NOAA Technical Memorandum OAR GSD-60
https://doi.org/10.7289/V5/TM-OAR-GSD-60

Assessment of Graphical Turbulence Guidance, Global (GTG-G), Part 1

March 2018

Arlene Laing
Matthew S. Wandishin
Joan E. Hart
Melissa A. Petty

Earth System Research Laboratory
Global Systems Division
Boulder, Colorado
March 2018

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

NOAA Technical Memorandum OAR GSD-60

Assessment of Graphical Turbulence Guidance, Global (GTG-G), Part 1

Arlene Laing[1]
Matthew S. Wandishin[2]
Joan E. Hart[2]
Melissa A. Petty[1]

[1] Cooperative Institute for Research in the Atmosphere (CIRA) and NOAA/ESRL/GSD

[2] Cooperative Institute for Research in Environmental Sciences (CIRES) 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.

UNITED STATES DEPARTMENT OF COMMERCE

Wilbur Ross
Secretary

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Benjamin Friedman
Acting Under Secretary for Oceans And Atmosphere/NOAA Administrator

Office of Oceanic and Atmospheric Research

Craig N. McLean
Assistant Administrator

TABLE OF CONTENTS

LIST OF FIGURES

EXECUTIVE SUMMARY

The Quality Assessment Product Development Team (QA PDT) was tasked with assessing a global version of the Graphical Turbulence Guidance (GTG-G) as part of the transition process to operations. The GTG-G was developed by the Turbulence Product Development Team in the Research Applications Laboratory at the National Center for Atmospheric Research (NCAR/RAL), under sponsorship from the FAA's Aviation Weather Research Program (AWRP). The study verified the GTG-G using a variety of observation sets including pilot reports (PIREPs), aircraft reports (AIREPs), aircraft sensor measurements of Eddy Dissipation Rate (EDR), and derived equivalent vertical gust (DEVG).

This evaluation of GTG-G provides a preliminary understanding of GTG-G performance, using current measures of the World Area Forecast Center (WAFCs) for turbulence forecasts, namely, Receiver Operator Characteristics (ROC) curves. The Area under the ROC curve (AUC) of the GTG-G EDR forecasts was compared with that of the operational World Area Forecast System (WAFS) Clear Air Turbulence (CAT) potential. Forecast performance was examined for Moderate-or-Greater (MOG) turbulence only, using the EDR threshold for medium aircraft (0.2). The assessment period was 1 July through 31 October 2017.

The verification study found that:

- GTG-G forecast products generally outperformed the WAFS CAT potential (e.g., Figure 1).
- The performance gap between GTG-G and WAFS was greater for EDR comparisons than for PIREP comparisons (section 5).
- Forecasts exhibited better skill when compared with EDR than when compared with PIREPs.
- Performance was similar when compared within a ± 1.5-hr window around the forecast valid time than with ± 3-hr window (section 5).

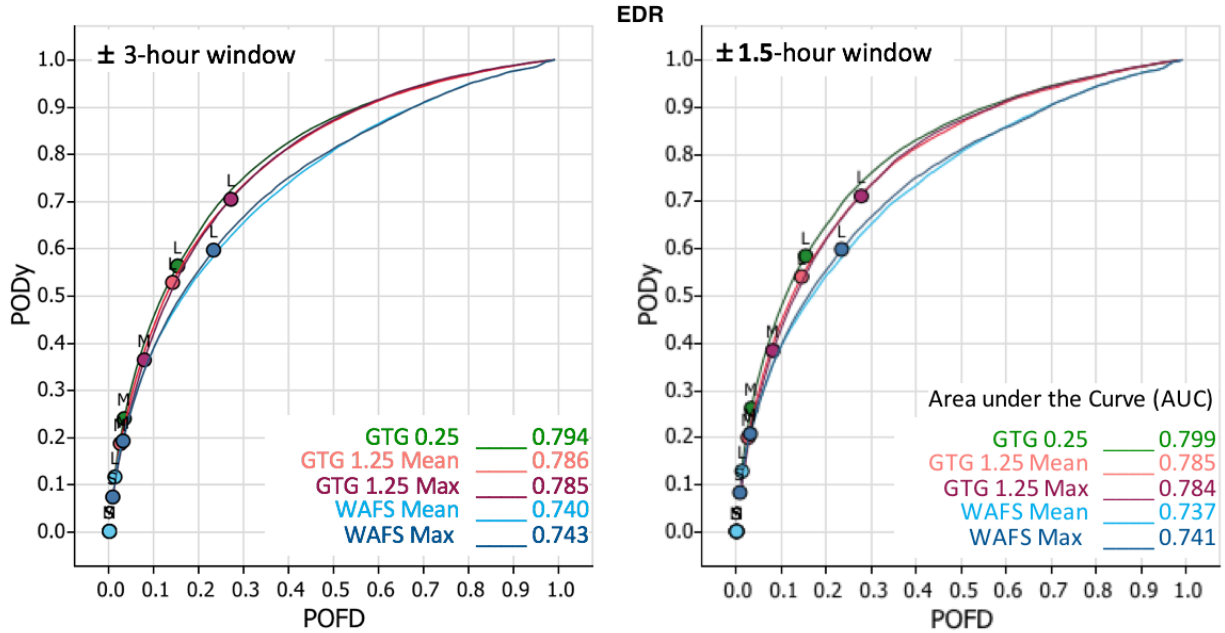

Figure 1. ROC curves for turbulence forecasts compared with EDR for the whole globe during the assessment period, using a ±3-hr window (left) and a ± 1.5-hr window (right). The AUC for each forecast product is color-matched with the corresponding ROC curve. The GTG-G forecasts are on 0.25° and 1.25° grids. The letters "L," "M," and "S" mark the forecast thresholds for "Light," "Moderate," and "Severe" turbulence, respectively.

The performance characteristics were broadly similar when evaluated with the stratifications listed below: although a few exceptions occurred with certain stratifications:

Geographic:

- Latitude Bands: Northern Hemisphere (NH) extratropics, tropics; Southern Hemisphere (SH) extratropics
- World Meteorological Organization (WMO) and UK Met Office (UKMO) areas (See Appendix A – WMO/Met Office Area Definitions for definitions): North Atlantic, North America, Asia, North Pacific, and Australia-New Zealand

Temporal:

- Issues (00, 06, 12, and 18 UTC)
- Lead times (12, 18, 24, 30, and 36 hours)

Vertical layers:

- 50-mb layers centered on six pressure levels/flight levels: 400 mb/FL240, 350 mb/FL270, 300 mb/FL300, 250 mb/FL340, 200 mb/FL390, and 150 mb/FL440.

The performance by geographic area generally reflects the global characteristics and there is little change in performance as a function of lead time or issuance (Figure 2, section 5.2).

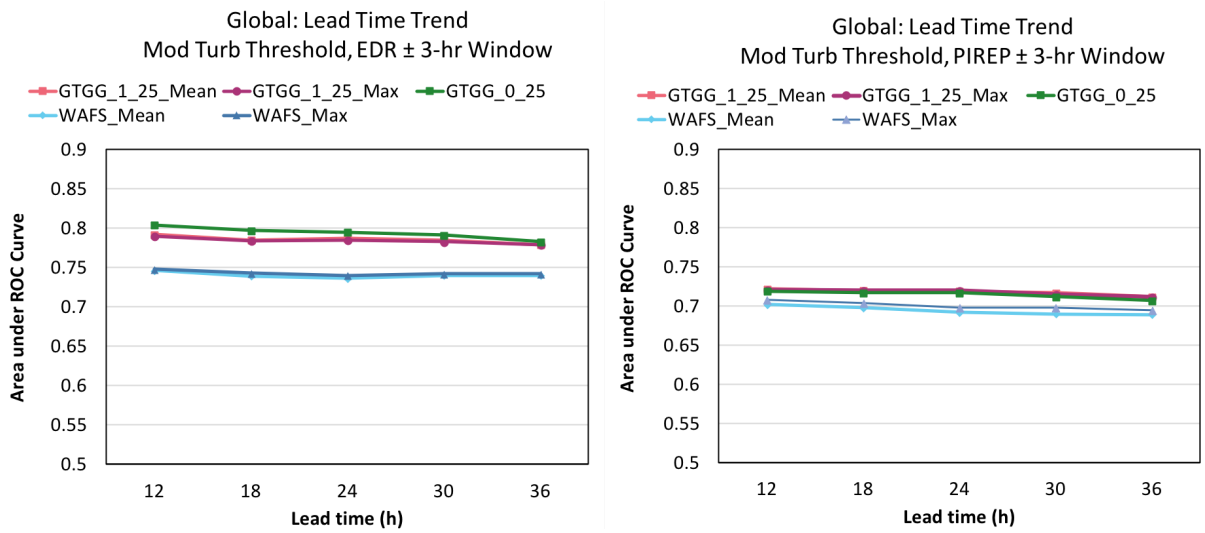

Figure 2. The AUC for the global set of observations for (left) EDR and (right) PIREP by lead time (h) within a ± 3-hour window.

When stratified vertically, the GTG-G Max-EDR AUC exceeded WAFS Max CAT potential except for the 200-mb layer where WAFS had a slight advantage over some NH sub-areas for PIREP comparisons (section 5.3).

Note that for the period of the assessment, the bulk of the observations were over the Northern Hemisphere (section 2.2) during the NH summer and autumn months. Those are seasons when overall model forecast skill is lower. With weaker and more poleward jet streams, CAT and MWT episodes are less frequent than during winter.

1 INTRODUCTION

To address the need for an automated turbulence forecasting tool, the Research Applications Laboratory at the National Center for Atmospheric Research (NCAR/RAL), under sponsorship from the FAA's Aviation Weather Research Program (AWRP) as the Turbulence Product Development Team (TPDT), has developed an automated turbulence forecasting system known as the Graphical Turbulence Guidance (GTG) system (Sharman et al. 2006). The GTG forecasting method ingests Numerical Weather Prediction (NWP) data as the basis for large-scale features of the atmosphere, from which small-scale turbulence forecast information is derived. The algorithm integrates a combination of several separate turbulence diagnostics, with each diagnostic weighted based upon agreement with available observations (i.e., PIREPs). The current operational GTG product is the version 3.0 (GTG3) disseminated on a 13-km grid covering CONUS (Sharman and Pearson 2017). In support of both the London and Washington WAFCs, the TPDT extended the GTG product to cover the entire globe. The GTG algorithm is run at the National Centers for Environmental Prediction (NCEP) as part of the post-processing of the Global Forecast System (GFS) model output. The product is then upscaled to both a 0.25° (~28 km) grid and a 1.25° (~140 km) grid, the latter in agreement with the current WAFS grid.

