P7.8

USING *IN SITU* EDDY DISSIPATION RATE (EDR) OBSERVATIONS FOR TURBULENCE FORECAST VERIFICATION

Agnes Takacs*, Lacey Holland, Robert Hueftle, Barbara Brown, Anne Holmes
Research Application Program, National Center for Atmospheric Research, Boulder, CO

1. INTRODUCTION

Turbulence forecasting algorithms are being developed by the FAA Aviation Weather Research Program (AWRP) Turbulence Product Development Team (TPDT). Because forecast verification has a critical role to play in the development of new forecast products, it is important to use the best observations possible for their evaluation. Currently, the verification of turbulence forecasts is based on pilot reports (PIREPs). Unfortunately, PIREPs are sporadic in space and time, provide only a subjective measure of the aircraft's response to turbulence, and include relatively few negative reports. These features make one of the most important parts of the verification process – matching observations and forecasts – somewhat unreliable. Automated observations of turbulence conditions will soon become operationally available in adequate numbers for use in verification studies. The *in situ* turbulence algorithm relates eddy dissipation rate (EDR) to aircraft vertical acceleration. Because the *in situ* EDR observations differ in many respects from PIREPs, it is important to evaluate characteristics of these reports and to consider possible approaches for their inclusion in forecast verification studies.

This study introduces statistical characteristics of the *in situ* EDR data based on the latest quasi-real-time quality controlled *in situ* observations. During a one-month sample period in February 2005, close to 1.5 million EDR observations were collected, but only about 36,000 PIREPs were available. In this paper, comparisons of the two data sets are presented and discussed. In addition, a verification methodology is described that is based on use of the *in situ* observations is described. Finally, turbulence forecast verification results are presented based on the use of each of these observational data sets. In particular, the Graphical Turbulence Guidance, version 2 (GTG2) was evaluated using PIREPs and EDR observations.

*Correspondent author address: Agnes Takacs, NCAR/RAP, Boulder, CO 80307-3000; email: agnes@ucar.edu

2. TURBULENCE OBSERVATIONS AND FORECASTS

2.1 Pilot reports (PIREPs)

Currently the best available real-time information concerning turbulence comes from pilot reports (PIREPs). PIREPs can be routinely obtained (e.g., at NCAR) through the National Weather Service (NWS) Family of Services (FOS) communication gateway. The textual messages are decoded to allow rapid retrieval and analysis of the most important parameters within the turbulence encounter report, including the date and time, latitude and longitude, altitude and severity (Thompson 1995). The severity is translated from a verbal description (e.g. smooth, moderate, severe, or extreme) to an integer scale 0-8, where 0 is smooth or null, and 8 is extreme. A natural concern with using PIREPs is the subjective nature of the intensities reported and the imprecise time and location of the reported encounter (Sharman et al. 2002a). Table 1 shows the frequency of PIREP turbulence intensities for February 2005. It can be seen that during this month pilots did not report any intensities of 7 or above, that is, severe/extreme turbulence events.

Table 1: Frequency of PIREP turbulence intensities in datasets for February 2005.

PIREP Turbulence Intensities	*PIREP Turb.Int. Scale*	*Count*	*Percent*
Null	0	9,611	26.26
Null/occasional light	1	0	0.00
Light	2	7,927	21.65
Light/occasional moderate	3	3,788	10.35
Moderate	4	14,321	39.12
Moderate/ occasional severe	5	481	1.32
Severe	6	474	1.30
Severe/occasional extreme	7	0	0.00
Extreme	Total	36,608	100.00

2.2 In situ Eddy Dissipation Rate (EDR)

To remove the subjective nature of turbulence PIREPs, the TPDT has developed a method that utilizes observations from commercial aircraft to create an automated turbulence reporting system. Two algorithms were developed to estimate the eddy dissipation rate (EDR) from on-board data. The first algorithm uses vertical acceleration sensor measurements on the aircraft and a mathematical model of the aircraft response to turbulence in order to estimate EDR values. The other method is based on the calculation of the vertical wind component. The accelerometer-based algorithm has been deployed on approximately 200 United Airlines B737 and B757 aircraft. Data from these aircraft already are available today for experimental studies.

The *in situ* EDR observations that are transmitted from the aircraft are composed of peak and median values provided approximately every minute. Both the peak and the median EDR values are binned into categories before they are downlinked from the aircraft, starting with a minimum EDR category of 0.05, with increments of 0.10 up to a maximum of 0.85. Due to the nature of these observations, the median and peak values are close to each other in cases of relatively continuous turbulence events, while the peak is usually much larger than the median for discrete events. (For more details and recent information about the observations and the EDR estimation methods see Cornman et al. 2004.)

Table 2 shows the counts and percentages of EDR reports in each category for February 2005. As shown in the table, almost all of the median and peak EDR values fell into the 0.05 category, which represents no (or smooth/light) turbulence. Thus, frequencies in the other categories are quite small. The percentage of peak EDR observations in the higher categories is larger than the percentage of median EDR values in these categories. However, no EDR observations fell in the 0.85 category; similar results were found in other studies and for different time periods (Brown et al. 2000a, Takacs and Chen 2003, Takacs et al. 2004a).

Table 2: Frequencies of different categories of EDR values in datasets for February 2005.

EDR category	*Median EDR*		*Peak EDR*	
	Count	*Percent*	*Count*	*Percent*
0.05	1,293,474	*97.36*	1,246,608	*93.83*
0.15	28,655	*2.16*	59,900	*4.51*
0.25	5,293	*0.40*	14,828	*1.12*
0.35	933	*0.07*	4,945	*0.37*
0.45	143	*0.01*	1,559	*0.12*
0.55	19	*0.00*	462	*0.03*
0.65	7	*0.00*	146	*0.01*
0.75	0	*0.00*	76	*0.01*
Total	1,328,524	*100.00*	1,328,524	*100.00*

2.3 Turbulence forecast algorithm (GTG2.1)

Sharman et al. (2004) describes the turbulence algorithm developed by the AWRP's TPDT. The Graphical Turbulence Guidance 2 (GTG2) is the newest version of the GTGx. It expands the capabilities of GTG by providing turbulence predictions at both mid-levels (10-20,000 ft msl) and upper levels (≥20,000 ft). In addition, GTG2 incorporates some new turbulence diagnostics. Within GTG2, the mid- and upper-level forecasts are computed separately, and the forecasts are merged at the 20,000-ft boundary. This merging is necessary since it was found that (a) the best sets of turbulence diagnostics

(with respect to their ability to discriminate between Yes and No turbulence observations) differs between mid- and upper levels; (b) the optimum threshold values also differ; and (c) the number of available PIREPs is substantially smaller at mid-levels than at upper levels, so different PIREP time windows must be used in the two altitude regimes.

Given a set of turbulence diagnostics, the GTG combined them to derive an optimal integrated algorithm; this fitting and combination process is repeated as new model analyses and observations are obtained. The combination process is described in Sharman et al. (2002b).

