P1.5 CLEAR-AIR TURBULENCE NOWCASTING AND FORECASTING USING IN-SITU TURBULENCE DATA

Jennifer Abernethy* and Robert Sharman
Research Applications Laboratory
National Center for Atmospheric Research
Boulder, Colorado

1. INTRODUCTION

Pilots' ability to avoid turbulence during flight affects the safety of the millions of people who fly commercial airlines and other aircraft every year. Of all weather-related commercial aircraft incidents, 65% can be attributed to turbulence encounters, and major carriers estimate that they receive hundreds of injury claims and pay out "tens of millions" per year (Sharman et al., 2005). In order to change flight paths to avoid turbulence, air traffic controllers, airline flight dispatchers, and flight crews must know where pockets of it are expected to be. While there are turbulence forecasts available currently, both human and automated, none meet the Turbulence Joint Safety Implementation Team's (TJIST) recommended > 0.8 probability of moderate-or-greater (MOG) turbulence detection and > 0.85 probability of null turbulence detection. TJIST is comprised of representatives from the FAA, NASA, federal laboratories and end users, and all these groups are working to improve turbulence forecasting accuracy.

The turbulence forecasting difficulty is due to two main factors: (1) turbulent eddies at the scales that affect aircraft (~ 100m) are a microscale phenomenon and NWP models cannot resolve that scale, and (2) lack of objective observational turbulence data. The prior factor was able to be addressed during the past 50 years, because it was found that most of the energy associated with turbulent eddies at aircraft scales cascades down from larger scales of atmospheric motion (Dutton and Panofsky, 1970; Koshyk and Hamilton, 2001; Tung and Orlando, 2003). The turbulence problem became one of linking large-scale features resolvable by NWP models to the formation of aircraft-scale eddies. Numerous 'rules of thumb' empirical linkages, termed *diagnostics*, were developed by the National Weather Service and airline meteorologists. The forecast skills of these diagnostics depends on the forecaster (for manual forecasts) and diminish with lead time; none meet the TJIST recommendations, either alone or used together in any current implementation. The diagnostics' skills reflect researchers' imperfect understanding of the atmospheric processes involved.

The imperfect nature of the current diagnostics leads forecasters to depend, at least partially, on available turbulence observations. Currently, the only available observations are pilot reports (PIREPs), and they are the second factor contributing to the difficulty of turbulence forecasting (and forecast verification). PIREPs are sparse, aircraft-dependent, subjective assessments by pilots reported during flight. Sharman et al. (2005) shows that PIREP inaccuracy is not as large as once thought (Shwartz, 1996), however, the distribution of reports is not representative of the state of the atmosphere because most non-turbulent areas are not reported.

One major effort by the FAA's Aviation Weather Research Program (AWRP), some major airlines, and the National Center for Atmospheric Research's Research Applications Laboratory (NCAR/RAL) is the development of a better observational data source: *in-situ data* (Cornman et al., 1995; Cornman et al., 2004). In-situ data is turbulence observations recorded automatically every minute during flight by on-board software. It addresses many of the faults of PIREPs: it is aircraft-independent, objective, less sparse, and is designed to be used quantitatively. While the in-situ measurement and reporting system is still in its first and limited deployment, the authors feel the data can and should be used now to increase turbulence forecasting accuracy. Not only does it offer higher-resolution observations, but is also helps alleviate the inconsistent null turbulence-reporting issues with PIREPs (Takacs et al., 2005).

Under sponsorship from the FAA/AWRP, NCAR/RAL and NOAA's Forecast Systems Laboratory (NOAA/FSL), together forming the Turbulence Product Development Team (TPDT), developed the Graphical Turbulence Guidance (GTG) forecasting product, a completely automated turbulence forecasting system currently running operationally at the Aviation Weather Center (AWC) and available on the web at NOAA's Aviation Digital Data Service (ADDS) website: http://adds.aviationweather.gov/turbulence (Sharman et al., 2000; Sharman et al., 2002; Sharman et al., 2004; Sharman et al., 2005). Currently, GTG nowcasts and forecasts only clear-air turbulence (CAT), but is slated to include turbulence associated with convection in later releases. The NCAR/RAL team is researching how to integrate in-situ data into the newest release of the forecasting product, GTG2, as well as taking a fresh look at the design of a forecasting system based on in-situ data. Preliminary results from the GTG2 integration are presented here, as well as future research directions.

2. THE GTG ALGORITHM

GTG2 nowcasts and forecasts clear-air turbulence at both mid (10000ft-20000ft) and upper levels (20000ft-

*Corresponding author address: Jennifer Abernethy, University of Colorado, Department of Computer Science, 430 UCB, Boulder, CO 80309; email: aberneth@cs.colorado.edu

