Qualitative Reasoning About Small-Scale Turbulence in an Operational Setting

Jennifer Abernethy[1,2*] **Elizabeth Bradley**[1] **and Robert Sharman**[2]
[1]University of Colorado at Boulder
[2]National Center for Atmospheric Research
Boulder, CO

INTRODUCTION

The main challenges in predicting the weather are insufficient computational power and gaps in our understanding of the complex dynamics of atmospheric phenomena. There are comparatively straightforward solutions to these problems: enough teraflops, the right equations. But what happens when you have neither? This is the problem facing aviation turbulence forecasters, who are charged with the task of predicting turbulent conditions that would affect aircraft, but who have neither the computational resources to predict it explicitly nor a complete understanding of how to derive it accurately from available observation data. Yet, commercial and private aviation communities expect accurate, timely turbulence forecasts. The automated turbulence forecasting system currently funded by the Federal Aviation Administration's Aviation Weather Research Program (FAA/AWRP) and used by the National Oceanic and Atmospheric Administration's Aviation Weather Center (NOAA/AWC) integrates qualitative and quantitative reasoning about atmospheric conditions and observations to produce a forecast. This tool, called Graphical Turbulence Guidance (GTG), was developed by the National Center for Amospheric Research (NCAR) and NOAA's Global Systems Division (NOAA/GSD). This paper describes the structure and function of GTG and explores how to improve its turbulence forecasting using better data. Obviously, better data should improve a forecast. Because of the complexity of the software and the system, however, there are significant challenges involved.

The accuracy of turbulence forecasts is critically important; pilots' ability to avoid turbulence affects the safety of the millions of people who fly commercial and private aircraft every year. Although fatalities are low, 65% of all weather-related commercial aircraft incidents 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.* 2006). Turbulence can occur in thunderstorms, clouds, over mountains, near the ground, and even in clear air. Clear-air turbulence or CAT is particularly hard to avoid because it is invisible both to the eye and to radar. One seasoned pilot noted that CAT was his "greatest worry" when flying (Salby 2006). In order to change flight paths to avoid turbulence, air traffic controllers, airline flight dispatchers, and flight crews must know where CAT pockets are likely to be. The dynamical scales on which CAT appears, however, are far finer than those of any current weather model. And observations of the state of the system—reports radioed in by pilots who encounter CAT—are sparse and subjective. For these reasons, no currently available CAT forecast, either human or automated, meets the Turbulence Joint Safety Implementation Team's[*] recommended > 0.8 probability of moderate-or-greater (MOG) turbulence detection and > 0.85 probability of null turbulence detection.

The underlying physics that makes forecasting so hard is one of the "grand challenge" problems of computational science. Turbulence exhibits structure at all scales, all of which trade energy with one another in complicated ways, and numerical methods simply cannot keep up. Turbulent eddies at the scales that affect aircraft (~ 100m), for example, are a microscale phenomenon, but operational numerical weather prediction (NWP) models cannot resolve that scale. There has been some work on understanding how the energy associated with turbulent eddies at aircraft scales cascades down from larger scales of atmospheric motion (Dutton & Panofsky 1970; Koshyk & Hamilton 2001; Tung & Orlando 2003), but that understanding has not yet translated into new turbulence prediction algorithms.

Faced with simulations that are too coarse to truly resolve the behavior that is of interest, plus sparse, subjective observations reported by pilots, the NWP community applies an interesting mix of qualitative and quantitative reasoning in order to identify regions where aircraft-scale eddies are likely to form. The basic reasoning tool is a set of *diagnostics*: "rules of thumb" that reflect human experts' partial knowledge of the physics and their empirical observations over the years. These diagnostics are described in more detail in the following section. Though many of them have been quantified and formalized as technology improved, they are still not up to the forecasting task.

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

[*]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 imperfect nature of the mapping between diagnostics and turbulence leads forecasters to depend, at least partially, on available turbulence observations. Those, too, are inadequate to the task. Currently, the only available observations are qualitative observations reported by pilots (PIREPs). A pilot who encounters a pocket of CAT radios in a report of 'light' or 'moderate to severe,' for instance. These are sparse, aircraft-dependent, highly subjective assessments of turbulence. Pilots are not required to report turbulence at regular intervals, so there may be only one or two PIREPs per flight—if any. Pilots report the level of turbulence they experienced, which can vary for the same atmospheric conditions depending on aircraft size, design, and pilot experience. In addition, the distribution of reports is not representative of the state of the atmosphere because pilots rarely report non-turbulent areas. Further description of PIREPs and their limitations as a data source can be found in the third section of this paper.

Very recently, much better turbulence observation data —termed in-situ data (Cornman, Morse, & Cunning 1995; Cornman, Meymarris, & Limber 2004)—has become available. This data, which is described in the fourth section of this paper, is part of a major effort by the FAA, some major airlines, and the NCAR's Research Applications Laboratory (NCAR/RAL). In-situ data is recorded automatically every minute during cruise by on-board software. It addresses many of the faults of PIREPs: it is aircraft-independent, objective, and less sparse. While the in-situ measurement and reporting system is still in its first and limited deployment, we feel the data can and should be used now to increase turbulence forecasting accuracy. Not only does it offer higher-resolution observations, but it also helps alleviate the inconsistent null turbulence-reporting issues that arise with PIREPs (Takacs *et al.* 2005).