The verification analyses described in this report are based GTG2.1 forecasts from February 2005. These forecasts were issued at 1200, 1500, 1800, and 2100 UTC with lead times of 0, 3, 6, 9, and 12 hours and valid times between 1500 and 0000 UTC. These forecasts are evaluated using PIREPs and median- and peak EDR observations. Figure 1 shows an example of the turbulence forecasts overlaid with the above-mentioned observations.

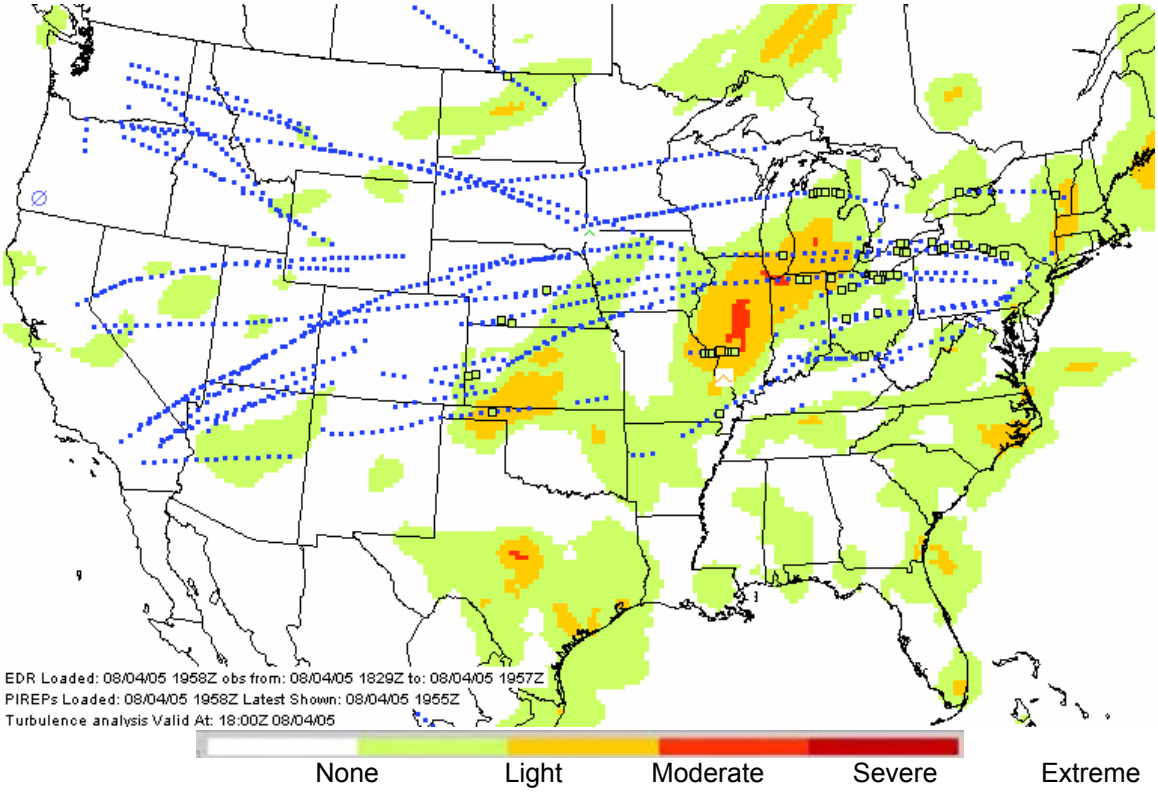


Figure 1: Example of a 6-h GTG2 turbulence forecast overlaid with the locations of PIREP and EDR observations; circles and carats: PIREPs, boxes: EDRs; blue color: no turbulence event observed (courtesy of TPDT).

3. Verification and comparison methods

This section summarizes the methods that were used to match forecasts and observations, and to match the two different types of observations with each other. The various verification statistics that were computed to compare PIREP and EDR observations and to evaluate the GTG2.1 forecasts are also described.

3.1 PIREP-EDR matching methods

The main target of this study is to show how *in situ* EDR observations can be used for verification of turbulence forecasts. The first step is to compare the two datasets, EDR values and PIREPs. Ideally, this comparison would only be performed with turbulence data observed on the same flight. However, since there are not enough data available from the same flights, in most cases observations from different flights are compared to each other. Three different time periods are included in these comparisons: an experimental dataset of peak EDR observations from 2001-2002, and median/peak observations for November 2003 and February 2005.

Overall, there are many fewer PIREPs than EDR observations. Therefore, when the observations came from different aircraft, the selection of matched pairs began with the identification of PIREPs. Next, EDR data were located within a region around the PIREP. This region was defined as a circle with a 40-km radius around the PIREP. In this matching process, a time window of ± 15 min, and an altitude window of ±4,000 ft (sometimes ±500 ft) were used. The PIREP turbulence intensity values were then paired with the EDR values in the circle in several different ways:

a) Each PIREP value was paired with the maximum of the median and the peak EDR values in the region;
b) The same PIREP intensity value was paired with each of the EDR values inside the circle;
c) Both a) and b) were repeated using a ±500 ft altitude range instead of ±4,000 ft.

3.2 Forecast-observation matching methods

In previous evaluations (e.g., Brown et al. 2000b, c, and d; Mahoney et al. 2001, Brown et al. 2002, Takacs et al. 2004b) only PIREPs were used for turbulence forecast verification. Each PIREP was associated with forecasts at the nearest eight forecast grid points (four surrounding grid points at two vertical levels). Specifically, the post-analysis verification system matched the PIREP to the most extreme forecast value among the four surrounding grid points (Brown et al. 2002, Takacs et al. 2004c). A time window of ±1 hour around the model valid time is typically used to evaluate the algorithm forecasts. In this study the method of the post analysis system is used. In particular, the forecast value associated with each observation is the maximum value at the four closest grid points. To match the EDR observations to the forecasts, a similar matching method was used. In particular, the EDR observations were converted into the PIREP format and then used in the same way as PIREPs in both the matching process and in the statistical computations.

3.3 Statistical verification methods

The statistical verification methods used to evaluate the performance of GTG2.1 and to show the results of the comparison of the two different datasets (using EDR values as forecasts) are consistent with the methods used in previous studies described by Brown et al. (1997, 2002). More details on the general concepts underlying verification of turbulence forecasts can be found in Brown and Mahoney (1998). These methods are briefly described here.

Turbulence forecasts and observations are treated here as dichotomous (i.e., Yes/No) values. The algorithm forecasts are converted to a variety of Yes/No forecasts by application of several thresholds for the occurrence of turbulence. The thresholds used for GTG2.1 are:

0.030, 0.060, 0.125, 0.200, 0.250, 0.312, 0.375, 0.437, 0.500, 0.562, 0.625, 0.750, 0.875.

The basic verification approach makes use of the two-by-two contingency table (Table 3). In this table, the forecasts are represented by the rows, and the columns represent the observations. The entries in the table are the joint distribution of forecasts and observations.

Table 3: Contingency table for evaluation of dichotomous (Yes/No) forecasts. Elements in the cells are the counts of forecast-observation pairs.

