 2013 (JFM), and a summer period, 1 July – 30 September of 2013 (JAS).

2.1 FORECASTS

2.1.1 GTG3

The spatial and temporal attributes of the GTG3, GTG2.5, and G-AIRMETs are outlined below.

Table 2.1: Attributes of the GTG3, GTG2.5, and G-AIRMETs.

	GTG2.5	GTG3	G-AIRMET
Issues	Every hour	Every hour	3,9,15,21
Leads	0,1,2,3,6,9,12	1,2,3,6,9,12,15,18 (JFM) 0-9,12,15,18 (JAS)	0,3,6,9,12
Altitudes	10,000–45,000ft, 1000-ft increments	100ft; 1000–50,000ft, 1000-ft increments	0–45,000 ft

2.1.2 G-AIRMET

The G-AIRMET is represented in the Binary Universal Form for the Representation (BUFR) of meteorological data, formatted in a time-series depiction of aviation hazards occurring with occasional or greater frequency throughout the conterminous U.S. and adjacent coastal waters (Murphy, 2010), and is a forecast for moderate or greater turbulence covering an area of at least 3000 mi^2. The G-AIRMET is issued four times per day (0300, 0900, 1500, and 2100 UTC) with forecast leads every 3 hours out to 12 h and from altitudes at the surface to 45,000 ft.

GTG is a gridded product whereas the G-AIRMETs are human-generated polygons. The mechanics and approaches will account for these forecast differences. Additionally, G-AIRMETs include amendments and corrections. Amendments to the G-AIRMETs were excluded from this evaluation.

2.2 OBSERVATIONS

2.2.1 VOICE PILOT REPORTS (PIREPS)

PIREPs are reported irregularly at the pilot's discretion and include a subjective assessment of many meteorological variables including the existence/absence of turbulence and a subjective measure of the turbulence intensity. Included in the turbulence reports are the location, altitude or range of altitudes, type of aircraft, air temperature, and intensity of turbulence (NWS 2007). Additionally, PIREPs include optional pilot remarks that are sometimes used to identify the source of the encountered turbulence, e.g., mountain waves.

2.2.2 *IN SITU* MEASUREMENTS

EDR is the International Civil Aviation Organization (ICAO) standard for automated reporting of turbulence from commercial aircraft. The values are derived from *in situ* measurements from a number of United Airlines (UAL) 737 and 757 and Delta Airlines (DAL) 737 and 767 aircraft. The derivation and reporting methods are different between the two airlines.

For the UAL aircraft, on-board equipment measures and reports vertical accelerations of the aircraft. 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 in 0.1-width bins. Due to equipment sensitivity during ascent/descent stages of flight, EDR observations below 20,000 ft are not utilized (Cornman et al.

2004).

EDR values from DAL aircraft are computed directly from the vertical wind measurements. Reports consist of "heartbeat" reports issued every 15 minutes after takeoff, and "triggered" reports, issued whenever one of the following three conditions are met:

1. A single peak EDR value > 0.18
2. Three out of six peak EDR values > 0.12
3. Four out of six mean EDR values > 0.08

Triggered reports provide the previous six minutes of EDR values, while reports triggered by either of the first two conditions also include the six minutes following the initial trigger. Between explicit reports, the aircraft location is interpolated for each minute and assigned a value of zero. All values are reported in 0.02-width bins.

2.3 STRATIFICATIONS

Performance results were stratified spatially, temporally, and according to certain turbulence intensity thresholds.

ALTITUDE BINS

Results are aggregated into the following altitude ranges:

Stratification	
Near-surface	0 – 9999 ft
Low	10000 – 19999 ft
Middle	20000 – 29999 ft
High	30000 – 50000 ft

Note that PIREPs and DAL EDR data are available for all altitude bins; UAL EDR data are usable only above 20000 ft.

TEMPORAL STRATIFICATION

Forecast performance is stratified by lead times. Also, GTG3 performance in winter months (JFM) will be compared against the performance in the summer months (JAS).

GEOGRAPHIC STRATIFICATION

GTG3 performance is examined across four geographic regions, defined as shown in the Fig. 2.1.

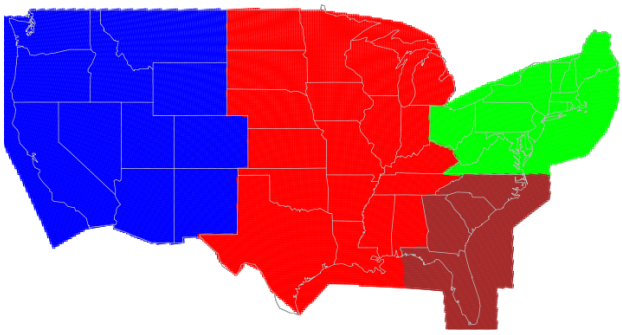

Figure 2.1: Map of the geographic regions.

INTENSITY STRATIFICATIONS

Forecast performance is also examined across a range of intensity thresholds ranging from Light to Severe turbulence (EDR values: 0.1 – 0.4, PIREP values: 1 – 5), with special attention paid to thresholds in the range of moderate turbulence (EDR: 0.18 – 0.24; PIREP: 3, "light-to-moderate").

3 APPROACH

The evaluation consists of five primary assessment areas:

1. Comparison of the distribution of values between GTG2.5 and GTG3 and between the forecasts and observations
2. Comparison of GTG3 with GTG2.5
3. Comparison of GTG3 with the G-AIRMET forecasts
4. Overall performance and accuracy of the GTG3 across the whole domain as well as in specific temporal and spatial sub-domains
5. Performance of the mountain-wave component of GTG3

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 the forecasts are paired with the observations using a neighborhood approach, described below.

4 METHODS

A variety of verification approaches are employed in this assessment. They are described in the following subsections.

4.1 GTG FIELD CHARACTERISTICS

The makeup of the GTG fields is first evaluated using value-based distributions. Distributions were generated for each forecast product: bins for GTG range from 0 to 1.0 using a bin size of 0.01.

Distributions for G-AIRMETs are not computed given that they are binary fields. Distributions for the observations are done according to the precision of the data (0.02 for DAL EDR, 0.1 for UAL EDR, severity categories for PIREPs).

4.2 Forecast-Observation Pairing Techniques

To enable forecast comparisons and evaluation of quality, forecasts and observations are matched spatially and temporally using the following mechanics.

4.2.1 Eddy Dissipation Rate (EDR)

Consecutive non-null turbulence reports are combined to form a single turbulence event; nearby turbulence events are then merged if the gap between them is five (5) minutes or less. The five-minute gap is based on FAA reporting regulations that define intermittent turbulence as turbulence occurring at least 1/3 of the time (i.e., two turbulence reports/[two turbulence + four null reports] = 1/3). Similar to the forecast neighborhood approach, the intensity of an event is defined to be the maximum-observed value within the event.

Null-events are defined as contiguous 15-minute segments in which no individual peak-EDR value exceeds the light-turbulence threshold (0.1). The 15-minute threshold is based on the ICAO requirement to report turbulence, in the absence of significant events, every 15 minutes.

4.2.2 PIREPs

Roughly 10% of reports include a non-zero depth in which turbulence was encountered; these will be treated as a single turbulence event (the height of which spans the depth of the report) in the same manner as the EDR-based turbulence events.

Differences in the reporting of turbulence from the two observation data sets requires that each set is used independently of the other. That is, statistics will be computed separately for PIREP and EDR observations.

For both EDR and PIREPs, once the observations have been consolidated into events, neighborhoods are constructed (described below) around each event. Only forecast data from grid cells lying within the observed-event neighborhoods will be considered in the evaluation.

4.2.3 Gridded-Forecast Neighborhood Approach

As in previous evaluations, a neighborhood approach is used to match forecasts and observations. 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 4.1), 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.

The forecast neighborhood consists of all forecast grid points located within a 65-km diameter centered on the observations at the model level nearest the altitude of the observations, plus one model level above and below. For events consisting of multiple reports, the neighborhood around

the event is the union of the neighborhoods around each individual observation. The 65-km horizontal neighborhood was determined from the power-spectra-derived resolution of the forecast data ($5\Delta x \approx 65$ km, where Δx is the resolution of the grid, or 13km). That is, power spectra were computed from the mid-level (between roughly 750- and 300- mb) kinetic energy fields from the RAP model analyses for the month of January 2011. The power spectra show a distinct loss of power around 65 km; i.e., the model is not capable of capturing variability at scales smaller than $5\Delta x$ due to implicit numerical diffusion in the dynamical core.

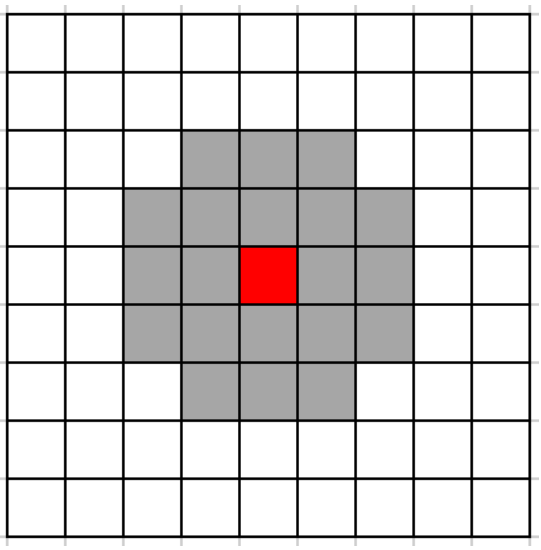

Figure 4.1: Example 65-km diameter gridded forecast neighborhood (shaded) around a given observation (red). Grid spacing is 13 km

The maximum forecast value within the neighborhood is taken to be the representative forecast intensity

4.2.4 Mountain-wave Component

Two approaches are used to evaluate the MW component of GTG3: The MW is judged directly against PIREPs explicitly associated with mountain waves in the remark section of the PIREP. The MW component also will be judged indirectly by comparing the performance of GTG3 against all turbulence observations, with and without the MW component. That is, looking beyond just explicit MW observations, the added value of the MW component in an overall sense will be assessed.

4.2.5 Associating Observations to G-AIRMETs

For determining whether an observation is inside a G-AIRMET, the following criterion was used. If any part of an observed turbulence event is inside a G-AIRMET, the entire event is considered to be within the advisory volume. Nearly all observed events are either entirely inside or entirely outside a 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.

4.3 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.

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

- Mountain-wave PIREPs: If the PIREP intensity meets or exceeds the event threshold and PIREP remark identifies event as being associated with mountain waves, it is considered a MW event.

Note that non-events are not limited to explicit nulls, but rather include all categories less than the event category. The exception to this is with the MW events. The absence of a MW remark does not guarantee that the encountered turbulence is caused by something other than mountain waves. Therefore, only explicit Null PIREPs are used to identify false alarms for the MW component of GTG3.

4.4 EVALUATIONS

Terminology and score definitions are first provided for reference in the subsequent sections:

MOG	Moderate-or-Greater Turbulence
POD (= PODy)	Probability of Detection: proportion of all observed events that are correctly forecast to occur, in this case, of detecting turbulence at a specific threshold
POFD (= 1 – PODn)	Probability of False Detection: proportion of all observed non-events that are mistakenly forecast to be events, in this case, detecting turbulence less than the specified threshold
PSS	Peirce Skill Score (aka True Skill Score, TSS): POD – POFD; Skill relative to an unbiased random forecast; Provides a measure of the product's ability to separate 'yes' events from 'no'
RMSE	Root-mean-square-error: typical distance between forecast and observed values
AUC	Area Under the Receiver Operating Characteristic (ROC) Curve: measure of ability of forecast to correctly distinguish between events and non-events ROC curve—the set of (POFD, POD) pairs as the forecast threshold is varied)
% Volume:	The percent of possible volume (the forecast domain) that is covered by the forecast

4.4.1 GTG EVALUATION

Due to the non-systematic nature of the verification data set (PIREPs, and even EDR, since planes will avoid known areas of turbulence when possible), the "yes" observations and "no" observations must be treated separately (Carriere et al. 1997). As a result, it becomes inappropriate to compute several common statistics that would otherwise be computed and analyzed (e.g. Critical Success Index, Bias, and False Alarm Ratio). The rationale for this is well documented by Brown and Young (2000) and Carriere et al. (1997).

The association of the GTG product to observations as described in section 4.2 yields the following contingency table:

Hit:	forecast = yes; obs = yes
False alarm:	forecast = yes; obs = no
Miss:	forecast = no; obs = yes
Correct no:	forecast = no; obs = no

where 'yes' signifies that the forecast or observation equals or exceeds a given threshold, and 'no' signifies that the forecast or observed value is less than the threshold. POD, POFD, and PSS are computed from the contingency table. Varying the forecast threshold for a given observation threshold produces a set of POD and POFD pairs, which form a ROC curve.

4.4.2 GTG3 COMPARED TO G-AIRMET

G-AIRMETs are, by definition, forecasts of MOG turbulence. Therefore, the contingency table is defined as:

Hit:	MOG observation inside a G-AIRMET
False alarm:	Less-than-moderate observation inside a G-AIRMET
Miss:	MOG observation outside a G-AIRMET
Correct no:	Less-than-moderate observation outside a G-AIRMET

The G-AIRMET contingency-table statistics POD, POFD, and PSS are then compared to the GTG3 contingency-table statistics as determined above.

4.4.3 GTG3 AS SUPPLEMENT TO G-AIRMET

In this study, we provide a complementary view of GTG3 performance by considering its contribution as a supplement to G-AIRMETs. Inside a G-AIRMET, where MOG turbulence is predicted, GTG3 disagreement can potentially lower false alarm rates by reducing forecast volume, with the goal being to reduce the forecast volume without missing too many of the MOG observations captured by the G-AIRMET. Outside a G-AIRMET, where MOG turbulence is not predicted, GTG3 disagreement can potentially reduce the likelihood of encountering a turbulence event without drastically increasing forecast volume, with the goal being to capture as many of the missed MOG observations as possible without unduly increasing the number of false alarms.

As mentioned in section 4.2.3, when making comparisons to observed events, the neighborhood approach is used for the GTG algorithms, but in comparing G-AIRMET to observed events, the 'in or out' metric described above is used.

4.5 Sensitivity to Verification Methodology

The calculated scores are sensitive to the various choices in methodology, including: the size of the forecast neighborhood, the representative value from neighborhood (e.g., max, mean, nearest neighbor), observation sampling (e.g., every report, regular subsample, event-based), and even whether one uses greater than (>) or greater than or equal to (>=) for the event threshold. In addition, the aforementioned choices can affect the two forecast products differently. For example, compared to the nearest-neighbor approach, the neighborhood approach used in this assessment yields higher ROC areas for both GTG2.5 and GTG3, but the difference between the ROC curves for the two forecast products is reduced. Regardless of the methodology employed, the overall story is unchanged.

5 Results

5.1 GTG Product Characteristics

Before looking at the verification scores, it is useful to examine characteristics of the fields themselves, specifically distributions of the forecast values. Figure 5.1 shows distributions of turbulence intensities from the 1-h forecasts from both GTG2.5 (blue) and GTG3 (red), along with the ratio of the two distributions (green). GTG3 produces a noticeably different distribution than the current operational version, with a narrower spread of values and the most frequently occurring values shifting to the right, away from the lowest intensities. As a result, GTG3 is more likely than GTG2.5 to produce intensities between 0.06 and 0.2, but less likely to produce the lowest and highest intensities. For both products, the distribution of intensities changes very little with lead time (Fig. 5.2).