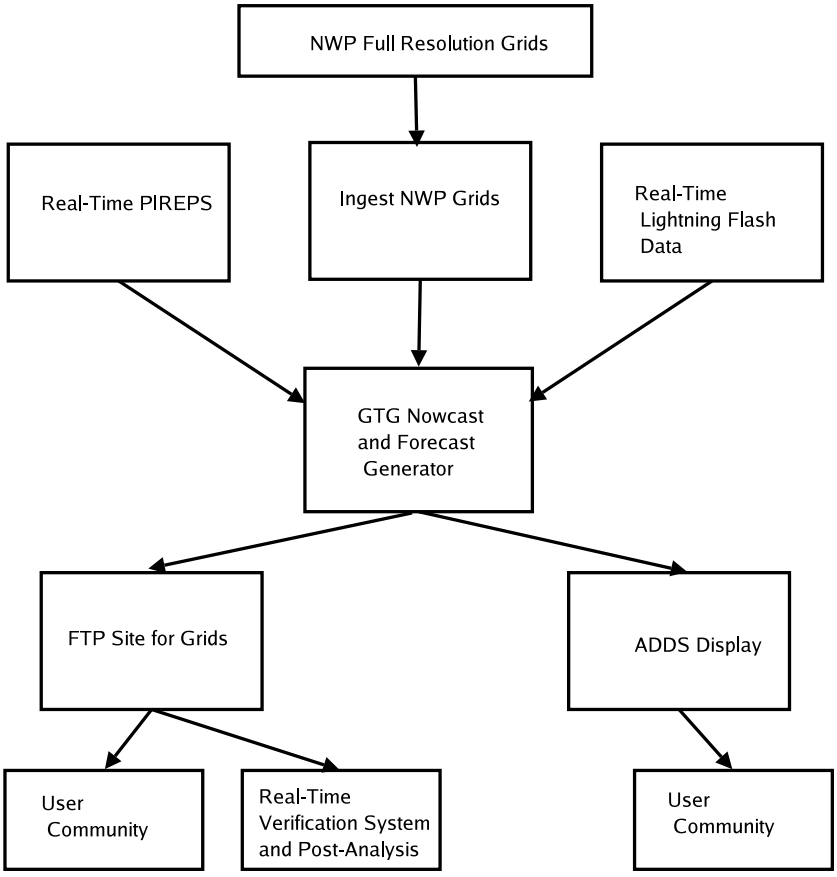


FIG. 1: The GTG forecasting system and its inputs and outputs. Adapted from Sharman et al. (2005).

45000ft[*]) in order to provide guidance for large aircraft in cruise and short, regional flights that do not reach upper levels.

The GTG system uses multiple diagnostics for forecasting (see Sharman et al. (2005) for a description of the diagnostics). While GTG is not the only forecasting system to use multiple diagnostics together, it is the only one to combine them dynamically at forecast time. The set of diagnostics is different for mid-levels and upper levels due to the forecast skill of different diagnostics for certain mechanisms of turbulence creation. GTG combines these values dynamically at every forecast hour using a fuzzy logic algorithm that weights each diagnostic according to agreement with current turbulence observations (PIREPs) to produce a turbulence forecast. If there are not enough PIREPs at the time of the forecast, climatologically-derived weights are used. Thus, GTG can handle the expected variations in observational data.

Figure 1 is a schematic of the GTG forecasting system. GTG uses the National Center for Environmental Prediction's (NCEP) operational Rapid Update Cycle (RUC) weather model output (20km resolution) as a representation of large-scale atmospheric processes. Pilot reports and cloud-to-ground lightning flash data from the National Lightning Detection Network (to filter out turbulence reports near convection) are used as observational data inputs. The majority of the processing takes place in the 'GTG Nowcast and Forecast Generator' box.

The GTG nowcast and forecast generator is detailed in Sharman et al. (2005), but briefly works as follows. Every hour, GTG receives RUC model output files, NLDN lighting data, and PIREP data. From the RUC model output variables in the analysis-time file, GTG calculates values for ten diagnostics for upper-levels and nine diagnostics at mid-levels. Each diagnostic value D is calculated for each RUC grid point.

The diagnostic values and observation data (PIREPs) both must be mapped to a common value range for comparison. For each diagnostic D, the values are mapped to a turbulence scale using a set of established thresholds for the diagnostic. Thresholds were derived from a one-year study of 18Z 6-hr forecasts and represent the medians of the raw diagnostic values corresponding to each major PIREP turbulence category.

[*]Mid and upper levels are flight pressure altitudes or flight levels, which are isobaric surfaces corresponding to a particular geopotential altitude according to the U.S. Standard Atmosphere.

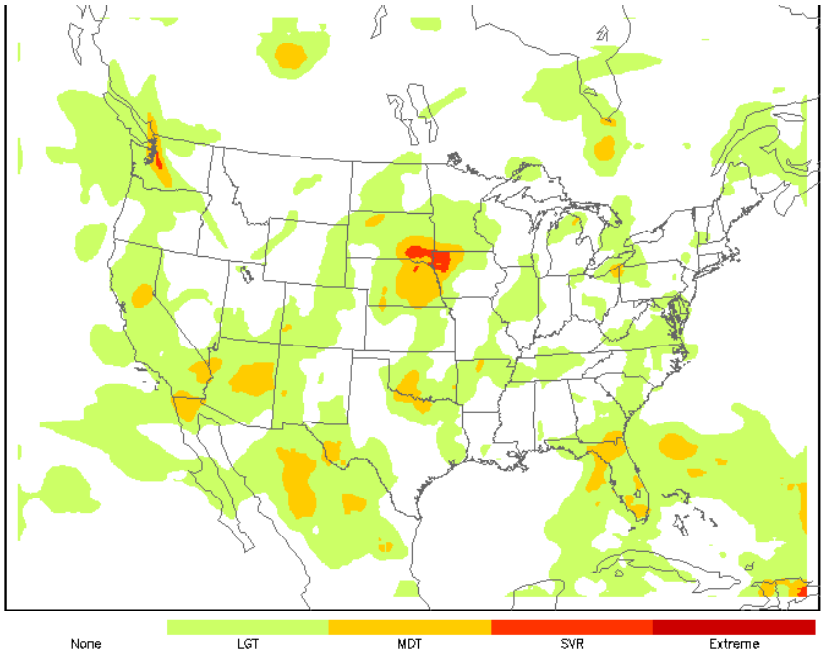


FIG. 2: A sample of the GTG 6-hour forecast for the CONUS available on the Aviation Digital Data Service site, http://adds.aviationweather.gov/turbulence.