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

Table 4 lists the verification statistics used in this evaluation. PODy and PODn are the primary verification statistics used for the evaluation of GTG2.1 and the EDR values when they are used as forecasts and for comparisons of EDR values and PIREPs. Together, PODy and PODn measure the ability of the forecasts to discriminate between (or correctly categorize) Yes and No turbulence observations. This discrimination ability is summarized by the True Skill Statistic (TSS), which frequently is called the Hanssen-Kuipers discriminant statistic (Wilks 1995). Note that it is possible to obtain the same TSS value for a variety of combinations of PODy and PODn. Thus, it is always important to consider both PODy and PODn, as well as TSS.

The relationship between PODy and 1-PODn for different algorithm thresholds is the basis for the verification approach known as "Signal Detection Theory" (SDT). For a given algorithm, this relationship can be represented by the curve joining the 1-PODn, PODy points for different algorithm thresholds. The resulting curve is known as the "Relative Operating Characteristic" (ROC) curve in SDT. The area under this curve is a measure of overall forecast skill (e.g., Mason 1982).

Table 4: Verification statistics used in this study.

Statistic	***Definition***	***Description***	***Interpretation***	***Range***
PODy	YY/(YY+NY)	Probability of Detection of Yes observations	Proportion of Yes observations that were correctly forecasted	0-1 Best: 1 Worst: 0
PODn	NN/(YN+NN)	Probability of Detection of No observations	Proportion of No observations that were correctly forecasted	0-1 Best: 1 Worst: 0
FAR	YN/(YY + YN)	False Alarm Ratio	Proportion of Yes forecasts that were incorrect	0-1 Best: 0 Worst: 1
TSS	PODy + PODn – 1	True Skill Statistic; Hanssen-Kuipers discrimination	Level of discrimination between Yes and No observations	-1 to 1 Best: 1 No skill: 0
Curve Area	Area under the curve relating PODy and 1-PODn	Area under the curve relating PODy and 1-PODn (i.e., the ROC curve)	Overall skill (related to discrimination between Yes and No observations)	0 to 1 Best: 1 No skill: 0.5

Table 4 includes the False Alarm Ratio (FAR), a statistic that is commonly computed from the 2x2 table. Brown and Young (2000e) showed that due to the non-systematic nature of PIREPs, it is not appropriate to compute FAR when evaluating forecasts using PIREPs as the verifying observations. The EDR observations have some attributes that are similar to the PIREPs. They are non-systematic, not entirely independent, and they do not cover the entire air space as covered by the forecasts. Consequently, observations are not available for every grid point. However, due to the large number of observations, there is a much higher likelihood that a relevant and consistent subset of grid points will have associated EDR observations at specified times than is the case for PIREPs. Therefore, computation of FAR might be appropriate if we use the *in situ* EDR observations for verification.

Due to the characteristics of PIREPs and the *in situ* EDR observations, the verification statistics (e.g., PODy and PODn) should not be interpreted in an absolute sense, but can be used for comparisons among algorithms and forecasts. Moreover, PODy and PODn should not be interpreted as probabilities, but rather as proportions of PIREPs that are correctly forecast. This statement applies to the statistics based on EDR observations as well.

4. Comparisons of turbulence observations

Comparisons of the PIREP turbulence intensity values and *in situ* EDR observations are shown in this section. Although there are many uncertainties and problems with this kind of comparison, mainly due to the number of possible errors in both the EDR and PIREP values (Cornman et al. 2004), these comparisons are needed in order to be able to interpret the results of turbulence forecast verification analyses based on the two different datasets.

As mentioned earlier, these comparisons would be best made with turbulence observations from the same flight. However, since not enough data of that type are available, for most comparisons observations from different, but nearby flights are compared to each other.

4.1 Charecteristics of PIREP and EDR frequency distributions

From the same flights, more than 800 PIREPs and peak EDR observations (not quality controlled) were available for this study (2001-2002). The other two periods included in the investigation were November 2003 and February 2005, when quality controlled median and peak EDR observations could be paired with PIREPs; however, during these time periods, the PIREPs and EDR values generally were not from the same flights. It is well known from previous studies that there are positioning and timing differences between PIREP and EDR observations even if they are from the same flights (personal communication with Sharman and Cornman 2005). This fact has encouraged us to use observations from different flights close enough to each other in space and time, the conditions that we applied are described in Section 3.1. Matching was based on the ±4,000 ft altitude restriction.

Table 5 shows the joint frequencies of the peak EDR and PIREP pairs for the same flights (2001-2002) while Tables 6 and 7 show the same statistics for the two one-month periods for different flights.

Table 5: Frequencies of pairs of peak EDR values and PIREP turbulence intensities for the same flights (2001-2002).

PIREP int.	***Peak EDR observations***								
	0.05	**0.15**	**0.25**	**0.35**	**0.45**	**0.55**	**0.65**	**0.75**	Σ/%
0	538	36	5	3	0	0	1	1	584 *67.8%*
2	129	81	18	3	0	2	0	1	234 *27.2%*
3	6	21	8	3	0	0	0	1	39 *4.5%*
4	0	3	1	0	0	0	0	0	4 *0.5%*
Σ/%	673 *78.2%*	141 *16.4%*	32 *3.7%*	9 *1.0%*	0 *0.0%*	2 *0.2%*	1 *0.1%*	3 *0.4%*	861 *100%*

Table 6: Frequencies of pairs of peak EDR values and PIREP turbulence intensities for different flights (November 2003).

PIREP int.	***Peak EDR observations***						
	0.05	**0.15**	**0.25**	**0.35**	**0.45**	**0.55**	Σ/%
0	27	4	0	0	0	0	31 *31.6%*
2	16	4	2	0	0	0	22 *22.5%*
3	7	3	0	0	0	0	10 *10.2%*
4	18	7	5	2	0	1	33 *33.7%*
5	0	2	0	0	0	0	2 *2.0%*
Σ *%*	68 *69.4%*	20 *20.5%*	7 *7.1%*	2 *2.0%*	0 *0.0%*	1 *1.0%*	98 *100.0%*

Table 7: Frequencies of peak EDR values and PIREP turbulence intensities for different flights (February 2005).

PIREP int.	***Peak EDR observations***						
	0.05	**0.15**	**0.25**	**0.35**	**0.45**	**0.55**	Σ/%
0	11	2	0	1	0	0	14 *15.1%*
2	20	7	0	1	0	0	28 *30.1%*
3	4	5	1	1	0	0	11 *11.8%*
4	21	3	7	3	5	0	39 *41.9%*
5	0	0	1	0	0	0	1 *1.1%*
Σ *%*	56 *60.2%*	17 *18.3%*	9 *9.7%*	6 *6.4%*	5 *5.4%*	0 *0.0%*	93 *100.0%*

Table 8: Frequencies of median EDR values and PIREP turbulence intensities for different flights (February 2005).