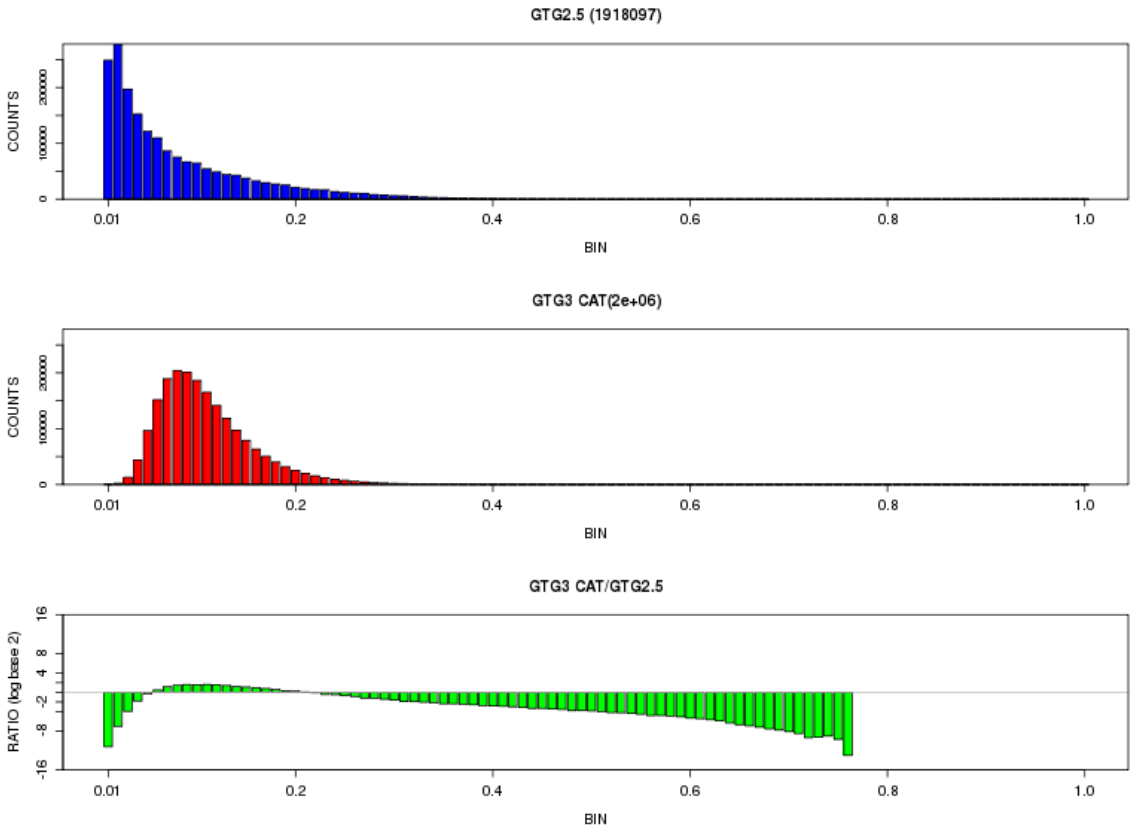

Figure 5.1: Distribution of GTG2.5 (blue), GTG3 (red), and the ratio of GTG3 to GTG2.5 (green) for 1-h lead forecasts. The ratio is expressed in $\log_2$, i.e., -2 denotes a ratio of 1/4.

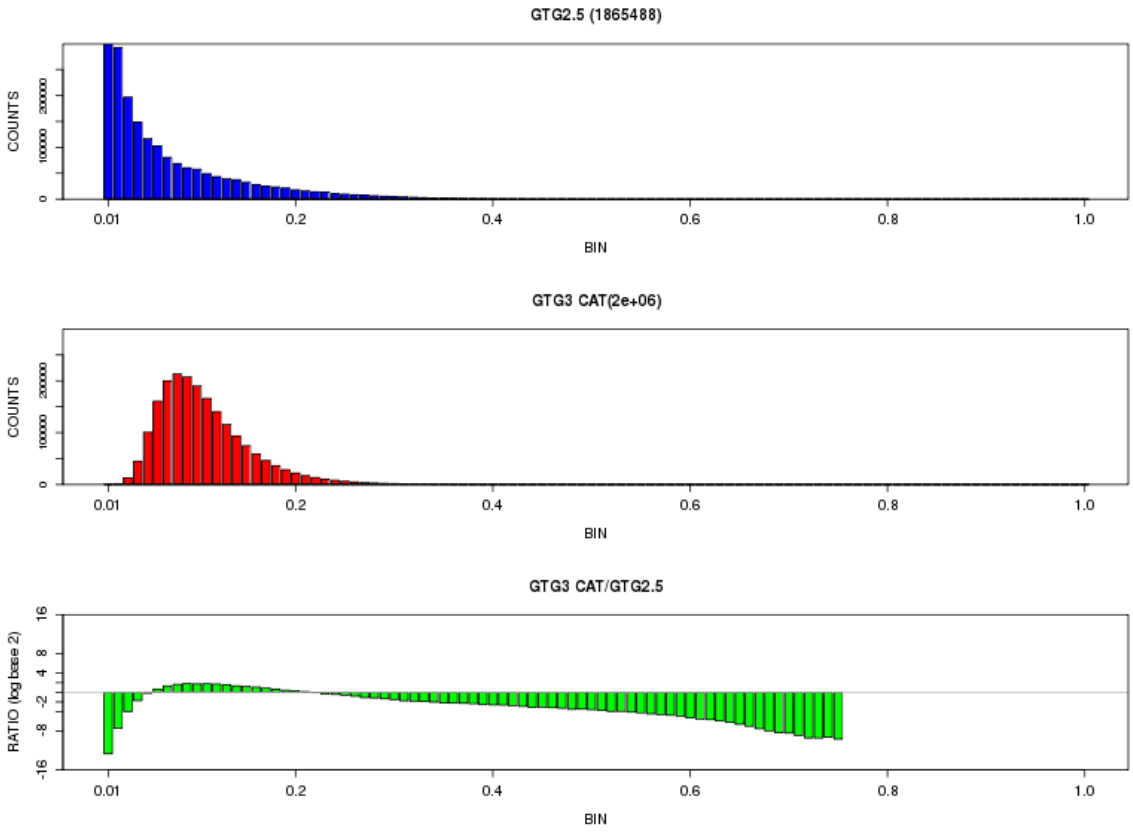

Figure 5.2: As in Fig. 5.1, but for 12-h lead forecasts.

Figure 5.3 shows a heatmap (i.e., a plot of the density of a scatterplot) of the association between GTG3 and GTG2.5 forecast values. Looking horizontally across the plot shows the range of GTG3 values that are paired with a given value of GTG2.5 forecasts. Similarly, looking vertically shows the range of GTG2.5 values paired with a given value of GTG3 forecasts. All points above the diagonal show pairs for which GTG2.5 values are greater than the GTG3 values and all points below the diagonal show pairs for which GTG3 values are greater than the GTG2.5 values. The black '+' marks indicate the most common pairings of GTG2.5 and GTG3 values within each GTG3 bin. For lower intensity forecast values, below about 0.15, GTG3 produces higher intensities than GTG2.5, while for forecasts in the moderate or greater range, GTG2.5 tends to produce higher forecast values. This is consistent with the difference between the two forecast distributions presented above. However, Fig. 5.3 demonstrates that it is not a simple one-to-one remapping where a forecast of 0.25 from GTG2.5 becomes a forecast of 0.2 from GTG3. Rather there is a fairly wide range of values in the pairings. Again, consistent with the narrower distribution of GTG3 intensities, the range of GTG3 values paired with a given GTG2.5 value is substantially smaller than vice versa.

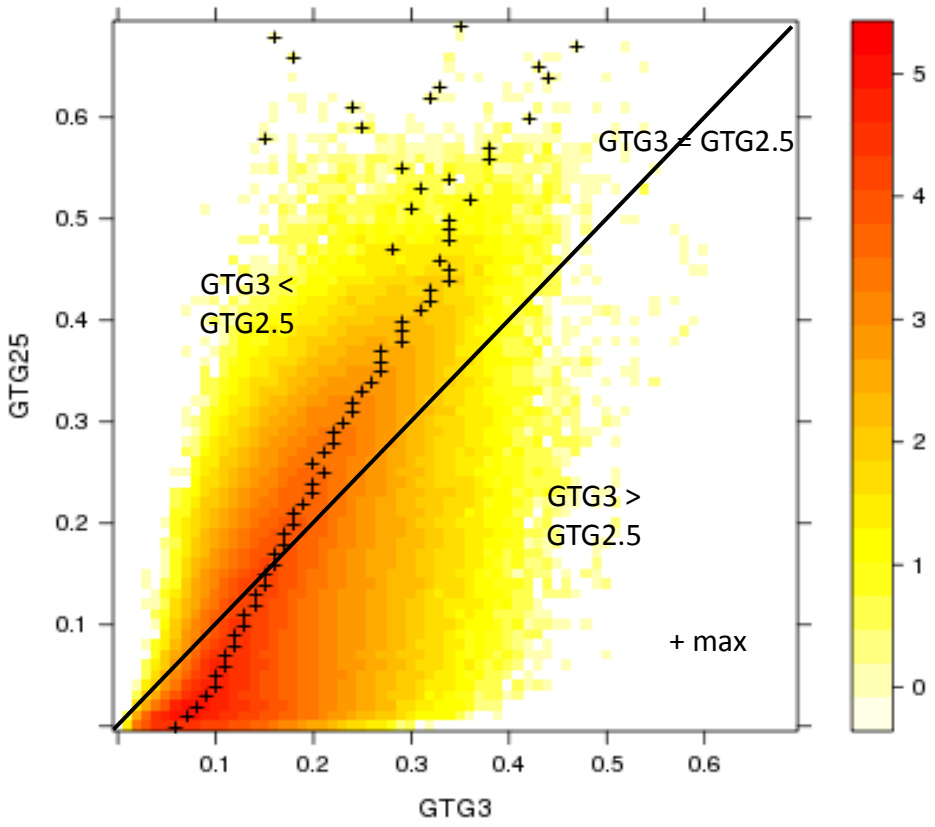

Figure 5.3: Heat map of the scatterplot of GTG3 and GTG2.5 forecast values. Scale is log10. '+' marks show the location of the most commonly paired GTG3 – GTG2.5 values as a function of GTG3 intensity.

The most likely cause of this shift is a change in how the raw output from the GTG algorithm is translated into forecast EDR values. This final step is now based on new, empirically derived, functions believed to more accurately represent the distribution of turbulence in the atmosphere (Sharman 2014, personal communication). Unfortunately, the true distribution of turbulence in the atmosphere is unknown. Aircraft fitted with EDR-measuring equipment provide a much better sample than is available from PIREPs alone, but those measurements are available only where the planes are flying, and pilots will avoid turbulence, especially stronger turbulence, when possible. Thus, one can expect EDR reports to undersample turbulence events in general.

As noted in section 2.2.2, the DAL reports fall into three categories: heartbeat, triggered, and interpolated. The anecdotal experience of any flier suggests that, for most locations, the atmosphere is largely quiescent so the full set of reports is expected to be dominated by the interpolated null reports. The predetermined, regular heartbeat reports (every 15 minutes) provide an unbiased sampling of turbulence along a given flight path, but are sensitive to the underreporting bias of the selection of the flight path. The triggered reports provide a sample of turbulence within and surrounding a turbulent event, possibly providing an oversampling of atmospheric turbulence.

Figure 5.4 shows the distribution of the intensity of EDR reports from DAL aircraft as a function of altitude layer and report type. The difference in the distribution of intensities in these categories can be seen by the percentage of reports within the lowest (0.0 – 0.02) bin. For all reports (Fig 5.4a, black line), including the interpolated zeros, almost 95% of all intensities are in the zero bin. The percentage drops to 82% for the heartbeat reports and only 25% for the triggered reports. If one combines the heartbeat and triggered categories, 56% of reports are in the zero bin. Not only do the heartbeat reports (Fig. 5.4b) consist of a large number of zeros, but there are very few larger turbulence values; almost all heartbeat reports have intensities less than about 0.1 (with the exception of the near-surface, 0 –10000 ft layer; blue line). In addition to being subject to the choice of flight path around known turbulent areas, the 15-minute reporting period is substantially larger than most turbulent events (cf. Wandishin et al. 2011 Fig. 5.10) and so the pseudo-random sampling of the heartbeat reports can be expected to miss many events.

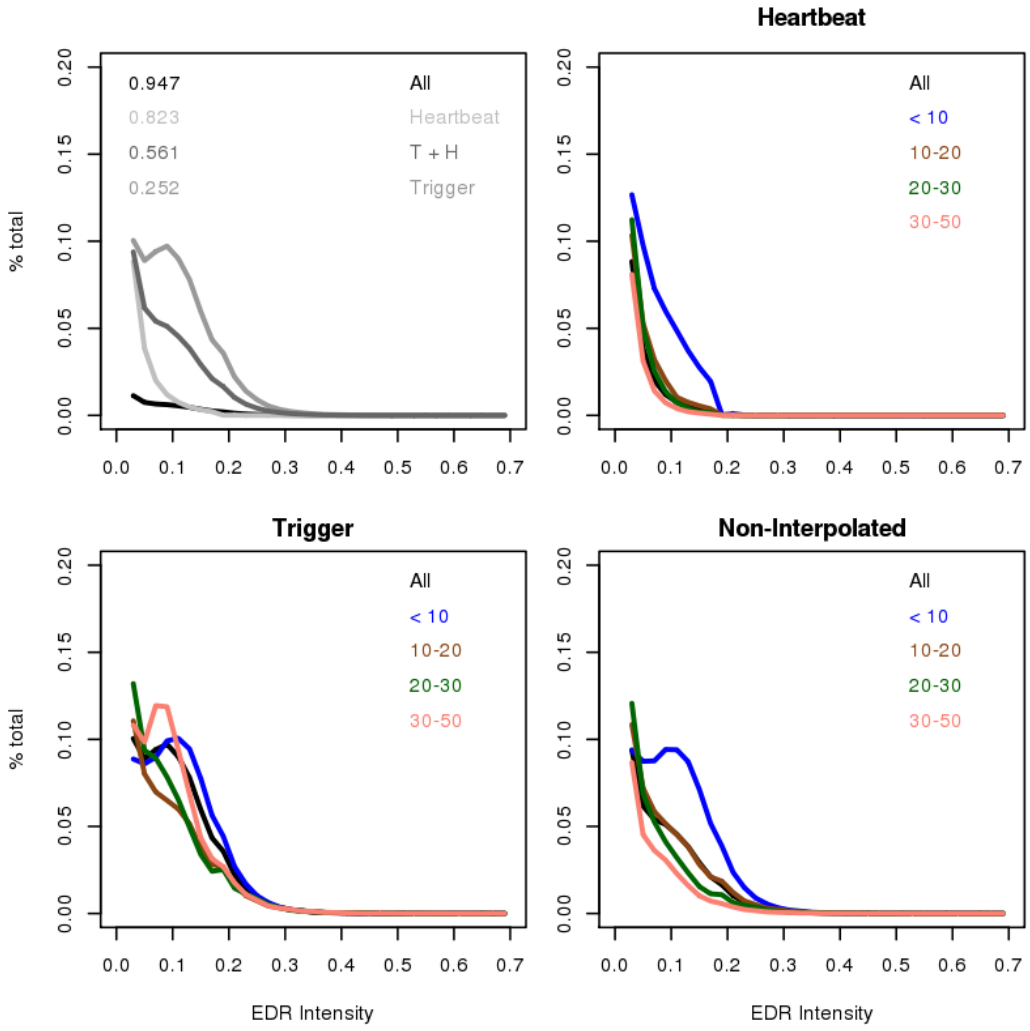

Figure 5.4: Distributions of DAL EDR value for different reporting types (a), and for different altitude layers (colored lines) for the heartbeat (b), trigger (c), and non-interpolated (d) categories. Percentages of intensities within the zero bin for each reporting type is shown in (a).