The ideas in this paper are related to work by several groups in the QR community. Orodonez & Zhao (2000) found climatological features using a spatial aggregation framework. Although we are also working on climatology, our goal is not to find large-scale features, and we do not use geometric reasoning. Rather, we are trying to integrate existing qualitative and quantitative data to produce a *local* forecast of a small-scale feature (turbulence). Geometric reasoning may well play a role in our future work, as there may be useful and interesting patterns in CAT distributions—alone or in conjunction with climatological features like the jet stream or geographical features like the Rocky Mountains. Like Yip (1995), we are reasoning qualitatively about coherent features in fluid systems; unlike Yip, we are working with coarse data, local scales, and detailed forecasts. The scale difference is key in our work. Oishi & Ikebuchi (1996) predict small-scale rainfall by solving for its three main predictors qualitatively. Turbulence predictors are large-scale, and we are solving for them quantitatively before interpreting their results qualitatively—at smaller scales. Note that our starting point is an existing numerical simulation (weather prediction) model, and our goal is to integrate that into a qualitative forecasting framework, not to redo the simulation or modeling in a qualitative or semi-qualitative manner. One of our fundamental issues is the transformation of quantitative measurements into qualitative values, which has been treated at length in the QR literature (e.g. Yamasaki *et al.* (1998)). We face sparse data problems similar to those treated in Struss, Sachenbacher, & Dummert (1997), but our system is infinite-dimensional and we want not only to diagnose, but also to forecast. Like Guglielmann & Ironi (2004), GTG uses fuzzy logic in the modeling process, but again, its target system is not finite dimensional.

CLEAR-AIR TURBULENCE DIAGNOSTICS

A clear-air turbulence diagnostic is a simple turbulence model (equation) derived from qualitative expert knowledge based on experience or from basic physical principles. Through the years when forecasts were done manually, forecasters developed "rules of thumb" about what atmospheric conditions typically indicated turbulence. These rules of thumb were an attempt to link the large-scale meteorological data that was available and the micro-scale CAT that was the subject of the forecast (Hopkins 1977). Forecasters later quantified these rules, creating CAT *diagnostics*. For instance, a major cause of CAT is the Kelvin-Helmholtz instability: when gravity waves become steep and unstable, they may break into a chaotic motion (Dutton & Panofsky 1970). This typically happens in areas of strong vertical shear[*] and low local Richardson number (Ri, the ratio of static stability and wind shear). Strong vertical shear overcomes what little stability exists, enabling the wave to break. Thus many qualitative CAT diagnostics concern shears and Ri. There are many different diagnostics linking a large-scale condition to small-scale turbulence. Their predictive power varies, depending upon the large-scale condition that they represent and how directly it is linked to turbulence.

Forecasters use these diagnostics by mapping their values to different turbulence severity levels. As an example, low Ri indicates high turbulence. Early on, forecasters determined some unofficial thresholds to quantify the severity of turbulence that corresponded to a given diagnostic value—"Ri< 0.25 ↔ moderate or greater turbulence," for example (Dutton & Panofsky 1970). In this way, forecasters took their qualitative knowledge about large-scale atmospheric conditions and their relationship to small-scale turbulence, quantified it in the form of diagnostic equations, then interpreted the results using thresholds to produce a qualitative forecast. The GTG forecasting system does exactly the same thing. Its authors used several years' worth of PIREPs to develop threshold values for each diagnostic that map to different levels of PIREP turbulence severity. This allows the diagnostics to work neatly with the qualitative PIREP observations in the GTG system.

QUALITATIVE OBSERVATIONS OF TURBULENCE

Until very recently, the only observation data about aircraft-scale turbulence came from pilots' reports. According to the

[*] The difference in velocity between horizontal layers

Table 1: Turbulence intensity values used in the Pilot Reporting system and their definitions. Taken from Shwartz (1996).

Value	Intensity	Definition
0	None	No turbulence is present
1	Light	Loose objects remain at rest
2	Light-moderate	
3	Moderate	Unsecured objects are dislodged; occupants feels definite strains against seatbelts and shoulder straps
4	Moderate-severe	
5	Severe	Occupants thrown violently against seatbelts; momentary loss of aircraft control; unsecured objects are tossed about
6	Severe-extreme	
7	Extreme	Aircraft is tossed violently about, impossible to control; may cause structural damage
9	Missing	No mention of turbulence

FAA Pilot Reporting System guidelines, a pilot must immediately report when and where s/he encounters turbulence, icing, or any other hazardous condition during flight. S/he must also respond with a hazardous condition report when queried by ground crew. A PIREP consists of a flight number, general location and time, and an assessment of the hazardous condition. In the case of a turbulence encounter, the assessment is a number from 0-7 that represents the qualitative severity of turbulence experienced by the pilot. The PIREP number system is shown in Table 1.

The pilot reporting system was not designed to be used for scientific data analysis; it was intended to keep Flight Service Stations and Air Route Traffic Control Centers aware of in-route conditions so that they might relay the information to other flights in the area (Shwartz 1996). PIREPs have been used quantitatively in verifications and forecasts simply for the lack of any better data with which to verify diagnostics' predictions or study the phenomenon of turbulence itself (Shwartz 1996). There are several specific properties of PIREPs that make them undesirable as a data source and limit their use in understanding and forecasting turbulence: aircraft dependence, accuracy of location and time reported, and sparseness and distribution of reports. For turbulence reports specifically, the regulations state (in Shwartz (1996)): "The degree of the turbulence intensity is determined by the pilot." Not only is a PIREP the pilot's subjective assessment, but the same amount of atmospheric turbulence can have different effects on aircraft of different sizes, causing conflicting reports anywhere from 25% to 32% of the time (Shwartz 1996; Sharman *et al.* 2002a).

Location and time accuracy of PIREPs is hindered by the rules of the reporting system. Pilots usually report the cities or aviation map (NAVAID) locations they are closest to when they experience the turbulence. Most NAVAID locations are kilometers apart, however, which leaves a large uncertainty in the location field of a PIREP. There may also be a time lag between the encounter and the report. A particularly thorny problem is 'en route' reports: the turbulent event could have been located anywhere in the flight path. 16.6% of PIREPs containing turbulence intensities give 'en route' observations (Shwartz 1996).