PIREP int.	***Median EDR observations***						
	0.05	**0.15**	**0.25**	**0.35**	**0.45**	**0.55**	Σ/%
0	13	0	1	0	0	0	14 *15.1%*
2	24	3	1	0	0	0	28 *30.1%*
3	7	4	0	0	0	0	11 *11.8%*
4	23	12	3	1	0	0	39 *41.9%*
5	0	1	0	0	0	0	1 *1.1%*
Σ *%*	67 *72.0%*	20 *21.5%*	5 *5.4%*	1 *1.1%*	0 *0.0%*	0 *0.0%*	93 *100.0%*

Characteristic variations can be seen among the frequencies shown in these tables for the cases of the same and different flights. First, in 2001-2002 (Table 5) pilots did not report higher than moderate (4) turbulence intensities, although some EDR values indicated higher intensities (0.55-0.75). Second, there is a monotonic decrease in the frequencies of the EDR values paired with PIREPs (last rows in Tables 5, 6, and 7); in contrast, the frequencies of PIREP-based turbulence intensities have a secondary maximum for moderate (4) intensity (last columns in the tables). Since moderate or greater turbulence is likely to occur more frequently in deep clouds, a possible explanation for this secondary maximum is that there were more in-cloud than clear-air turbulence reports during these time periods; it also is possible that this result simply reflects the reporting practices of the pilots (Wolff and Sharman 2004). Because this signal is not seen in the EDR frequencies, the latter explanation seems most likely.

Table 8 shows the frequencies of median EDR and PIREP pairs for different flights in February 2005 (matching was based on the ±4,000 ft altitude restriction).

Table 8 indicates a higher frequency of null (0.05) values for the median EDR than peak EDR (Table 7). Only 28% of the median EDR values show other than no turbulence, while there were 39.8% in case of peak values. For the lower category of EDR values, a fairly large proportion (34%) of the associated PIREPs indicated moderate or greater turbulence. In contrast, when the EDR category was greater than 0.05, the PIREPs almost always indicated a turbulence intensity greater than null.

4.2 Comparison of PIREPs and median/peak EDR observations

For this comparison, various verification statistics shown in Table 4 were computed. The cells of the contingency table (Table 3) are filled with counts of PIREP/EDR pairs, where EDR observations are treated as the forecasts. Table 9 shows the statistical results for different time periods (based on the ±4,000 ft altitude restriction). The 2003 and 2005 analyses were restricted to the layer between 10,000 and 46,000 ft. The layer used for the 2001-2002 analyses is not known. Median EDR observations are not available for this time period.

Table 9: Statistics for median and peak EDR observations (treated as forecasts) for different time periods (threshold: EDR > 0.05) compared to PIREPs (treated as observations).

	No. of cases	*PODy*	*PODn*	*FAR*	*TSS*
Median EDR					
2003 November; different flights; 10-46,000 ft	98	0.149	0.903	0.231	0.052
2005 February; different flights; 10-46,000K ft	93	0.316	0.929	0.038	0.245
Peak EDR					
2001-2002; same flights	861	0.513	0.921	0.245	0.434
2003 November; different flights; 10-46,000 ft	98	0.388	0.870	0.133	0.258
2005 February; different flights; 10-46,000 ft	93	0.430	0.785	0.081	0.215

For these computations the maximum of both the median and peak EDR values in the matching region were selected to match to each PIREP. These statistics are fairly good for observations from the same flights; both PODy and TSS are relatively large. The results are not nearly as good for comparisons based on observations from different flights, but there is a slight improvement from 2003 to 2005 in PODy, PODn, and FAR, and in TSS for the median EDR values. For peak EDR, PODy and FAR improve somewhat, but PODn and TSS degrade slightly between the 2003 and 2005 datasets.

The sample sizes in Table 9 are fairly small due to the fact that each PIREP has been matched to only a single EDR value in the matching region. The number of cases can be increased greatly when all EDR values in the matching area (i.e., the 40-km circle and altitude range surrounding the PIREP) are used (however, it should be noted that many of the pairs considered in this way are not independent). Statistical results for pairs identified using different matching methods can be seen in Table 10.

Table 10: Statistics for median and peak EDR observations (treated as forecasts) for February 2005 (threshold: EDR > 0.05) compared to PIREPs (treated as observations), with matching done in three different ways.

	No. of cases	*No. of YY*	*No. of NN*	*PODy*	*PODn*	*FAR*	*TSS*
Median EDRs							
One PIREP/max EDR – ±4,000 ft	93	25 *26.8%*	13 *14.0%*	0.316	0.929	0.038	0.245
One PIREP/all EDR – ±4,000 ft	1379	314 *22.7%*	215 *15.6%*	0.282	0.808	0.140	0.090
One PIREP/all EDR – ±500 ft	602	189 *31.4%*	90 *15.0%*	0.391	0.756	0.133	0.147
Peak EDRs							
One PIREP/max EDR – ±4,000 ft	93	34 *36.6%*	11 *11.8%*	0.430	0.785	0.081	0.215
One PREP/all EDR – ±4,000 ft	1379	542 *39.3%*	166 *12.0%*	0.487	0.624	0.156	0.111
One PIREP/all EDR - ±500 ft	602	292 *48.5%*	66 *10.9%*	0.605	0.555	0.154	0.160

The overall results show better agreement between EDR values and PIREPs when only one EDR value (the maximum) is selected to match to the PIREP within the circle. The most likely explanation for this result is the frequent occurrence of 0.05 EDR values close to the PIREP. Although the ±4,000 ft altitude range seems quite large, this difference in statistics exists even if the altitude range is only ±500 ft. Further investigation is necessary to find the best approach for pair selection. The method should be based on differences between EDR values and PIREPs in time and space (x, y, and z), and in the turbulence intensities.

Since GTG2.1 predicts turbulence for both mid- and upper levels, the same comparisons between EDRs and PIREPs were also computed for the different layers, that is, for 10-20,000 ft and 20-46,000 ft. The overall statistics seem to be better for the mid-level, which was not expected. However, due to the small number of cases, mainly for upper levels, the results cannot be considered to be reliable. Therefore, neither of these results is shown.

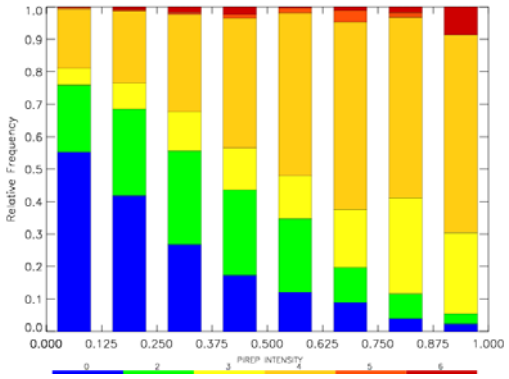
The comparison of PIREP and EDR values does not show perfect agreement. However, the results do indicate that the two types of observations have some level of correspondence. It is particularly notable that the FAR values are quite small even when the PODy values are moderate. Thus, the two types of observations are reasonably comparable for use in turbulence forecast verification studies.