In contrast to the heartbeat reports, the triggered category (Fig. 5.4c) possesses a secondary peak of intensities around 0.1 and a non-negligible number of reports up to around 0.3. Again, in contrast to the heartbeat category, the distribution of triggered reports higher in the atmosphere closely resembles the distribution in the near-surface layer. Because of the interactions with the surface and well-mixed character of the boundary layer, the near-surface layer is expected to be more turbulent than approximately frictionless, more stratified higher layers of the atmosphere. This transition is captured somewhat in the triggered reports with the low (brown) and middle (green) layers containing more low intensities.

If one considers the heartbeat and triggered categories together, that is, the set of all non-interpolated reports (Fig. 5.4d), the highly turbulent near-surface layer again stands out compared to higher layers, but those higher layers contain a non-negligible number of reports up to intensities of about 0.25.

The GTG3 distribution most closely resembles the set of triggered EDR reports while GTG2.5 more closely resembles the combined set of non-interpolated reports. It must be emphasized again, that at this time, there is no agreed-upon true distribution of atmospheric turbulence.

SUMMARY

The distribution of intensity values from GTG3 is substantially different from that of GTG2.5, with a narrower range of values and a peak shifted to the right (from around 0 for GTG2.5 to near 0.08 for GTG3). On average the result is that for Light turbulence forecasts, GTG3 tends to produce larger values than GTG2.5, while for MOG forecasts, GTG3 tends to produce lower values than GTG2.5. The GTG3 distribution most closely resembles that of triggered DAL EDR, while the GTG2.5 distribution is most similar to that of the combined triggered and heartbeat reports.

5.2 GTG3 COMPARED TO GTG2.5

Before examining more specific measures of forecast performance, the overall performance of GTG3 and GTG2.5 are explored. Figure 5.5 plots the root-mean-squared-error (RMSE) of GTG3 (red) and GTG2.5 (blue) as a function of the forecast value for 1-h lead forecasts from the winter season. (Results for the summer season are similar; not shown). The RMSE is computed for bins of forecast values of 0.02 width. For example, all forecasts with values between 0.2 and 0.22 are grouped together with the RMSE computed just for that set of forecasts and plotted. The lines then connect the individual RMSE scores over the range of forecast value bins. Only DAL EDR reports are used in this calculation to avoid the large bin width (0.1) of the UAL EDR data. Despite the notable difference in the distributions of the two products, they yield similar scores with GTG3 possessing slightly lower RMSE values for forecast values between about 0.15 and 0.45. For that range of forecast values, GTG3 is in slightly better agreement with observations than is GTG2.5.

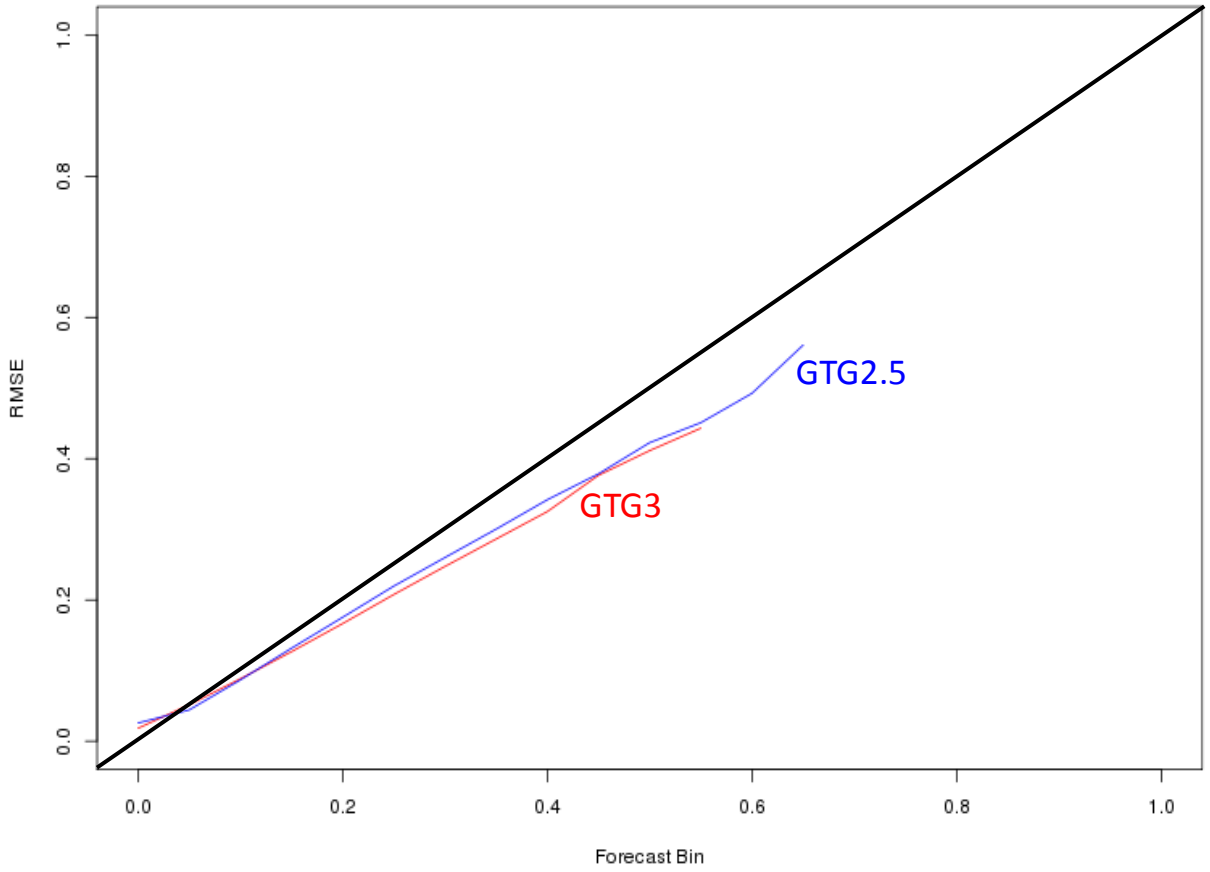

Figure 5.5: Root-mean-squared-error (RMSE) as a function of forecast intensity for GTG3 (red) and GTG2.5 (blue). The black line indicates where the RMSE value is equal to the forecast value, itself.

ROC curves plot POD as a function of POFD 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 under 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 AUC, 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 larger forecast thresholds.

One feature of ROC curves is that they are (mostly) independent of calibration. That is, the ROC curve depends on the extent to which forecast scores associated with events tend to be greater than forecast scores associated with non-events and not on the accuracy of the forecast scores themselves. Forecasts systems that are identical with the exception of calibration will lead to ROC curves that trace the same path, but the various forecast thresholds will be located at different points along the curve. The ROC curve presents a variety of potential decision thresholds for a

particular event and the likelihood of false alarms and missed events associated with each threshold; the appropriate threshold for a particular user is determined by that user's sensitivity to missed events and false alarms. Users more sensitive to false alarms will be more interested in points on the left side of the ROC curve (lower POFD) while users more sensitive to missed events should focus more on the top portion of the curve (higher POD).

GTG3 (red) and GTG2.5 (blue) produce very similar ROC curves for the 0.1 (light turbulence) and 0.2 (moderate turbulence) EDR thresholds (Fig. 5.6), but with a slightly larger AUC for GTG3. The larger area comes mostly from the top half of the curves. For POFD values greater than about 0.25, the GTG3 curve is higher than the GTG2.5 curve, especially for the 0.1 EDR light turbulence events (Fig. 5.6a). Note also the location of the forecast thresholds along the ROC curves. The 0.2 forecast threshold points (squares) are nearly co-located. The points marking the 0.1 (circles) and 0.3 (triangles) forecast thresholds have greater separation.

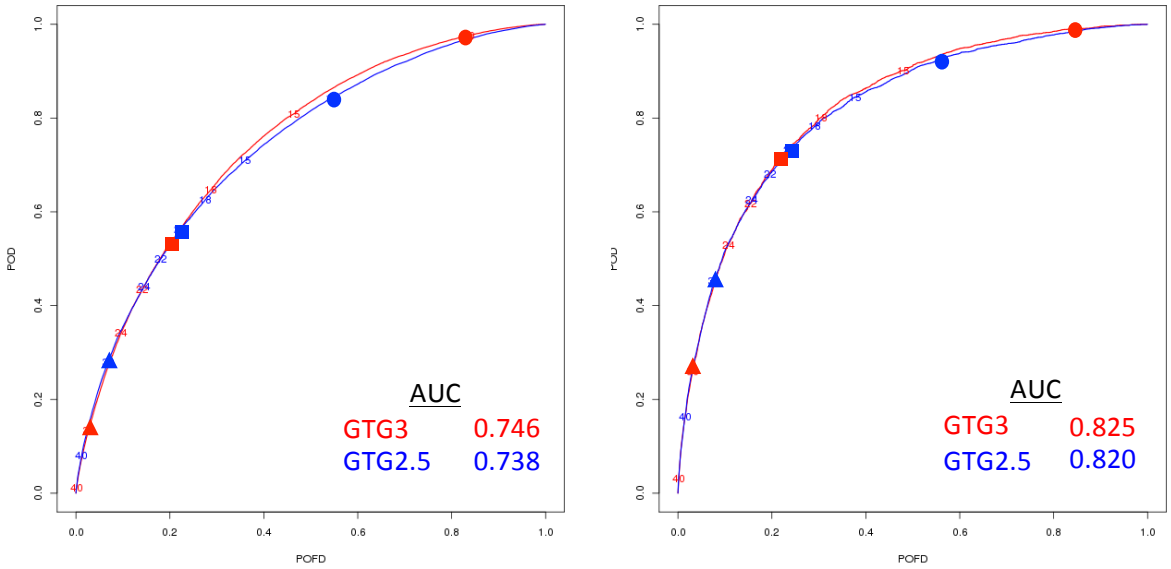

Figure 5.6: Receiver Operating Characteristic (ROC) curves for GTG3 (red) and GTG2.5 (blue) for the 0.1 (a) and 0.2 (b) EDR thresholds. Area under the ROC curve (AUC) shown in the bottom right corner. Markers show the location of specific thresholds along the curves: 0.1 (circle), 0.2 (square), 0.3 (triangle).

Once more, this can be explained by the differences in forecast distributions and manifests in plots of accuracy and skill scores. Both turbulence products have nearly the same proportion of forecast values greater than 0.2. This means that the forecast volumes will be nearly identical, resulting in very similar plots of POD, POFD, and PSS (Fig. 5.7). A much larger proportion of the GTG3 forecasts are greater than 0.1 than is the case for GTG2.5; therefore, the forecast area for GTG3 will be much larger than for GTG2.5, leading to few missed events but many more false alarms (Fig. 5.8). For the 0.3 forecast threshold, the longer tail of the GTG2.5 distribution means that this situation is reversed, the area of GTG2.5 forecasts is larger resulting in more captured events and more false alarms (Fig. 5.9).

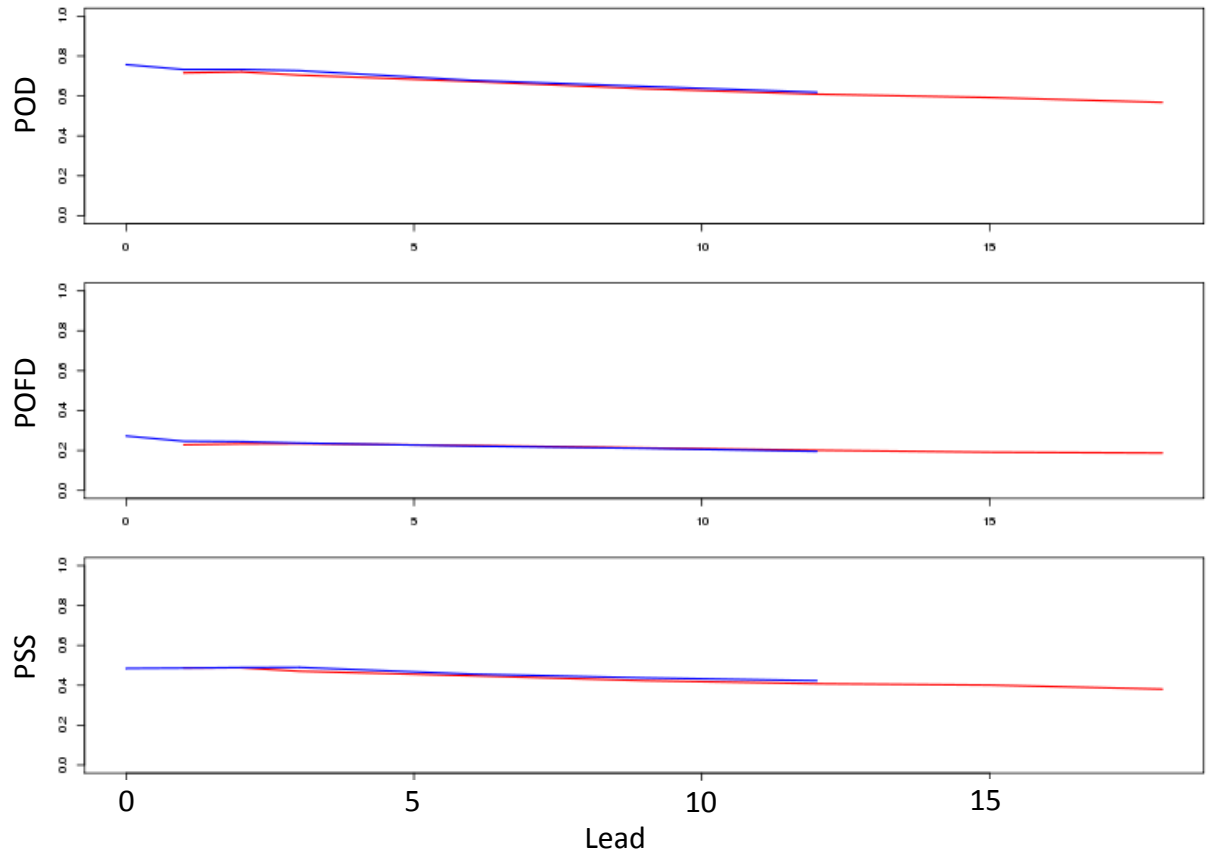

Figure 5.7: Performance measures for 0.2 EDR threshold as a function of lead time for GTG3 (red) and GTG2.5 (blue) forecasts.

