7.1 FORECASTERS' EVALUATION OF THE INTEGRATED TURBULENCE FORECAST ALGORITHM (ITFA), WINTER 2003

Matthew Kelsch*
Cooperative Institute for Research in Environmental Sciences (CIRES)
University of Colorado/NOAA Research-Forecast Systems Laboratory
Boulder, Colorado

Chris Fischer
Cooperative Institute for Research in Environmental Sciences (CIRES)
University of Colorado/NOAA Research-Forecast Systems Laboratory
Boulder, Colorado

Jennifer L. Mahoney
NOAA Research-Forecast Systems Laboratory
Boulder, Colorado

1. INTRODUCTION

Aviation forecasters at the National Oceanic and Atmospheric Administration (NOAA) Aviation Weather Center (AWC) and United Airlines (UAL) participated in a subjective evaluation of automated turbulence forecasts during the period 23 January through 1 April 2003. The forecasts were generated by the Integrated Turbulence Forecast Algorithm (ITFA) know operationally as the Graphical Turbulence Guidance (GTG). ITFA was developed by the National Center for Atmospheric Research (NCAR) and uses Rapid Update Cycle (RUC) grids along with a suite of algorithm diagnostics and observations to produce turbulence forecasts (Sharman et al., 1999).

The subjective evaluations were gathered via an electronic questionnaire, similar to those used for evaluations in previous years (Mahoney et al., 2002). Corresponding objective verification scores are generated and displayed through the NOAA Forecast Systems Laboratory (FSL) Real-Time Verification System (RTVS) accessible at http://www-ad.fsl.noaa.gov/fvb/rtvs/.

Details on the ongoing objective verification of aviation impact variables are not presented here, but can be found at Brown et al. (2002). Section 2 provides a brief summary of the winter 2003 subjective evaluation with more details in Kelsch et al. (2003). A comparison between the objective verification scores and the subjective evaluations of ITFA is described in section 3. An overall summary is in Section 4.

2. THE SUBJECTIVE EVALUATION

Thirteen forecasters from AWC and UAL completed 76 questionnaires as part of the winter 2003 subjective evaluation of ITFA. The questionnaire focused on ITFA's performance with respect to turbulence area coverage, altitude coverage, intensity, and overall performance considering the area, altitude, intensity, and timing. The pie charts in Figs. 1-3 show the forecasters' assessment of ITFA with respect to area, altitude, and intensity of the forecasts. Figure 4 shows the forecasters' assessment of the overall performance, classified in four qualitative categories: excellent, good, fair, and poor.

Forecasters rated the overall performance of ITFA as *good* or *excellent* in a little over half the cases evaluated. In those cases most were rated as *about right* for area coverage, altitude coverage, and intensity.

In cases where the forecasters rated the overall performance of ITFA as fair or *poor*, there was a tendency for area coverage to be rated *too large* and the altitude coverage to be rated *too broad*. In these cases with less desirable overall performance the forecasters generally rated intensity as either *too severe* or *too light*, but without a strong tendency toward either category.

There was a tendency for regions east of the Rockies to receive proportionally higher *good* or *excellent* assessments than in the western areas. More details can be found in Kelsch et al. (2003).

*Corresponding author address: Matthew Kelsch, NOAA/FSL FS5, 325 Broadway, Boulder, CO 80305; email: kelsch@ucar.edu

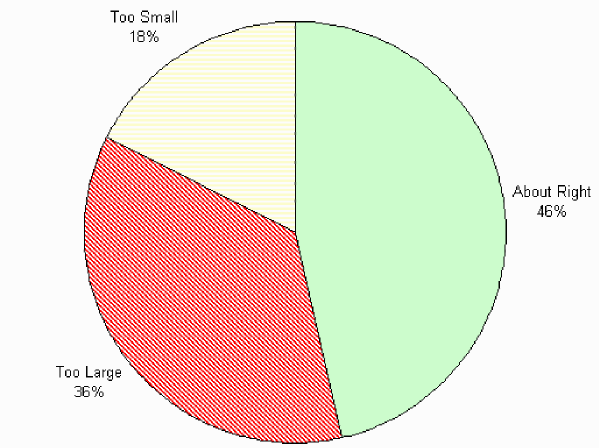

Figure 1. Forecasters' assessment of ITFA area coverage.