This document presents the results of an initial evaluation that was designed to align with the current WAFS verification and provide an early look at forecast performance to the WAFS

turbulence forecasts users. The assessment was coordinated in consultation with NCEP, NCAR, the Aviation Weather Center (AWC), and the United Kingdom Met Office (UKMO).

The details about the forecast and observation data used in the assessment are given in sections 2.1and 2.2, respectively. The stratifications of the assessment are given in section 3, the verification methods are described in section 4, and the results are presented in section 5.

2 DATA

This section describes the turbulence forecast and observation data that is used for the assessment. The time period for the GTG-G assessment encompasses July through October 2017.

Because of a gap in GTG-G data during 24 July to 25 August 2017, the original data collection end-date extended from 30 September to 31 October 2017. The GTG-G data for that period was later backfilled with model runs using code that had undergone minor changes but with the original GTG configuration used for the rest of the assessment period. The backfilled GTG-G data gap period was analyzed separately from the rest of the data period.

Note that the assessment period encompassed the NH summer and autumn, when overall model forecast skill is lower and NH jet streams are weaker and more poleward, while the SH has stronger more equatorward jet streams (Figure 3).

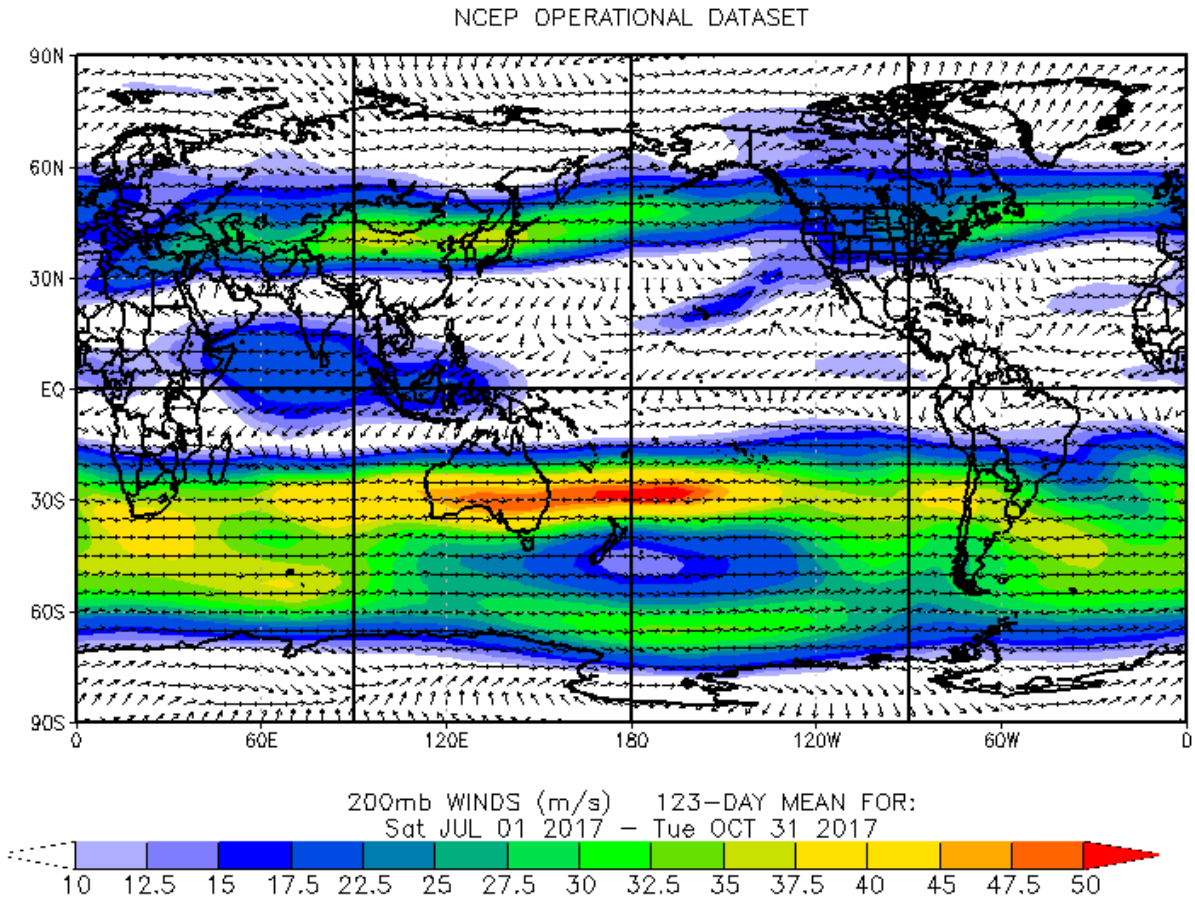

Figure 3. The mean wind velocity at 200 mb, from the NCEP operational analysis, for the assessment period 1 July to 31 October 2017.

2.1 Forecast Products

2.1.1 GTG-G

The GTG-G is calculated on the native GFS grid (~13 km) and then upscaled to both a 0.25° and 1.25° resolution, the latter matching the WAFS grid. The output of the algorithm is Eddy Dissipation Rate (EDR; $m^{2/3}s^{-1}$), with values ranging from 0.0 to 1.0 (Sharman et al. 2006 and Sharman and Pearson 2017). The underlying GTG-G algorithm is the same as GTG3, but employs fewer diagnostics than are used in GTG3. The algorithm is applied to output from the GFS model, which is run at six-hourly synoptic intervals (0, 6, 12, 18 UTC). GTG-G output takes into consideration both clear-air turbulence and mountain wave-induced turbulence (Kim et al. 2018).

2.1.2 World Area Forecast System (WAFS)

The WAFS is a 1.25° gridded product of clear-air turbulence (CAT) potential produced using the Ellrod TI1 Index (Ellrod and Knapp, 1992), which combines vertical wind shear and deformation. The CAT forecasts from both WAFCs are linearly scaled to form a CAT potential. Note that the CAT forecasts from WAFC-London include a mountain wave turbulence diagnostic that is detectable in the maximum turbulence. This product is issued four times a day (0, 6, 12, 18 UTC) with lead times

ranging from 6 to 36 hours with three-hour intervals. The CAT potential is reported at six pressure levels/flight levels (400 mb/FL240, 350 mb/FL270, 300 mb/FL300, 250 mb/FL340, 200 mb/FL390, and 150 mb/FL440).

The WAFS forecasts are harmonized between the two WAFCs by taking the higher value of the two forecasts for the maximum CAT potential in each grid box. The mean CAT potential is the mean of the means of the two WAFC forecasts for each grid box. The basic characteristics of the WAFS and GTG-G are summarized in Table 1.

Table 1. Comparison of the characteristics of GTG-G and WAFS.

	GTG-G	**WAFS CAT Potential**
Model from which forecasts are derived	Global Forecast System (GFS)	GFS (WAFC Washington) and Unified Model (WAFC London)
Algorithm basis	Multiple turbulence indices	Ellrod Index
Type of turbulence predicted	CAT and MWT	CAT (WAFC London includes MWT)
Grid spacing	0.25° and 1.25°	1.25°
Vertical spacing	25 hPa	Selected flight levels
Issuance	0, 6, 12, 18 UTC	0, 6, 12, 18 UTC
Lead time	6-36 h, every 6 h	6-36 h, every 3 h

2.2 Observational Products

2.2.1 Pilot Reports / Air Reports (PIREPs/AIREPs)

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. The global PIREP coverage of turbulence reports for the assessment time period is given in Figure 4.

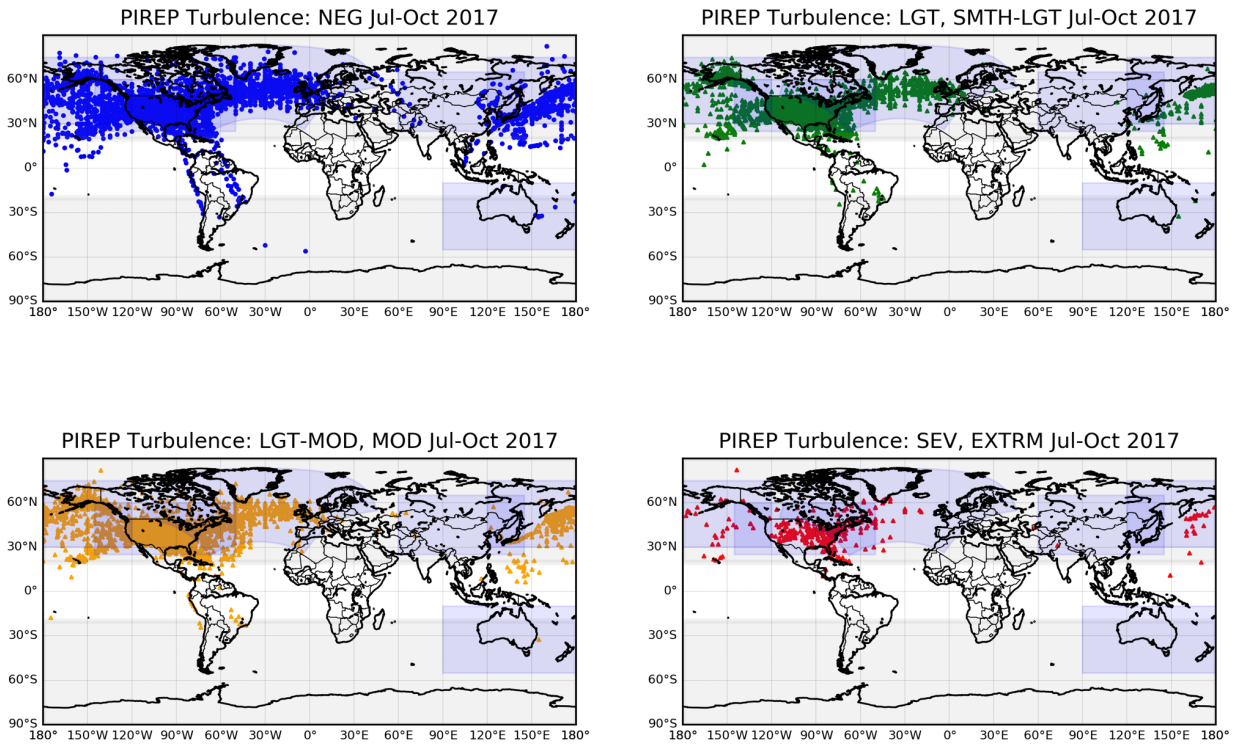

Figure 4. The global coverage of the PIREP turbulence reports at or above FL200 for the assessment time period. The blue shaded regions mark specialized areas for WMO and UKMO Evaluations; the grey-scaled areas mark latitudinal zones for the tropics and extratropics, the areas are defined in Appendix A – WMO/Met Office Area Definitions.