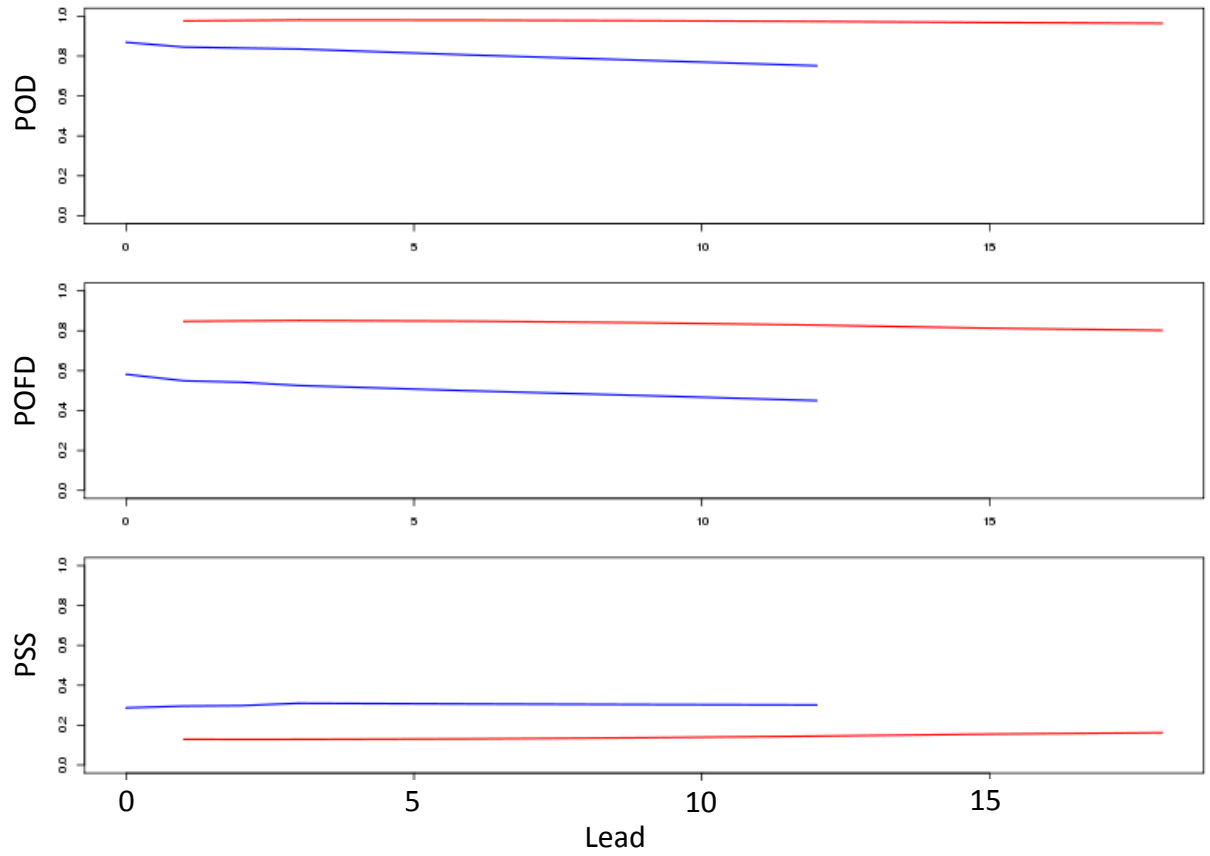

Figure 5.8: As in Fig. 5.7, but for 0.1 threshold.

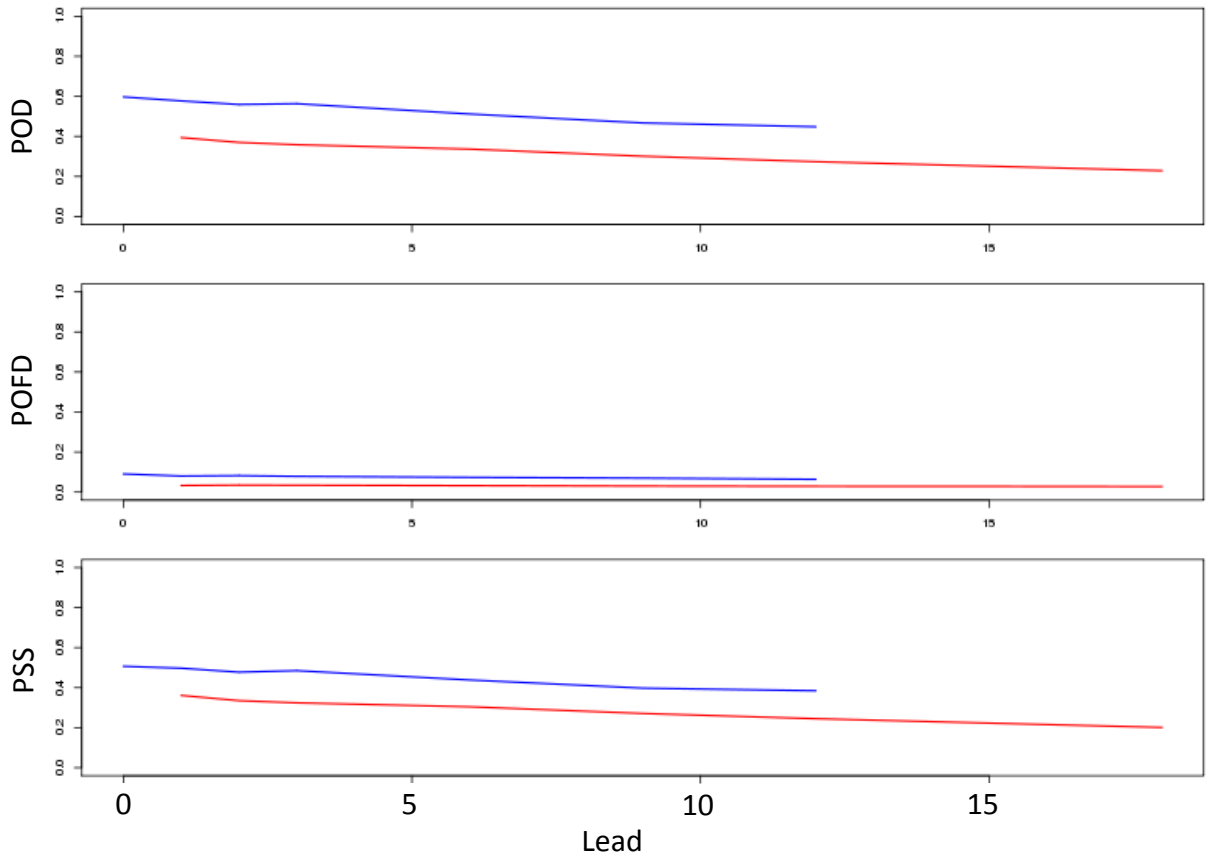

Figure 5.9: As in Fig. 5.7, but for 0.3 threshold.

Recall that the ROC curve is insensitive to forecast calibration. Skill (PSS) and accuracy (POD, POFD) calculations, however, are very sensitive to calibration. As a result, the skill for a GTG3 forecast of 0.1 observed EDR using the 0.1 forecast threshold is less than the skill for a GTG2.5 forecast using a 0.1 forecast threshold. Recall, as well, in Fig. 5.6 that the ROC curve for GTG3 is higher than the curve for GTG2.5 for much of the curves. Therefore, one can choose a different forecast threshold for GTG3 to achieve greater skill. Figure 5.10 shows the POD, POFD, and PSS for light turbulence (0.1), as shown in Fig. 5.8. Additionally, the performance of GTG3 using the 0.18 forecast threshold is also included (thick red line). Choosing a larger threshold leads to a smaller forecast area, which, in this portion of the ROC curve, reduces the POFD much more than the POD, resulting in improved skill. Calibration will be discussed further in section 5.6.

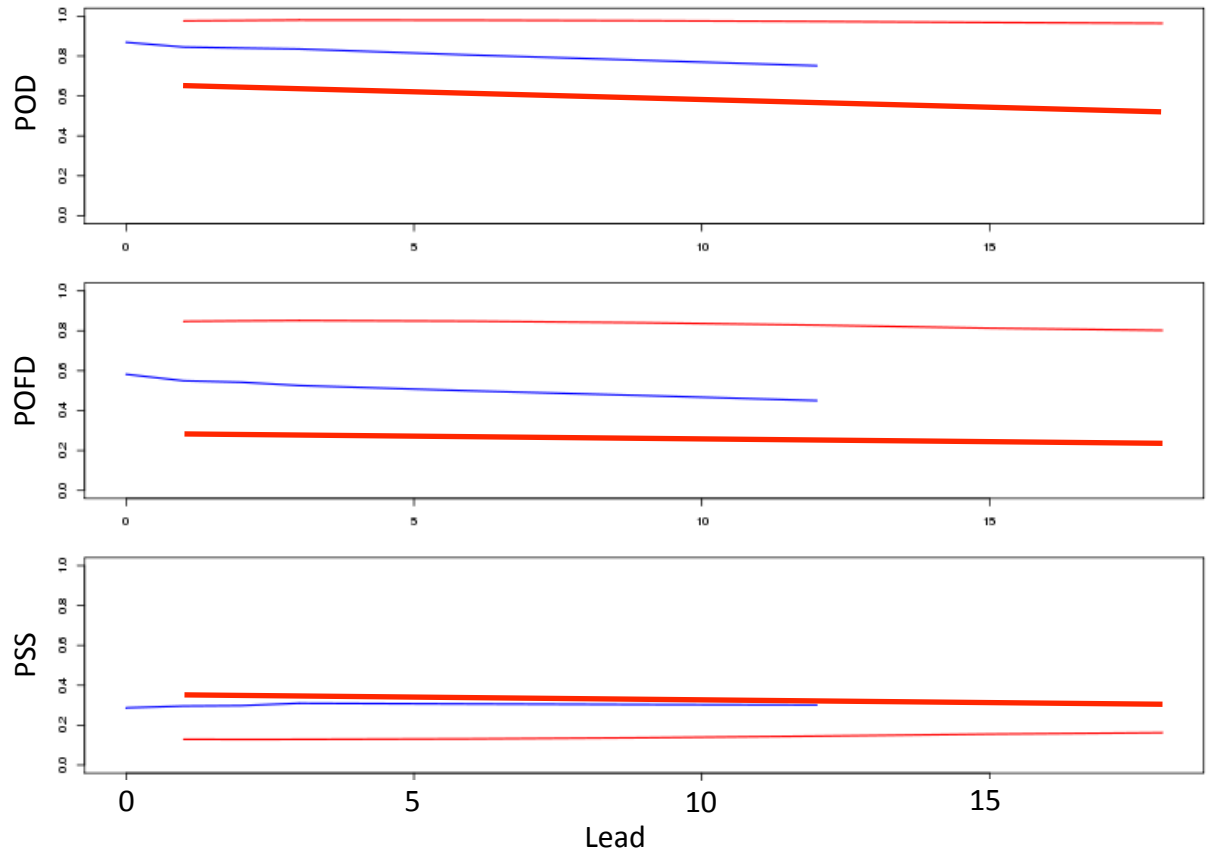

Figure 5.10: As in Fig. 5.9, but with performance of GTG3 > 0.18 forecast added (thick red line).

Summary

As measured by the area under ROC curves, GTG3 and GTG2.5 have similar ability to distinguish events from non-events, but GTG3 outperforms GTG2.5 consistently for all observed thresholds. Similarly, the distance between forecast values and observed values, as measured by RMSE scores, is slightly but consistently better for GTG3. When the forecast threshold is constrained to match the observed threshold (i.e., no calibration), GTG3 is more skillful than GTG2.5 for only a small range of thresholds; however, this range can be expanded with proper calibration.

5.3 GTG3 AND G-AIRMET

5.3.1 GTG3 COMPARED TO G-AIRMETS

The performance of GTG relative to G-AIRMETs is considered by examining the proportion of events captured (POD) relative to volume of the forecasts (Fig. 5.11). Because G-AIRMETs are forecasts of MOG turbulence, observed events are defined by MOG PIREPs or EDR ≥ 0.18. The average volume of a G-AIRMET is nearly equivalent to the volume of GTG3 (or GTG2.5) exceeding the 0.18 threshold, but GTG3 captures many more events: POD ≈ 0.75 for GTG3 compared to POD ≈ 0.45 (0.55) for G-AIRMETs measured against EDR (PIREPs). Alternatively, if the POD of the G-

AIRMETs was acceptable, the volume of the forecast required to achieve that POD is reduced by a third with the GTG3 forecast (using a threshold between 0.22 and 0.24).

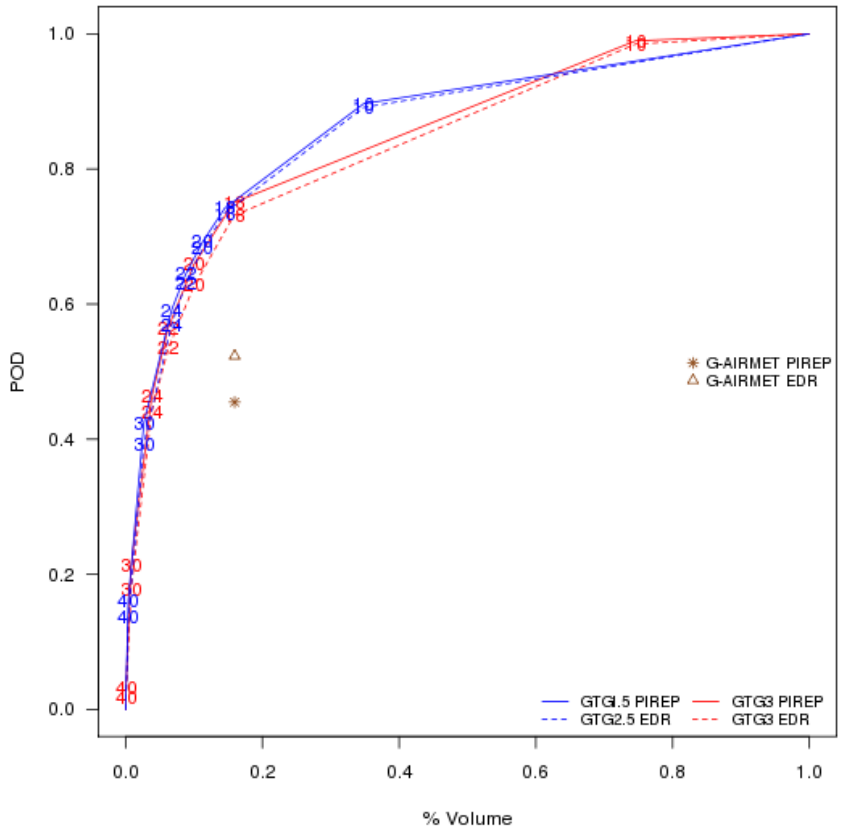

Figure 5.11: POD as a function of the volume of the forecast for GTG3 (red), GTG2.5 (blue), and G-AIRMETs (symbols) verified against EDR (dashed) and PIREPs (solid). Numbers along the curves mark various forecast thresholds (number equals threshold * 100) at their associated POD and volume. For G-AIRMETs, forecasts verified against EDR denoted by the triangle and forecast verified against PIREPs denoted by the star. Observation thresholds are 0.18 for EDR and 3 for PIREPs.

Like GTG3, G-AIRMETs may extend down to the surface; therefore a comparison is available for the Near-surface layer (Fig. 5.12). When verified against EDR reports, GTG3 (dashed line) performs very similarly to the G-AIRMET (triangle) in this layer, with the G-AIRMET nearly matching the performance of a GTG3 > 0.19 forecast. When verified against PIREPs, however, GTG3 (solid) captures 20% more events than the G-AIRMETs (star). There are two reasons for the smaller gap in the performance of GTG3 relative to the G-AIRMET in this layer compared to higher layers, when verified by EDR: 1) the slight reduction in the forecast volume of the G-AIRMETs; and 2) the lower POD of the GTG3 forecasts for the Near-surface layer.

