10.2 RESULTS FROM THE NCAR INTEGRATED TURBULENCE FORECASTING ALGORITHM (ITFA) FOR PREDICTING UPPER-LEVEL CLEAR-AIR TURBULENCE

Robert Sharman*, Claudia Tebaldi, Jamie Wolff, and Gerry Wiener
National Center for Atmospheric Research, Boulder, CO USA

1. Introduction

Commercial (Part 121/129), air taxi (Part 135), and general aviation (GA - Part 91) encounters with turbulence continue to be a source of occupant injuries, and in the case of GA, fatalities and loss of aircraft. According to a recent MCR Federal survey of NTSB accident data (1983-1997), turbulence contributed to 664 accidents leading to 609 fatalities (mostly GA), 239 serious and 584 minor injuries, for an estimated average annual societal cost of $134 M. Although fatalities related to commercial airline turbulence encounters are almost nil (only one in this time period), turbulence-related injuries still account for a significant fraction (about 30%) of all weather related Part 121/129 incidents. The average number of air carrier turbulence-related injuries according to the NTSB records is about 45 per year, but these are of course only those that were actually reported. The actual number is probably higher: one major carrier reported almost 400 turbulence encounters leading to injuries over a 3 year period; another estimated about 200 turbulence-related customer injury claims per year. Costs to the airlines are difficult to establish, but one major air carrier estimated it pays out "tens of millions per year" for customer injuries, and looses about 7,000 days in employee injury-related disabilities. The vast majority of air carrier turbulence incidents occur above 20,000 ft, where passengers are more likely to be unbuckled. The MCR report also estimated that only about 30% of these upper-level incidents were forecast based on previous turbulence pilot reports (PIREPs) or valid AIRMETs. Hence better upper-level turbulence forecasts should substantially reduce injuries to passengers and crew.

Over the last four years NCAR/RAP and NOAA/FSL, under sponsorship from the FAA Aviation Weather Research Program (AWRP), have been developing a completely automated turbulence forecasting system for upper-level (≥20,000 ft) turbulence. According to the NTSB records, over half of the encounters at these levels are in the so-called clear-air turbulence (CAT) category. This forecasting system, called ITFA (Integrated Turbulence Forecasting Algorithm) concentrates on nowcasting and forecasting CAT above 20,000 ft. Convective sources of turbulence are not explicitly accounted for. Using PIREPs for verification, a time history of ITFA performance for a specified threshold value of moderate or greater (MOG) turbulence, is given in Fig. 1 in the form of the true skill statistic (TSS). The Aviation Weather Center manually-derived AIRMETs TSS are included for comparison. These were derived using NOAA/FSL's Real-time Verification System (RTVS) (Brown, et al., 2000), which derives daily statistics for probabilities of detection of MOG events (PODY) and null events (PODN), where an event is a valid PIREP. By this measure the two techniques

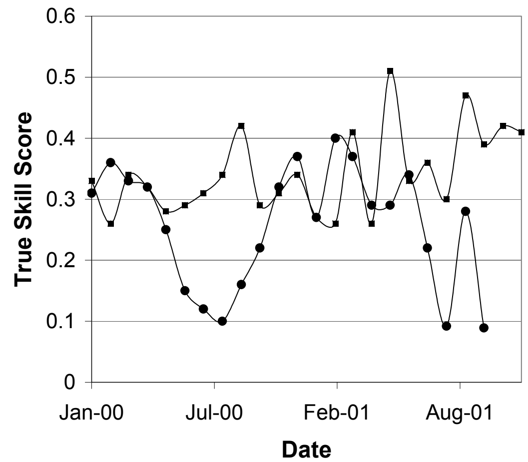

Figure 1. Time series of true skill statistic (TSS) for ITFA thresholded at 0.15 (■) and AIRMETs (●) Jan 2000 – Dec 2001.

