SRH SSD 2006-06
12/2006

Technical Attachment

An Analysis of Aviation Verification Statistics for WFO Nashville

Mark A. Rose
WFO Nashville, Tennessee

1. Introduction

Aviation verification statistics were generated over a two-year period for two forecast sites within the county warning area of WFO Nashville, Tennessee. The statistics were used to measure forecaster performance versus the Global Forecast System (GFS) model guidance. While the summary statistics suggested little overall difference between the performance of the forecasters and the GFS, the relatively small gains achieved by the model guidance during the most frequent Visual Flight Rules (VFR) conditions masked the relatively large improvements achieved by the forecasters for the remaining flight categories. When flight categories other than VFR were considered, it became apparent the forecasters added substantial value to the GFS model guidance.

2. Methodology

Verification data for the Terminal Aerodrome Forecasts (TAF's) issued by WFO Nashville for the Nashville (KBNA) and Crossville (KCSV) airports (Fig. 1) were computed for the two-year period January 1, 2004 through December 31, 2005. The study included 5,431 individual TAF's. The AVN Verify program (Available online at http://www.srh.noaa.gov/srh/cwwd/msd/sram/support/avnverify.html) was used to generate the verification data.

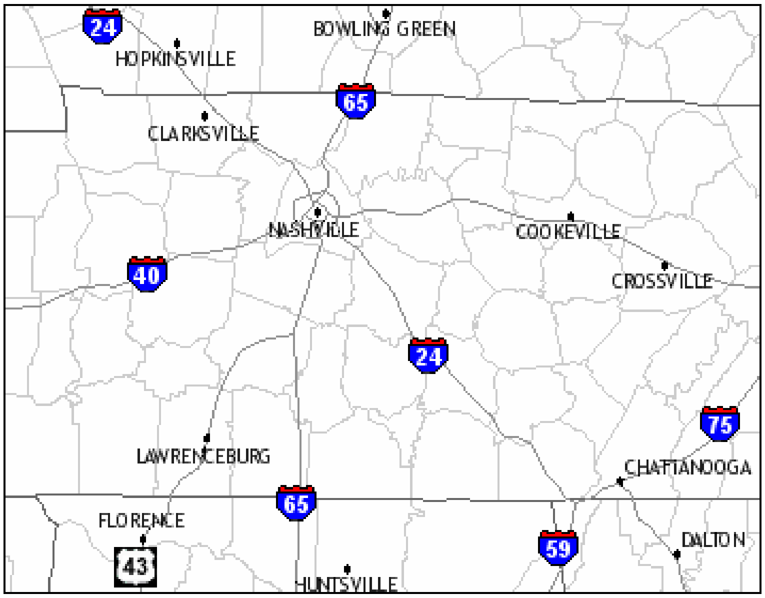

Figure 1. Map of central Tennessee and surrounding area showing the location of the two airports, Nashville (KBNA) and Crossville (KCSV), used in this study.

The generally accepted flight rule categories, as specified in National Weather Service Instruction 10-813, were used in this study (Table 1).

Table 1. Flight categories used in this study based on the ceiling and visibility criteria specified in National Weather Service Instruction 10-813.

Flight Category	Ceiling & Visibility
VFR (Visual Flight Rules)	greater than 3,000 feet and 5 statute miles
MVFR (Marginal Visual Flight Rules)	1,000 to 3,000 feet and/or 3 to 5 statute miles
IFR (Instrument Flight Rules)	500 to 900 feet and/or 1 to 2¾ statute miles
LIFR (Low Instrument Flight Rules)	200 to 400 feet and/or ½ to ¾ statute mile
VLIFR (Very Low Instrument Flight Rules)	0 to 100 feet and/or 0 to ¼ statute mile

3. Results

a. Summary statistics

The summary statistics (Table 2) showed little difference between the performance of the field forecasters and that of the model guidance. However, as will be shown, when the data were separated by flight rule category, in every category other than VFR, the forecasters substantially improved on the model guidance. Because the more critical categories occur much less frequently that the VFR category (Fig. 2), the value added by the field forecasters during those more important weather conditions is not obvious in the summary statistics.