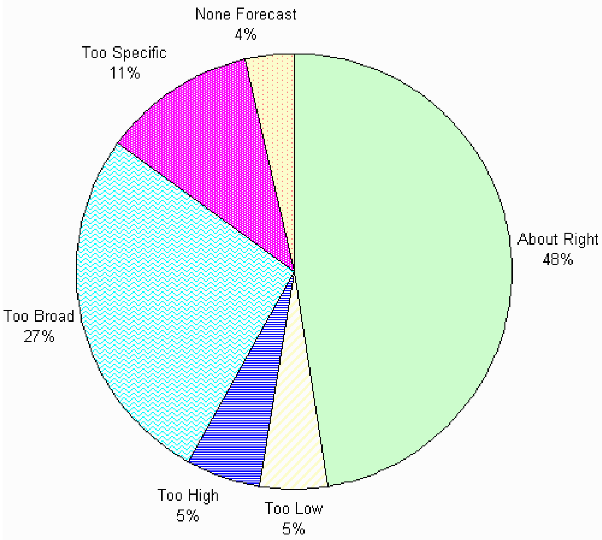

Figure 2. Forecasters' assessment of ITFA altitude coverage.

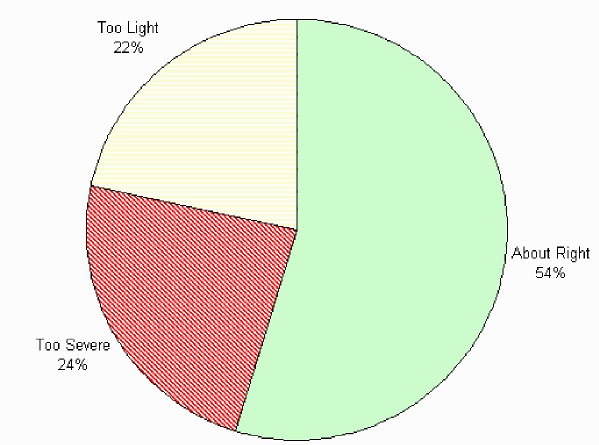

Figure 3. Forecasters' assessment of ITFA intensity forecasts.

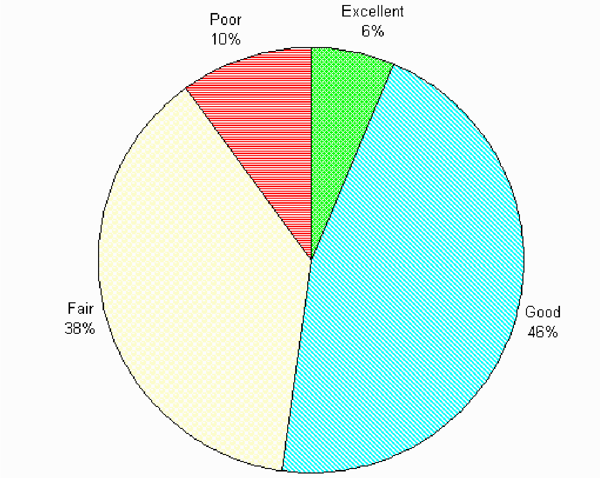

Figure 4. Forecasters' assessment of ITFA *Overall Performance* considering the area, altitude, intensity, and timing.

3. COMPARING SUBJECTIVE WITH OBJECTIVE

An important part of the winter 2003 ITFA evaluation was to compare the results of the forecasters' subjective evaluation with the objective verification scores available through RTVS. RTVS presents standard scores such as Probability of Detection of a "Yes" observation (PODy), Probability of Detection of a "No" observation (PODn), and the True Skill Statistic (TSS). These scores are derived from the 2X2 Yes-No contingency table and are described in more detail in Brown et al. (2002). PODy scores that consider only moderate or greater intensity reports are MOG PODy. In this dataset the PODn showed very little variability, but the MOG PODy varied quite a bit. Since the TSS varied in much the same way as the MOG PODy, this study presents just the MOG PODy and PODn scores for comparison with the subjective evaluations.

The forecasters' assessment of ITFA's overall performance (shown in Fig. 4) is compared with the objective scores. However, the objective scores are numerical and the subjective overall performance rating is a qualitative assessment with four categories: *excellent*, *good*, *fair*, and *poor*. Therefore, the subjective assessments were converted to a numerical scale similar to the POD scores of 0.00 to 1.00 with 1.00 representing the most favorable score. The conversion of the subjective overall performance was done as follows: *excellent*=1.00, *good*=0.67, *fair*=0.33, and *poor*=0.00. For all cases evaluated during the winter 2003 period, the average score for the overall performance of ITFA (per forecast region) is shown in Fig. 5. Note that central and eastern regions were generally rated better than western areas. The highest average score was 0.58 (Boston region) and the lowest was 0.39 in the northern Great Basin and northern Rockies (Salt Lake City North region).