The five thresholds (T1,T2,T3,T4,T5) correspond to null, light, moderate, severe and extreme turbulence categories in PIREPs. Using the thresholds, raw diagnostic values are mapped to the range $0 \leq D \leq 1$. T1(null) is 0, T2(light) is .25, T3(moderate) is .5, T4(severe) is .75, and T5(extreme) is 1. Thus, for a diagnostic, raw values below T1 are scaled to 0, values between T1 and T2 are scaled linearly to 0-.25, values between T2 and T3 are scaled linearly to .25-.5, and so on.

Pilot reports (PIREPs) for 90 minutes before and after the analysis time are linearly mapped from a range of $0 \leq p \leq 8$ to a range of $0 \leq p \leq 1$ to enable direct comparison to mapped D_n values. PIREP locations are mapped onto the RUC grid that has been interpolated to flight level. If there is more than one PIREP located in a single RUC grid cell, the pilot report with the highest turbulence intensity is used and the others in that grid cell are ignored. PIREPs coincident to lightning reports (currently, within 20 minutes and 50km) are ignored in order to isolate clear-air turbulence reports from reports of turbulence from other sources.

Each remaining PIREP is matched by location with ten diagnostic values for upper-levels or nine diagnostic values for mid-levels. The diagnostics are then scored according to their agreement with PIREPs. If both the PIREP and a diagnostic are above or below a certain threshold – for GTG2, the threshold used is 0.375 – they are considered in agreement. The threshold corresponds to a moderate PIREP turbulence intensity report, thus defining the class separation between null reports and Moderate or Greater (MOG) reports. Counts of observation and diagnostic agreement (correct and incorrect classifications by each diagnostic) are tallied in a contingency table for each diagnostic. The Probability of Detection (POD) of a MOG event, POD-Yes (PODY) is the fraction of correct MOG classifications out of all MOG observations. Likewise, POD-No (PODN) is the fraction of correct null classifications out of all null observations. From PODN and PODY values, the diagnostic's True Skill Score(TSS) is calculated:

$$\mathrm{TSS} = \mathrm{PODY} + \mathrm{PODN} - 1 \tag{1}$$

Low levels of atmospheric turbulence are expected at any given time (see Section 3). Therefore, it is important to include the volume of forecasted MOG turbulence when scoring a diagnostic. Both f_{MOG} and TSS are used to cal-

culate the score for the diagnostic:

$$\phi_n = \left(\frac{TSS + 1.1}{1 + Cf_{MOG}^{0.25}} \right) \quad (2)$$

Currently, $C = 1$. From the n scores in step 6 for D_n, weights are formed:

$$W_n = \frac{\phi_n}{\sum_{m=1}^{n} \phi_m} \quad (3)$$

subject to $\sum_{m=1}^{n} W_m = 1$ .If 100 reports cannot be found in the full time window, the default weight vector (climatologically-derived weights) is used for this step instead of calculating a new weight vector. The diagnostics are combined into a weighted sum to form the GTG combination. This is done for every grid point (i,j,k) using the weights derived in 3.

$$GTG(i, j, k) = \sum_{m=1}^{n} W_m D_{m,i,j,k} \quad (4)$$

This sum is GTG's turbulence nowcast for the analysis time. This weight vector is then applied to each RUC model forecast output (3,6,9,12 hour forecasts) to produce turbulence forecasts for each forecast time. An example of GTG's output, made into an image by the Aviation Digital Data Service (ADDS), is shown in Figure 2.

3. IN-SITU TURBULENCE MEASUREMENT AND REPORTING SYSTEM

In-situ turbulence measurements are data recorded by special software on commercial aircraft during flight. This measurement and reporting system was developed at NCAR under FAA sponsorship in order to augment or replace PIREPs with a data source that has more precise location and intensity data. Insitu measurements use existing aircraft equipment and are reported using existing communications networks. Detailed coverage of in-situ data methods can be found in Cornman et al. (2004) and Cornman et al. (1995).

The in-situ-derived turbulence metric is the eddy dissipation rate (EDR), $\epsilon^{\frac{1}{3}}$. EDR is recognized as an objective measure of atmospheric turbulence intensity (Panofsky and Dutton, 1984). Two methods to estimate $\epsilon^{\frac{1}{3}}$ onboard aircraft were developed: the accelerometer-based method and the vertical wind-based method. Both are aircraft-independent measurements, and both result in approximately the same turbulence measurements. As an example, the accelerometer-based method uses aircraft vertical acceleration data to estimate eddy dissipation rate through an aircraft vertical-acceleration response function describing how a particular aircraft responds to gusts. The response function considers the vertical motion and pitch of the aircraft, various wing lift forces, etc., and can be mathematically modeled or obtained from the manufacturer or simulation studies. Currently, only the accelerometer-based method is in use, in United Airlines 737 and 757 aircraft. Southwest Airlines and Delta Airlines are scheduled to use the wind-based method when the system is deployed in their aircraft, which is expected to happen by the end of the year.