5. Verification of GTG2.1 based on PIREPs and EDR observations

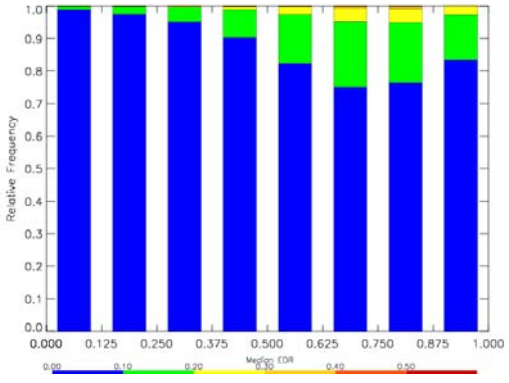
In this section, the EDR observations are treated as PIREPs and used for verification of the GTG2.1 turbulence forecast algorithm. In particular, the EDR values were converted to PIREP format, assigned intensity values and matched to the forecasts values in the same manner as used for PIREPs (e.g., see Takacs et al. 2004c). PIREPs are also used to evaluate the algorithm, and the results based on the two datasets are compared. All results shown in this section are based on the February 2005 data set.

5.1 Characteristics of GTG2.1 and PIREP/EDR frequency distributions

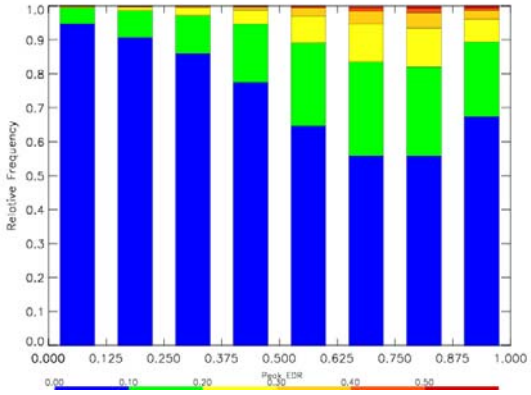
The frequency distributions of PIREPs and median/peak EDR observations are shown for mid- and upper levels, respectively, in Figs. 2 and 3. It can be seen that the relationship between observations and forecasts varies more smoothly for PIREPs than for the EDR values. This result is primarily due to the large number of EDR observations in the 0.05 category, indicating no or only smooth/light turbulence. According to Frehlich and Sharman (2004) the ratios of null and moderate and greater PIREPs do not correctly reflect the actual distribution of turbulence intensities in the free atmosphere, where the air is predominantly nonturbulent at aircraft scales. Therefore, the GTG2.1 – EDR distribution of intensity seems likely to be more realistic than the PIREPs.



(a) PIREPs

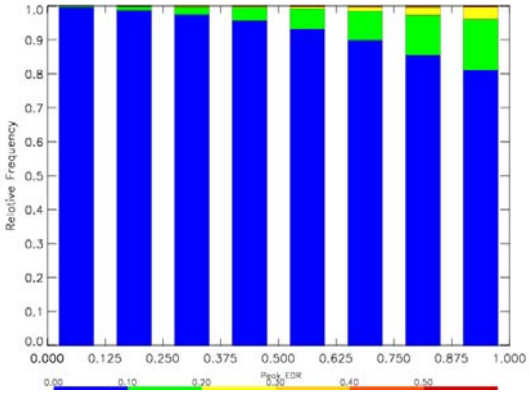


(b) Median EDR



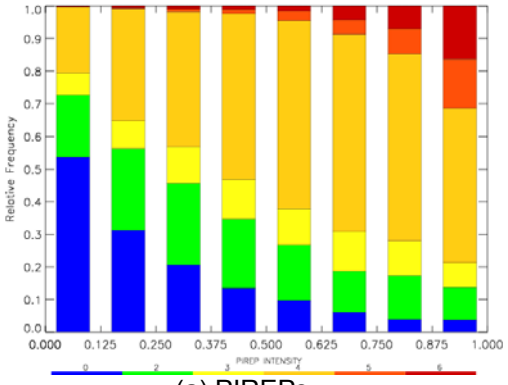
(c) Peak EDR

Figure 2. Relative frequency distributions of PIREPs, Median EDR, and Peak EDR categories according to values of matched GTG2.1, for mid-levels.

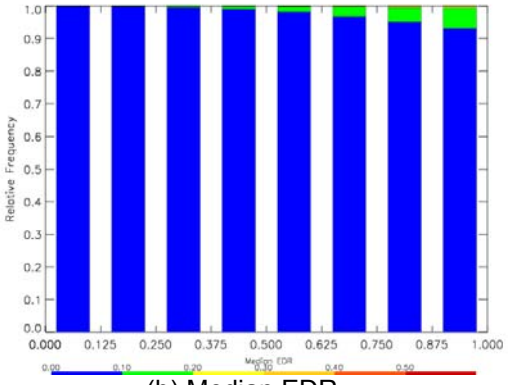


(c) Peak EDR

Figure 3. As in Fig 2, for upper levels.



(a) PIREPs



(b) Median EDR

From these distributions it can also be seen that in February 2005 more turbulence events were reported by EDRs in mid-levels than upper levels. Similar results were found in the PIREP/EDR comparison for both PIREPs and EDRs.

5.2 Verification statistics based on PIREPs and median/peak EDR observations

This evaluation used GTG2.1 forecasts for February 2005, issued at 1200, 1500, 1800, and 2100 UTC with lead times of 0, 3, 6, 9, and 12 hours and valid times between 1500 and 0000 UTC. For the overall statistics, all lead times and issue times were combined and evaluated together; this combination is reasonable because there are no significant differences in statistics among forecasts for the different issue and lead times. In addition, statistics for the analysis time and the 6-hr lead time are presented with the other statistics in Table 11 for one GTG2.1 threshold (0.125).

Table 11: Verification statistics for GTG2.1 for different lead times based on PIREPs and median/peak EDR observations for February 2005, 10-46,000 ft, and one GTG2.1 threshold (0.125).

	No. of cases	*PODy*	*PODn*	*FAR*	*TSS*	*ROC curve area*
PIREPs						
All lead/ issue times	36,365	0.887	0.380	0.200	0.267	0.634
Analysis	7,619	0.915	0.348	**0.197**	0.263	0.632
6 hr	7,262	0.888	0.387	0.199	0.275	0.638
Median EDRs						
All lead/ issue times	3,641,302	0.912	0.465	0.998	0.377	0.689
Analysis	746,415	**0.945**	0.463	0.988	**0.408**	**0.704**
6 hr	731,462	0.908	0.473	0.987	0.381	0.691
Peak EDRs						
All lead/ issue times	3,641,302	0.848	0.471	0.957	0.319	0.660
Analysis	746,415	0.887	0.428	0.958	0.315	0.658
6 hr	731,462	0.845	**0.479**	0.957	0.324	0.662

The false alarm ratio (FAR) determined using the EDR values indicates a very large proportion of the Yes forecasts were incorrect, and is much larger than the corresponding value calculated using PIREPs. This result is expected based on the frequency distribution of forecasts and observations shown earlier, and is due to the very large proportion of EDR values in the first (0.05) category.