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 Delta Airlines (DAL) and United Airlines (UAL) 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 to 20 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 >= 0.06

Triggered reports provide the previous six minutes of EDR values (at one-minute resolution), 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. The global coverage of EDR reports for the assessment time period is shown in Figure 5.

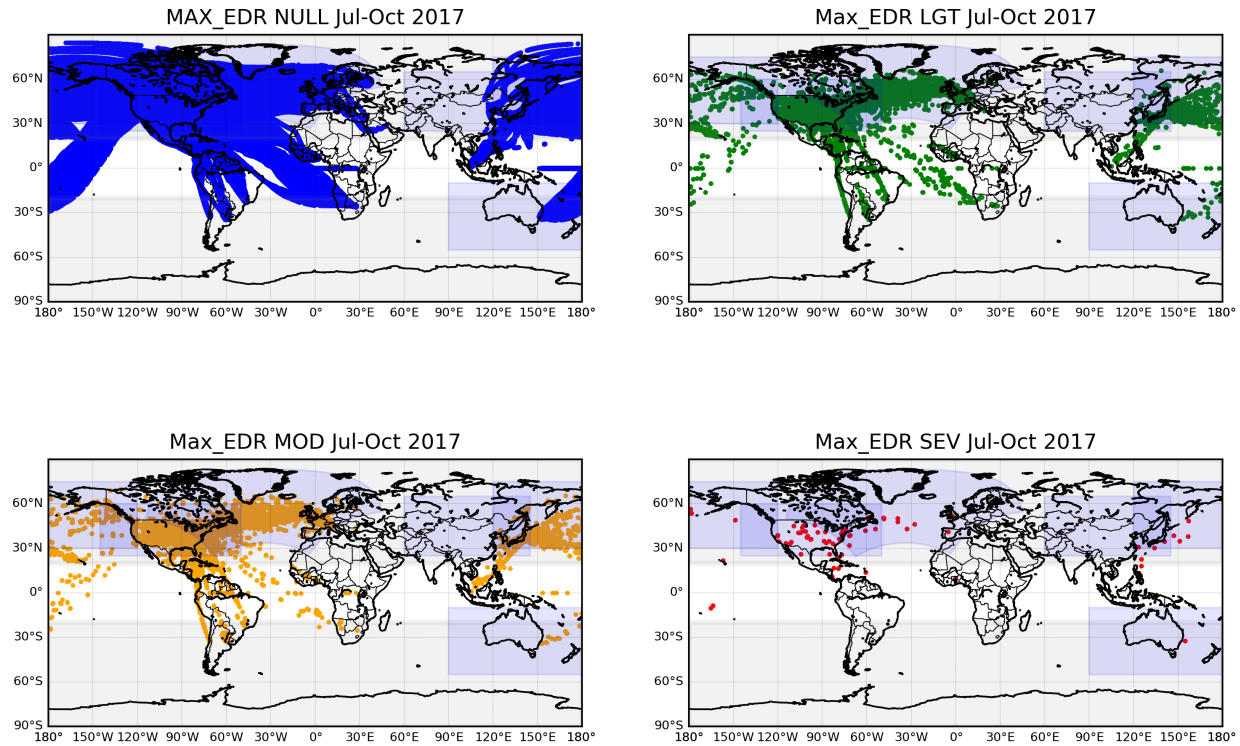

Figure 5. Same as Figure 4, except for the global coverage of the EDR reports above FL200 for the assessment period.

2.2.3 Aircraft Meteorological Data Reporting (AMDAR)

Among direct turbulence observations, the most prominent is the Aircraft Meteorological Data Reporting (AMDAR) turbulence data. The number of airlines contributing to AMDAR reports has grown steadily over the last two decades, reaching roughly 700,000 reports, daily. However, turbulence reports make up only a small fraction of that tally and nearly 85% of all AMDAR reports are from US-based airlines (WMO 2015a,b). Another challenge presented by the AMDAR data is that the non-US-based airlines use a different calculation of turbulence than EDR. Instead, airplane vertical displacements are combined with aircraft-specific information (e.g., speed, fuel levels, aircraft type, etc.) to compute the Derived Equivalent Vertical Gust (DEVG; Stickland 1988). In the absence of DEVG and EDR calculations from the same aircraft, either a less robust calibration between the two reports must be used, or verification should be performed separately for each data set. Still, with DEVG being reported from European and Australian airlines, they provide complementary coverage to the US-based EDR reports.

With the global coverage of AMDAR DEVG turbulence reports (Figure 6), particularly in the Southern Hemisphere where EDR reports are few, we anticipated using the DEVG to supplement

the EDR and PIREPs. However, our investigation of the AMDAR DEVG uncovered problematic observations over the Australia-New Zealand area, e.g., different algorithms for 737 and 747 aircraft, as well as anomalously high report values for a subset of flight legs for certain aircraft, differing by orders of magnitude from other reports on similar routes close in time. With no systematic means of identifying those problematic reports, it was determined that the AMDAR-DEVG would not be used as a primary truth set to verify the GTG-G. Details of the AMDAR-DEVG data analysis and use in verification are presented in Appendix B – AMDAR-DEVG Analysis and results.

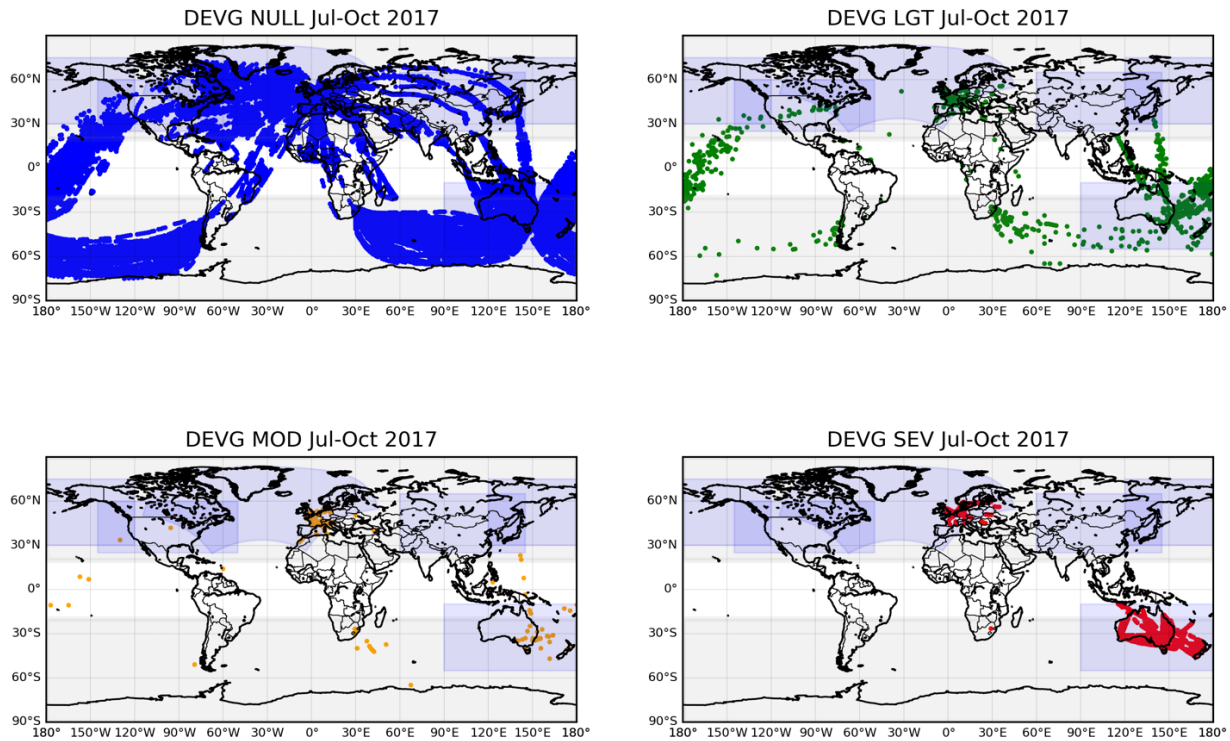

Figure 6. Same as Figure 4 except for the global coverage of the AMDAR-DEVG turbulence reports above FL200 for the assessment period.

3 STRATIFICATIONS

Evaluations were performed for several geographic, temporal, and turbulence intensity stratifications, as outlined below.

3.1 INTENSITY STRATIFICATION (THRESHOLDS)

Forecast performance was examined for Moderate-or-Greater (MOG) turbulence (Table 2) which is considered a critical significant weather forecast threshold (ICAO 2007) and is also used currently in verifying the WAFS turbulence forecasts. While many overseas flights use heavy aircraft, medium-sized aircraft also fly these routes (along with domestic route), and so MOG turbulence is defined using the medium aircraft EDR threshold of 0.2 (Table 2).

Table 2. Thresholds for turbulence intensity categories. GTG-G is an EDR forecast and uses the EDR thresholds. The WAFS forecast is for turbulence potential, not intensity, but threshold values were determined by matching the frequency distribution of the EDR.

PIREP	EDR/GTG	WAFS CAT Potential	
	Medium Aircraft	Max	Mean
Null		4	1.5
Light	0.15	7	3.7
Moderate	0.2	10	5.5
Severe	0.44	33	15

3.2 GEOGRAPHIC STRATIFICATION

Verification was performed for two types of geographical stratifications: (i) latitude bands and (ii) areas defined for verification of deterministic numerical weather prediction by the WMO (WMO 2016) and the UKMO. The areas are specified below and illustrated in Figure 7.