EDR data is reported once a minute except during takeoff and landing, when data is reported more frequently depending on rate of altitude change. Each in-situ data report is a location (latitude, longitude, altitude) and a set of statistics about various turbulence levels calculated from a number of EDR measurements taken onboard during that minute. The set of statistics are the median eddy dissipation rate (medEDR) and the maximum eddy dissipation rate (maxEDR). Reporting just these two fields reduces transmission costs while still providing a way to distinguish between discrete and continuous turbulence events. The medEDR is the median value of a time series. The maxEDR value is the 95% value of the time series; as a protection measure against erroneous data, peak values are not used. Due to transmission costs, both values are binned into 1 of 8 bins, and each possible pair of maxEDR/minEDR values for a minute is mapped to a single 8-bit character and then downloaded off the aircraft as the EDR data for that minute. The number of bins was limited by the available character sets, but a newer version of the algorithm now in development compresses the EDR data to enable more bins and thus a higher resolution of data.

Currently, in-situ data is being downloaded from 89 United Airlines 757 aircraft. The software is installed on 96 757s and 101 737s. A snapshot of currently available in-situ data is shown in Figure 4. Only a fraction of the null reports are plotted, for clarity.

In-situ data is thought to better reflect the actual amount of turbulence in the atmosphere (Dutton, 1980; Sharman et al., 2005). Figure 3 shows that over 99% of in-situ reports are reports of null turbulence. At any time, at most 0.01% of the atmosphere at upper levels should contain MOG turbulence. In contrast, about half of PIREPs report null turbulence, 27% report light, 17% report moderate and 1% report severe; thus, pilots substantially underreport the null events. In-situ data overcomes this uncertainty by reporting data every minute during flight.

4. THE USE OF IN-SITU DATA IN GTG

The planned release schedule necessitated that the incorporation of in-situ data into GTG2 remain quite simple. Our trials involve replacing PIREPs with in-situ data and how to best merge both data sources together. In addition, we've investigated issues in forecast verification using both in-situ and PIREP data as verification data.

4.1 *Forecast Verification*

This investigation followed the verification practices of the TPDT team, covered in (Takacs et al., 2004; Brown and Young, 2000; Brown et al., 1997), but explained briefly here. First, a forecast is verified against

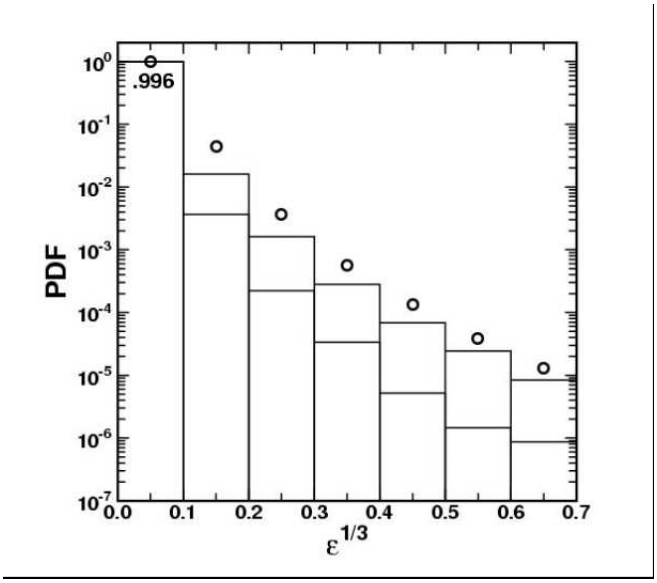


FIG. 3: Taken from Sharman et al. (2005). Distribution of binned $\epsilon^{\frac{1}{3}}$ median (lower bar) and peak, i.e. 95th percentile, (upper bar) values from United Airlines 757 aircraft over a three month time period using the accelerometer-based method described fully in Cornman et al. (1995); Cornman et al. (2004). The open circles are estimates of the distribution based on an assumed lognormal distribution with parameters derived from the RUC20 model (Frehlich and Sharman, 2004). The difference may reflect the ability of commercial air carriers to successfully avoid turbulence.

observational data; a 6-hour forecast at 12 UTC, for instance, has a valid time of 18 UTC and would be verified against observations from 18 UTC. Forecast points are matched with observations by location as described in Section 2. As the primary verification metric, we use the Receiver Operating Characteristic Curve (ROC) curve: the class separation threshold between null and MOG turbulence is varied over a range of 0 to 1, producing a curve of PODY/PODN pairs for that diagnostic (or group of diagnostics, as in the GTG forecast). The curve measures the ability of a forecast algorithm to discriminate between MOG and null turbulence observations. When the threshold is near 0, PODY will be high because almost every observation and almost every forecast value is classified as MOG. The high PODY value reflects the high level of agreement between the two. When the threshold is near 1, PODN will be high and PODY will be low by the same logic. Higher PODY-PODN combinations over the range of thresholds – producing a larger area under the ROC curve - implies greater classification skill. An area under the curve (AUC) of .5 implies no skill (no greater than chance) and an AUC of 1 implies perfect forecasting/classification skill. Background on the use of the ROC curve and AUC as a discrimination metric can be found in (Mason, 1982; Hanley and McNeil, 1982; Marzban, 2004; Kharin and Zwiers, 2003). Observations in the 'light' turbulence category ($0 < p < 0.375$) are generally left out of the verification process due to their higher level of uncertainty, however, they were included in some results below because (1) the first in-situ intensity bin is thought to include light turbulence and (2) it is unknown which bin(s) fully capture light turbulence intensity .

4.2 *Forecasts Using Only In-situ Data*

The first attempt at incorporating in-situ data into GTG2 strived for simplicity: GTG used in-situ data as an observational data source, replacing PIREPs (see Figure 1), but the GTG algorithm remained unchanged. Peak (95%) in-situ turbulence intensities were used instead of median intensities in order to have more non-null turbulence data points available for the GTG forecast. Despite the large difference in turbulence intensity distributions between the two data sources, GTG's scoring algorithm (see Section 2) is robust enough to handle both.