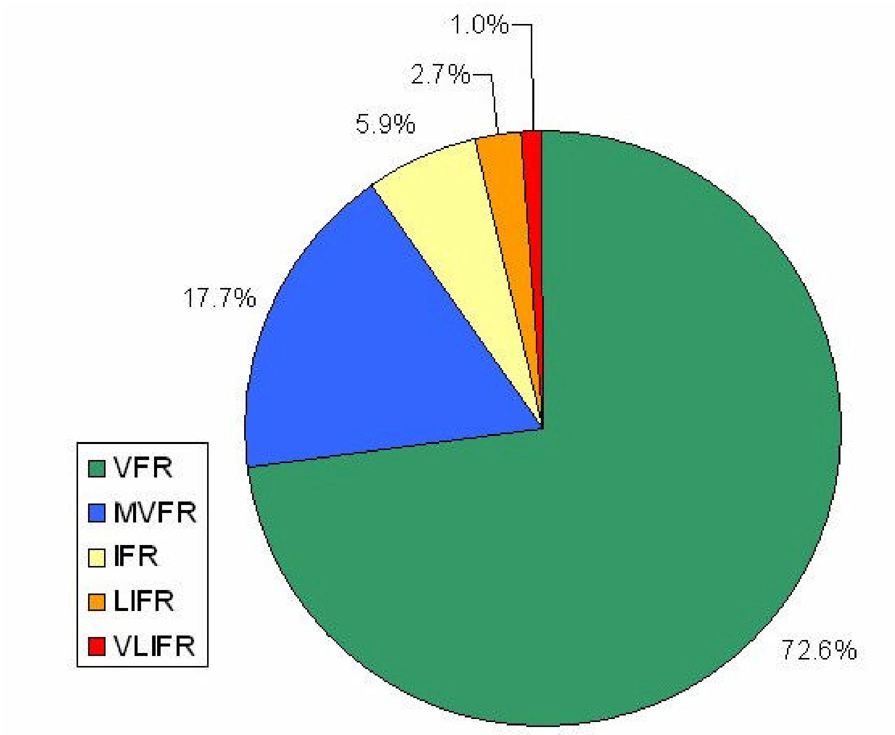

Figure 2. Frequency of occurrence of the various flight rule categories at the two airports used in this study during the two-year period January 1, 2004 through December 31, 2005.

Table 2. Verification of first- and second-period TAFs for Nashville and Crossville, Tenn. airports, January 1, 2004 – December 31, 2005. Comparisons are made separately for ceiling, visibility and appropriate flight category.

Para-meter	0-12 hours						12-24 hours					
	Percent of time TAF correct	Percent of time model correct	Percent of time TAF out-performs model	Percent of time model out-performs TAF	TAF busts	Model busts	Percent of time TAF correct	Percent of time model correct	Percent of time TAF out-performs model	Percent of time model out-performs TAF	TAF busts	Model busts
Ceiling	61	57	23	17	13	17	55	56	21	19	17	19
Visibility	77	81	10	12	6	6	73	79	10	14	7	6
Flight Category	73	71	17	14	5	7	68	69	16	15	6	9

b. Stratification by flight category and weather type.

When the data were stratified by flight category, meteorological element, and by the two TAF periods (Table 3), the skill of the field forecasters becames apparent. Except for the VFR category, the forecasters significantly improved upon the model guidance in every other category.

The AVN Verify program can also be used to compute performance statistics according to the type of weather. In this study, statistics for four of the five weather types calculated by AVN Verify were considered. (Freezing precipitation was not included due to its rarity – only a cumulative sixteen hours of observed freezing precipitation in two years at both TAF sites combined.) The weather types analyzed in this study were thunderstorms, rain, fog (which includes light fog, dense fog, and haze), and snow.

Table 3. Comparison between first- and second-period TAFs and model guidance for Nashville and Crossville, Tenn. airports, January 1, 2004 – December 31, 2005. Comparisons are made separately for each flight category and various weather types.