- Latitude Bands:
  - Northern Hemisphere (20˚N–90˚N)
  - Tropics (20˚N–20˚S)
  - Southern Hemisphere (20˚S–90˚S)

- WMO/UKMO Areas (See Appendix A – WMO/Met Office Area Definitions for definitions):
  - North Atlantic (Area 2)
  - North America (Area 141)
  - Asia (Area 143)
  - North Pacific (Area 145)
  - Australia / New Zealand (Aus/NZ)

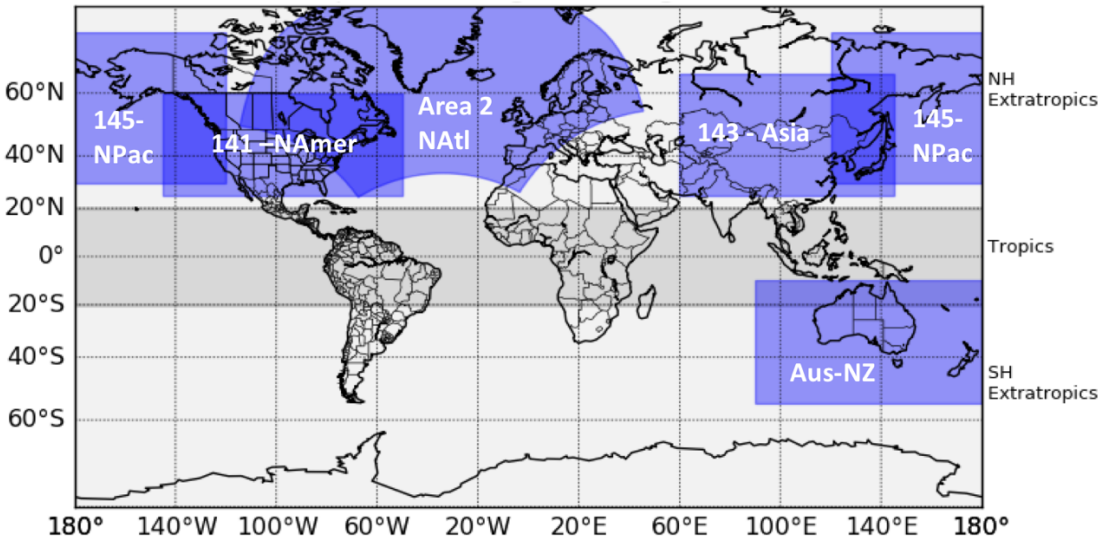

Figure 7. The global observational coverage for the assessment period overlaid by the WMO/UK MET areas (blue). The dark grey distinguishes the tropics from the extratropical latitude bands.

3.3 Temporal Stratification

The skill and characteristics of GTG-G were examined by issue (00, 06, 12, and 18 UTC) and by lead times (12, 18, 24, 30, and 36 hours).

3.4 Vertical Stratification

Verification was performed for flight levels at 24 kft and above and stratified based on WAFC levels and the associated layers used by the UKMO for WAFS turbulence verification:

- 400 mb → [425, 375 mb] or FL240 → [FL225, FL255)
- 350 mb → [375, 325 mb] or FL270 → [FL255, FL285)
- 300 mb → [325, 275 mb] or FL300 → [FL285, FL320)
- 250 mb → [275, 225 mb] or FL340 → [FL320, FL365)
- 200 mb → [225, 175 mb] or FL390 → [FL365, FL415)
- 150 mb → [175, 125 mb] or FL440 → [FL415, FL475].

4 Methods

The WAFS turbulence products served as baseline products against which to compare the GTG-G. As indicated in the introduction, this assessment mimics the official WAFS turbulence verification. The GTG-G and WAFS forecasts were verified against the observational products listed in section 2.2 to objectively quantify their performance. The performance comparison quantifies the added value, if any, by GTG-G over the current turbulence guidance products.

The following subsections outline the methods in more detail.

4.1 FORECAST PERFORMANCE

The forecast performance was determined by pairing each observation to the forecast in horizontal, vertical, and temporal space, then defining the yes/no events for the paired forecast and observation, and, finally, evaluating the skill given the agreement between the pairs (i.e., the contingency table).

4.1.1 FORECAST-OBSERVATION PAIRING TECHNIQUES

All the observational data used in this assessment were reported as point data with an associated turbulence value (Intensity, EDR, or DEVG), latitude/longitude coordinate, vertical location, time, and other supplemental information such as the quality of the observation, etc. A point-to-grid matching was done to create forecast-observation pairs, as follows:

- Temporally, forecasts were matched to observations that occurred within two temporal blocks: (i) a ± 3-hour time window around the forecast valid time, and (ii) a ± 1.5-hour window around the forecast valid time.
- Vertically, all observations are mapped to the nearest forecast level. GTG-G provides forecasts every 25 mb, but only the levels corresponding to the WAFS turbulence forecasts were used in this evaluation.
- Horizontally, the 1.25˚ GTG-G grid was produced by upscaling the 0.25˚ GTG-G grid, creating two upscaled grids: the mean and max of the constituent finer grid data points.
- When multiple measurements occurred within a forecast grid box, the maximum value among the included measurements was matched to the grid-box value, consistent with the current approach used by the WAFC. PIREPs were treated in a manner similar to the other observing platforms, even though it is known that uncertainties exist in the timing and location of PIREPs (Pearson and Sharman 2013).
- For all grids, the forecast value resulted from bi-linear interpolation of the four corner points of each grid box to the location of the maximum observation.

4.1.2 EVALUATIONS

Terminology and definitions of the statistics used in this assessment are:

MOG	Moderate-or-Greater Turbulence
POD	Probability of Detection: proportion of all observed events that are correctly forecast to occur, in this case, of detecting turbulence at or above a specific threshold
POFD	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
ROC	Receiver Operating Characteristic: curve made up of (POFD, POD) pairs as the forecast threshold is varied, where the line along the diagonal means no skill
AUC	Area Under the (ROC) Curve: measure of ability of forecast to correctly distinguish between events and non-events, where a larger AUC implies better performance.

The association of the GTG-G product to observations 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 and POFD will be 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. Those evaluations were conducted for the 0.25° GTG-G and the 1.25° GTG-G resolutions to understand the impact of grid resolution on performance.

5 RESULTS

For the globe as a whole, the Area under the ROC curve (AUC) for the GTG-G forecasts exceeds that of the WAFS forecasts for both the maximum and mean turbulence and within the ± 3-hour and ± 1.5-hour windows (Figure 8). Those values account for all layers, all issuances, and all lead times.

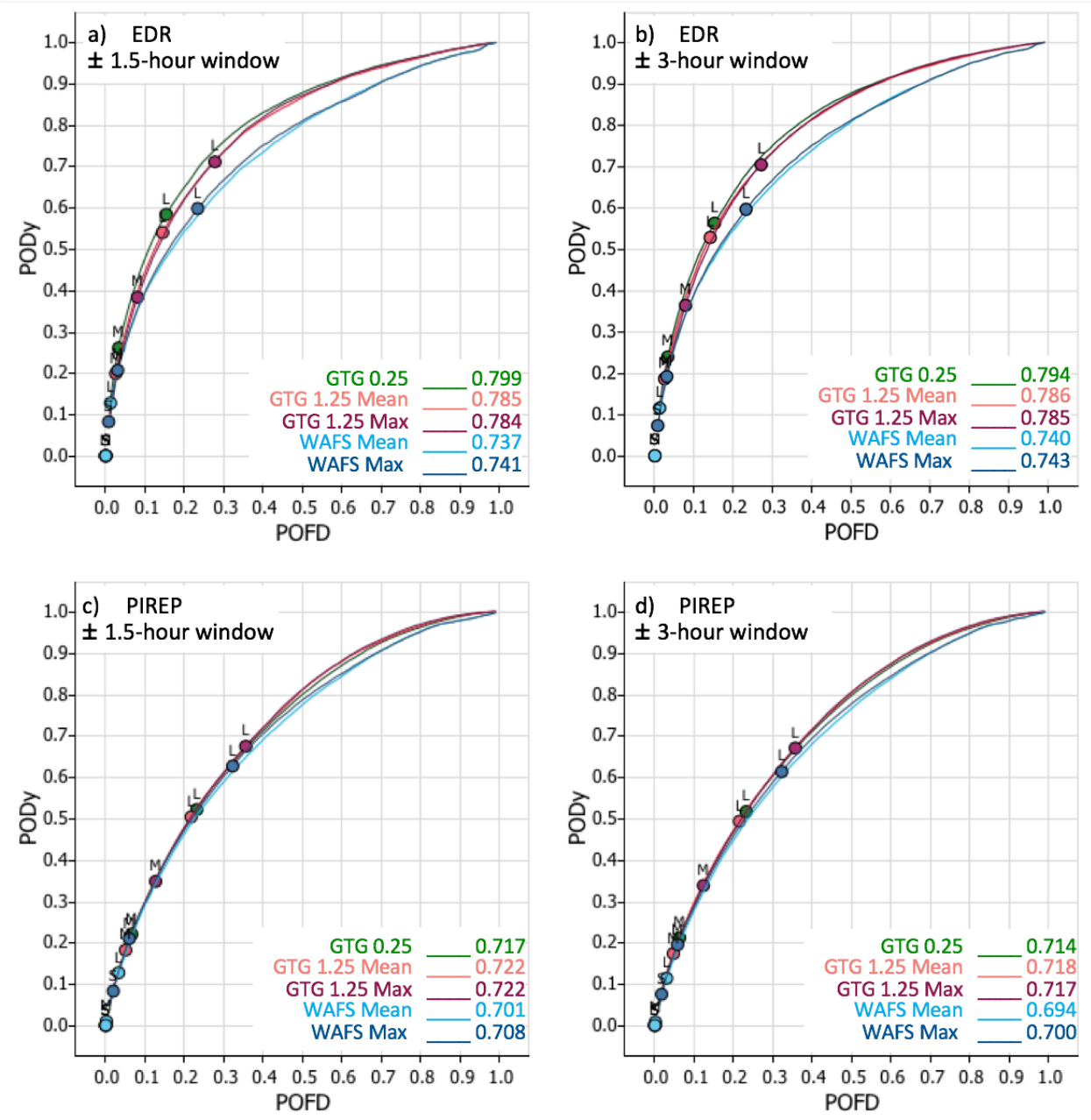

Figure 8. The ROC curves for the globe (includes all issuances, lead times, and all layers) compared with EDR (upper) and PIREP (lower). On the left are ROC curves for the ± 1.5-hour and on the right are the ± 3-hour window.

For all forecasts, the AUC is lower for PIREP comparisons than with EDR comparisons. For example, within the ±1.5-hr window, the AUC for the GTG-G Maximum EDR on the 1.25 °-grid (hereafter GTG-G 1.25 Max) was 0.784 compared with 0.722 for PIREP comparisons (Figure 8a, c). Also clear from Figure 8 is that the difference between the AUC for GTG-G and WAFS is smaller for PIREP comparisons than for EDR comparisons. Interestingly, the AUC is slightly lower for GTG-G 0.25° than the lower-resolution GTG-G 1.25 ° forecasts when compared with PIREPs, perhaps because location errors in PIREPs are more likely to affect a higher-resolution forecast.