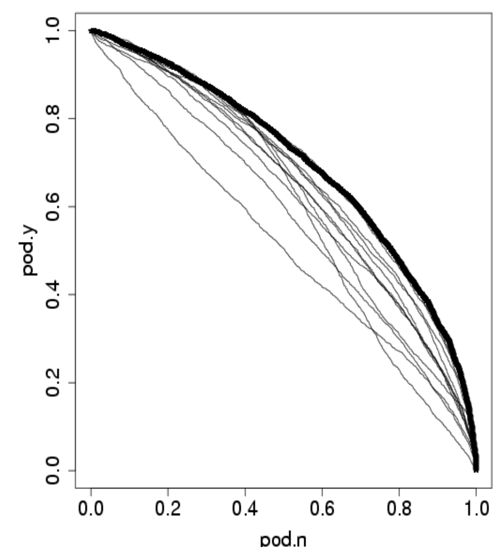

Figure 2. ITFA PODY vs. PODN performance (heavy line) compared to several individual diagnostic performances (thin lines) for 0 hr forecast (i.e., analysis).

* *Corresponding author address*: Robert Sharman, NCAR/RAP, P. O. Box 3000, Boulder CO 80307; email: sharman@ucar.edu

(manual and automated) give similar performance. Figure 2 shows the current ITFA performance relative to individual diagnostic performance on a PODY vs. PODN plot, where (1,1) would be perfect performance, for a range of MOG thresholds. Over most of the range, the ITFA combination gives superior results.

The ITFA procedure has been described elsewhere (Sharman et al. 2000), but briefly, it involves the following four-steps:

(1) Derive a set of turbulence diagnostics from numerical weather prediction (NWP) model output. The current ITFA calculates some 31 turbulence diagnostics relevant to upper-level CAT. For a list of these see the Appendix in Sharman et al. (2000). However, some of these have been found to perform poorly overall, so are computed but not used in further processing.

(2) Determine a mapping of diagnostic values to a turbulence potential. This involves establishing thresholds for each diagnostic that define a set of minimum values for a turbulence category (light, moderate, severe) on a 0-1 scale.

(3) Score the performance of each individual diagnostic against the available observations, also mapped to a 0-1 scale. Currently, the only available observations are in the form of pilot reports (PIREPs) which NCAR receives routinely from NOAA's Family of Services. Soon we expect to be receiving quantitative in-situ measurements of turbulence for ACARS equipped aircraft (see Cornman, et al., 1995, for a description) that will significantly improve the quality and quantity of our observational database.

(4) Combine the different diagnostics as a weighted sum with the weights determined dynamically from the results of step 3.

Each of these steps has involved research into optimal methods for accomplishing that step, and the remainder of this paper will summarize some of the results of that research.

2. Diagnostics and thresholding

As mentioned in the introduction, currently 31 different diagnostics of CAT are computed within ITFA. These are primarily intended to diagnose regions of high turbulence potential due to the presence of upper-

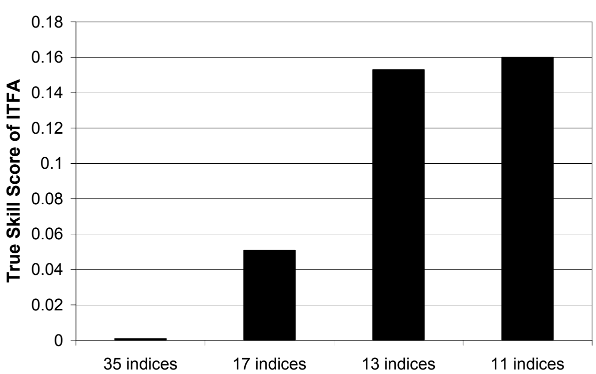

Figure 3. TSS of ITFA as a function of the number of separate diagnostics used.