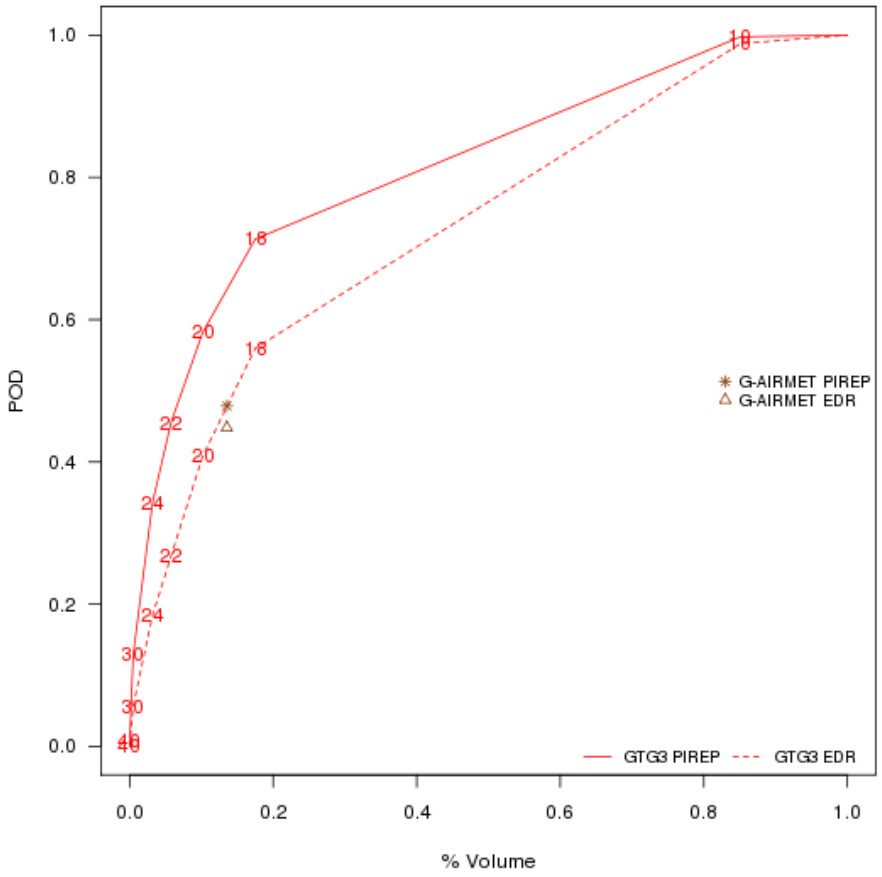

Figure 5.12: As in Fig. 5.11, but for the Near-surface layer. GTG2.5 is not plotted because it does not provide forecast information below 10 kt.

It is possible that the sensitivity of the performance of the GTG3 forecasts in the near-surface layer to the type of verifying observation is tied to the different spatial distributions of the EDR and PIREPs in this layer (Fig. 5.13). PIREPs are spread more evenly across the country compared to EDR reports, which are concentrated around major airports. A slight translation error in the forecast area will have a larger impact when judged against intermittent, highly-clustered reports compared with more smoothly distributed observations.

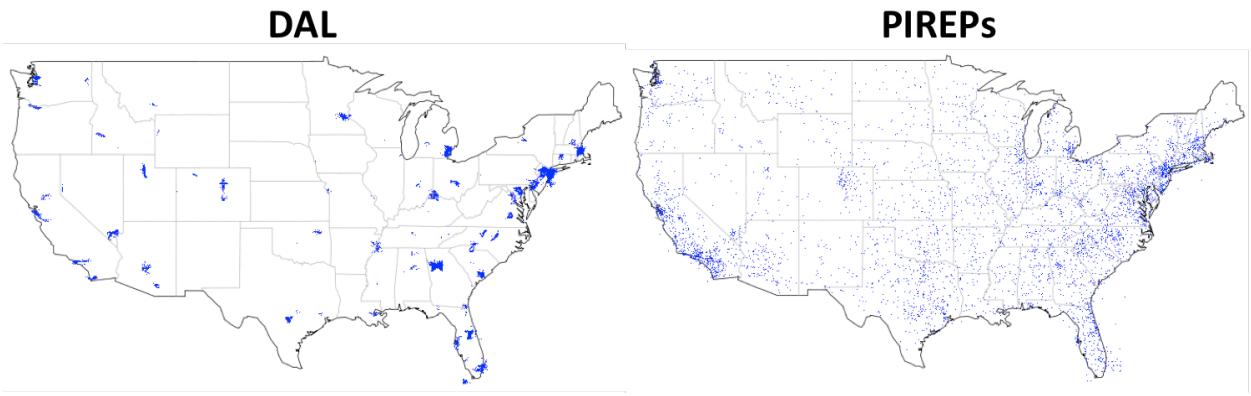

Figure 5.13: Spatial distribution of EDR and PIREPs in the Near-surface layer for the period 1 Jan – 31 Mar 2013.

5.3.2 GTG3 AS A SUPPLEMENT TO G-AIRMETS

A complementary view of GTG3 performance considers its contribution as a supplement to G-AIRMETs. Inside a G-AIRMET, where MOG turbulence is predicted, GTG3 disagreement can potentially lower false alarm rates by reducing forecast volume, with the goal being to reduce the forecast volume without missing too many of the MOG observations captured by the G-AIRMET. Outside a G-AIRMET, where MOG turbulence is not predicted, GTG3 disagreement can potentially reduce the likelihood of encountering a turbulence event without drastically increasing forecast volume, with the goal being to capture as many of the missed MOG observations as possible without unduly increasing the number of false alarms.

Measured against PIREPs and using the 0.18 forecast threshold, GTG3 and GTG2.5 perform similarly (Fig. 5.14). Inside G-AIRMETs, both forecasts reduce false alarms by about half while still capturing 80–90% of all the MOG PIREPs. Outside G-AIRMETs, the forecasts capture about 60% of MOG PIREPs missed by the G-AIRMETs at the cost of including roughly 20% of the non-MOG PIREPs.

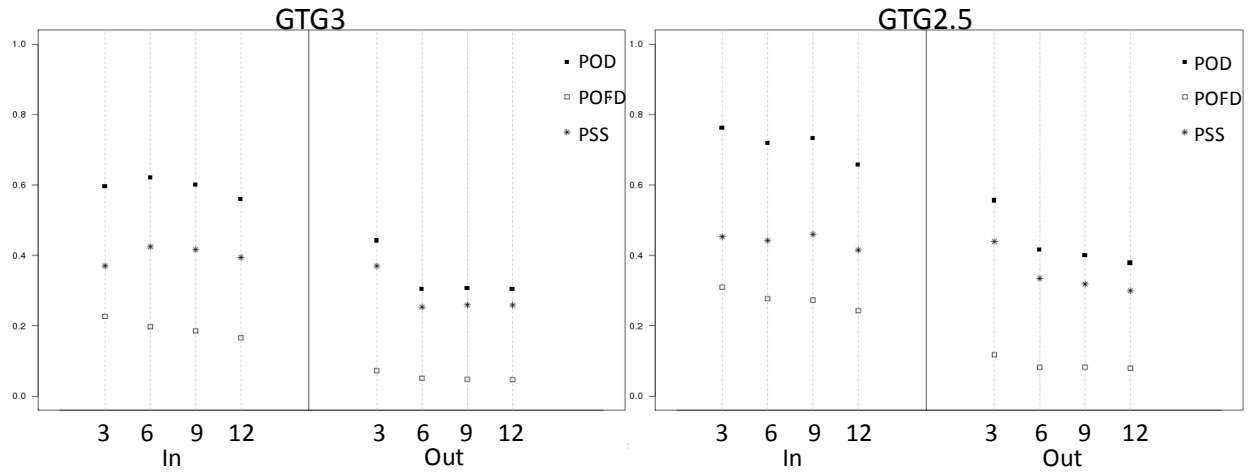

Figure 5.14: POD (filled squares), POFD (hollow squares), and PSS (stars) measured against PIREPs inside (left panel) and outside (right panel) of G-AIRMETs for GTG3 (top) and GTG2.5 (bottom) for 3-, 6-, 9-, and 12-h leads, using a forecast threshold of 0.18 and observed threshold of 3.

Increasing the forecast threshold to the high end of the moderate range (0.24) reduces false alarms, but leads to an increase in missed events (Fig. 5.15). The drop in POD and POFD is more pronounced for GTG3 than for GTG2.5, but the skill remains very similar between the two. GTG3 captures 60% of MOG turbulence inside a G-AIRMET (compared with 85% for the 0.18 threshold) while reducing false alarms by 80% (compared with 50% for the 0.18 threshold). The skill of both GTG3 and GTG2.5 is very similar for either threshold inside G-AIRMETs, but the skill outside of G-AIRMETs decreases as the forecast threshold is raised from 0.18 to 0.24. The results presented here use PIREPs as the verifying observations; similar results are found when using EDR observations (not shown).

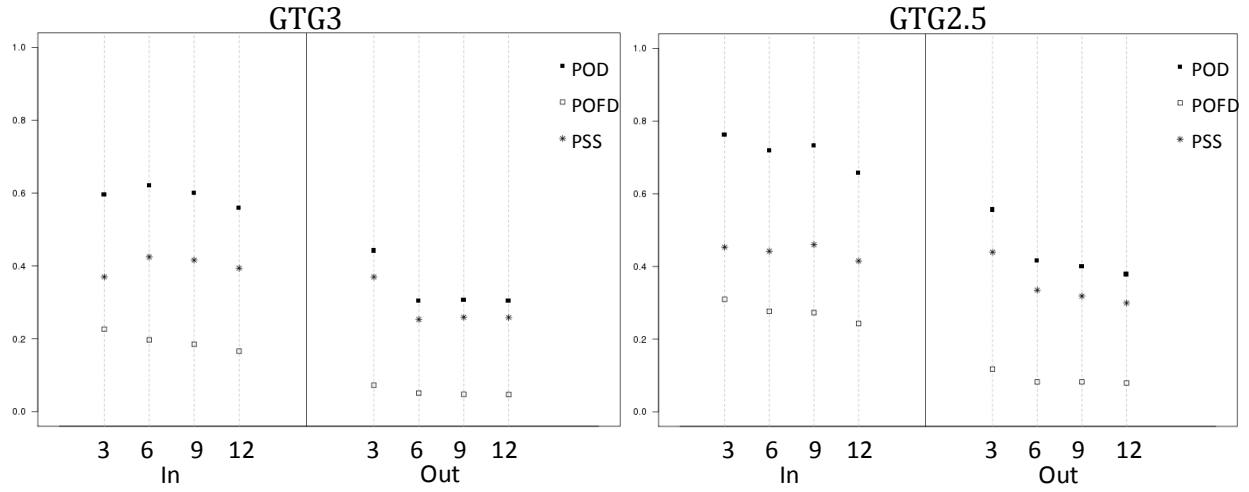

Figure 5.15: As in Fig. 5.14, but for a 0.24 forecast threshold.

SUMMARY

GTG3 is able to capture many more MOG events than G-AIRMET forecasts using the same forecast volumes. Alternatively, using a different threshold, GTG3 captures the same number of events as G-AIRMETs using only one-third of the forecast volume. For the Near-surface layer, the gap between GTG3 and G-AIRMETs is reduced, but GTG3 is still superior. Used as a supplement to G-AIRMETs, GTG3 can reduce the number of false alarms within a G-AIRMET by half while still capturing 80–90% of all MOG turbulence events. GTG3 is able to capture 60% of all MOG events found outside G-AIRMETs, while including only 20% of the non-MOG events.

5.4 OVERALL GTG3 PERFORMANCE

This section provides a summary of the general performance characteristics of GTG3. Unless otherwise stated, results are for the winter period.

Figure 5.16 gives the skill and accuracy of GTG3 for a range of thresholds from Light to Severe as measured against PIREPs, UAL EDR, and DAL EDR. Because the UAL EDR data is not considered trustworthy below 20,000 ft, only the middle and high layers are considered here. Also, for all of the results in this section, the forecast threshold is constrained to match the observed threshold (the impact of forecast calibration is discussed in section 5.6). POD and POFD decrease rapidly as the threshold increases: 90 to nearly 100% of all turbulence at the 0.1 threshold is captured by GTG3 while 10% or less of the severe turbulence exceeding 0.4 is captured. False detections drop from around 80% for light turbulence to nearly 0% for the 0.4 threshold. In contrast, the skill peaks for the moderate 0.2 threshold, with a PSS of 0.4 to 0.5, depending on the observation type. While there is some variability in the performance of GTG3 according to the observation type, no clear pattern emerges.

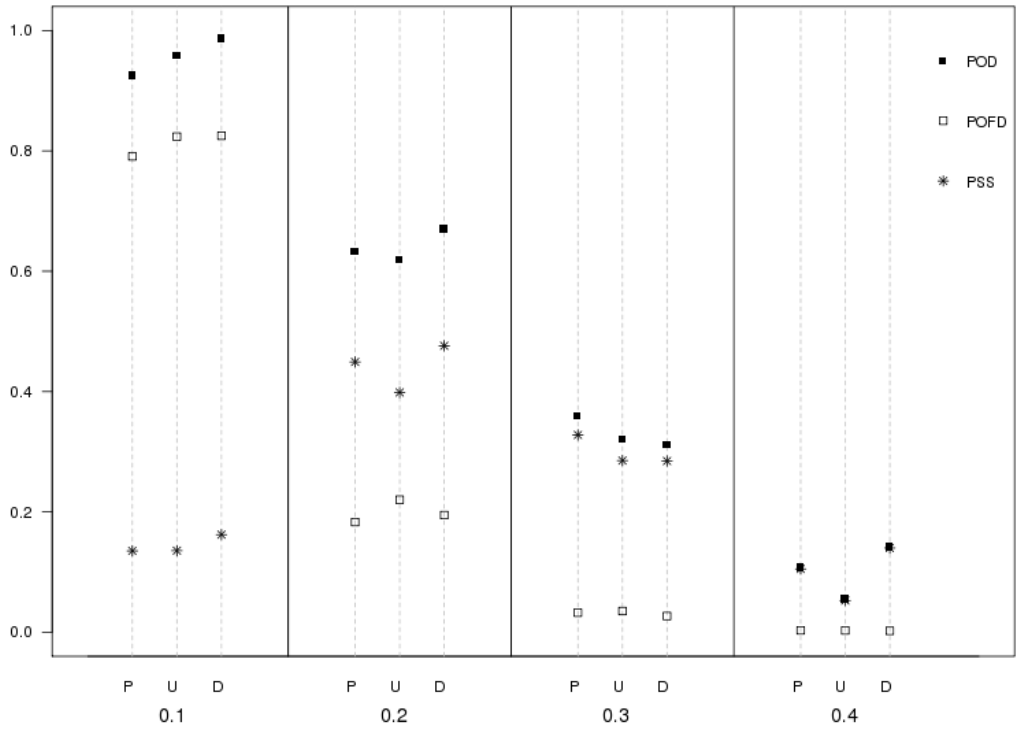

Figure 5.16: POD (filled squares), POFD (hollow squares), and PSS (stars) measured against PIREPs UAL EDR, and DAL EDR (indicated by the P, U, and D, respectively, along the bottom of the plot) for GTG3 (left) for the forecast thresholds 0.1, 0.2, 0.3, 0.4. A PIREP threshold of 1 is used for the 0.1 forecast threshold, 3 for the 0.2 threshold, and 5 for the 0.3 and 0.4 thresholds.

When the range of thresholds is focused in the range of moderate turbulence (0.18 – 0.24), the skill of GTG3 remains nearly unchanged, apart from a small decline for the highest threshold (Fig. 5.17). As in the previous figure, however, the POD and POFD are more sensitive to a change in the threshold, with POD dropping from 0.75 to 0.45 as the threshold is increased. In other words, while the skill is consistent for the various moderate thresholds, the missed-event/false-alarm tradeoff associated with each threshold is not. Different users with different sensitivities to these two types of forecast errors will want to use different thresholds to define moderate turbulence.