This simple approach had two main drawbacks. First, we ignored any additional knowledge about the turbulence we could have derived from the in-situ data (i.e., an estimate of the size of the turbulent area from a cluster of non-null readings, whether the encounter was discrete or continuous by examining both the median and the peak reading). Second, the scaling of diagnostics essentially is 'calibrated' for PIREPs, and in-situ data is on a different scale with no clear mapping to PIREP turbulence intensities as of yet (work on this question is ongoing).

Despite these drawbacks, the results were positive. When the forecasts were verified against in-situ data, forecasts over three winter months of 2004 and 2005 showed similar or improved forecasting accuracy of in-situ data over PIREPs. Due to uncertainty in how in-situ data turbulence intensities map to PIREP turbulence in-

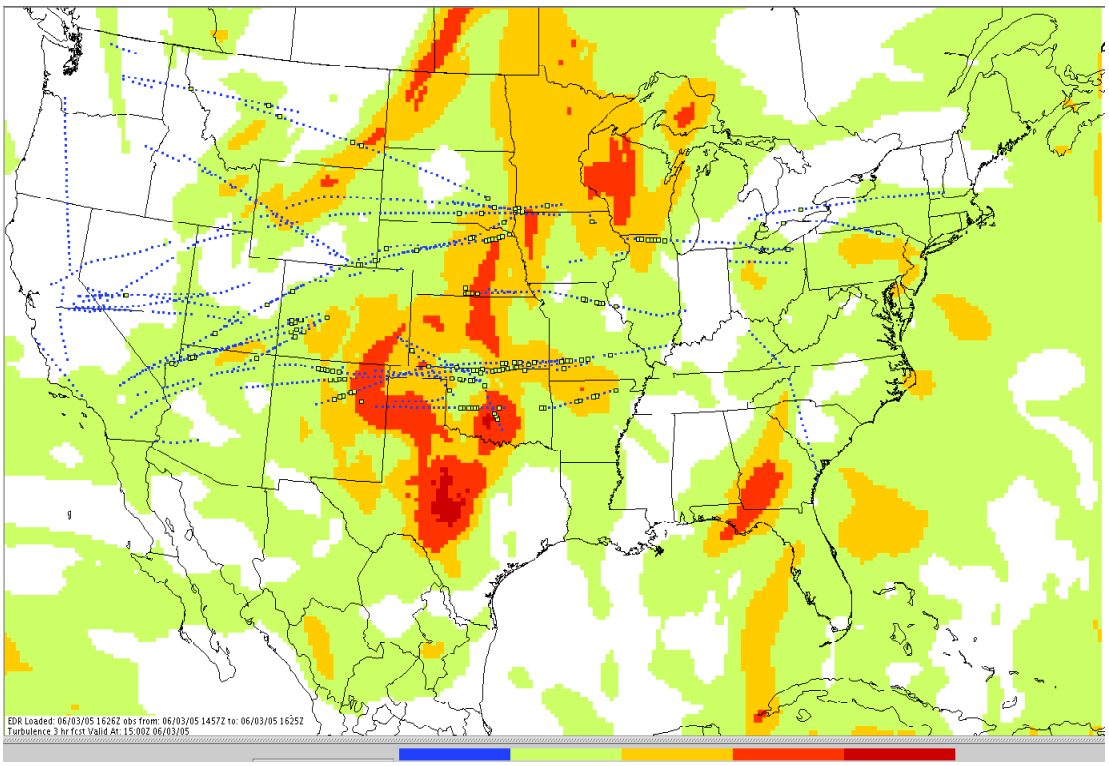


FIG. 4: A snapshot of the in-situ data currently available from United 757 aircraft. The vast majority of data points are reports of null turbulence, so only a sample of those are plotted. Every report $> .05$ is represented by a square. The GTG PIREP-based forecast is shown in the background.

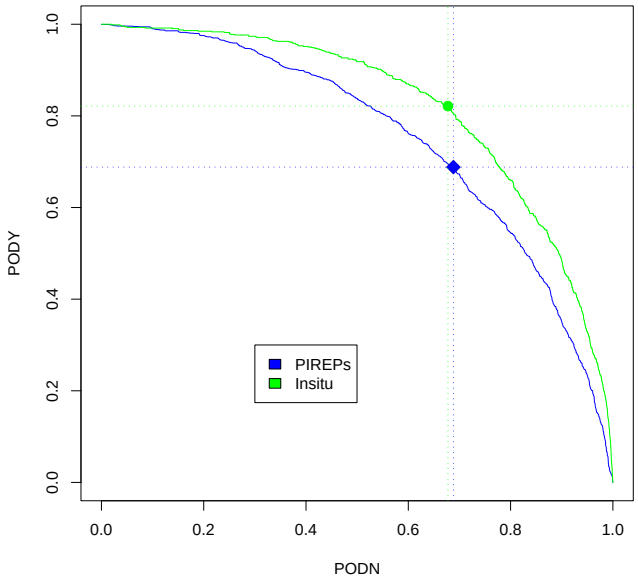


FIG. 5: ROC curves for four winter months (2004-2005) of mid-day 6-hour forecasts made with PIREPs only (blue) and in-situ data only (green). In-situ data was scaled using mapping 1. The in-situ forecast shows increased forecast skill. Here, all PIREP or in-situ intensities were included in the verification (i.e., light reports were not excluded). The AUCs for the PIREP forecast and in-situ forecasts were 0.753 and 0.821, respectively. (The analysis-time (nowcast) AUCs were 0.781 and .878, respectively.) The large point on each curve marks the highest (PODN,PODY) pair and can help identify the optimal threshold.