While verification was conducted for the GTG-G forecasts at 0.25° and 1.25° and the WAFS mean and maximum, the remainder of the results will focus mainly on the GTG-G 1.25° Max and the WAFS Max. The emphasis is for consistency with current forecast grids, WAFC guidance, and user familiarity. The WAFCs recommend use of the maximum CAT forecasts for planning, while the mean is used to ascertain the confidence in the maximum turbulence forecast (ICAO 2012).

5.1 Geographic Stratification

The global performance characteristics are also observed when the forecasts are stratified by latitudinal zones and areas (Figure 9). Among the latitudinal zones, GTG-G forecasts performed best when compared with EDR observations over the SH extratropics, which had relatively few observations but experienced winter and spring during the assessment period. The WAFS skill is especially poor over the tropics (overall and relative to GTG-G) for both EDR and PIREP comparisons, where the latter had very few "Yes" reports of MOG turbulence.

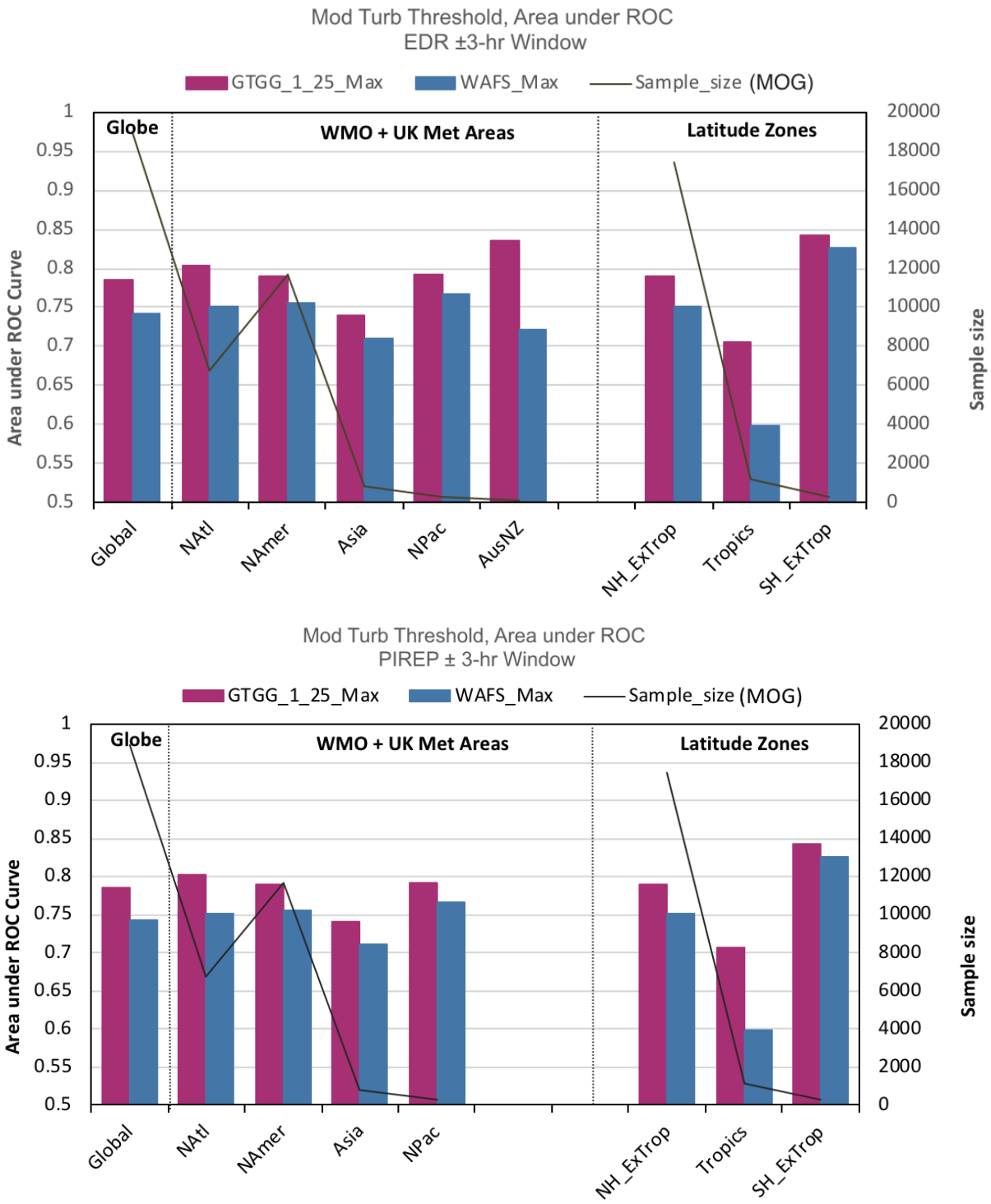

Figure 9. The AUC for specialized area and zones forecasts compared with EDR (upper) and PIREP (lower) for GTG-G Maximum EDR and WAFS Maximum CAT Potential over the ± 3-hour window. The sample sizes are for MOG = "Yes".

For the sub-areas with sufficient sample size, the AUC was relatively similar compared with EDR. On average, both sets of forecasts performed best over the North Pacific when compared with

PIREPs (Figure 9). The finding of maximum performance over the North Pacific is similar to the verification results of the WAFC-London for WAFS CAT potential forecasts, verified with Global Aircraft Data Set (GADS) DEVG data.

5.2 Temporal Stratification

Generally, the issuance times had minimal impact on the performance of the forecasts. The 00 and 12 UTC had a slightly greater AUC, which may be a reflection of those issuance times having better model forecasts due to the assimilation of soundings. The better performance of the GTG-G relative to WAFS and the difference in skill between the AUC for the EDR comparison and the PIREP comparison is also evident here (Figure 10). For PIREP comparisons, the GTG-G AUC exceeds WAFS only slightly for all issues.

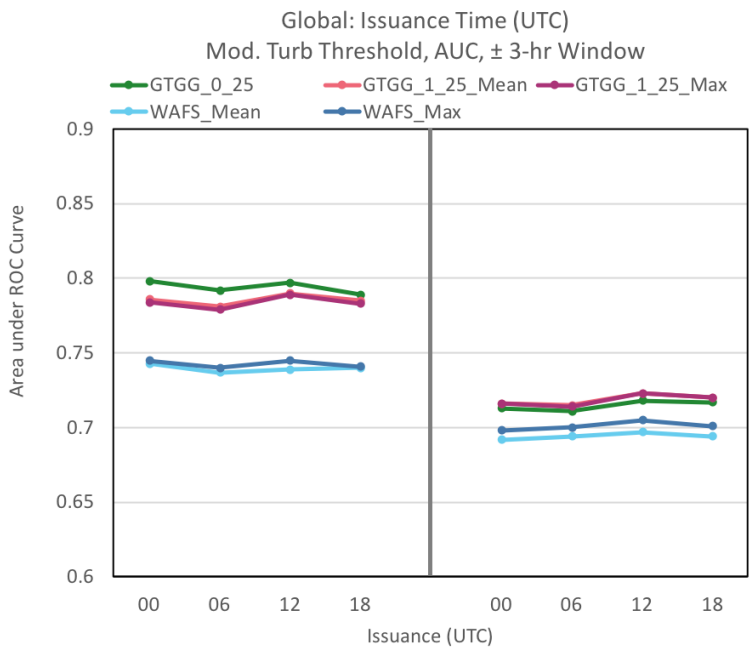

Figure 10. The AUC when stratified by issue times for the global set of observations; compared with EDR (left) and PIREP (right) for the ± 3-hour window.

The lead time also has little influence on the performance of the forecasts over most sub-areas (Figure 11), although the 12-hr lead time was best. Forecasts over Asia and Australia-New Zealand show a decreasing trend with lead times compared with EDR, but note that those two sub-areas had smaller sample sizes (fewer MOG "yes" observations) compared with other regions

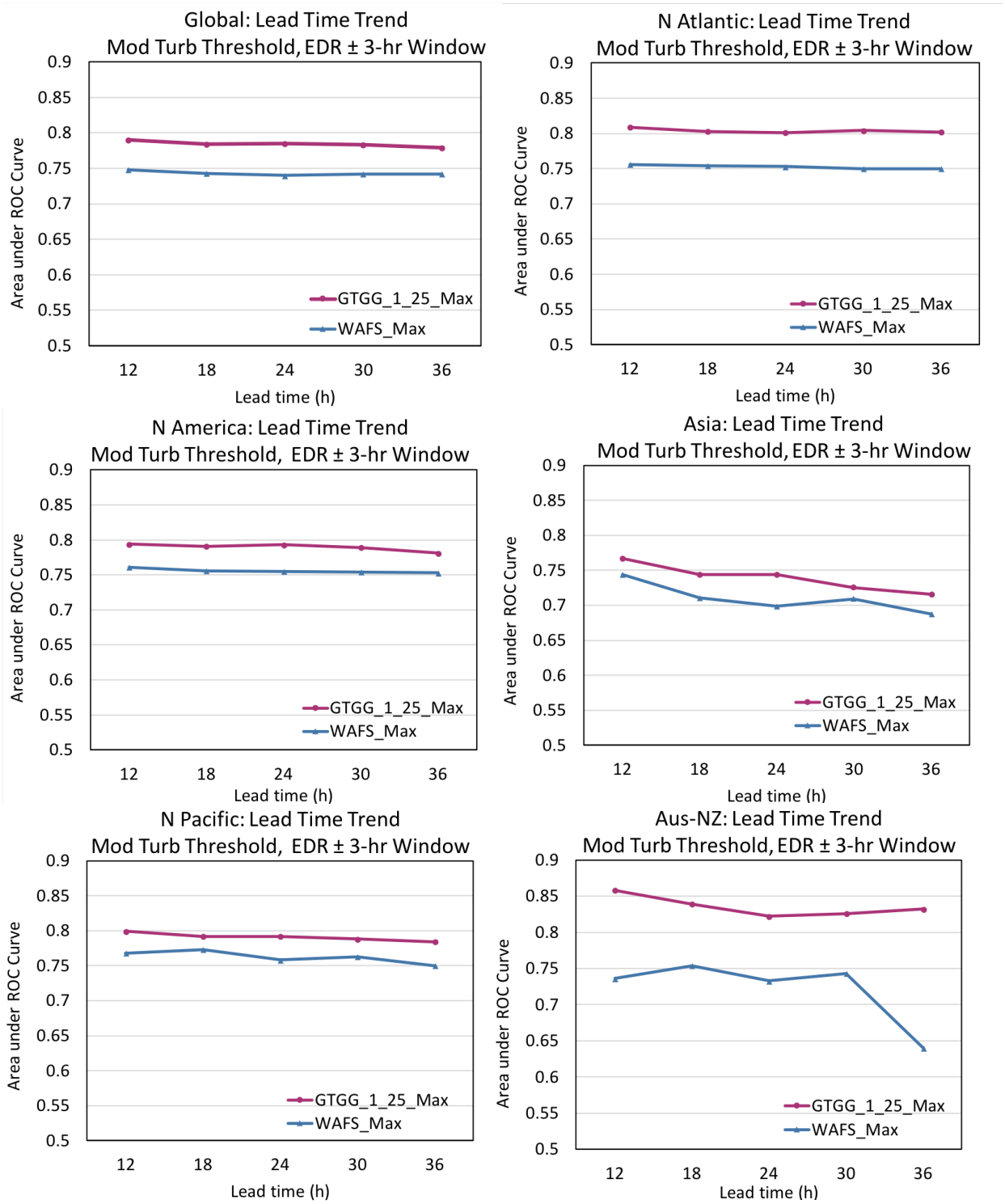

Figure 11. The AUC for lead times from 12 hr to 3 hr compared with EDR within the ± 3-hour window for various geographic regions.