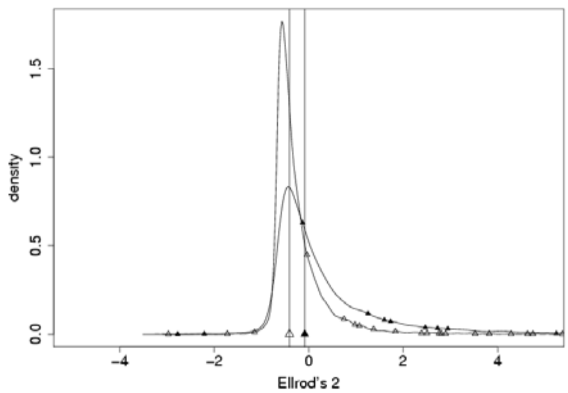

Figure 4. Normalized densities of TI2 corresponding to MOG (▲) and null (∆) PIREPs. Vertical lines are median values.

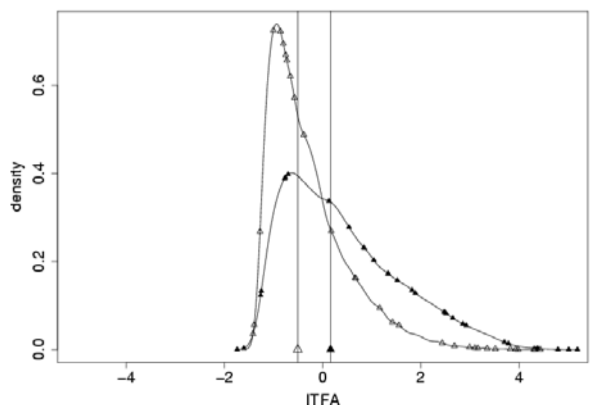

Figure 5. Same as Fig.4 but for ITFA using 16 diagnostics.

level fronts and jet streams. Other classes of upper-level turbulence diagnostics are under development e.g., at FSL and NASA (Kaplan, et al., 2000) that should lead to better diagnoses of turbulence associated with sharp upper-level ridges and other regions of highly unbalanced flow. An example of the effect of the number of indices used for a fixed MOG threshold of 0.375 on ITFA performance is shown in Fig. 3. In general, we have found fewer indices is better, and the current version uses 11 diagnostics in the actual ITFA combination.

Part of the difficulty with any diagnostic is its lack of ability to discriminate between different levels of turbulence intensity, again based on comparisons to available PIREPs. Figure 4 demonstrates the discrimination power for a common turbulence diagnostic TI2, due to Ellrod and Knapp (1992). Here the two curves represent the conditional density of values of TI2 corresponding to 1) null and 2) MOG PIREPs. Ideally, the two curves should not overlap, but in fact the overlap is substantial, so that the median values, drawn as vertical lines, are not widely separated. In these normalized coordinates, the degree of discrimination can be inferred from the distance between the medians of the two curves.

Almost all individual diagnostics show very small separations between the medians, in fact most are worse than TI2 by this measure, but the ITFA combination is better in this regard (Figure 5).

Based on figures like Figure 4 we are able to estimate threshold values for each of the diagnostics. These can of course be adjusted seasonally, or they can be adjusted dynamically with each new forecast. Dynamic mapping can also be performed by assuming those diagnostic values that are in the upper say 97% of the range of computed values correspond to MOG turbulence values. This seems to give more robust results than statically derived (constant) thresholds. For example, in one case study the TSS of ITFA using static thresholds applied to 11 indices was about 0.02, but with the 97% dynamic thresholding criterion the TSS increased to about 0.15, and with a 95% dynamic threshold the TSS increased to about 0.16.