It is important to investigate the performance of the turbulence forecasting algorithm for all turbulence intensities and for only the moderate or greater (MOG) events. Table 12 shows the results for *ALL* and *MOG* events for the GTG2.1 analysis (0-hr forecast). MOG events are defined in two ways, with different EDR thresholds used for each approach.

The results in Table 12 indicate that GTG2.1 performs better for MOG turbulence events than for the ALL cases. If the EDR thresholds are selected differently, with all non-0.05 EDR values used as a Yes observation, the results are even better. It should be noted that the EDR threshold selected for MOG2 is not arbitrary. In particular, when the EDR values are transmitted from the aircraft, they are binned into categories starting from 0.05, with increments of 0.10. Thus, the first bin essentially includes EDR values between 0 and 0.10. As a result, many of the lower intensity turbulence events are categorized as 0.05, that is, no turbulence. If more categories were available for binning the lower turbulence intensities, a more appropriate bin could be paired with higher forecast and PIREP values (personal communication with Cornman and Sharman 2005).

The results in Table 12 also indicate that the measured forecasting performance of GTG2.1 is better when EDR observations are used as the verifying observations rather than PIREPs.

Stratifications for mid- and upper-level GTG2.1 forecasts (analysis and all lead times combined) are shown in Table 13.

Table 12: Statistics for GTG2.1 analysis (0-hr forecast) for *ALL* and *MOG* turbulence events (no. of PIREPs: 7,619, Med/Peak EDRs: 746,415).

Thresholds	*No. of YY*	*PODy*	*PODn*	*FAR*	*TSS*	*ROC curve area*
PIREPs						
ALL GTG2.1 ≥ 0.125 PIREP intensity > 0	5,186	0.915	0.348	**0.197**	0.263	0.632
MOG1 GTG2.1 ≥ 0.375 PIREP intensity ≥ 3	2,033	0.629	0.684	0.406	0.313	0.656
Median EDRs						
ALL GTG2.1 ≥ 0.125 Median EDR > 0.05	5,331	**0.945**	0.463	0.988	0.408	0.704
MOG1 GTG2.1 ≥ 0.375 Median EDR ≥ 0.25	3,698	0.656	0.848	0.968	0.504	0.752
MOG2 GTG2.1 ≥ 0.375 Median EDR > 0.05	5,331	**0.945**	0.850	0.954	**0.796**	**0.898**
Peak EDRs						
ALL GTG2.1 ≥ 0.125 Peak EDR > 0.05	18,027	0.887	0.428	0.958	0.315	0.658
MOG1 GTG2.1 ≥ 0.375 Peak EDR ≥ 0.25	10,604	0.522	0.855	0.909	0.387	0.694
MOG2 GTG2.1 ≥ 0.375 Peak EDR > 0.05	18,027	0.887	**0.865**	**0.845**	0.752	0.876

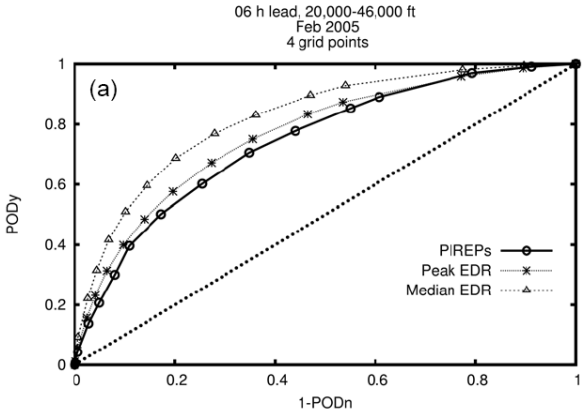
Table 13 shows that the verification statistics for GTG2.1 are better for the upper levels and for comparisons based on the EDR observations. In addition, the results in Table 13 indicate that GTG2.1 performance only degrades slightly for mid-level forecasts. Large differences in FAR are associated with the type of observation used for the evaluation (i.e., PIREPs vs. EDR values). As mentioned earlier, FAR may be a useful measure of performance when no-turbulence observations are available in adequate numbers. However, we do not suggest its use without further investigation.

Table 13: Verification statistics for GTG2.1 forecasts for mid- and upper levels.

	No. of cases	*PODy*	*PODn*	*FAR*	*TSS*	*ROC curve area*
PIREPs						
All 10-46,000 ft *MOG1*	36,365	0.887 *0.887*	0.380 *0.827*	0.200 *0.065*	0.267 *0.714*	0.634 *0.857*
All 20-46,000 ft *MOG2*	24,245	0.894 *0.894*	0.394 *0.813*	**0.176** ***0.062***	0.288 *0.707*	0.644 *0.854*
All 10-20,000 ft *MOG2*	12,085	0.871 *0.871*	0.359 *0.845*	0.248 *0.072*	0.230 *0.716*	0.615 *0.858*
Median EDRs						
All 10-46,000 ft *MOG2*	3,641,302	0.912 *0.912*	0.465 *0.863*	0.998 *0.912*	0.377 *0.775*	0.689 *0.888*
All 20-46,000 ft *MOG2*	3,276,166	**0.928** ***0.928***	0.470 *0.865*	0.993 *0.974*	**0.398** ***0.793***	**0.699** ***0.897***
All 10-20,000 ft *MOG2*	341,161	0.898 *0.898*	0.430 *0.855*	0.934 *0.782*	0.328 *0.753*	0.664 *0.877*
Peak EDRs						
All 10-46,000 ft *MOG2*	3,641,302	0.848 *0.848*	0.471 *0.869*	0.957 *0.847*	0.319 *0.717*	0.660 *0.859*
All 20-46,000 ft *MOG2*	3,276,266	0.870 *0.870*	**0.475** *0.869*	0.971 *0.894*	0.345 *0.739*	0.673 *0.870*
All 10-20,000 ft *MOG2*	341,161	0.815 *0.815*	0.447 ***0.872***	0.981 *0.541*	0.262 *0.687*	0.631 *0.844*

5.3 ROC curves

In this section, the results of the GTG2.1 performance evaluations based on PIREPs and median/peak EDR observations are summarized using ROC curves for mid- and upper level forecasts. In the ROC diagrams, the individual points on the algorithm curves represent particular thresholds used to create Yes/No forecasts. More skillful forecasts are represented by curves that are located closer to the upper left corner of the diagram.