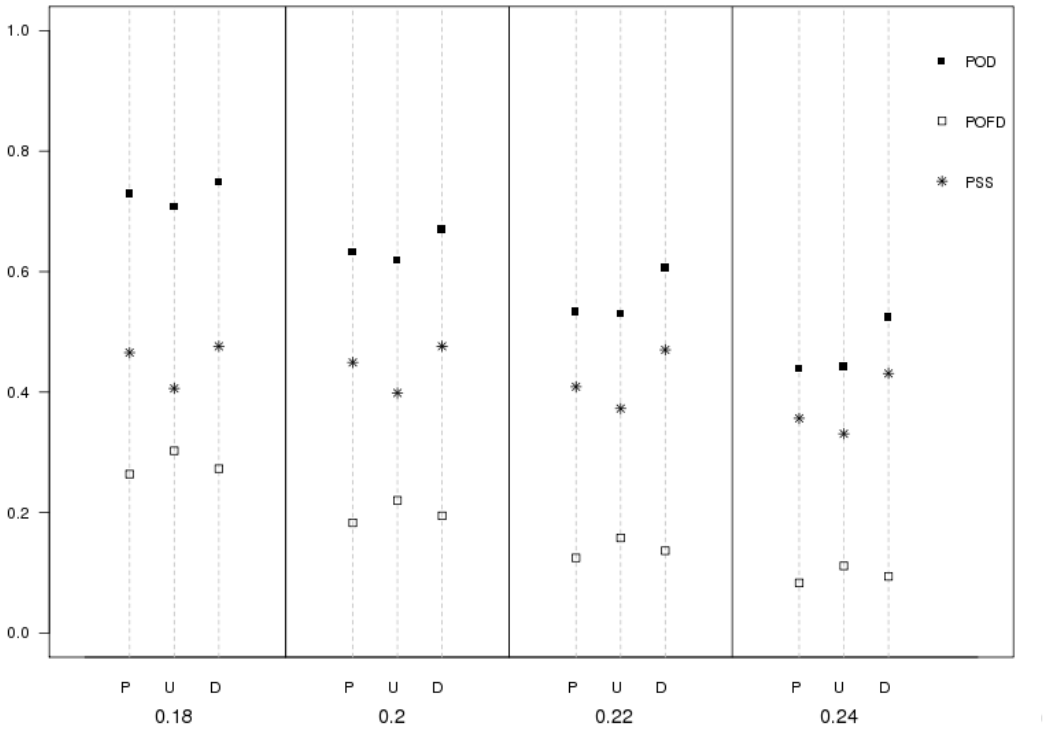

Figure 5.17: As in Fig. 5.16, but for forecast thresholds of 0.18, 0.2, 0.22, and 0.24.

The performance of GTG3 declines for the summer period (Fig. 5.18) as compared to winter; the proportion of missed events as much as doubles for the summer. With an absence of strong jet dynamics and a larger role of convectively-induced turbulence in the summer, this decline in performance is expected.

For comparisons across geographic regions and altitude layers, only a summary is provided here. The relevant plots can be found in the appendix. The definition of the altitude bands and a map of the regions can be found in section 2.3. The performance in the Central region is most similar to the overall performance while false alarms are most prevalent in the West region. GTG3 is most skillful in the Northeast region, especially when verified against DAL EDR, with the increase in skill coming from a higher POD, while skill is lowest in the Southeast, especially when verified against UAL EDR. The coverage within the regions varies across observation platform and this may account for the variations in performance when verifying against the different observation types.

When using PIREPs or UAL EDR, GTG3 is most skillful in the High layer, while skill against DAL EDR peaks in the Middle layer. Against PIREPs, however, GTG3 is least skillful in the Middle layer. Against DAL EDR, GTG3 skill is lowest in the near-surface layer. POD and POFD tend to be highest in the Middle and Low layers against both PIREPs and EDR. As discussed in section 5.3.1, the difference in the performance of GTG3 against PIREPs or DAL EDR may be a function of their different geographical distributions in this layer (see Fig. 5.13).

SUMMARY

The skill of GTG3 peaks in the Moderate range (~0.2), dropping off quickly for stronger or weaker turbulence (when calibration is not considered). For thresholds within the moderate range (0.18 – 0.24), skill is fairly steady, but the probability of detection and false detection decrease substantially as the threshold increases. Consequently, the optimal threshold is determined by a user's particular sensitivity to false alarms and missed events.

5.5 MOUNTAIN-WAVE (MW) TURBULENCE

One of the important new features of GTG3 is the mountain-wave (MW) component. The diagnostics within the MW component are computed separately from the CAT diagnostics, reflecting the different character of turbulence resulting from flow over a protruding obstacle. The MW component will be evaluated in two ways: as a distinct MW forecast and together with the core CAT component.

5.5.1 DIRECT EVALUATION

To evaluate the MW component separately, only explicit MW turbulence observations are used. Because EDR observations do not convey any information about the source of the turbulence, they are not included in this evaluation. For PIREPs, all reports that specify that the turbulence is a result of mountain waves are used and all null turbulence reports are included. Table 5.1 shows the number of explicit MW PIREPs by season and by severity. The stronger character of MW turbulence is revealed in the fact that there are 2.5 times as many MOG reports as there Light reports. Also, while MW turbulence can exist in the summer months, it is predominantly a winter phenomenon.

Table 5.1: Counts of mountain-wave (MW) turbulence PIREPs by season and severity.

	Winter	**Summer**
Light	1369	401
MOG	3228	733

For explicit MW turbulence, the POD is calculated for Light-or-Greater (LOG; 0.1 forecast threshold) and MOG (0.22 forecast threshold) categories (Table 5.2). For the LOG category both the MW component (99%) and the CAT component (97%) capture nearly all MW turbulence in winter and a very large proportion in summer (MW, 88%; CAT, 93%). However, for stronger MW turbulence, the MW component clearly outperforms the CAT component (70% vs. 40%).

Table 5.2: POD of explicit mountain-wave PIREPs for the mountain-wave (MW) and clear-air turbulence (CAT) components of GTG3.

	Winter		Summer	
	MW	**CAT**	**MW**	**CAT**
LOG	0.994	0.971	0.888	0.933
MOG	0.704	0.398	0.375	0.404

Since the absence of a MW specification is not sufficient to negate the possibility of MW turbulence, only explicit null reports can be used to determine a POFD (Table 5.3). For the LOG category, false alarms are fairly common for both but the MW component clearly produces fewer false alarms. This is particularly true in summer, when the MW component produces 60% fewer false alarms than the CAT component. For the MOG category, false alarms are exceedingly rare; fewer than 1% of all null observations are associated with a MOG forecast and POFD scores are much lower for the MW component.

Table 5.3. POFD of explicit null PIREPs for the mountain-wave (MW) and clear-air turbulence (CAT) components of GTG3.

	Winter		Summer	
	MW	**CAT**	**MW**	**CAT**
LOG	0.596	0.776	0.287	0.740
MOG	0.053	0.083	0.0098	0.045

5.5.2 INDIRECT EVALUATION

In addition to the direct evaluation above, the overall improvement to the forecast product can be assessed, as well. First, Figure 5.18 shows the MW domain and the number of times per grid column that the MW component produces a higher-intensity forecast than the CAT component; that is, the number of times in which the addition of the MW component leads to a different forecast. The MW component is much more active in the Intermountain West than in over the Appalachians. The greatest frequency is over the Canadian Rockies where the MW component changes nearly one-quarter of all forecasts in the winter. Regardless of the threshold examined, the performance of a forecast consisting of the maximum of the forecast from CAT and MW components, within the MW domain, is nearly identical to the performance of the CAT component alone (Fig. 5.19). One possible explanation for the lack of an impact of the MW component on GTG3 is that the number of non-MW turbulence events simply overwhelms the MW events, so that the impact is not discernable. Another possibility is that when the MW component does produce a stronger forecast than the CAT component, the difference in intensities is not, in general, large enough to change the performance.

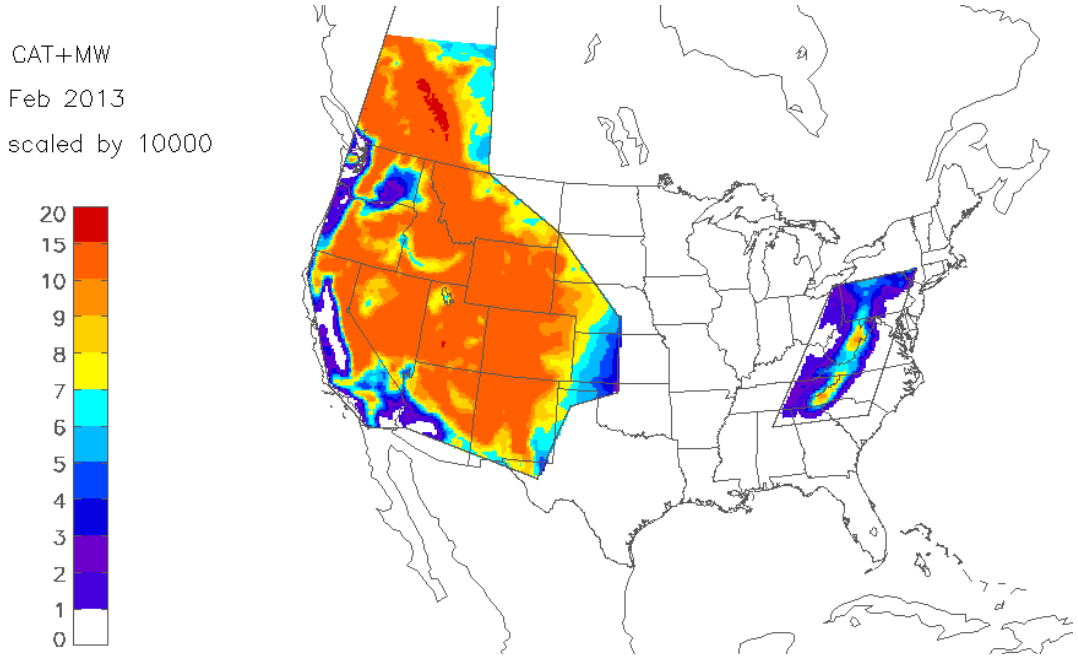

Figure 5.18: Map showing the number of times within each grid column that the MW component produces a higher-intensity forecast than the CAT component. Black lines denote the MW domain.

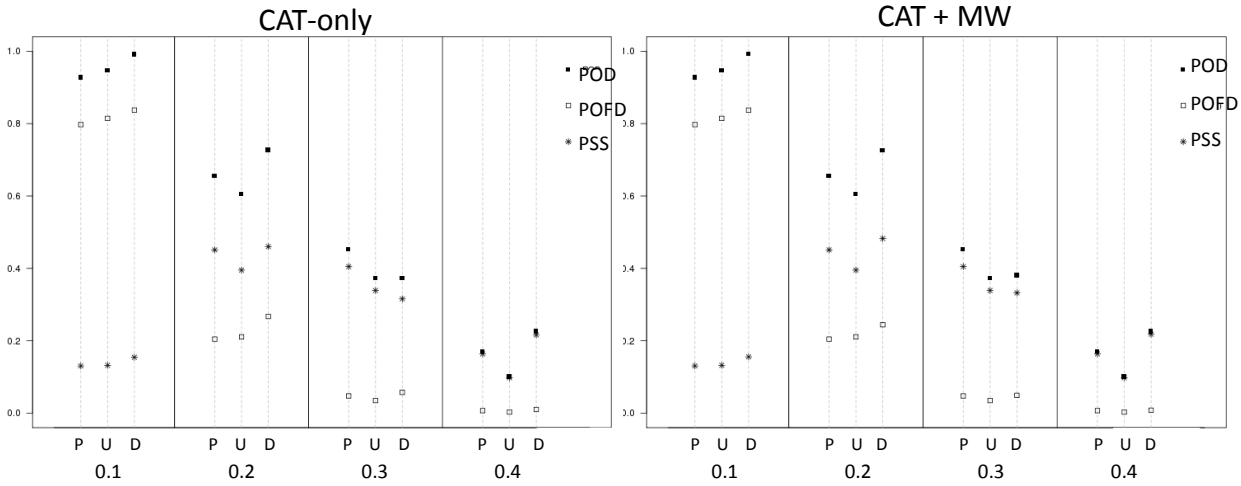

Figure 5.19: As in Fig. 5.13, but for the CAT only (left) and the combined CAT and MW forecasts (right).

SUMMARY

The GTG3 MW component captures 99% of LOG explicit MW PIREPs and 70% of MOG MW events in winter. False alarms are relatively high for Light-or-greater events (~ 60%) but nearly non-existent (5%) for MOG events. When all observed turbulence events within the MW domain are considered (i.e., not only the explicit MW PIREPs), the performance of GTG3 using the CAT component combined with the MW component is nearly identical to the performance of the CAT component alone. In other words, the GTG3 MW component successfully identifies MW events without negatively impacting the overall GTG3 performance.

5.6 FORECAST CALIBRATION

Section 5.1 showed that the performance of GTG3 could be improved through the selection of a different forecast threshold. Due to the substantial change in the forecast distribution between GTG2.5 and GTG3, a direct comparison of the forecasts is less than straightforward. Calibration typically involves adjusting the forecast distribution toward a known observed distribution (or reasonable estimate thereof). As discussed previously, the true distribution of turbulence in the atmosphere is not known. An alternative calibration approach is to plot the performance of the forecasts as a function of the climatological quantiles instead of fixed thresholds. However much this transformation resembles the (unknown) true distribution, it should treat the two forecasts equally.

The method for this calibration is to convert the forecast distributions to cumulative distribution functions (Fig. 5.20) and then, for a large number of quantiles, locate the forecast threshold from each product that corresponds to this quantile. In the schematic shown in Fig. 5.20, the 50th percentile (as many forecast values occur greater than this value as occur less than it) is achieved by using a threshold of 0.1 from forecast A and a threshold of 0.06 from forecast B.

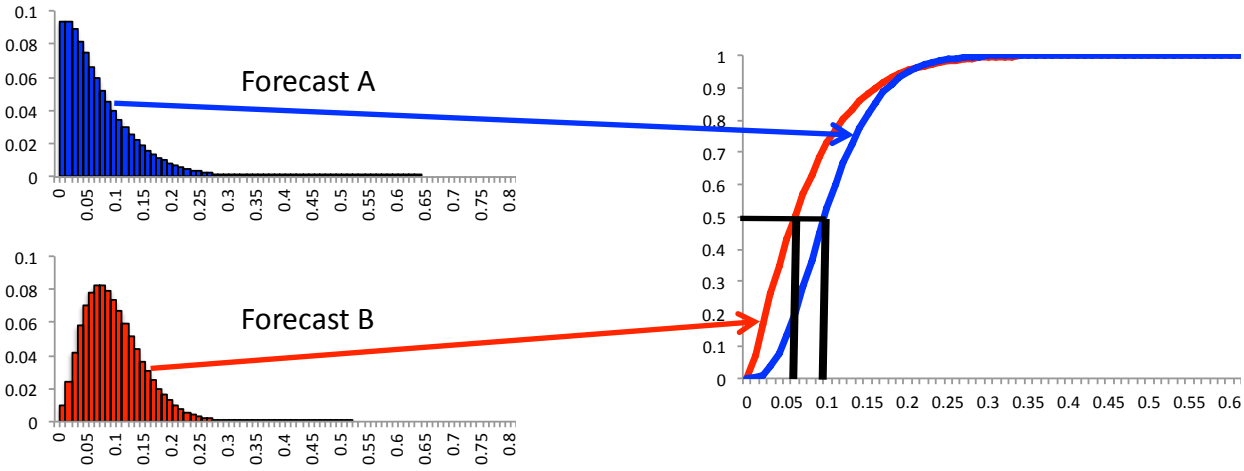

Figure 5.20: Schematic of calibration technique. Forecast intensity distributions are converted to cumulative distributions. The intensity thresholds are then mapped to climatological values, e.g., in the schematic the 50th percentile values are 0.1 for Forecast A and 0.06 for Forecast B. Forecast performance can then be plotted as a function of the climatological quantiles.