PIREPs are a spatially sparse data set. Not only is aircraft coverage of the atmosphere at any one time infinitesimal, but most PIREPs do not contain turbulence information. At best, a few hundred per hour in the contintental U.S. are usable —still far fewer than needed for a comprehensive assessment of the current atmospheric conditions at all common flight altitudes. This includes all sources of turbulence, not just CAT. In fact, PIREPs do not specify the type of turbulence, so researchers must compare PIREP location with data about lightning strikes to separate probable convective (thunderstorm) turbulence from CAT.

The amount of PIREP information available at any given time varies widely by time of day, season, altitude and geographical region (Shwartz 1996). The majority of reports are during the day, along major airline routes, and at the typical cruising altitudes of the two distinct groups of pilots: general aviators, who fly up to 18000ft, and commercial airline pilots, who cruise at 30000 to 35000ft. CAT varies geographically, occuring more in the southwest and west than in the northeast. It also varies seasonally, occuring more frequently in the continental U.S. in the winter, when the jet stream (and its associated winds and weather patterns) moves down to lower latitudes.

Pilots rarely file turbulence reports when flying is smooth, so PIREPS are not really representative of the state of the atmosphere (Dutton 1980; Sharman *et al.* 2006). The vast majority of turbulence reports should be 'smooth' or 'null' turbulence, but because pilots rarely feel it necessary to report the absence of hazardous conditions, null reports make up only half of all turbulence reports. Lack of PIREPs, then, may mean the area has a null turbulence level, or just that the pilot did not report the turbulence. Additionally, the area may not be along a flight path, or there might not be many flights using that flight path. Worse yet, it is believed that severe turbulence is probably underreported because of the maintenance checks required for the aircraft after a severe report (Sharman 2005). All of these factors make it difficult to interpret the PIREPs data set.

For all of these reasons, it is the general consensus in the aviation research community that PIREPs should not be used quantitatively (for example, (Shwartz 1996; Brown & Young 2000; Sharman 2005)). This data is sparse and irregular, its accuracy most likely varies, and its distribution does not reflect the distribution of turbulence in the atmosphere. As stated above, PIREPs have been the only source of observation data for researchers, and so they have been forced to use them for turbulence research. Recently, however, NCAR researchers have developed and deployed a new observation system, the In-situ Turbulence Reporting System, that addresses the subjectivity and many of the irregularities of PIREPs.

IN-SITU DATA: SEMI-QUANTITATIVE OBSERVATION DATA

In-situ turbulence measurements are sensor data that is recorded by special software on commercial aircraft during flight. These measurements use existing avionics data and are reported using existing communications networks. Detailed coverage of in-situ data methods can be found in Cornman (1995; 2004).

An in-situ measurement is a measurement of the eddy dissipation rate (EDR) around an aircraft. Eddies are irregular currents of air, and the rate at which eddies break down is recognized as a good measure of atmospheric turbulence intensity (Panofsky & Dutton 1984). EDR is fundamentally different than PIREPs because it *objectively* records the effects of turbulence. It can be estimated from accelerometer or vertical wind data. Both methods yield approximately the same aircraft-independent data. We focus here on data from the accelerometer method, which uses aircraft vertical acceleration data to estimate EDR. This can be converted to an aircraft-independent form using the inverse of an "aircraft vertical-acceleration response function," which describes how a particular aircraft responds to gusts[*]. Currently, in-situ measurements of EDR are being gathered from 197 United Airlines aircraft (101 737s and 96 757s). Several other airlines will deploy the system in the coming year. An example of this data is shown in Figure 1.

EDR data is reported once a minute in cruise and more frequently during takeoff and landing, depending on rate of altitude change. Each in-situ data report is a location triple (latitude, longitude, altitude) and a median and peak (95th percentile) EDR reading from measurements taken over the corresponding minute. Reporting just these two intensity fields reduces transmission costs while still providing a way to distinguish between discrete and continuous turbulence events. The two EDR fields are binned, and each possible pair of median/peak values for a minute is mapped to a single 8-bit character. These data are then downloaded off the aircraft to the ground using the Aircraft Communication and Reporting System (ACARS) network. This binning turns otherwise continuous quantitative observation data into a set of discrete values that are cognate to the PIREP intensity levels.

In-situ data reflects the actual state of the atmosphere (Dutton 1980; Sharman *et al.* 2006). Meteorologists' experience indicates that at any time, at most 1% of the atmosphere at upper levels should contain MOG turbulence. Corroborating this, over 99% of in-situ reports are reports of null turbulence, as shown in Figure 2. In contrast, about half of PIREPs report null turbulence, 27% report light, 17% report moderate and 1% report severe. Unlike pilots, who substantially underreport the null events, sensors are objective and therefore provide a more realistic status of atmospheric turbulence.

Understanding In-Situ Data

While in-situ data accuracy as a measure of atmospheric turbulence has been verified (Cornman, Morse, & Cunning 1995; Cornman, Meymarris, & Limber 2004), researchers are still uncertain about how to interpret the data qualitatively, in terms of the turbulence intensity levels reported by pilots. At what EDR value, for instance, does a 757 aircraft become hard to control? Binning of the data may further complicate the interpretation. The in-situ value bins were somewhat arbitrary, so there is no guarantee that they correspond to qualitative levels of turbulence such as moderate or severe.

In an effort to understand the qualitative features of in-situ data, and with an eye towards integrating it—instead of PIREPs—into GTG, we compared fourteen months of in-situ data to PIREPs from the same flights coincident in time and location[*]. This comparison is not trivial. While a turbulent event may result in single PIREP observation, there may be a number of in-situ observations proportional to the duration of the event. Moreover, in-situ data contains both a median and peak EDR reading each minute. When turbulence is light over several minutes with one big jolt midway, does the pilot report the jolt (severe), or her overall impression of the event (moderate)? The way pilots report turbulence may vary depending on the pilot and situation.