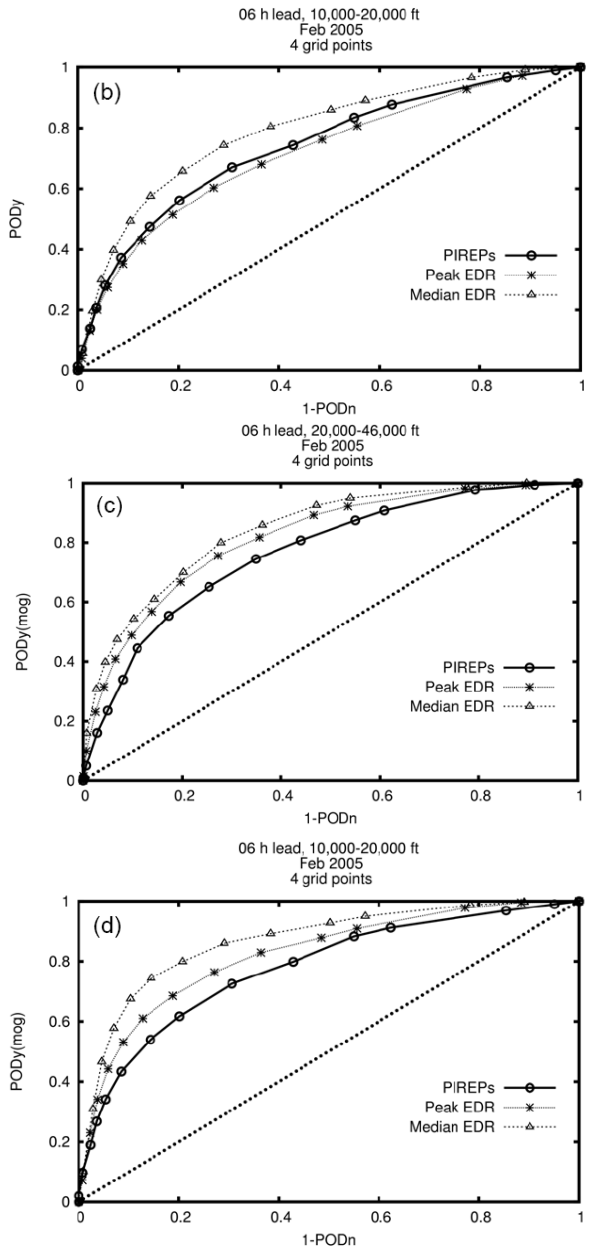


Figure 4. ROC diagrams for 6-hr GTG2.1 forecasts with verification based on PIREPs and EDR values, for mid- and upper levels, and for ALL and MOG1 turbulence events; (a) ALL, 20-46,000 ft, (b) ALL, 10-20,000 ft, (c) MOG1, 20-46,000 ft, and (d) MOG1, 10-20,000 ft.

Figure 4 shows the results of 6-hr GTG2.1 forecast performance for the ALL and MOG1 turbulence events (defined in Section 5.2), and for mid- and upper levels. It can be seen that, in general, the results are better if the comparison is based on the median EDR observations than PIREPs. With one exception, this is true for the peak EDRs as well. In the ALL case at mid levels, the PIREP-based comparison shows slightly better performance than the peak EDR-based comparison. However, for MOG events the results based on both EDR look notably better than those based on PIREPs.

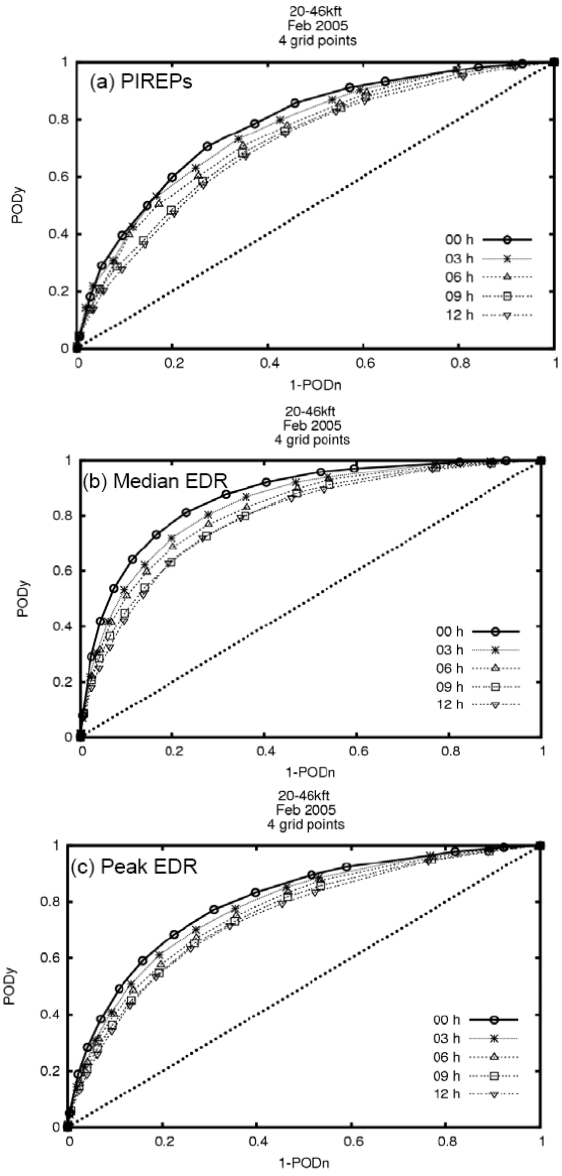


Figure 5: ROC diagrams for GTG2.1 for upper levels, ALL turbulence events for different lead times, stratified by datasets used for the evaluation; (a) PIREPs, (b) Median EDRs, (c) Peak EDRs

Figure 5 shows the performance of GTG2.1 for upper levels, ALL turbulence events, and different lead times. These diagrams show similar results for GTG2.1 performance based on different datasets. ROC areas, and thus skill, decrease with increasing lead time in all cases. However, the best results are obtained by using the median EDR values for verification for all lead times, followed by the peak EDRs, with somewhat less skill indicated by the results based on PIREPs. Most importantly, all these comparisons show similar patterns.

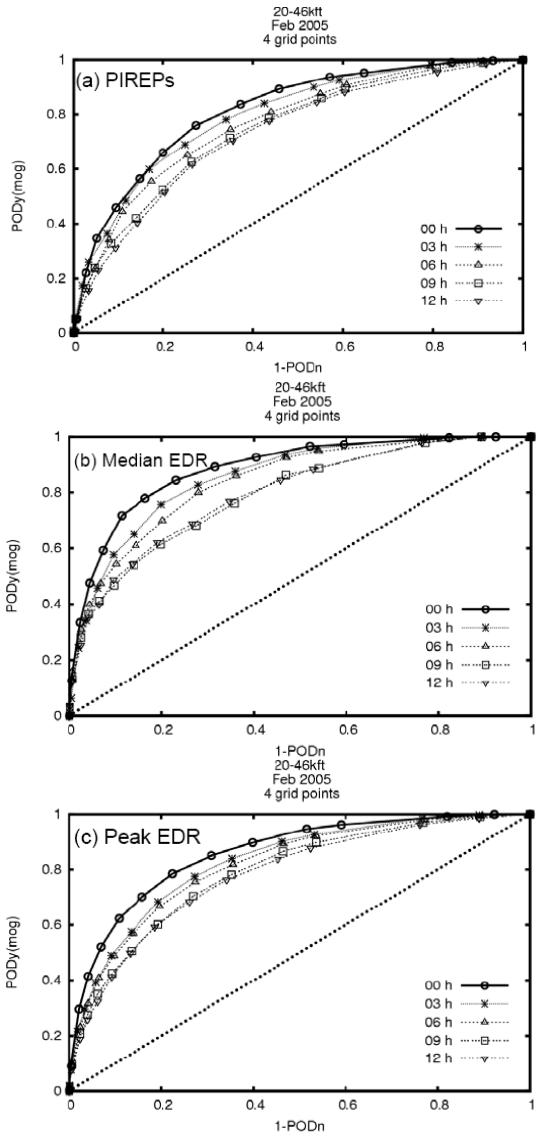


Figure 6: As in Fig 5, for MOG1 events

In the case of MOG1 turbulence events (Fig. 6), the results are similar to those for ALL turbulence events, with somewhat greater skill indicated for comparisons based on each of the datasets. These results are in good agreement with previous findings in this study.