3. Scoring and optimizing strategies

Once the diagnostic thresholds are established, either statically or dynamically, the next step is to determine how well the diagnostic can account for current observations. To do this PIREPs are collected within some time window of the NWP analysis time, typically 60 to 90 minutes, and at each location corresponding to a PIREP the thresholded diagnostic value is compared to the PIREP intensity value. Considering the PIREP as "truth", diagnostics that agree well with the PIREP are given a higher "score" than those diagnostics that do not agree well. For example if diagnostic A at the location of a moderate intensity PIREP was above its threshold value for moderate turbulence the score assigned would be relatively high, whereas if diagnostic B was below the light intensity threshold, the score assigned would be relatively lower. The score can be determined in a number of ways, for example by forming the difference |PIREP intensity - diagnostic intensity| - squared or unsquared, and summed over all PIREPs or by computing PODY,N statistics. The difference scoring has finer granularity but the PODY,N scoring provides a more consistent measure of overall performance. We have not found substantial differences in overall performance between these two methods.

Once each diagnostic has been scored relative to every other diagnostic, it still remains to combine the diagnostics in some manner consistent with the relative scores. Methods for optimizing the information from the suite of diagnostics is an ongoing area of research. The current optimization strategy uses a simple weighted sum of the indices with the weights inversely proportional to the minimum squared difference score of each index computed from the latest observation time. This simple strategy seems to work well, although other methods such as logistic regression (e.g., McCullagh, 1983) or neural nets (e.g., Cheng and Titterington, 1994) show slightly better performance (see Figure 6). Experiments with training, where more than one observation period is used to establish the scores and weights does seem to be beneficial. Figure 7 shows the effect of training on logistic regression performance for the same data as was used to construct Figure 2. Increasing the number of training sets (observational times) displaces the PODY-N curve upward so that performance is enhanced for all threshold values.

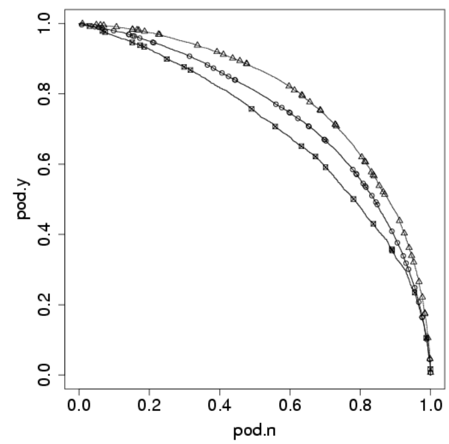

Figure 6. Effect of POD performance on the optimization strategy used. □ ITFA, ○ logistic regression, Δ neural nets, based on average of 3-12 hr forecasts average.

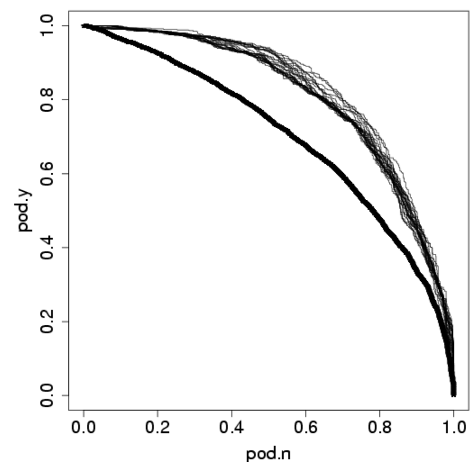

Figure 7. The effect of training on POD logistic regression performance. Lower curve: no training, upper curves: increasing amounts of training. All forecast times are used (0-12 hrs).