To compare these very different observations, we defined a turbulent event to be a set of consecutive peak EDR readings of the second bin or higher. We (somewhat arbitrarily) used the highest peak EDR value of that set to be the representative 'in-situ value' for the event. This method ignored much of the knowledge about a turbulent event that is inherent in in-situ data, but it enabled simple comparison. Table 2 shows the results. In the vast majority of turbulence events captured by in-situ data, the pilot made no turbulence report within the comparison radius. When a report was made, a slight positive correlation can be seen between PIREPs and

[*] This function considers the vertical motion and pitch of the aircraft, various wing lift forces, etc.

[*] i.e., the report(s) made by a pilot near the time that the in-situ software onboard was automatically recording elevated EDR values: within 40km, five minutes and 1000 ft.

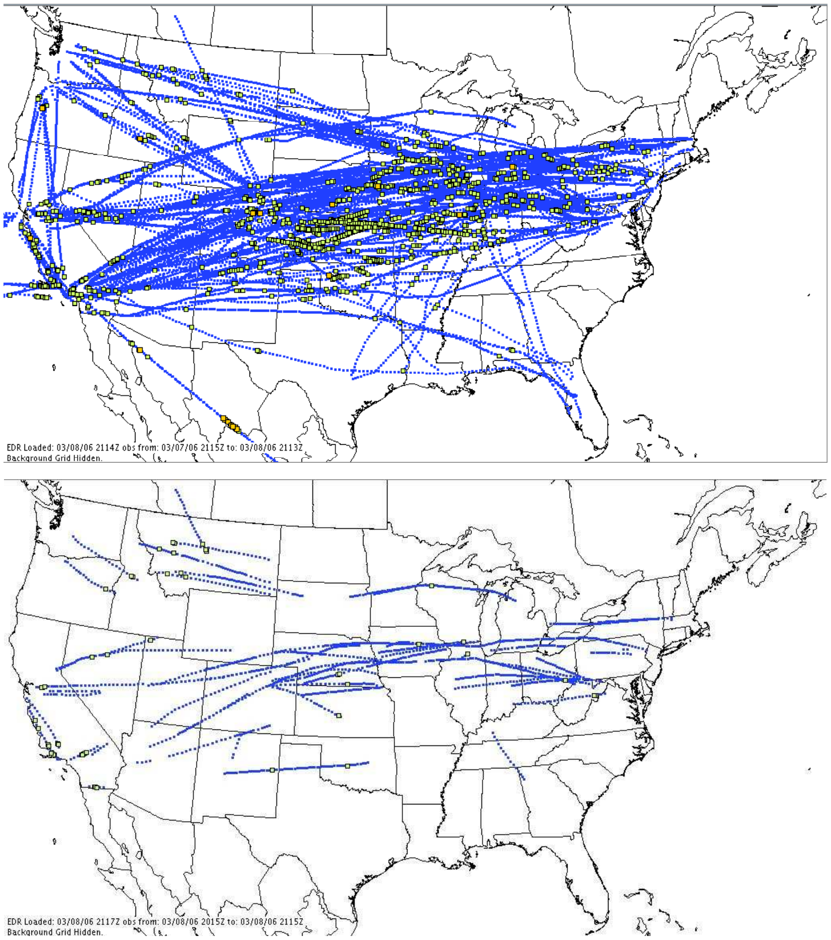

EDR Loaded: 03/08/06 2114Z obs from: 03/07/06 2115Z to: 03/08/06 2113Z
Background Grid Hidden.
EDR Loaded: 03/08/06 2117Z obs from: 03/08/06 2015Z to: 03/08/06 2115Z
Background Grid Hidden.

Figure 1: An example of the in-situ data currently available from United 757 aircraft over 24 hours (top) and from one mid-day hour (bottom). The dots are reports of null turbulence. Squares represent high turbulence levels.

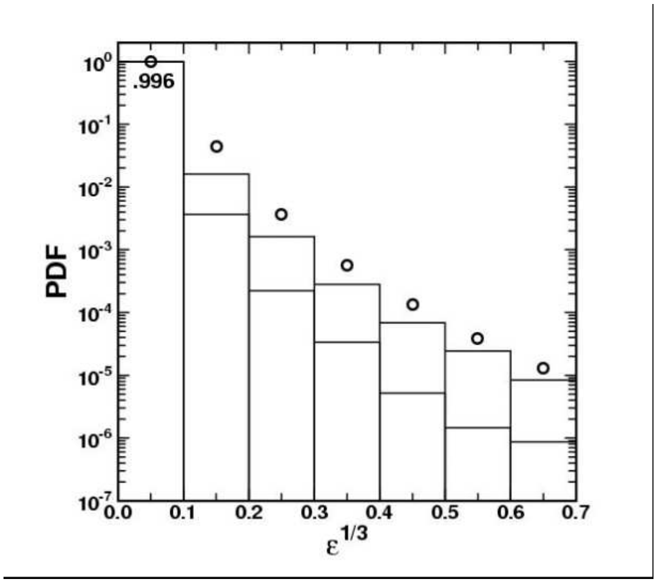


Figure 2: Taken from Sharman *et al.* (2006). This figure shows the probability distribution function (PDF) of observed EDR values ($\epsilon^{\frac{1}{3}}$) in each in-situ bin, both median (lower bar) and 95th percentile (upper bar). This distribution was created from United Airlines 757 aircraft over a three-month time period using the accelerometer-based method described fully in Cornman (1995; 2004). The open circles are estimates of the true lognormal distribution of turbulence in the atmosphere based on the RUC20 weather model (Frehlich & Sharman 2004). The fact that the observed EDR distribution is different than the estimated true distribution of turbulence for upper bins may reflect the ability of commercial air carriers to avoid turbulence during flight.