6. Conclusions and discussion

This report has considered several aspects of the use of *in situ* EDR observations for the verification of turbulence forecasts. Basic statistical characteristics of PIREPs and the *in situ* EDR data have been examined. Comparisons of different datasets, observations and forecasts have been described. Verification results using PIREPs and median/peak EDR observations have also been shown. The results of the study suggest the following conclusions:

- The *in situ* EDR observations are very frequent and provide wide (though not complete) coverage over the continental United States. Nearly all of the EDR values fall in the lowest category (0.05), which includes observations of null and at least some light turbulence; higher categories of EDR (0.25 or greater) are relatively rare.
- The *in situ* EDR observations were compared to PIREP turbulence intensities. These comparisons were encouraging, but should be repeated using data for longer time periods, different seasons and preferably with observations from the same flights. The matching method used in this report should also be further developed.
- The GTG2.1 turbulence algorithm was evaluated using two different datasets, PIREPs and *in situ* EDRs. The verification results are comparable. Skill estimates based on these datasets are in reasonably good agreement with previous results based on PIREPs.
- GTG2.1 shows somewhat more skill when evaluated using *in situ* EDR observations than with PIREPs. The differences appear to be largest in comparisons of the scores based on median EDR vs. PIREPs.

- False Alarm Ratio (FAR) results indicate much worse performance of GTG2.1 when this statistic is computed using EDR values rather than PIREPs. This result is expected due to the EDR's more realistic, large number of zero turbulence observations (and the unrealistically low number of no-turbulence PIREPs). Use of FAR for verification of turbulence forecasts is not recommended without further investigation.
- The EDR observations can be transitioned to the Global System Division (GSD) of the NOAA Earth System Research Laboratory (ESRL) Real Time Verification System (RTVS) for turbulence verification using an approach that is similar to the approach that is used for PIREPs.

This study demonstrates the suitability of *in situ* EDR observations for use in turbulence forecast verification studies. These verification studies can provide more useful information to users if they are based on multiple independent datasets. These newly developed observations, along with PIREPs, have the capability to meet these requirements.

Acknowledgements

This research is in response to requirements and funding by the Federal Aviation Administration. The views expressed are those of the authors and do not necessarily represent the official policy and position of the U.S. Government.

References

Brown, B.G., G. Thompson, R.T. Bruintjes, R. Bullock, and T. Kane, 1997: Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Weather and Forecasting*, **12**, 890-914.

Brown, B.G. and J.L. Mahoney, 1998: Verification of Turbulence Algorithms. Report, Available from B.G. Brown (bgb@ucar.edu).

Brown, B.G., J.L. Mahoney, R. Sharman, J. Vogt, and J. Henderson, 2000a: Use of automated observations for verification of turbulence forecasts. Report to the FAA. Available from B.G. Brown (bgb@ucar.edu).

Brown, B.G., J.L. Mahoney, R. Bullock, J. Henderson, and T.L. Fowler, 2000b: Turbulence Algorithm Intercomparison: 1998-99 Initial Results. NOAA Technical Memorandum OAR FSL-25, 64 pp.

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

Brown, B.G., J.L. Mahoney, J. Henderson, T.L. Kane, R. Bullock, and J.E. Hart, 2000d: The turbulence algorithm intercomparison exercise: Statistical verification results. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 Sept., American Meteorological Society (Boston), 466-471.

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

Brown, B.G., J.L. Mahoney, R. Bullock, M.B. Chapman, C. Fischer, T.L. Fowler, J.E. Hart, and J.K.Henderson, 2002: Integrated turbulence forecasting algorithm (ITFA): Quality assessment report. Report to the FAA; available from B.G. Brown (bgb@ucar.edu).

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. *11th Conf. of Aviation, Range and Aerospace Meteorology, Hyannis, MA.*

Frehlich, R. and R. Sharman, 2004: Estimates of turbulence from numerical weather prediction model output with applications to turbulence diagnosis and data assimilation. *Mon. Wea. Rev.,* **132**, No. 10, pp. 2308–2324.

Mahoney, J.L., B.G. Brown, R. Bullock, C. Fischer, J. Henderson, and B. Sigren, 2001: Turbulence Algorithm Intercomparison: Winter

2001 Results. Report, submitted to the FAA Weather Research Program. Available from J.L. Mahoney (Jennifer.Mahoney@noaa.gov).

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

Sharman, R., J. Wolff, T. Fowler, and B. Brown, 2002a: Climatologies of upper-level turbulence over the Continental U.S. and Oceans. *10th Conf. of Aviation, Range and Aerospace Meteorology, Portland, OR.*

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

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

Takacs, A., and C. Chen, 2003: Preliminary results of using *in situ* eddy dissipation rate (EDR) observations for turbulence verification. Report, submitted to the Federal Aviation Administration Weather Research Program (FAA/AWRP); available from A. Takacs (agnes@ucar.edu).

Takacs, A., L. Holland, and A. Holmes, 2004a: Using *in situ* eddy dissipation rate (EDR) observations for turbulence forecasts verification. Report, submitted to the Federal Aviation Administration Weather Research Program (FAA/AWRP); available from A. Takacs (agnes@ucar.edu).

Takacs, A., L. Holland, M. Chapman, B. Brown, J. Mahoney, and C. Fischer, 2004b: Graphical turbulence guidance 2 (GTG2): Quality Assessment Report. Report, submitted to the FAA Weather Research Program; available from B. Brown (bgb@ucar.edu) and A. Takacs (agnes@ucar.edu).

Takacs, A., L. Holland, M. Chapman, B. Brown, T. Fowler, A. Holmes, and J. Mahoney, 2004c: graphical Turbulence Guidance 2 (GTG2): Results of the 2004 post analyses. Report, submitted to the FAA Weather Research Program; available from A. Takacs (agnes@ucar.edu).

Thompson, G., 1995: Pilot Report text decoding software. Copyright © 1995, UCAR/NCAR; available from G. Thompson (gthompsn@ucar.edu).

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

Wolff, J.K. and R. Sharman, 2004: Climatological study of aircraft turbulence versus cloud cover based on 3 years worth of data. *11th Conf. of Aviation, Range and Aerospace Meteorology, Hyannis, MA.*