Category	0-12 hours		12-24 hours	
	TAF better	Model better	TAF better	Model better
VFR	6	14	7	16
MVFR	44	12	41	11
IFR	50	15	45	17
LIFR	38	14	26	19
VLIFR	46	27	55	15
Thunderstorms	36	44	24	56
Rain	40	16	32	21
Fog	33	24	32	24
Snow	28	36	16	31

The differences between the forecasters and model guidance were someshat mixed when weather types were considered. The model guidance consistently outperformed the forecasters in predicting thunderstorms, especially in the second twelve hours of the TAF. The data reflect the forecasters' tendency to over-forecast convection (as will be discussed below). The model guidance also showed slightly better skill over the forecasters in predicting snow, again most notably in the second twelve hours of the TAF. With respect to both rain and fog, however, the forecasters were several percentage points better than the model guidance.

c. Probability of Detection and False Alarm Ratios

Probability-of-Detection (POD) statistics show much the same result. The model guidance showed slight improvement over the forecasters during VFR weather in both the 0-12 hour and 12-24 hour TAF periods, while the forecasters showed improvement over model guidance in every other flight category for both periods, except during LIFR conditions in the 12-24 hour period (Table 4). In many cases, the forecaster POD was 10-20% greater than the model POD.

Table 4. Comparison of first- and second-period Probability of Detection data for Nashville and Crossville, Tenn. airports, January 1, 2004 – December 31, 2005. Comparisons were made separately for each flight category and various weather types.

Category	0-12 hours		12-24 hours	
	TAF	Model	TAF	Model
VFR	81	87	77	86
MVFR	57	31	50	27
IFR	39	20	30	18
LIFR	36	28	21	27
VLIFR	22	12	10	7
Thunderstorms	39	4	30	3
Rain	71	24	66	22
Fog	57	20	49	19
Snow	48	1	29	1

As might be expected, the False Alarm Ratios (FAR) mirror the PODs, except the forecasters improved upon the model guidance for all periods and in all categories, including VFR, except for MVFR conditions (Table 5).

Likewise, both the forecasters and model guidance showed increasing FAR as flight categories decrease. In fact, only during VFR conditions were FARs below 50%. During IFR, LIFR, and VLIFR conditions, the model guidance FARs were close to 80%, while the forecaster's FARs were generally in the 60-70% range.

Both the forecasters and model PODs generally decreased with deteriorating flight categories. With respect to VFR conditions, forecaster and model PODs were on the order of 80%, but dropped considerably for the more restrictive flight categories. Forecasters substantially improved upon the model guidance, as measured by the POD and FAR, in the MVFR, IFR, and VLIFR flight categories.

Table 5. Comparison of first- and second-period False Alarm Ratios for Nashville and Crossville, Tenn. airports, January 1, 2004 – December 31, 2005. Comparisons are made separately for each flight category and various weather types.

Category	0-12 hours		12-24 hours	
	TAF	Model	TAF	Model
VFR	8	18	11	18
MVFR	59	60	66	64
IFR	70	77	77	79
LIFR	59	77	69	83
VLIFR	63	79	72	89
Thunderstorms	93	98	95	99
Rain	62	86	69	87
Fog	53	69	59	73
Snow	58	87	70	93

Ironically, the largest forecaster FARs occurred during IFR conditions, with the FARs for the LIFR and VLIFR flight categories less than those for the IFR category. The opposite was true for the model guidance, which had largers FARs for the more restrictive categories. Again, forecaster PODs were greater than those for model guidance, except for LIFR conditions in the 12-24 hour period of the TAF. *That forecasters combine consistently higher PODs with consistently lower FARs in weather conditions less than MVFR indicates the forecasters dominance during meteorological conditions that demand the greatest precision*.

Both PODs and FARs were computed for forecast weather types. With respect to the PODs, the forecasters handily improved upon the model guidance for all four weather elements across all TAF periods. Rain was the most accurately forecast weather element by both the forecasters and model guidance, followed by fog. The first period TAF PODs for snow were slightly higher than those for thunderstorms, but the reverse was true for model guidance. However, across-the-board, model guidance PODs for both thunderstorms and snow were less than 10%.

The forecasters also showed consistent improvement over model guidance in FARs. However, the margin of improvement was not as pronounced as with PODs. Notably, neither the forecasters nor model guidance had FARs less than 50% for any weather element during any period of the TAF, with the FARs for thunderstorms for both the forecaster and model guidance across the entire TAF period exceeded 90%.