Table 2: PIREPs and in-situ data recorded concurrently on the same flights from August 2004 - November 2005. There were not enough severe, severe/extreme and extreme PIREPs to make a comparison. No light PIREPs matched any turbulent events captured in in-situ data.

In-Situ Value	Null PIREP	Light PIREP	Light/Moderate PIREP	Moderate PIREP	Moderate/Severe PIREP
0.15	35(12.7%)	0	154 (55.8%)	71 (25.7%)	16 (5.8%)
0.25	3(8.1%)	0	13 (35.1%)	16 (43.2%)	5 (13.5%)
0.35	0	0	3 (25%)	6 (50%)	3 (25%)
0.45	0	0	1 (33.3%)	2 (66.7%)	0

in-situ data. Binning of the in-situ data may be obscuring some of the correlation. We found more of a correlation than did (Takacs *et al.* 2005), primarily because more data was available for our comparison.

As the base of in-situ data grows and we explore other methods of comparison, we may be able to show more definitive correlations between the two data sets.

Using In-Situ Data to Understand PIREP Errors

The in-situ data also allowed us to begin quantifying the errors in PIREPs. To do this, we looked only at time and location, since the intensity comparisons above yielded such vague results. We compared PIREPs and in-situ data from the same flights, as above, and found that the median error in a PIREP-reported location was 50km and the median error in report time was three minutes. This discrepancy is over twice the resolution of the NWP model (20km) used to compute diagnostics. This new result calls the whole PIREP-based forecasting framework into question, and reinforces our intuition that effectively integrating in-situ data into GTG should improve turbulence forecasts.

Compared to PIREPs, in-situ data is more objective, more accurate, more plentiful, and more representative of turbulence distribution. It is clearly a much better set of observation data for use in turbulence research and forecasting. We believe that in-situ data not only allows for more-accurate forecasting, but also expands the options for forecasting algorithms. In the next sections, we describe the currently deployed version of the GTG forecasting system, which was designed and tuned to the PIREP data set, and its attendant limitations. We then discuss how to integrate in-situ data into that framework.

THE GTG ALGORITHM

Under sponsorship from the FAA/AWRP, NCAR/RAL and NOAA/GSD, together forming the Turbulence Product Development Team (TPDT), developed the Graphical Turbulence Guidance (GTG) forecasting product, a completely automated CAT forecasting system currently running operationally at the NOAA/AWC and available on the web at NOAA's Aviation Digital Data Service (ADDS) website(**?**; Sharman, Wiener, & Brown 2000; Sharman *et al.* 2002b; 2004; 2006). GTG forecasts CAT at both mid (10000ft-20000ft) and upper levels (20000ft-45000ft) in order to provide guidance for both large aircraft and short, regional flights that do not reach upper levels.

GTG is a qualitative forecasting system that compares diagnostic values to observation values and uses that information to produce a turbulence forecast. GTG weights each diagnostic based on its binary classification agreement with the current observation data and combines the weighted diagnostics using fuzzy logic to produce the forecasted turbulence intensity. For the final output display, each forecasted intensity value is binned into one of five qualitative bins representing null, light, moderate, severe and extreme turbulence. An example of the output is shown in Figure 3.

GTG uses multiple diagnostics because each has imperfect performance in predicting CAT, and together they can better account for the multiple causes of CAT in the atmosphere. 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. This combination makes GTG the most accurate CAT forecasting system to date (Sharman *et al.* 2006).

The core GTG algorithm is detailed in Sharman *et al.* (2006), but briefly works as follows. Every hour, GTG receives weather model data files (National Center for Environmental Prediction's Rapid Update Cycle (RUC) model at 20km resolution) and PIREP data. From the model output variables in the analysis-time (current conditions for a certain hour) file, GTG calculates values for n diagnostics for upper-levels and mid-levels. Currently, $n = 10$. Each diagnostic value $\mathsf{D_i}$ is calculated for each RUC model grid point.

The diagnostic values and observation data (PIREPs, in the case of the current GTG) both must be mapped to a common value range for comparison. As mentioned earlier, the value range for each diagnostic $\mathsf{D_i}$ is mapped to a $0 \le \mathsf{D_i^*} \le 1$ scale using a set of established thresholds for the diagnostic, corresponding to major PIREP categories. Usually, thresholding is not linear; for instance, some diagnostic values increase exponentially as turbulence increases, or 90% of the value range indicates null turbulence with the remaining 10% split among higher intensities.

PIREPs are mapped linearly from a range of $0 \le \mathsf{p} \le 7$ (Table 1) to a range of $0 \le \mathsf{p}^* \le 1$ to enable direct comparison to $\mathsf{D_i^*}$ values. PIREPs coincident to lightning reports from the National Lightning Detection Network[*] are ignored in order to isolate CAT reports from reports of turbulence related to thunderstorms.