tensities, as mentioned above, we tried several different mappings. The mappings are shown in Table 1. Most mappings resulted in roughly equivalent forecasting accuracy to PIREP forecasts. However, two mappings did show some improvement in forecast skill. Mapping 1 in Table 1 produced a larger AUC (when including lights) than did the equivalent PIREP forecast (0.821 vs. 0.753, respectively). As an example, Figure 5 shows the ROC curves for that trial. However, when excluding lights, mapping 1 did not perform significantly better than the equivalent PIREP forecast. When the raw in-situ values (ranging from .05 to .75) were unchanged, we also saw an improvement in forecasting accuracy. When excluding light turbulence, the AUC for the original PIREP forecast is .80075 (verified by PIREPs); the AUC for the in-situ forecast is .8399 (verified by in-situ data). When including light turbulence, the PIREP forecast AUC dropped to .753 (as seen above) but the in-situ forecast AUC remained almost exactly the same. When the same in-situ forecast was verified against PIREPs, the AUC lowered to .7866. When the PIREP forecast was verified against in-situ data (excluding lights since PIREP light turbulence intensity value is known), the AUC ranged from .755 to .81 (depending on which in-situ mapping was used). We found from the trials that adjustments in mapping – effectively interpreting the in-situ values as different levels of turbulence intensity – significantly affected the resulting forecast. Additionally, it appears that the type of verification data used affects the forecast's perceived skill.

Another way to look at a forecast's discrimination skill is the difference in the medians of the probability density functions of null and MOG turbulence categories. Figure 6 plots the probability density functions of null and MOG turbulence categories for forecasts using only PIREPs and for forecasts using only in-situ data, respectively. The forecast using in-situ data has a larger difference in the medians of the two categories than does the PIREP forecast (.27 vs. .22 , respectively), indicating a better discrimination skill.

4.3 *Forecasts Using Both PIREPs and In-situ Data*

The next step in investigating simple incorporation of in-situ data into GTG2 was to use both PIREPs and in-situ data as observational data inputs. In each of these trials, the PIREP data points were scaled linearly between 0 and 1 as in the original GTG algorithm, and in-situ data points were scaled between 0 and 1 according to one of several different mappings. The mappings are shown in Table 1 (Mapping 1 was not used because it did not perform well when lights were excluded, and it was not seen as providing the best interpretation of in-situ data). The AUCs for each trial are shown in Table 2.

In these trials, we verified each forecast against PIREPs, in-situ data, and against both sources merged together. The AUCs were slightly lower when the forecasts were verified with both sources. The only improvement in forecast skill was shown when the 'no-scaling' forecast was verified against in-situ data only.

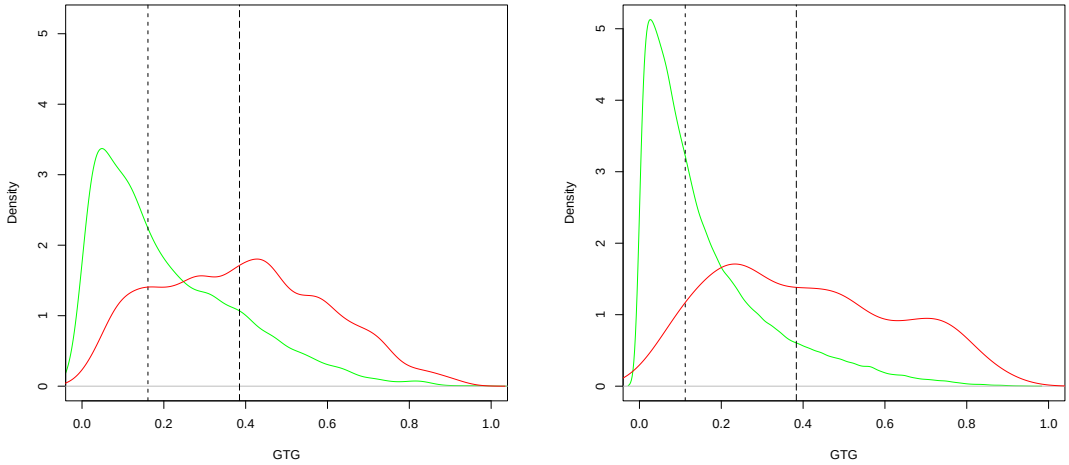
A further analysis of the data sources and the GTG algorithm can shed light the results thus far. First, PIREPs and in-situ data do not agree 100% (see Section 3); observations of each type from the same flight can vary in location and time, PIREPs can disagree with each other about 15% of the time, and the proper mapping of intensities between the two is still being investigated. Thus, we expect that some observations coincident in space and time might contradict each other (this has been seen anecdotally), and this would lower the forecast's verification score. The PIREPs and in-situ data together, when treated as a single observational data set, essentially has a higher error rate than either type of data alone. The diagnostics typically are smooth both horizontally and vertically (Sharman et al., 2005), so contradictory observations in neighboring RUC grid cells can cancel out a diagnostic's positive forecast skill score. We also acknowledge that current interpretation of in-situ turbulence intensity (mappings) could be treating some cases of PIREP/in-situ agreement as contradictions.