Comparisons with PIREP produced nearly flat AUC trend lines with lead times for most sub-areas with sufficient sample size (Figure 12). For the North Atlantic, the AUC was slightly higher at the

30-hr lead than other lead times. Forecasts over Asia had few matching observations and poor skill, especially WAFS.

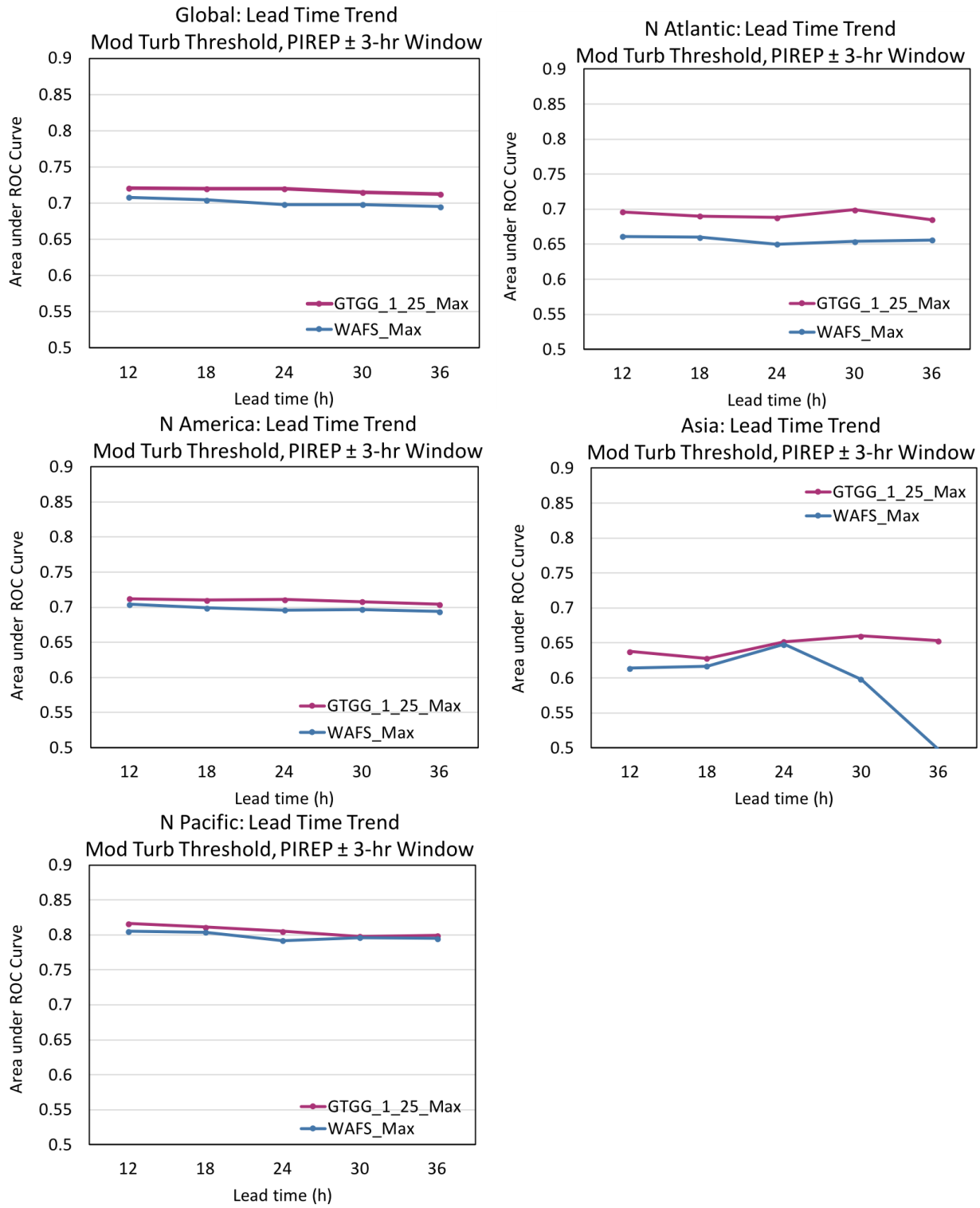

Figure 12. The AUC for lead times from 12 hr to 36 hr compared with PIREP within the ± 3-hour window for various geographic regions. The AUS-NZ plot is not shown due to lack of sample events.

5.3 VERTICAL STRATIFICATION

Consistent with the previous stratifications, the GTG-G Max-EDR AUC exceeded that for WAFS Max CAT potential for all vertical layers globally and the AUC was lower for the PIREP comparisons. The differences between the GTG-G and the WAFS were greater for EDR comparisons than for PIREP (Figure 13).

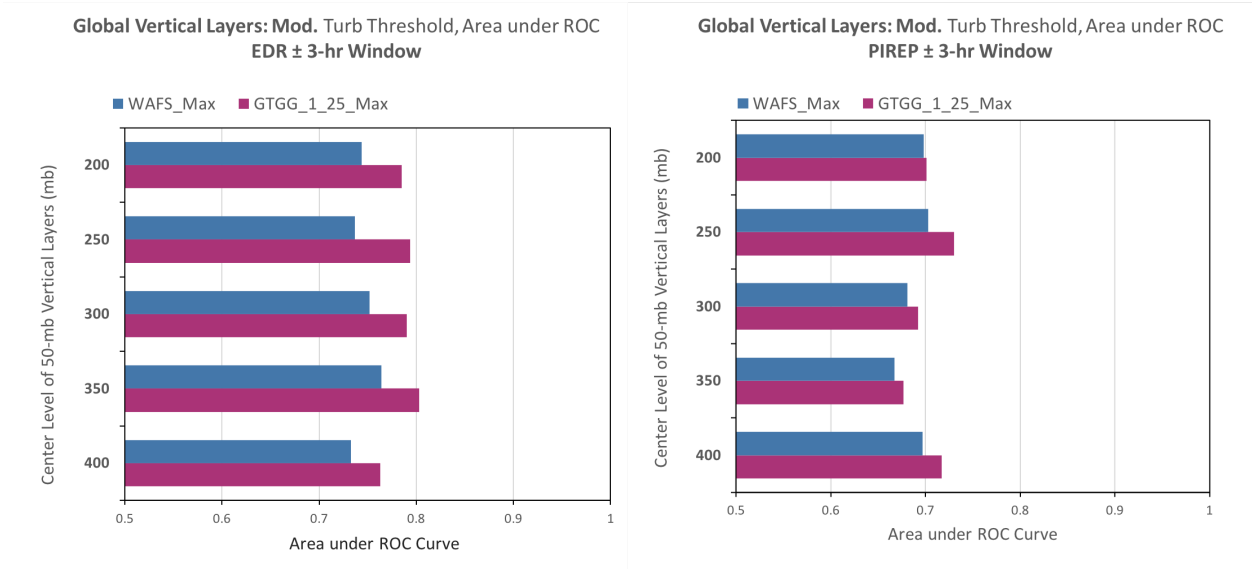

Figure 13. The AUC for the global set of observations for EDR (left) and PIREP (right) for the ± 3-hour window.

However, the WAFS AUC was slightly greater than GTG-G for PIREP comparisons within the 200-mb layer (FL390 – used for long-haul flights) over North America and the North Pacific (Figure 14). Some areas had an insufficient sample size from which to draw a definitive conclusion.

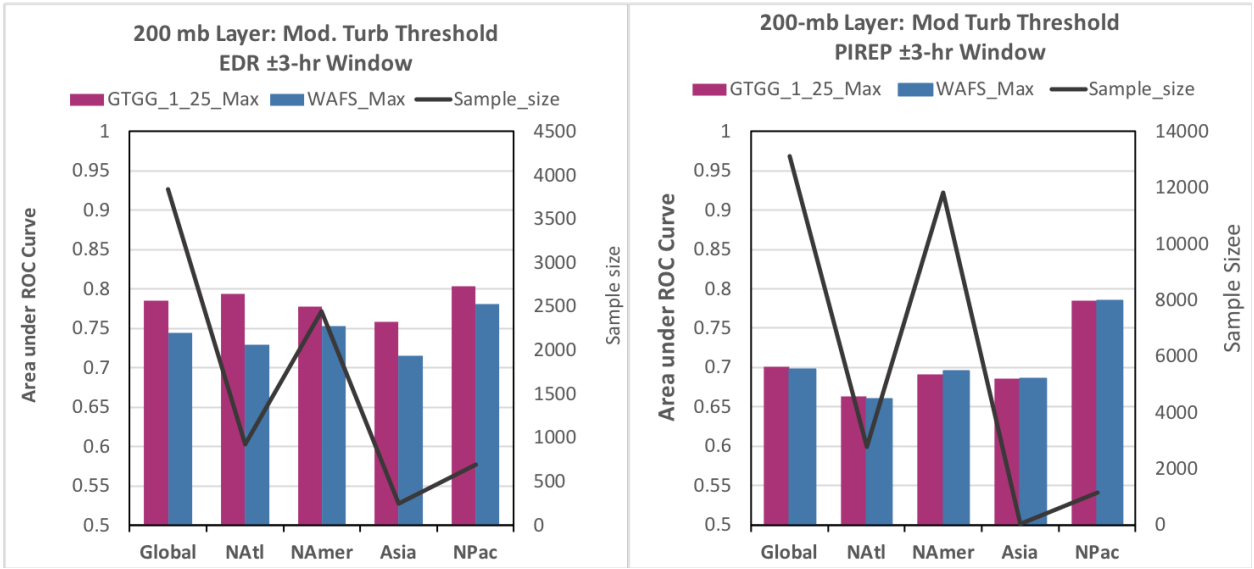

Figure 14. The AUC for the 200-mb layer/FL390 for EDR comparison (left) and PIREP (right) comparison.