Each remaining PIREP is matched by grid point location with the ten diagnostic values. Here, turbulence is divided into two categories: null and moderate-or-greater (MOG). The dividing threshold corresponds to the scaled value of a moderate PIREP turbulence intensity report. 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 counts, a diagnostic's True Skill Score(TSS) is calculated:

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

Low levels of atmospheric turbulence are expected at any given time. Therefore, it is important to measure the volume of forecasted MOG turbulence ($\mathsf{f_{MOG}}$) for each diagnostic and penalize those that forecast large amounts of MOG turbulence. Both $\mathsf{f_{MOG}}$ and TSS are used to calculate the score for each diagnostic:

$$\phi_\mathsf{i} = \left(\frac{\mathsf{TSS + 1.1}}{\mathsf{1 + Cf_{MOG}^{0.25}}} \right) \tag{2}$$

Currently, the constant C is equal to 1. C and the exponent 0.25 are fudge factors used for tuning. From the n scores for the set D of n diagnostics, weights are formed as follows:

$$\mathsf{W_i} = \frac{\phi_\mathsf{i}}{\sum_{\mathsf{m=1}}^{\mathsf{n}} \phi_\mathsf{m}} \tag{3}$$

subject to $\sum_{\mathsf{m=1}}^{\mathsf{n}} \mathsf{W_m} = 1$. 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).

$$\mathsf{GTG\,(i,j,k)} = \sum_{\mathsf{m=1}}^{\mathsf{n}} \mathsf{W_m D^*_{m,i,j,k}} \tag{4}$$

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.

GTG can produce a forecast in less than seven minutes, which makes it usable in a real-time operational forecast setting. However, it has two main limitations. First, it produces only one forecast over the whole U.S., ignoring regional variations in turbulence conditions and therefore diagnostic performance. Second, GTG was developed using PIREPs for verification. Not only was the algorithm's performance tuned with PIREP data, but the diagnostics and their thresholds were also developed using PIREPs for verification. We believe that turbulence forecasting can be improved by the careful incorporation of in-situ data. The following sections describe our first attempts at exploiting this better data set to improve aviation turbulence forecasting.

[*] Currently, within 20 minutes and 50km

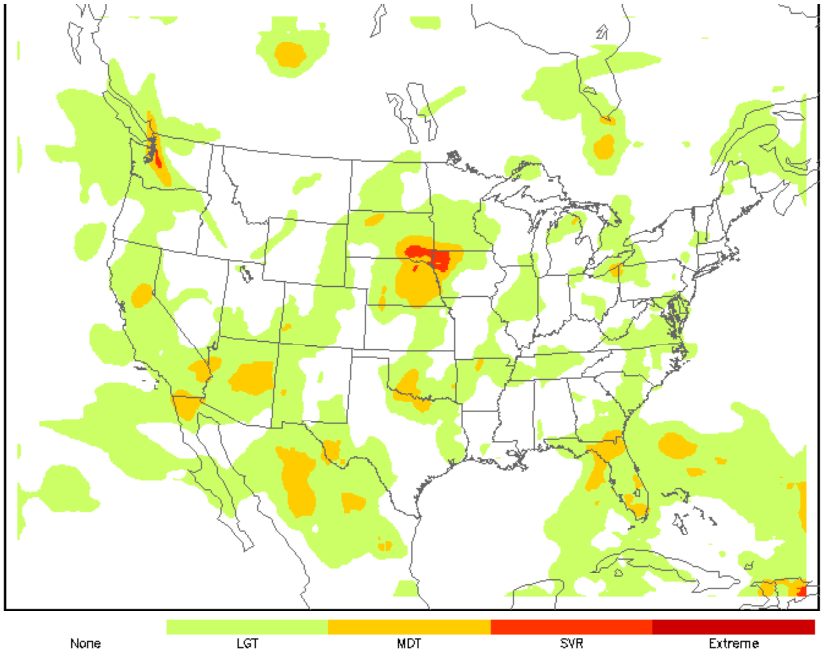


Figure 3: A sample of the GTG 6-hour forecast available on the Aviation Digital Data Service site, http://adds.aviationweather.gov/turbulence. Turbulence levels are color-coded by category: light, moderate, severe and extreme.

THE USE OF IN-SITU DATA IN GTG

We began by producing GTG forecasts using as observation data inputs a) PIREPs, b) in-situ data, or c) both, and verifying the forecasts' accuracies with one or both data sets. We then examined mechanisms in GTG that explained the verification results. Our method of forecast verification is described in the next section.

Performance metrics

It is not trivial to assess the accuracy of a forecast because we do not know the 'truth;' we must use available observation data, however flawed or irregular. To assess forecast performance, we followed the verification practices of the TPDT team, covered in Takacs *et al.* (2004), Brown & Young (2000), and Brown *et al.* (1997), which are explained briefly here. A 6-hour forecast initialized 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 the previous section. As the primary verification metric, we use the Receiver Operating Characteristic Curve (ROC) curve. To construct a ROC curve, we vary the value or threshold at which the null and MOG turbulence classes are separated over a range of 0 to 1, producing a curve of PODY/PODN pairs for the GTG forecast. Each point on the curve is the forecast's classification accuracy for a certain threshold. Figure 4 is an example of a ROC curve. The curve measures how well a forecast algorithm discriminates between MOG and null turbulence observations. Higher PODY-PODN combinations over the range of thresholds – producing a larger area under the ROC curve – implies greater classification accuracy. An area under the curve (AUC) of .5 implies an accuracy no greater than chance. Background on the use of the ROC curve and AUC as a discrimination metric can be found in (Mason 1982; Hanley & McNeil 1982; Marzban 2004; Kharin & Zwiers 2003).

Forecasts Using Only In-situ Data

Our first attempt at incorporating in-situ data into GTG strived for simplicity: we used in-situ data as an observational data source, replacing PIREPs, but left the GTG algorithm 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.