Indeed, the data shown in Tables 3-5 underscore the importance of analyzing aviation verification statistics categorically in order to gain the most complete picture of the forecasters' ability to add value to the model guidance.

d. Critical Success Index

The difficulty in forecasting thunderstorms can best be illustrated by computing the Critical Success Index (CSI) for such forecasts. The CSI is calculated by dividing the number of correct forecasts by the sum of the correct forecasts, missed events, and false alarms (Schaefer 1990). With respect to thunderstorms in the 0-12 hour TAF period, there were 444 hours during which this weather element was carried by METAR observations. (This number is twice the total hours during which thunderstorms were carried by METAR observations. Since TAF's are issued every six hours, each hour of observed weather is counted twice by AVN Verify in each 12 hour period. For example, a thunderstorm that occurred during the ninth hour of one TAF would have also occurred during the third hour of the following TAF.)

Since the forecaster POD for thunderstorms during the 0-12 hour period of the TAF was 39%, the hours of correct forecast were 173 (39% of the 444 hours of observed thunderstorms), with the number of hours of missed events 271. The AVN Verify program showed that false alarm hours totaled 2,141. Thus, the forecaster's CSI for thunderstorm forecasting in the first twelve hours of the TAF was 0.07.

The model guidance POD was 4%, meaning the hours of correct forecasts were 18 (again, 4% of the 444 hours of observed thunderstorms). The hours of missed events were therefore 426. The AVN Verify program showed that false alarm hours totaled 1,090. Thus, the model guidance's CSI for thunderstorm forecasting in the first twelve hours of the TAF was 0.01.

But these data suggest a paradox. The CSIs for both thunderstorms and snow indicate the model guidance outperforms the TAF more often than the TAF outperforms the model guidance, even though the TAF PODs and FARs were better than those of the model guidance.

To use thunderstorms during the first twelve hours of the TAF as an example, the answer to this paradox lies in the high number of false alarm hours, which were much greater for the forecasters (2,141) than for the model guidance (1,090). Each hour of false alarm issued by the forecaster coincident with a null forecast by model guidance gives the model guidance one hour during which it outperforms the TAF. Thus, even though the forecaster had 173 hours of correct forecasts compared to just 18 hours for model guidance, this discrepancy (155 hours) was unable to overcome the difference (1,051) in false alarm hours between the forecaster and model guidance.

This discrepancy becomes even wider for thunderstorms during the second twelve hours of the TAF. Here, the forecasters made 132 hours of correct forecasts, compared to 13 by model guidance. But the forecasters also issued 2,700 false alarm hours, compared to just 914 by model guidance. Thus, even though the forecasters outperformed the model guidance in POD (30% to 3%) and FAR (95% v. 99%), the model guidance still outperformed the forecasters 56% of the time, whereas the forecaster improved upon the model guidance just 24% of the time.

e. Use of TEMPO and PROB30 groups.

Finally, AVN Verify also calculates forecaster performance in TEMPO and PROB30 groups. The TEMPO group verification statistics with respect to ceiling, visibility, and weather element are shown in Figure 3. Use of TEMPO implies a probability of expectation of greater than 50% (NWSI 10-813), meaning that, ideally, TEMPO groups should verify more than 50% of the time. Unfortunately, this is not the case. The output statistics show that TEMPO groups hurt the TAF more often than they improved it, indicating a routine over-use of TEMPO groups.