5.4 August Gap Analysis

When the GTG-G performance for the gap period was compared that of the WAFS, it was found that the AUC was lower during the gap period, but GTG-G still outperformed WAFS CAT Potential forecasts (Figure 15).

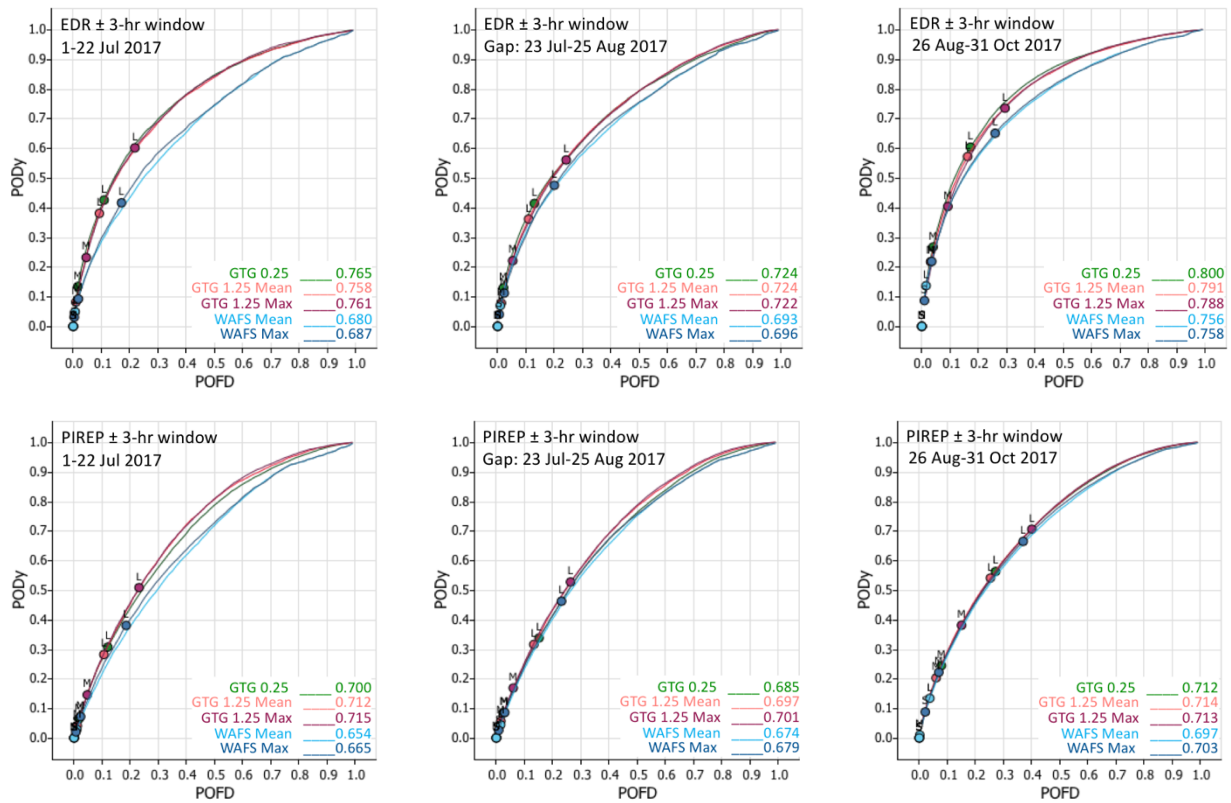

Figure 15. ROC curves and AUC for the period before the gap (left), during the gap (middle), and after the gap (right) for comparison with EDR (upper) and PIREP (lower).

A seasonal signal is also seen in the forecast performance, with better performance for the period that includes autumn than for July, more marked for EDR comparisons (Figure 15, upper panel) than for PIREP comparisons (Figure 15, lower panel). Seasonal differences are not surprising since weather prediction models are generally better at forecasting synoptic-scale waves, which dominate the cool season, than forecasting warm season weather, which is more mesoscale and convective (Langland and Maue 2012).

6 Conclusions

Performance of the global version of the GTG (GTG-G) was evaluated for the period 1 July to 31 October 2017 using EDR and PIREP observations and compared to the performance of the

operational WAFS global turbulence forecasts. With the aim of consistency with current WAFS verification, forecast performance was measured using ROC curves and the Area under the ROC curve (AUC) for Moderate-or-Greater (MOG) turbulence. The EDR threshold for medium aircraft (0.2) was used as the MOG threshold.

The evaluation found that GTG-G forecast products generally outperformed the WAFS CAT potential for both the maximum and mean attributes. The performance gap between GTG-G and WAFS was greater for EDR comparisons than for PIREP comparisons. Forecasts exhibited better skill when compared with EDR than when compared with PIREP.

Those performance characteristics remained broadly similar when various spatial and temporal stratifications were applied. When stratified by latitudinal zones, GTG-G forecasts performed best when compared with EDR observations over the SH extratropics, where it was winter and spring—winter is marked by stronger and more equatorward jet streams and more episodes of CAT and MWT. For the sub-areas, the forecasts performed best over the North Pacific, similar to results of the WAFC long-term verification, which used GADS DEVG observations. Stratification by lead had minimal impact on performance, with the 12-hr lead being best overall. The forecast issue time also had negligible influence on the performance although the AUC was greatest for 1200 UTC. For vertical layers, the performance differences were broadly similar to the other stratifications, even though some variations occurred for more narrow stratification, such as the 200-mb/FL390 layer where the WAFS performed slightly better for PIREP comparisons over some sub-areas.

It is worth noting that the period of the assessment was NH summer and early autumn, a period when global model forecast skill is lowest (Langland and Maue 2012). Since the vast majority of the observations were in the NH, the results are indicative of periods with weaker and more poleward jet streams and, hence, fewer CAT and MWT episodes than during winter. Results are likely to change for periods of evaluation that include the NH winter, when stronger jet streams are coincident with the bulk of the turbulence observations.

7 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.

8 References

Cornman, L. B., G. Meymaris, and M. Limber, 2004: An update on the FAA Aviation Weather Research Program's in situ turbulence measurement and reporting system. Preprints, 11th Conf. on Aviation, Range, and Aerospace Meteorology, Hyannis, MA, Amer. Meteor. Soc., 4.3. [Available online at https://ams.confex.com/ams/pdfpapers/81622.pdf.]

Ellrod G.P. and D. I. Knapp, 1992. An objective clear air turbulence forecasting technique: verification and operational use. Weather Forecast. **7**: 150–165.

ICAO, 2007: Meteorological service for international air navigation. Annex 3 to the Convention on International Civil Aviation, 16th Ed., ICAO Rep., 187 pp

ICAO, 2012: Guidance on the Harmonized WAFS Grids for Cumulonimbus Could, Icing and Turbulence Forecasts, Version 2.5.

Kim, J.-H. R. D. Sharman, C. Batholomew, M. Strahan, J. W. Scheck, J. C. H. Cheung, and P. Buchanan 2018: Global Graphical Turbulence Guidance (G-GTG) for World Area Forecast System (WAFS) Upgrade. *8th AMS Conference on Transition of Research to Operations*. 7-11 Jan 2018. Austin, TX.

Langland, R. H. and R. N. Maue, 2012: Recent Northern Hemisphere mid-latitude medium-range deterministic forecast skill. *Tellus,***64A,** 17531.

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.

Pearson, J. M., and R. Sharman, 2013: Calibration of in situ eddy dissipation rate (EDR) severity thresholds based on comparisons to turbulence pilot reports (PIREPSs). *16th Conference on Aviation, Range, and Aerospace Meteorology*. 5-10 January 2013, Austin, TC, Amer. Met. Soc.

Sharman, R. D., and J. Pearson, 2017: Prediction of energy dissipation rates for aviation turbulence. Part I: Forecasting non- convective turbulence. *J. Appl. Meteor. Climatol.*, 56, 317–337

Sharman, R., C. Tebaldi, G. Wiener, and J. Wolff, 2006: An Integrated Approach to Mid- and Upper-Level Turbulence Forecasting. *Weather and Forecasting*. **21,** 268-287.

Stickland, J.J., 1998: As assessment of two algorithms for automatic measurement and reporting of turbulence from commercial public transport aircraft. Bureau of Meteorology Rep, to the ICAO METLINK Study Group, 42 pp + appendices.

WMO, 2015a: WMO AMDAR Observing System Newsletter. **9,** April 2015, WMO, Geneva, Switzerland.(https://sites.google.com/a/wmo.int/amdar-news-and-events/newsletters/volume-9-april-2015)

WMO, 2015b: Aircraft-based Observations Data Statistics. (http://www.wmo.int/pages/prog/www/GOS/ABO/data/ABO_Data_Statistics.html)

WMO, 2016: Surface Verification. Commission for Basic Systems OPAG on DPFS, Meeting of the CBS (DPFS) Expert Team on Operational Weather and Forecasting Process and Support. Montreal, Canada, 09-13 May 2016. (https://www.wmo.int/pages/prog/www/DPFS/Meetings/ET-OWFPS_Montreal2016/documents/Doc-4-1_SurfaceVerification.doc)

9 Appendix A – WMO/Met Office Area Definitions

WMO Area 141 - N America

North Limit = 60.00
South Limit = 25.00
West Limit = 215.00
East Limit = 310.00

WMO Area 143 – Asia

North Limit = 65.00
South Limit = 25.00
West Limit = 60.00
East Limit = 145.00

Australia / NZ

North Limit = -10.00
South Limit = -55.00
West Limit = 90.00
East Limit = 180.00

WMO area 145 – N Pacific

North Limit = 75.00
South Limit = 30.00
West Limit = 120.00
East Limit = 240.00

N Atlantic (Polar Stereographic Projection)

North Limit = 50.1419716
South Limit = 25.3162384
West Limit = 243.1301
East Limit = 356.7594975

10 APPENDIX B – AMDAR-DEVG ANALYSIS AND RESULTS

10.1 DATA INVESTIGATION

As discussed in section 2.2.3, the AMDAR-DEVG dataset was analyzed as a potential "truth" set for verifying turbulence forecasts, particularly because of its coverage of the SH where PIREP and EDR reports are much less frequent than over the NH. The AMDAR DEVG data were obtained from the quality-controlled NOAA/ESRL/GSD ACARS archives (https://amdar.noaa.gov/qc-info.html). The data used in the assessment met the following additional criteria:

- Were flagged as "good" based on the AMDAR Data descriptor indicator
- Observations were from aircraft reporting DEVG
- Reports were from flight levels exceeding 20000 ft.