Since GTG scoring amounts to simple binary classification agreement, the issue of scaling (really, intepreting) in-situ bin values became one of choosing the bin value that

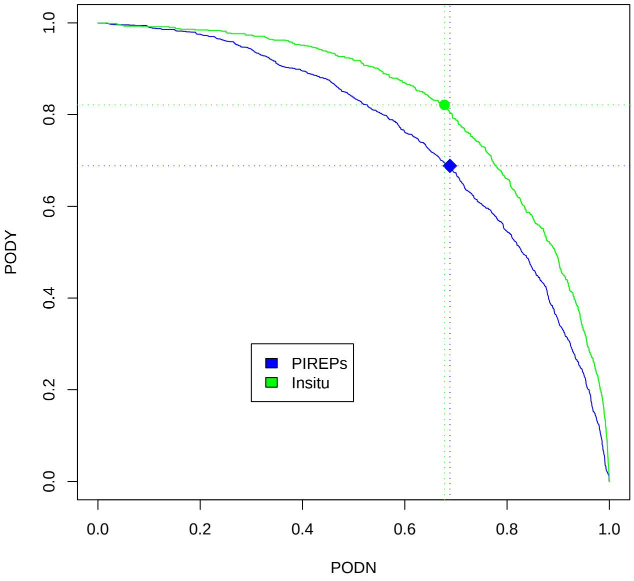
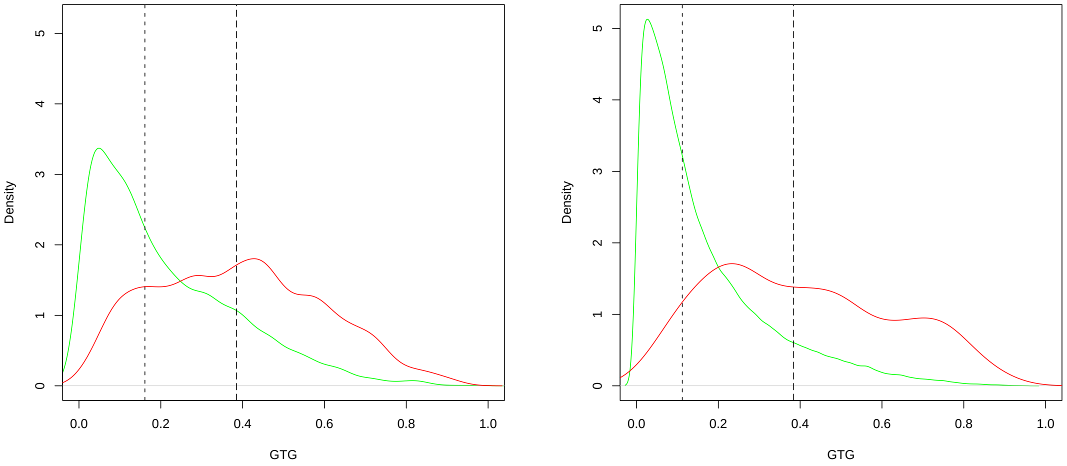


Figure 4: This ROC curve plots the probability of correctly detecting MOG turbulence against the probability of correctly detecting null turbulence for a range of MOG thresholds. The data is from four winter months (2004-2005) of mid-day 6-hour forecasts made with PIREPs only (blue, lower curve) and in-situ data only (green, upper curve), including 'light' reports. Note that the area under the in-situ forecast curve is larger, indicating higher forecast accuracy. The AUCs for the PIREP forecast and in-situ forecasts were 0.753 and 0.821, respectively. The large point on each curve marks the highest (PODN,PODY) pair and can help identify the optimal threshold.



(a) GTG using PIREP data in forecasts

(b) GTG using in-situ data in forecasts

Figure 5: Probability density curves of GTG forecast (diagnostic combination) values coincident with null reports (green, leftmost curve in each figure), and those values coincident with MOG reports (red, rightmost curve in each figure), 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 ability.

separated the classes. Currently in GTG, a PIREP of intensity 3 or higher (scaled to 0.375 in GTG) is classified as MOG and less than 3 is classified as non-MOG (null). Substituting in-situ data as the observational data input into GTG, we found interesting results when $\mathsf{MOG} \geq 0.25$, the third bin, and when $\mathsf{MOG} \geq 0.45$, the fifth bin. Forecasts showed improvement in both cases (AUC of 0.821 and 0.839, respectively) over the same time-period forecasts using PIREPs (AUC 0.753). Figure 4 shows the ROC curve for the $\mathsf{MOG} \geq 0.25$ trial. PIREP forecast accuracy is sensitive to the inclusion of 'light' observations (PIREP intensity 1) in the verification data set because pilots differ in their assessments of light turbulence more than in any other category (Shwartz 1996); in these trials, the PIREP AUC improved to 0.801 when excluding 'lights'. To see if in-situ forecasts had the same sensitivity, we did additional verification on in-situ forecasts by excluding 'light' observations: in the former case, excluding all second-bin in-situ observations and in the latter case, excluding all second, third and fourth bin observations. We found that the in-situ forecast accuracy only varied in the former case (AUC lowered to near that of PIREP forecast's 0.801). The former case's variability shows that the second bin observations buoyed the overall score, since the third and fourth bin observations were probably misclassified as MOG (as indicated by the latter case).

For further comparison, we validated each type of forecast with the other type of observation data. The PIREP forecast had an AUC ranging from 0.755 to 0.810 depending on the in-situ class separation threshold, which is comparable to its forecast accuracy verified by PIREPs (with and without 'lights'). The in-situ forecast verified by PIREPs had an AUC of 0.786, indicating disagreement between the data sets and reflecting the far lower number of null observations in the PIREP data set.

Another way to assess a forecast's accuracy is to measure its ability to correctly discriminate between MOG and null turbulence. Figure 5 plots the probability density functions of the GTG forecasted intensities at points of observed null and MOG turbulence for forecasts using only PIREPs and for forecasts using only in-situ data. The medians of each curve are marked with vertical lines. The forecasts using in-situ data have a larger difference between the medians of the two categories than does the PIREP forecast (0.27 vs. 0.22 , respectively), indicating a better discrimination ability. This confirms our intuition that in-situ data improves forecasts.