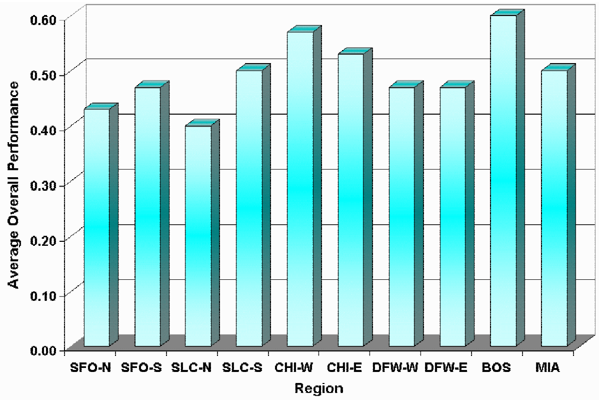

Figure 5. Average "overall performance" scores by region. The Y-axis is the average score (with 0.00=poor and 1.00=excellent). The X axis shows the regions which are from left to right: San Francisco North (SFO-N, San Francisco South (SFO-S), Salt Lake City North (SLC-N), Salt Lake City South (SLC-S), Chicago West (CHI-W), Chicago East (CHI-E), Dallas West (DFW-W), Dallas East (DFW-E), Boston (BOS), and Miami (MIA).

The questionnaires allowed forecasters to complete evaluations for up to 10 forecast regions. The number of forecast regions actually chosen varied from one to seven per questionnaire. Thus, in the comparisons, the objective scores were computed for the specified regions and valid times chosen in the questionnaires. The results are shown in Figs. 6 and 7, which compare the average overall performance (subjective score) with the objective MOG PODy (Fig. 6) and PODn (Fig. 7). Ideally, a diagonal "best fit" line would go from lower left to upper right if there is a good positive correlation between the subjective and objective evaluations. In Figs. 6 and 7, there is a positive correlation, but it is weak, as evidenced by the shallow slope from lower left to upper right. Furthermore, there is a large amount of scatter for all values on the MOG PODy plot (Fig. 6).

It is important to remember that the MOG PODy values depend greatly on the number of PIREPs. Since the number of PIREPs is small for many of the individual evaluations, it is not surprising that there are large variations in the MOG PODy values.

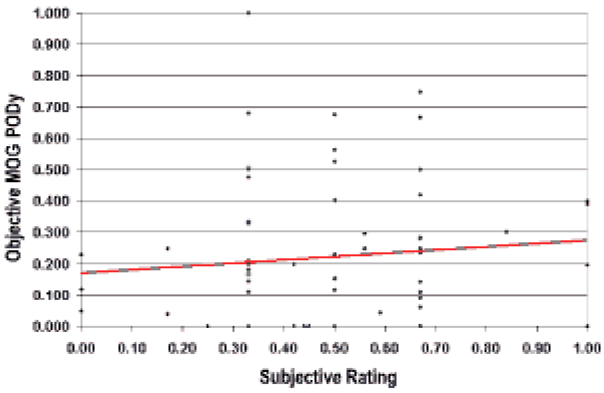

Figure 6. Objective MOG PODy scores versus forecasters' subjective overall assessment of ITFA for specified regions and times listed in the questionnaires. The subjective rating uses the numerical scale 0.00 (poor) to 1.00 (excellent). The bold line is the "best fit" line to the points.

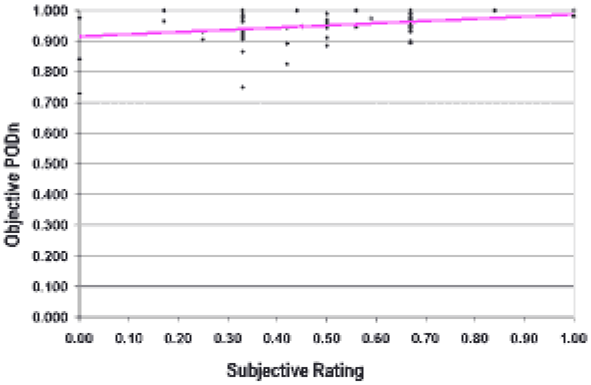

Figure 7. Same as Fig. 6, but with PODn.

A more detailed look at the individual cases suggests that the greatest discrepancies between the subjective and objective evaluations tend to occur on the less active days. This is likely to be due in part to the lack of PIREPs on such days.

Figures 8 and 9 show the MOG PODy and PODn versus the subjective evaluation for one-third of the more active turbulence days in the dataset. The "active days" were defined based on responses in the questionnaires. If the number of forecast regions chosen was at least the median for the whole dataset (3) and the number of *Yes* PIREPs per region was at least the median (8), then it was considered an active day.