During the analysis, it was found that the distribution of AMDAR-DEVG (Airline 7) over the Australia-New Zealand (Aus-NZ) area differed from other areas and from the EDR distribution (Figure B1).

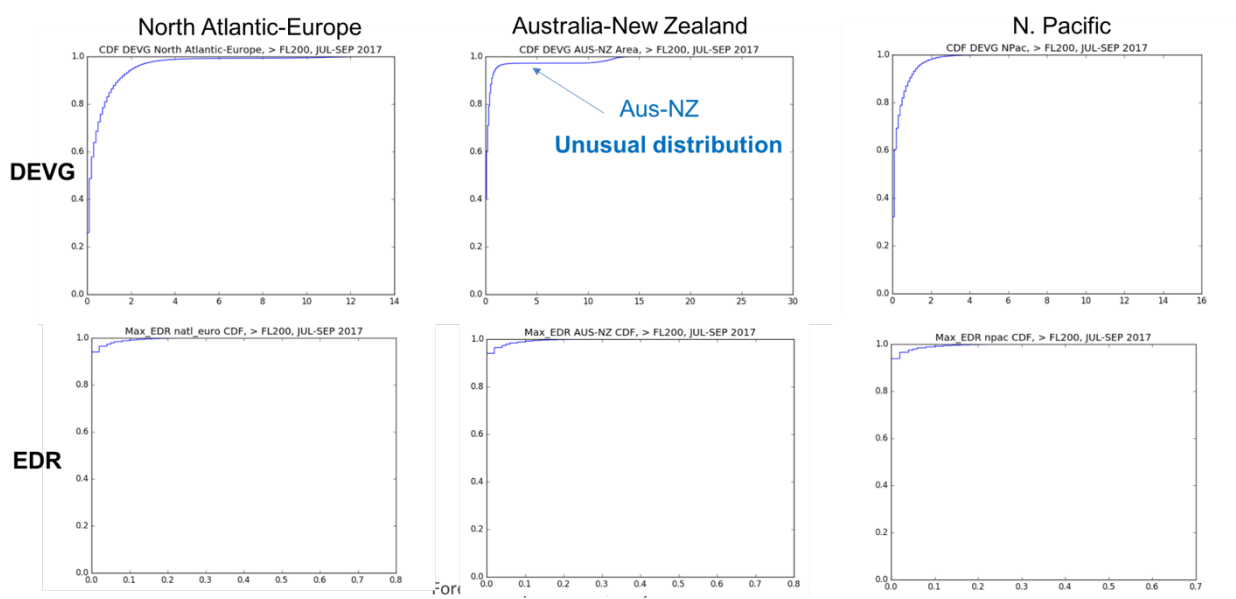

Figure B1. The cumulative distribution of AMDAR-DEVG (upper) and EDR (lower) for the North Atlantic and Europe (left), Australia-New Zealand (middle), and North Pacific (right) during July–September 2017.

The corresponding frequency distribution (Figure B2) showed very few reports within the moderate turbulence DEVG threshold (4.5 to 9 m s^{-1}) and many severe reports (> 9 m s^{-1}).

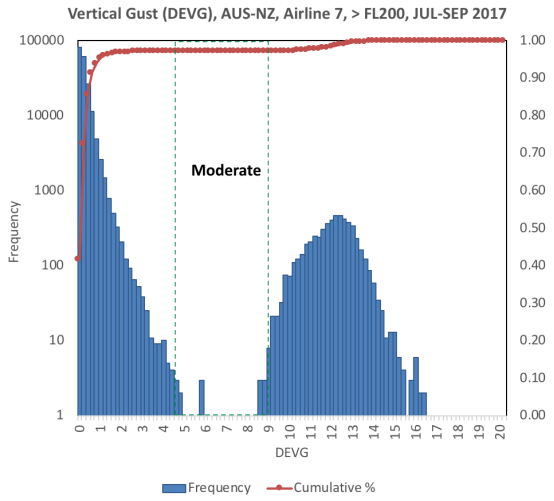

Figure B2. The FREQUENCY and cumulative distribution of AMDAR-DEVG Australia-New Zealand during July–September 2017.

After in-depth review of several of the anomalously high values over Aus-NZ, it was discovered that some flight legs of the same aircraft reported anomalously high values. Another feature of AMDAR-DEVG reports over Aus-NZ is the difference in reporting of DEVG: 747s, used for long haul flights, reported the highest observed since the last message, while 737s, used for regional flights, reported current/instantaneous values of DEVG at time of message. Given those uncertainties, it was decided that AMDAR-DEVG observations over Aus-NZ would be removed from the AMDAR-DEVG dataset, which resulted in the cumulative distribution shown in Figure B3, more similar to the regional distributions (e.g., Figure B1). The amended set of data was used for verification of the GTG-G and WAFS forecasts.

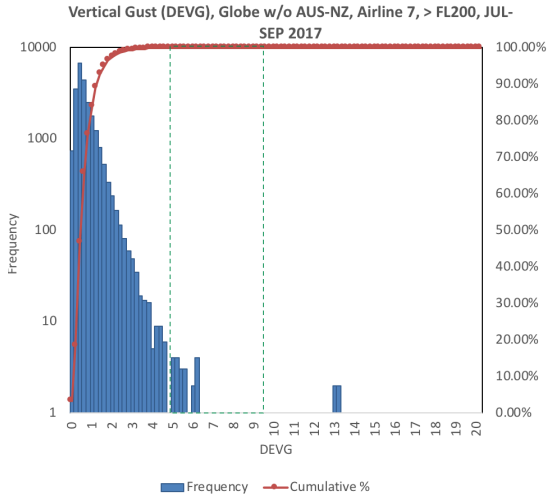

Figure B3. The frequency and cumulative distribution of Global AMDAR-DEVG without reports over Australia-New Zealand during July–September 2017.

10.2 Verification Using AMDAR-DEVG

When the amended global set of AMDAR DEVG were used to verify the turbulence forecasts, the results were not useful, even when stratified by latitudinal zones (Figure B4). Hence, the decision to use the EDR and PIREPs for verification.

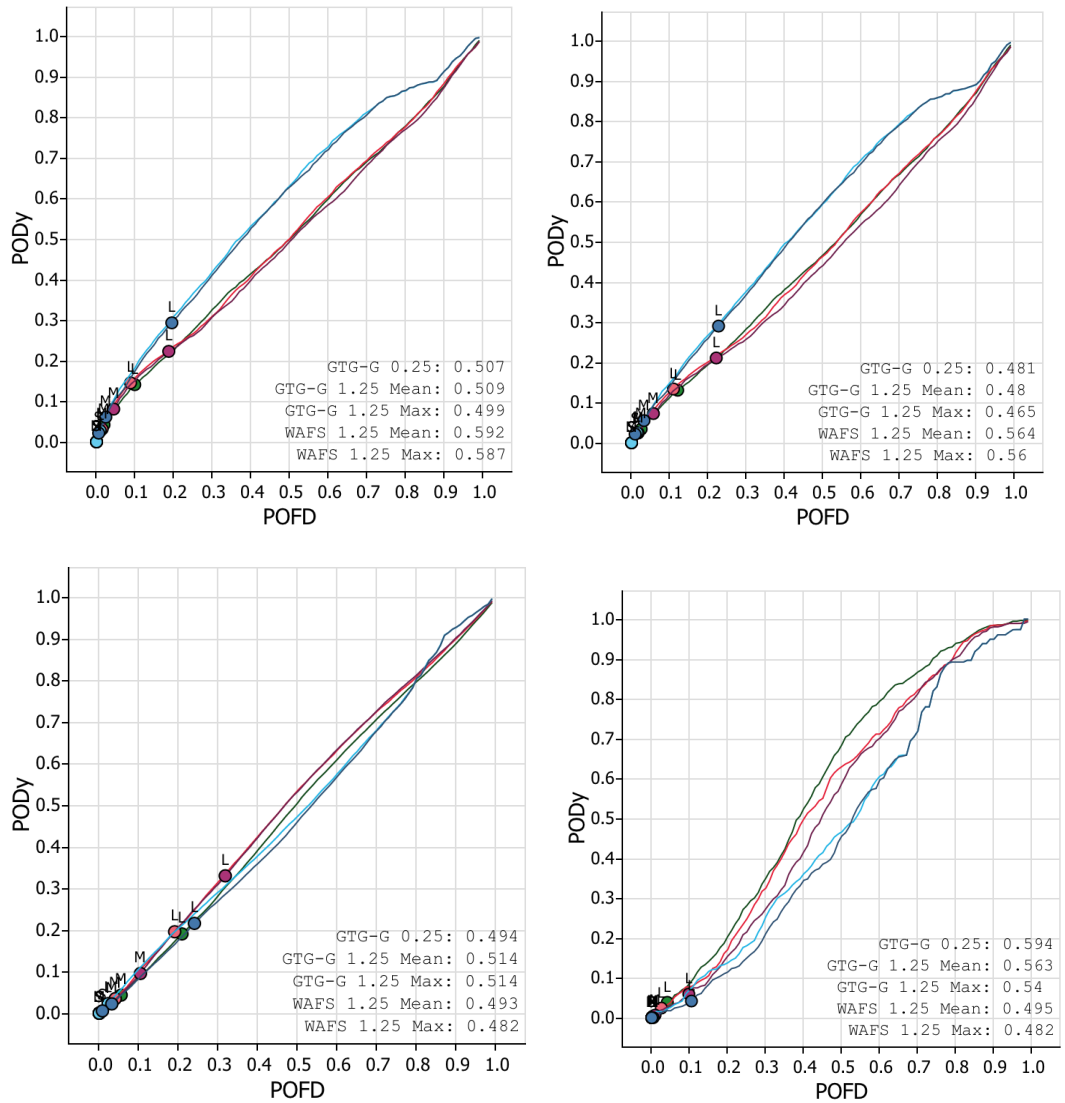

Figure B4. ROC Curves for turbulence forecasts compared with AMDAR-DEVG for (a) the whole globe, (b) the NH, (c) The SH, and (d) the Tropics. The letters "L," "M," and "S" mark the forecast thresholds for "Light," "Moderate," and "Severe" turbulence, respectively.