Combining Semi-Quantitative and Qualitative Data

Our next step was to combine the two types of observation data for both forecasting and verification. Using the class separation of $\mathsf{MOG} \geq 0.45$ for in-situ data and $\mathsf{MOG} \geq 3$ for PIREPs, the forecast AUC was 0.7845 when verified by both data sets. However, the AUC rose to 0.847 when verified by in-situ only and 0.803 when verified only by PIREPs. We believe the lower AUC for the combined verification is due to the contradictions in the data sets; the diagnostics typically are smooth both horizontally and vertically (Sharman *et al.* 2006), so contradictory observations in neighboring RUC grid cells can cancel out a diagnostic's positive forecast accuracy score. Additionally, the GTG algorithm gives equal weight to both the MOG and null forecasting accuracies 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. 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.

CONCLUSION

Turbulence is both a financial and human safety issue for aviation, but the scale of turbulence that affects aircraft cannot be resolved by current NWP models. Thus, forecasters have long had to predict turbulence qualitatively: using qualitative observation data (PIREPs), understanding the effects of large-scale conditions on turbulence formation, and interpreting a forecast as qualitative levels of turbulence. The most accurate turbulence forecasting system to date, GTG, automates this forecasting process by thresholding quantitative data and scoring its classification accuracy with qualitative observation data. With fuzzy logic those scores are translated into weights that are then used in a weighted sum of the diagnostics to produce a turbulence forecast.

In this paper we presented new results on the comparison of in-situ and PIREP data and the improvement of turbulence forecasts using in-situ data. Our comparison showed a positive relationship between PIREP and in-situ turbulence intensities, and we found the average location error of PIREPs to be 50km. We improved the forecasting accuracy of GTG by using in-situ data as an observation data source. Forecasts using in-situ data showed more forecasting skill and better categorical discrimination ability than did traditional GTG forecasts using PIREPs. PIREPs and in-situ data have complementary distributions but still contradict each other enough to negate any advantage of combining them for forecasting or verification. Further improvements in forecast accuracy are expected when the algorithm is adapted to reason with all the information that is available in in-situ data.

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., and Young, G. 2000. Verification of icing and turbulence forecasts: Why some verification statistics can't be computed using PIREPs. In *Preprints, American Meteorological Society Seventh Conference on Aviation, Range and Aerospace Meteorology*, 393–398.

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

Cornman, L.; Meymarris, G.; and Limber, M. 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*.

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

Dutton, J., and Panofsky, H. 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 Sharman, R. 2004. Estimates of turbulence from numerical weather prediction model output with applications to turbulence diagnosis and data assimilation. *Monthly Weather Review* 132:2308–2324.

Guglielmann, R., and Ironi, L. 2004. The need for qualitative reasoning in fuzzy modeling: Robustness and intepretability issues. In *Proceedings of the International Qualitative Reasoning Workshop*.

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

Hopkins, R. 1977. Forecasting techniques of clear-air turbulence including that associated with mountain waves. Technical Note No. 155 for the World Meteorological Organization.

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

Koshyk, J., and Hamilton, K. 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 and Forecasting* 19:1106–1114.

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

Oishi, S., and Ikebuchi, S. 1996. Inference of local rainfall using qualitative reasoning. In *Proceedings of the 11th International Qualitative Reasoning Workshop*.

Orodonez, I., and Zhao, F. 2000. STA: Spatio-temporal aggregation with applications to analysis of diffusion-reaction phenomena. In *Proceedings of the AAAI*.

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

Salby, M. 2006. personal communication.

Sharman, R.; Wolff, J.; Fowler, T.; and Brown, B. 2002a. Climatologies of upper-level turbulence over the continental U.S. and oceans. In *American Meteorological Society Ninth Conf. on Aviation, Range and Aerospace Meteorology*.

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

Sharman, R.; Wolff, J.; Wiener, G.; and Tebaldi, C. 2004. Technical description document for the graphical turbulence guidance product v2 (GTG2). Technical Report submitted to FAA for AWRP Turbulence PDT Project.

Sharman, R.; Tebaldi, C.; Wiener, G.; and Wolff, J. 2006. An integrated approach to mid- and upper-level turbulence forecasting. *Weather and Forecasting, accepted*.

Sharman, R.; Wiener, G.; and Brown, B. 2000. Description and verification of the NCAR integrated turbulence forecasting algorithm (ITFA). In *Proceedings of the 38th Aerospace Sciences Meeting and Exhibit*.

Sharman, R. 2005. personal communication.

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

Struss, P.; Sachenbacher, M.; and Dummert, F. 1997. Diagnosing a dynamic system with (almost) no observations: A case study in off-board diagnosis of the hydraulic circuit of an anti-lock braking system. In *Proceedings of the 12th International Qualitative Reasoning Workshop, Cortona, Italy and Working Papers of the Third IJCAI Workshop on Engineering Problems for Qualitative Reasoning, Nagoya, Japan*.

Takacs, A.; Holland, L.; Chapman, M.; Brown, B.; Mahoney, J.; and Fischer, C. 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.; Holland, L.; Hueftle, R.; Brown, B.; and Holmes, A. 2005. Using in-situ eddy dissipation rate (EDR) observations for turbulence forecast verification. Quality Assessment PDT Report to the FAA Aviation Weather Research Program.

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

Yamasaki, T.; Yumoto, M.; Ohkawa, T.; Komoda, N.; and Miyasaka, F. 1998. Transformation of quantitative measurements into qualitative values in stochastic qualitative reasoning for fault detection. In *Proceedings of the 13th International Qualitative Reasoning Workshop*.

Yip, K. 1995. Reasoning about fluid motion 1: Finding structures. In *Proceedings of the International Joint Conference of Artificial Intelligence*.