Second, the GTG algorithm gives equal weight to the MOG forecasting skill and the null forecasting skill of each diagnostic, regardless of the number of observations available in each category. PIREPs tend to dominate the MOG category while in-situ data dominates the null category. An analysis of the set of observational data used by GTG (as opposed to the superset of data available to be used) over a month of two mid-day forecasts where both PIREPs and in-situ data were available revealed that while there were far fewer PIREPs used for each forecast hour (medians of 5835 in-situ observations and 202 PIREP obervations), PIREPs were overwhelmingly the source of MOG turbulence observations for the scoring step of the GTG algorithm (median of 53 MOG PIREPs per hour and 0 in-situ MOG reports per hour. There were 23 MOG in-situ reports for the month). Likewise, the number of in-situ data null reports was much greater than the number of PIREP null reports used per forecast hour (medians of 5834 and 140, respectively). Thus, the effect of adding a much higher-resolution and more accurate data source is tempered by the way GTG computes a forecast. The majority of the in-situ-only forecast trial improvement shown above probably came from an improvement in the PODN scores of the diagnostics.

At this time, we believe is is best to verify a forecast against in-situ data only for two reasons. First, there are more observations available every hour. While over 99% are observations of null turbulence (Figure 2), it is accepted that the majority of the atmosphere is not turbulent (Sharman et al., 2005) and a low volume of turbulence is desired in turbulence forecasts (eqn. 2). Second, in-situ data is believed to be more accurate than PIREPs for reasons outlined in Section 3.

5. FUTURE DEVELOPMENT

Clearly, the above experiments are merely first attempts at examining the behavior of GTG with in-situ data. We have both short-term and long-term plans for the use of



(a) GTG using PIREP data in forecasts

(b) GTG using in-situ data in forecasts

FIG. 6: Probability density curves of GTG forecast (diagnostic combination) values coincident with null reports (green), and those values coincident with MOG reports (red), from forecasts made with each data source. The medians are marked with vertical lines. The forecasts using in-situ data have a larger difference in medians between the two distributions, indicating better discrimination skill.

in-situ data.

5.1 *Next Steps*

Our next step is to examine the performance of the diagnostics and their re-mapping thresholds with in-situ data. Preliminary work on this revealed that the rank order of best-performing diagnostics changed only slightly when in-situ data was used for verification instead of PIREPs. However, the individual forecasting performance of most diagnostics did vary. This result is not surprising, as most diagnostics were developed using PIREPs for verification.

Along with the diagnostic re-assessment comes the re-assessment of the diagnostic thresholds. Recall that the values of each diagnostic are remapped to a 0 to 1 scale, specifically developed for that diagnostic using PIREPs. A diagnostic has a certain range of possible values, and what values in that range correspond to moderate or severe turbulence can only be determined by comparison to observations. A preliminary look at diagnostic values compared to in-situ observations reveal that the current thresholds need to be adjusted, at least for moderate and severe turbulence intensities. Additionally, whether to have one set of diagnostics thresholds for both PIREPs and in-situ, or a specific set of thresholds for each, needs to be determined.

5.2 *Future Plans*

The team's longer-term plans anticipate a redesign of the GTG algorithm. This redesign could take many forms. Currently, there are several possibilities being investigated. First, with the much greater amount of in-situ data, regional forecasts should be possible. GTG currently makes one forecast for CONUS between 10000-20000ft and one for 20000-45000ft due to the limited number of PIREPs available per hour to score diagnostics, despite the knowledge that mechanisms of turbulence are different in different geographic regions and within the mid- and upper-level bands. With regional forecast areas, forecasts can be tailored to the specific turbulence profile of the area by the choice of diagnostics. Individual forecast regions may be defined by geography, such as the Rocky Mountains, by vertical level, or by atmospheric feature, such as a front or the jet stream. Algorithms exist to identify these features from NWP model variables (Hewson, 1998).

In-situ data is not only more plentiful, but it can give more information about the turbulent area than does a PIREP. Each in-situ report contains both median and peak intensity value. A report with a large difference between the median and peak intensities can indicate a discrete turbulence event (i.e., a single jolt or bump). Likewise, a report with similar median and peak intensities can indicate a continuous event, especially if several consecutive reports are alike. GTG currently does not have any way to consider this additional information in the scoring. How to use this additional information is also a question. Knowledge about a turbulent event could be used to make judgements on the accuracy of a neighboring PIREP, or to place some confidence level on the in-situ report (i.e., in most cases one discrete report is more likely to be erroneous than several consecutive reports of non-null turbulence) and factor that confidence level into the diagnostic scoring procedure.

Table 1: Mappings of raw in-situ intensity (binned) values to the value range used by GTG (0 to 1). One trial involved no change to the raw in-situ values ('no-scaling' trial).

In-situ raw value	Mapping 1	Mapping 2	Mapping 3
0.05	0	0	0
0.15	0.2	0.25	0.25
0.25	0.4	0.50	0.25
0.35	0.6	0.75	0.50
0.45	0.8	1.0	0.5
0.55	1.0	1.0	0.75
0.65	1.0	1.0	0.75
0.75	1.0	1.0	1.0

Parallel to the research on how to best use in-situ data is the choice of combinatorial algorithm to possibly replace the fuzzy logic algorithm currently used in GTG. The fuzzy logic algorithm was designed to be robust for the sparseness of PIREPs, but in-situ data is much more plentiful. Other algorithms such as logistic regression, neural networks and decision trees have been applied ot the CAT forecasting problem but required more training data than was available with PIREPs to outperform the fuzzy logic algorithm (Sharman et al., 2005). These algorithms as well as other machine learning techniques will be researched as possible replacements to the fuzzy logic algorithm, in the pursuit of a more accurate turbulence forecast.