The results as applied to GTG3 and 2.5 are shown in Fig. 5.21. The left panel contains the ROC curves for GTG3 and GTG2.5 for the 0.2 observed EDR threshold, along with plots of the PSS, as a function of the forecast threshold. The view shows GTG3 outperforming GTG2.5 over only a narrow range of forecast thresholds around 0.18. The right panel redraws the PSS curves as a function of the climatological quantiles. In this calibrated view, GTG2.5 still outperforms GTG3 for the lighter (left) half of the distribution, but GTG3 now is seen to outperform GTG2.5 for the stronger (right) half of the distribution.

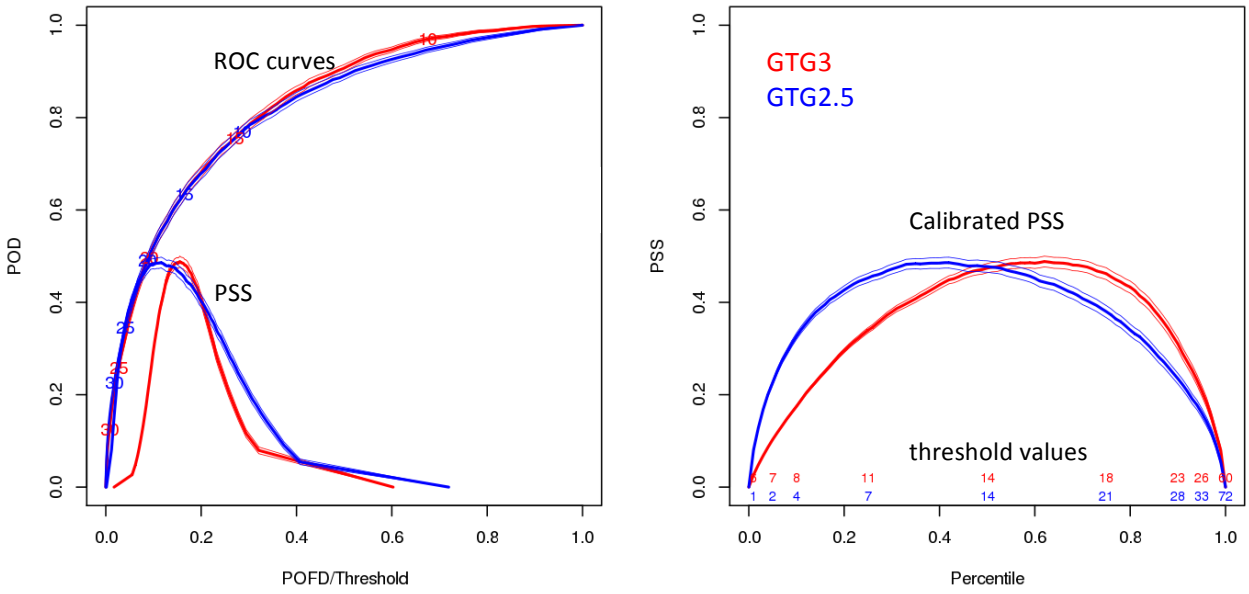

Figure 5.21: (Left panel) ROC curves of GTG3 (red) and GTG2.5 (blue) against 0.2 DAL EDR observations, overlaid with plots of PSS as a function of forecast threshold. (Right panel) PSS plotted as a function of the climatological quantiles of each forecast product. Forecast thresholds (number equals threshold * 100) corresponding to the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, ad 99th percentiles are show along the bottom. Thin lines represent the 5th and 95th percent confidence bounds.

SUMMARY

Without calibration, GTG3 is more skillful than GTG2.5 for only a narrow range of forecast thresholds. Using a simple calibration approach, GTG3 is superior to GTG2.5 for the stronger half of forecast thresholds.

6 SUMMARY

Version 3 of the Graphical Turbulence Guidance (GTG3) algorithm is intended to replace the current GTG2.5 algorithm currently being used for operational aviation turbulence decisions. Changes between GTG2.5 and GTG3 include: 1) an extension of the forecast domain down to 100 ft altitude (from 10,000 ft), 2) an increase in forecast leads from 12 to 18 hours, 3) the addition of an explicit mountain-wave turbulence component, and 4) an upgrade to the conversion of the raw algorithm output to EDR.

The assessment results are as follows:

The distribution of intensity values from GTG3 is substantially different from that of GTG2.5, with a narrower range of values and a peak shifted to the right (from around 0 for GTG2.5 to near 0.08 for GTG3). On average the result is that for Light turbulence forecasts, GTG3 tends to produce larger values than GTG2.5, while for MOG forecasts, GTG3 tends to produce lower values than GTG2.5. The GTG3 distribution most closely resembles that of triggered DAL EDR, while the GTG2.5 distribution is most similar to that of the combined triggered and heartbeat reports.

As measured by the area under ROC curves, GTG3 and GTG2.5 have similar ability to distinguish events from non-events, but GTG3 outperforms GTG2.5 consistently for all observed thresholds. Similarly, the distance between forecast values and observed values, as measured by RMSE scores, is slightly but consistently better for GTG3. When the forecast threshold is constrained to match the observed threshold (i.e., no calibration), GTG3 is more skillful than GTG2.5 for only a small range of thresholds; however, this range can be expanded with proper calibration.

GTG3 is able to capture many more MOG events than G-AIRMET forecasts using the same forecast volumes. Alternatively, GTG3 can capture the same number of events as G-AIRMETs using only one-third of the forecast volume. For the Near-surface layer, the gap between GTG3 and G-AIRMETs is reduced, but GTG3 is still superior. Used as a supplement to G-AIRMETs, GTG3 can reduce the number of false alarms inside a G-AIRMET by half while still capturing 80–90% of all MOG turbulence events. GTG3 is able to capture 60% of all MOG events found outside G-AIRMETs, while including only 20% of the non-MOG events.

The skill of GTG3 peaks in the Moderate range (~0.2), dropping off quickly for stronger or weaker turbulence (when calibration is not considered). For thresholds within the moderate range (0.18 – 0.24), skill is fairly steady, but the probability of detection and false detection decrease substantially as the threshold increases. Consequently, the optimal threshold is determined by a user's particular sensitivity to false alarms and missed events.

The GTG3 MW component captures 99% of Light-or-greater explicit MW PIREPs and 70% of MOG MW events in winter. False alarms are relatively high for Light-or-greater events (~ 60%) but nearly non-existent (5%) for MOG events. When all observed turbulence events within the MW domain are considered (i.e., not only the explicit MW PIREPs), the performance of GTG3 using the CAT component combined with the MW component is nearly identical to the performance of the CAT component alone. In other words, the GTG3 MW component successfully identifies MW events without negatively impacting the overall GTG3 performance.

ACKNOWLEDGMENTS

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA. The authors would like to thank the Turbulence Product Development Team for providing the forecast data that was needed for the evaluation.

REFERENCES

Brown, B.G., and G.S. Young, 2000: Verification of icing and icing forecasts: Why some verification statistics can't be computed using PIREPs. Preprints, 9th conference on Aviation, Range, and Aerospace Meteorology, Orlando, FL, Sep. 11-15, American Meteorological Society (Boston), 393-398.

Murphy, M. P., 2010: Product Description Document, Graphical Airman's Meteorological Advisory (G-AIRMET). Available at: http://aviationweather.gov/static/docs/gairmet/G-AIRMETPDD_2010.pdf

NWS, 2007: Aviation Weather Services, Advisory Circular AC 00-45F. U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, and U.S. Department of Transportation, 393 pp.

Wandishin, M.S., B.P. Pettegrew, M.A. Petty, and J.L. Mahoney, 2011: Quality Assessment Report: Graphical Turbulence Guidance, version 2.5. NOAA Technical Memorandum OAR GSD-39, 44 pp.

7 APPENDIX

7.1 GTG3 PERFORMANCE

Comparisons of the performance of GTG3 among the four geographic regions and among the four vertical layers were provided in section 5.4. The corresponding figures are provided here.

7.1.1 BY GEOGRAPHIC REGION

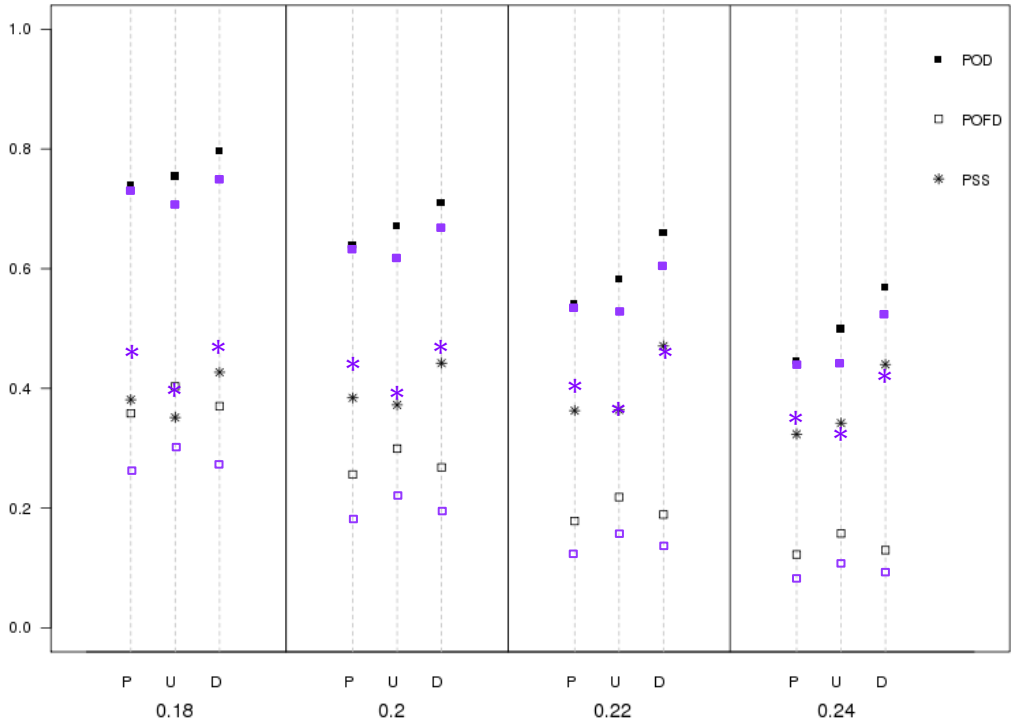

Figure 7.1: POD (filled squares), POFD (hollow squares), and PSS (stars) measured against PIREPs UAL EDR, and DAL EDR (indicated by the P, U, and D, respectively, along the bottom of the plot) for GTG3 (left) for the forecast thresholds 0.18, 0.2, 0.22, 0.24 for the West region. A PIREP threshold of 1 is used for the 0.1 forecast threshold, 3 for the 0.2 threshold, and 5 for the 0.3 and 0.4 thresholds. Purple markers indicate the scores for all regions (see Fig. 5.17).

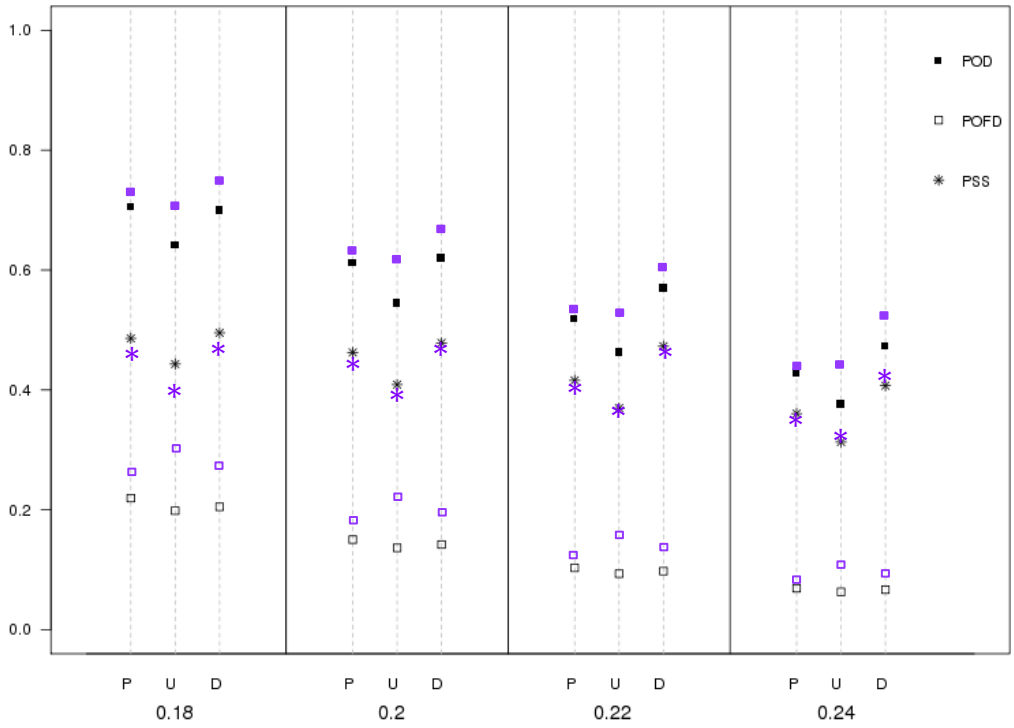

Figure 7.2: As in Fig. A.1, but for the Central region

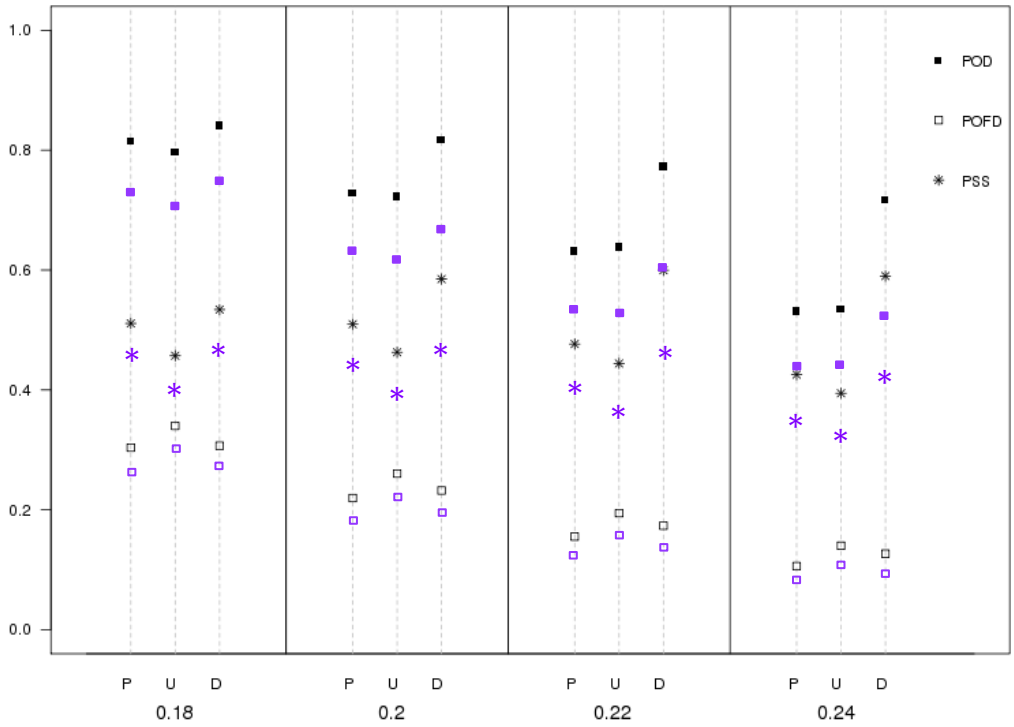

Figure 7.3: As in Fig. A.1, but for Northeast region.