Figure 8 shows a much stronger positive correlation between the objective and subjective numbers than did Fig. 6. One reason for this may simply be that the larger amount of observations leads to more robust and representative objective and subjective evaluations. Another reason is that on days with more organized areas of turbulence, the impact of isolated turbulence reports in smooth flying areas on the evaluation scores may be minimized.

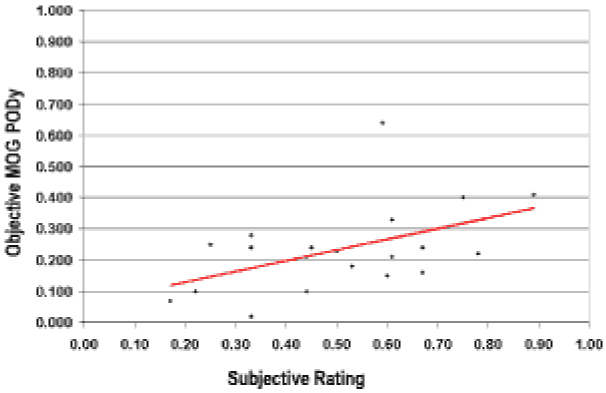

Figure 8. Same as Fig. 6, but for the active turbulence days as defined by the number of forecast regions chosen in the questionnaire (at least 3) and the number of *Yes* PIREPs per forecast region (at least 8).

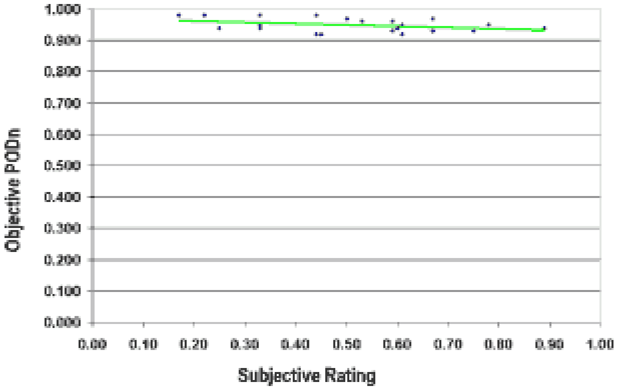

Figure 9. Same as Fig. 8, but for PODn.

4. SUMMARY

Subjective evaluations of ITFA were completed by forecasters at AWC and UAL through online questionnaires during 23 January–1 April 2003. The results provided a great deal of information regarding the important sources of turbulence and the performance of ITFA.

Overall, the results suggest that most of the turbulence as characterized by the forecasters was caused by the jet stream, and was forecast better in the east than in the west. ITFA forecasts were judged to capture the turbulence well about half the time. When ITFA did not perform well, forecasters suggested a general tendency for forecast areas to be too large and altitudes to be too broad or too specific rather than too high or too low. For the less favorable ITFA days, forecasters seemed split on whether the intensity was too severe or too light. Forecasts that were too severe did seem to coincide with the area being categorized as too large.

There was a positive correlation between the objective verification numbers and the forecasters assessment when the active days are considered. That is, if the forecasters said ITFA performed well, the verification numbers supported this. However, when all times and all regions are considered, the correlation between objective and subjective assessments was weaker. It appears that the low activity days can sometimes have objective numbers that suggest a very different performance than the forecasters indicate.

5. REFERENCES

Brown, B. G., J. L. Mahoney, R. Bullock, M. B. Chapmen, C. Fischer, T. L. Fowler, J. E. Hart, and J. K. Henderson, 2002: Integrated Turbulence Forecasting Algorithm (ITFA): Quality Assessment Report. Submitted to AWIT Technology Review Panel (available from B. Brown, bgb@rap.ucar.edu).

Kelsch, M., C. Fischer, and J. L. Mahoney, 2003: Forecaster assessment of upper level guidance from the Integrated Turbulence Forecast Algorithm (ITFA): a summary of results for the winter 2003 study. Submitted to the FAA Aviation Weather Research Program (available from Jennifer.Mahoney@noaa.gov).

Mahoney, J. L., T. Fowler, B. G. Brown, J. Braid, C. Fischer, M. Kay, and J. Wolff, 2002: Forecaster Assessment of Turbulence Algorithms: A Summary of Results for the Winter 2002 Study. Submitted to FAA Aviation Weather Research Program Leadership Team.

Sharman, R, C. Tebaldi, and B. Brown, 1999: An integrated approach to clear-air turbulence forecasting. *Preprints, Eighth Conference on Aviation, Range, and Aerospace Meteorology*, Dallas, TX, American Meteorological Society, 68-71.