6. CONCLUSION

In an effort to continue improving CAT turbulence forecasting accuracy, NCAR/RAL and the FAA have developed a new, more objective and higher-resolution method to measure atmospheric turbulence: *in-situ data*. This paper covers the first attempts at integrating in-situ data into the current turbulence forecasting product, GTG2, without making changes to the forecasting algorithm. Including in-situ data as an observational data source in GTG2 either alone or combined with PIREPs did not significantly reduce the forecast accuracy; in some cases, it increased the forecast accuracy. Forecast perceived skill depended on the data used for verification; combining PIREPs and in-situ data reduced forecast perceived skill because of the inconsistencies between the two data sources. A fundamental issue is the mapping between PIREP turbulence intensity and in-situ turbulence intensity, which is still being investigated so it is unknown exactly how to interpret in-situ intensities. Another main factor limiting performance increases is the diagnostic thresholding, which was developed for PIREPs. Preliminary work has begun to develop new thresholds for diagnostics using in-situ data.

GTG2 was designed when PIREPs were the only choice in observation data, and as such it does not take advantage of any of the additional turbulence information or resolution of in-situ data. Plans to investigate ways to take advantage of the accuracy and resolution of in-situ data, and alternatives to the fuzzy logic algorithm, were discussed.

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.

REFERENCES

Brown, B., G. Thompson, R. 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. and G. Young, 2000: Verification of icing and turbulence forecasts: Why some verification statistics can't be computed using pireps. In *Preprints, American Meteorological Society Ninth Conference on Aviation, Range and Aerospace Meteorology*, Orlando, FL., 393-398.

Cornman, L., G. Meymarris, and M. Limber, 2004: An update on the faa aviation weather research program's in situ turbulence measurement and reporting system. In *American Meteorological Society Eleventh Conf. on Aviation, Range and Aerospace Meteorology*, Hyannis, MA.

Cornman, L., C. Morse, and G. Cunning, 1995: Real-time estimation of atmospheric turbulence severity from in-situ aircraft measurements. *Journal of Aircraft*, **32**(1), 171–177.

Dutton, J. and H. Panofsky, 1970: Clear air turbulence: A mystery may be unfolding. *Science*, **167**(3920).

Dutton, M., 1980: Probability forecasts of clear-air turbulence based on numerical output. *Meteor. Mag.*, **109**, 293–310.

Frehlich, R. and R. Sharman, 2004: Estimates of turbulence from numerical weather prediction model

Table 2: Areas under the ROC curve (AUC) for trials of Jan 2005 mid-day GTG forecasts made with both PIREPs and in-situ data.

Mapping	Verified by PIREPs only	Verified by In-situ only	Verified by PIREPs and in-situ
'no-scaling'	0.8028	0.847	0.7845
'Mapping 1'	0.796304	0.7847	0.73546
'Mapping 2'	0.7965	0.80302	0.7345

output with applications to turbulence diagnosis and data assimilation. *Monthly Weather Review*, **132**, 2308–2324.

Hanley, J. and B. McNeil, 1982: The meaning and use of the area under the receiver operating characteristic (roc) curve. *Radiology*, **132**, 29–36.

Hewson, T., 1998: Objective fronts. *Meteorol. Appl.* 37–65.

Kharin, V. and F. Zwiers, 2003: On the roc score of probability forecasts. *J. Climate*, **16**, 4145–4150.

Koshyk, J. and K. Hamilton, 2001: The horizontal energy spectrum and spectral budget simulated by a high-resolution troposphere-stratosphere-mesosphere gcm. *Journal of Atmospheric Science*, **58**, 329–348.

Marzban, C., 2004: The roc curve and the area under it as performance measures. *Weather Forecasting*, **19**, 1106–1114.

Mason, L., 1982: A model for assessment of weather forecasts. *Austr. Met. Mag.*, **30**, 291–303.

Panofsky, H. and J. Dutton, 1984: *Atmospheric Turbulence: Models and Methods for Engineering Applications*. John Wiley and Sons, New York.

Sharman, R., C. Tebaldi, G. Wiener, and J. Wolff, 2005: An integrated approach to mid- and upper-level turbulence forecasting. *Weather and Forecasting, submitted*.

Sharman, R., G. Wiener, and B. Brown, 2000: Description and verification of the ncar integrated turbulence forecasting algorithm (itfa). In *Proceedings of the 38th Aerospace Sciences Meeting and Exhibit*, Reno, NV.

Sharman, R., J. Wolff, G. Wiener, and C. Tebaldi, 2002: Technical description document for the integrated turbulence forecasting algorithm (itfa). Technical Report submitted to FAA for AWRP Turbulence PDT Project.

Sharman, R., J. Wolff, G. Wiener, and C. Tebaldi, 2004: Technical description document for the graphical turbulence guidance product v2 (gtg2). Technical Report submitted to FAA for AWRP Turbulence PDT Project.

Shwartz, B., 1996: The quantitative use of pireps in developing aviation weather guidance products. *Weather and Forecasting*, **11**, 372–384.

Takacs, A., L. Holland, M. Chapman, B. Brown, J. Mahoney, and C. Fischer, 2004: Graphical turbulence guidance 2 (gtg2): Quality assessment report. Technical Report by the FAA Aviation Weather Research Program Quality Assessment Product Development Team.

Takacs, A., L. Holland, R. Hueftle, B. Brown, and A. Holmes, 2005: Using in situ eddy dissipation rate (edr) observations for turbulence forecast verification.

Tung, K. and W. Orlando, 2003: The $k^(-3)$ and $k^(-5/3)$ energy spectrum of atmospheric turbulence: quasi-geostrophic two-level model simulation. *Journal of Atmospheric Science*, **10**, 824–835.