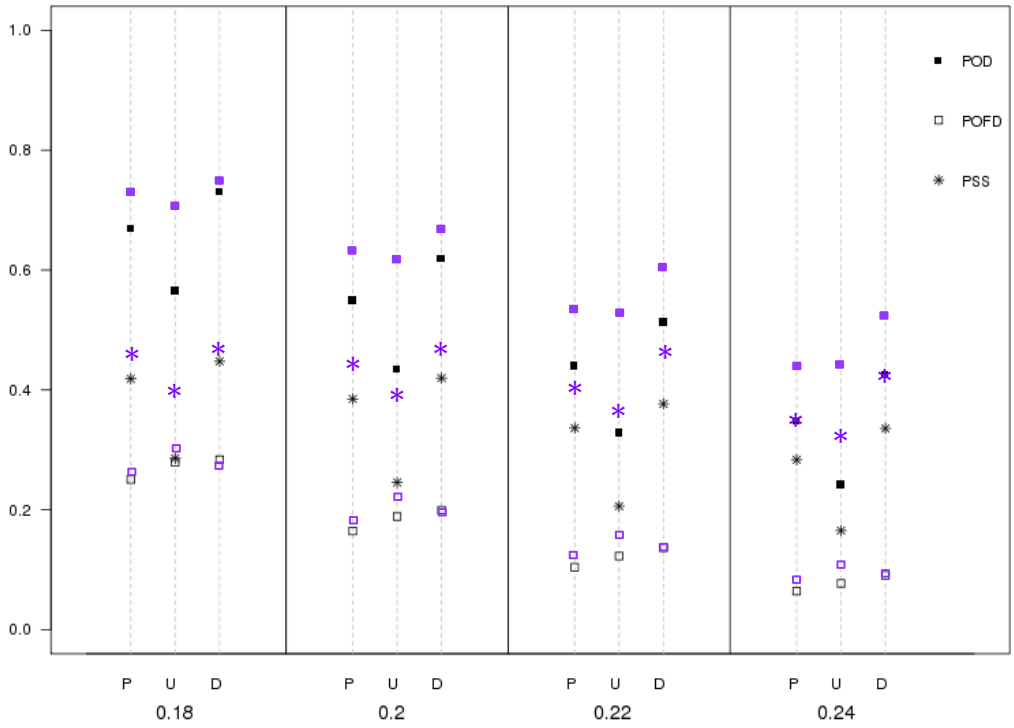

Figure 7.4: As in Fig. A.1, but for Southeast region.

7.1.2 GTG3 PERFORMANCE BY ALTITUDE

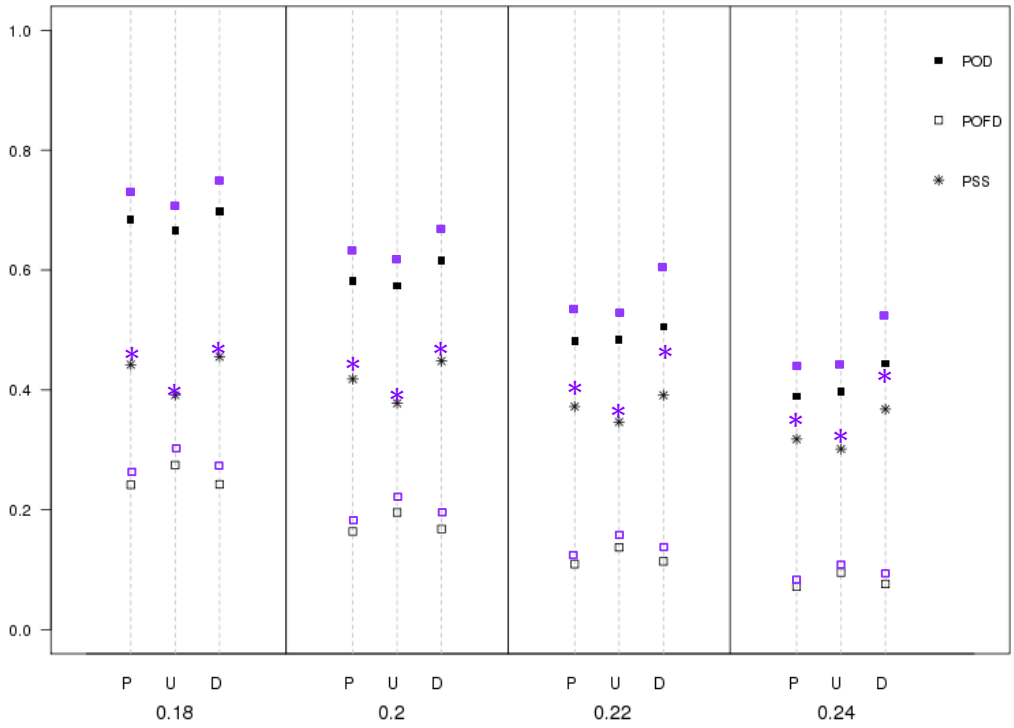

Figure 7.5: As in Fig. A.1, but for the High (30,000 – 50,000 ft) layer.

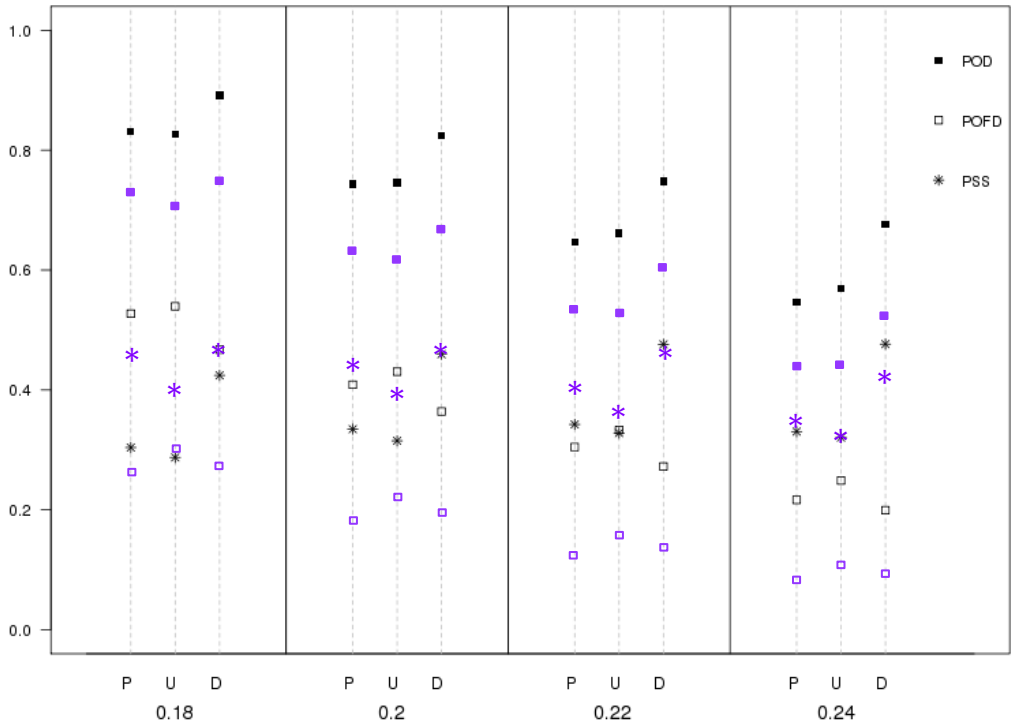

Figure 7.6: As in Fig. A.1, but for the Middle (20,000 – 29,999 ft) layer.

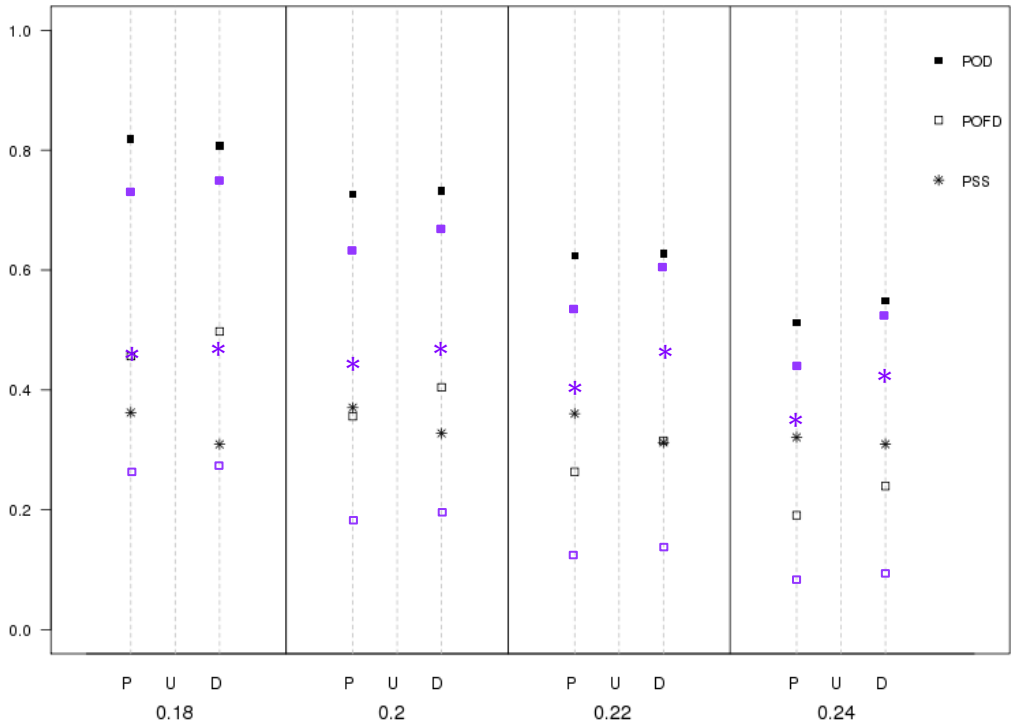

Figure 7.7: As in Fig. A.1, but for the Low (10,000 – 19,999 ft) layer. UAL EDR data are not reliable below 20,000 ft and so are not included.

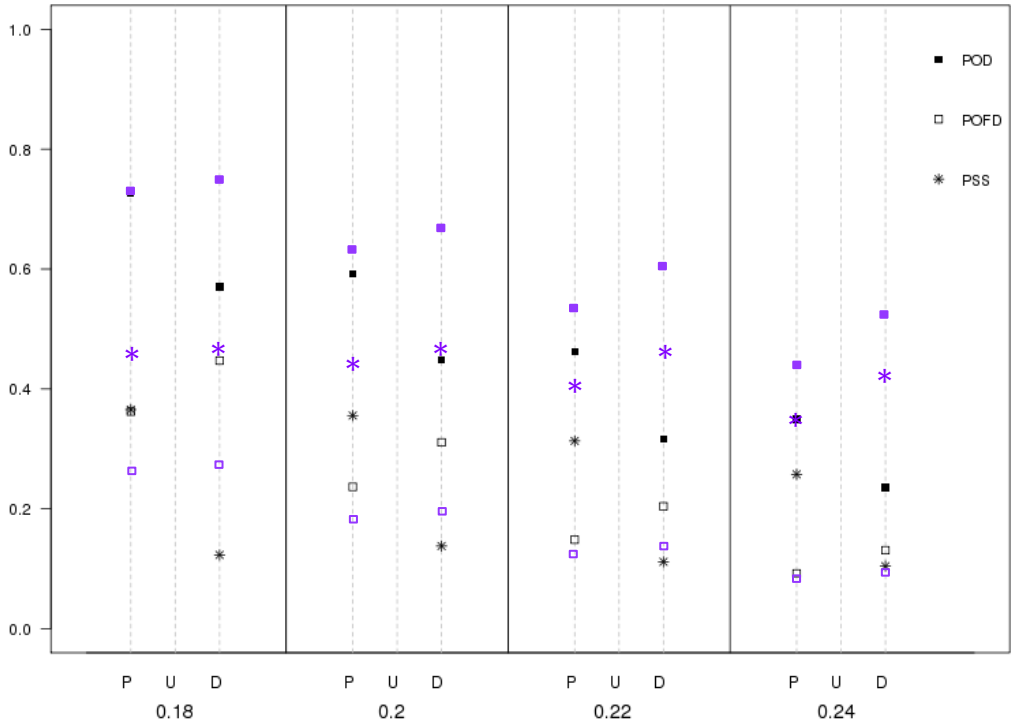

Figure 7.8: As in Fig. A.7, but for the Near-Surface (0 – 9999 ft) layer.

7.2 GTG3 Conditional Forecast Distributions

The distribution of forecast values conditioned on the observations (e.g., the distribution of forecast values from all forecast-observation pairs for which the observations are of Light turbulence; Fig. A.9) is the fundamental basis of ROC curves. The area under the ROC curve is related to the distance between the means of the event and non-event distributions. The greater the separation of the means, the more the forecasts are able to distinguish between events and non-events, and so the greater the AUC. Figure A.9 shows the distributions for Null (black), Light (blue), Moderate (red), and Severe (green) turbulence for both EDR (top) and PIREP (bottom). To construct a ROC curve for Moderate turbulence, one would need to combine the Null and Light curves to construct a non-event distribution, and combine the Moderate and Severe curves to form an event distribution. The Moderate curve is better separated from the Light and Null curves when using PIREP observations than when using EDR observations. This is confirmed by the AUC: 0.809 for EDR observations compared with 0.820 for PIREP observations (not shown).

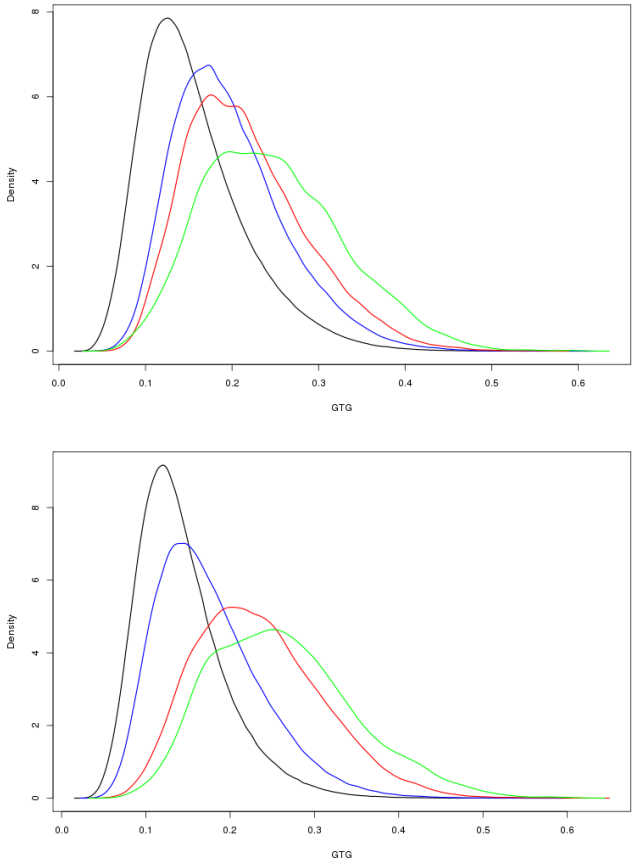

Figure 7.9: Distribution of GTG3 forecast intensities conditioned on the observational intensity (Null, black; Light, blue; Moderate, red; Severe, green) for EDR (top) and PIREP (bottom) observations. For EDR, Null: EDR <= 0.1, Light: 0.1 < EDR < 0.2, Moderate: 0.2 <= EDR < 0.3; Severe: EDR >= 0.3. For PIREPs, Null: PIREP = 0; Light: PIREP = 1,2; Moderate: PIREP = 3,4; Severe: PIREP >= 5.