4. Discussion

The ability to provide accurate aircraft-scale turbulence forecasts is hampered by several fundamental difficulties. First, the resolution of current numerical weather prediction (NWP) models (several 10s to 100 km roughly) is about two orders of magnitude too coarse to resolve aircraft-scale turbulence (roughly 100s m). Therefore, aircraft-scale turbulence diagnoses/predictions must be based on resolvable scale features. However, and this is the second difficulty, the performance of turbulence diagnostics is hampered by our current lack of understanding of the linkage between NWP observable scale features and aircraft-scale turbulence. An implicit assumption in all these diagnostics is that turbulence generating mechanisms have their origin at resolvable scales and the energy cascades down to aircraft scales, but it is unclear what the exact mechanism is that creates small scale motion from the larger scales. Third, even if it is true that aircraft-scale turbulence has its origins at the resolvable scales, the turbulence forecast system has all the inherent NWP errors associated with the resolvable scales. Fourth, it is not clear that the current suite of turbulence diagnostics is in fact capturing all the relevant information that the larger scale representations can provide. For example, turbulence associated with upper-level ridges is poorly modeled by the current suite of indices (e.g., Knox, 1997). Finally, there is the difficult matter of verification. In the ITFA system we are using PIREPs for tuning and verification. But the individual PIREP is subject to spatial and temporal errors, and is subjective in its intensity rating. Further, the PIREPs are variable in space and time, and in particular undergo a strong diurnal period (considerably fewer at night) making in difficult to perform consistent verifications over all time periods. As mentioned earlier, the quantitative automated in-situ reports (Cornman, et al., 1995) should eliminate most of the uncertainty associated with PIREPs but this is probably a couple of years away.

The following research areas are being pursued to obtain better overall performance within the ITFA framework:

- Better diagnostics. This is a continued research area in the many laboratories and universities. But any diagnostic must be judged by its overall performance, not just its performance on a few selected cases. In addition, information about when a particular diagnostic performs well could be used to dynamically modify its weight within the ITFA framework. But this situational dependence can only be assessed through careful case studies.
- Extension to lower altitudes. Most of the current diagnostics within ITFA implicitly attempt to detect/forecast jet stream/upper-level frontal turbulence. However, others are more general, e.g., the Richardson number, and these diagnostics might be extendable to altitudes below 20,000 ft, allowing a mid-level ITFA.
- Dynamic tests for discrimination. As shown earlier, the discrimination power of most diagnostics is poor on the average, and for particular cases, can be nil. The discrimination power could be assessed dynamically, and weights within ITFA adjusted accordingly.
- “Local” fits. Within the current ITFA framework, the best fit of diagnostics is attempted for the entire volume of atmosphere above some level. Better fits are probably attainable to subvolumes, which could be overlapped to give smooth transitions from one subvolume to another.
- Better optimization strategies. Although several optimization or weighting strategies have been tried, others are available and it may be that one of those methods leads to demonstrably better performance. Also, better methods may be derived for combining indices when several sets of indices are intended to describe one turbulence generation source, and another set describes a different generation source.

Acknowledgments. This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.

References

Brown, B.G., J. L. Mahoney, J. Henderson, T. L. Kane, R. Bullock, and J. Hart, 2000: The turbulence algorithm intercomparison exercise: Statistical verification results. Preprints, *Ninth Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL., 466-471.

Cheng, B. and D. M. Titterington, 1994: Neural Networks: A review from a statistical perspective. *Statistical Science*, **9**, 2-54.

Cornman, L. B., C. S. Morse, and G. Cunning, 1995: Real-time estimation of atmospheric turbulence severity from in-situ aircraft measurements. *J. Aircraft*, **32,** 171-177.

Ellrod, G. P. and D. L. Knapp, 1992: An objective clear-air turbulence forecasting technique: verification and operational use. *Wea. Forecasting*, **7**, 150-165.

Kaplan, M. L., Y-L Lin, A. J. Riordan, K. M. Lux, and A. W. Huffman, 2000: Observational and numerical simulation-derived factors that characterize turbulence accident environments. Preprints, *Ninth Conf. on Aviation, Range and Aerospace Meteorology*, 476-481. American Meteorological Society, Orlando, 466-471.

Knox, J. A., 1997: Possible mechanisms of clear-air turbulence in strongly anticyclonic flows. *Mon. Wea. Rev.*, **125,** 1251-1259.

McCullagh, P., 1983: Generalized linear models. Chapman and Hall, New York, 261 pp.

Sharman, R. D., G. Wiener, and B. Brown, 2000: Description and verification of the NCAR Integrated Turbulence Forecasting Algorithm (ITFA). AIAA Paper 2000-0493.