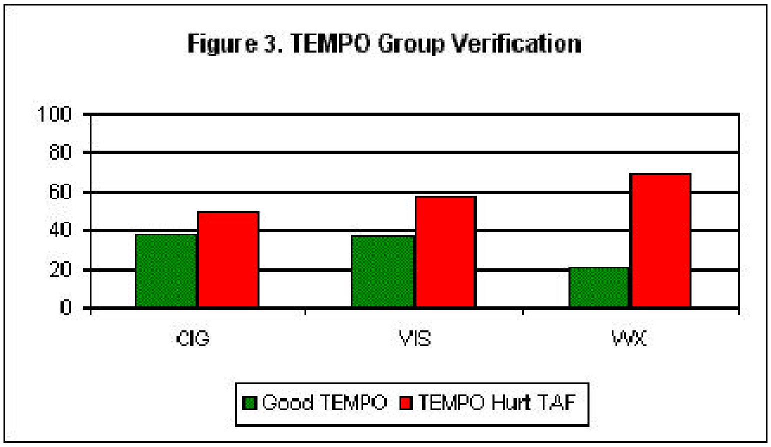

Figure 3. TAF TEMPO group verification. Frequency of occurrence (in percent) of the times the use of the TEMPO group improved the TAF vs the times the use hurt the TAF. Data have been stratified by ceiling (CIG), visibility (VIS) and weather type (WX), and are for the two airports used in this study during the two-year period January 1, 2004 through December 31, 2005

Likewise, the use of PROB30 groups within the TAF implies a 30% probability of occurrence. Once again, with five forecast elements sampled (thunderstorms, rain, snow, ceiling, and visibility), the verification scores are consistently well below the desired 30% threshold (Fig. 4). Thus, like TEMPO groups, AVN Verify shows a chronic over-use of the PROB30 groups as they are defined in NWSI 10-813.

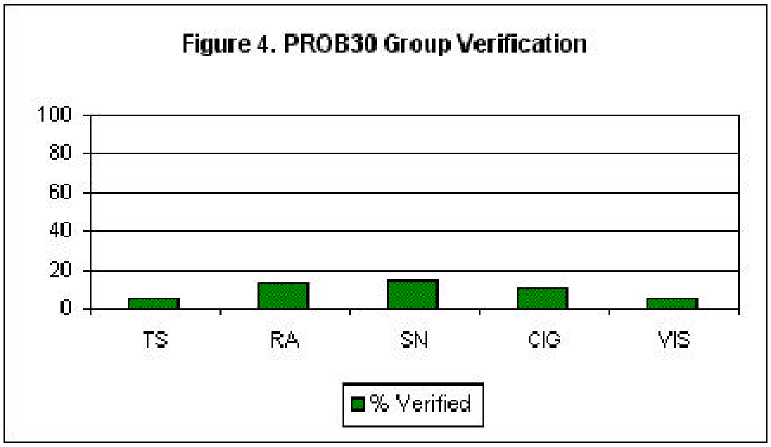

Figure 4. TAF PROB30 group verification. Percentage of time the various weather type, low ceiling, or obstruction to visibility occurred when the corresponding PROB30 group was included in the TAF. The abbreviations represent thunderstorm (TS), rain (RA), snow (SN), ceiling (CIG) and visibility (VIS) and are for the two airports used in this study during the two-year period January 1, 2004 throughDecember 31, 2005**.**

4. Concluding Remarks

It has been shown that general TAF verification statistics produced by the AVN Verify program, which show little, if any, improvement by the forecaster over model guidance, miss some crucial realities of aviation forecasting, at least for WFO Nashville.

For a large majority of meteorological scenarios, the forecasters added value to the model guidance. Although the model guidance outperformed the forecasters in the VFR category, in every other flight category (MVFR, IFR, LIFR, and VLIFR), the forecasters improved upon the model guidance, often by significant margins. Probabilities-of-Detection showed the forecasters showed improvement over model guidance in every flight category for both periods, except during LIFR conditions in the 12-24 hour period and VFR conditions in both periods. In many cases, the forecasters' PODs were 10-20% greater than the model's. The results based on False Alarm Ratios were similar.

Almost 73% of the observed weather was within the VFR category. When the summary statictics were computed, the relatively small gains achieved by the model guidance during the most frequent VFR conditions masked the relatively large improvements achieved by the forecasters for the remaining flight categories.

Performance between the forecasters and model guidance was somewhat more mixed when weather types are considered. The model guidance consistently outperformed the forecasters in predicting thunderstorms, especially in the second twelve hours of the TAF. These statistics reflect the forecasters' tendency to over-forecast convection. The model guidance also showed slight improvement over the forecasters in predicting snow, again most notably in the second twelve hours of the TAF. With respect to both rain and fog, however, the forecasters showed a substantial margin of improvement over model guidance.

With respect to the PODs, the forecasters handily improved upon the model guidance for all four weather elements across all TAF periods. Across-the-board model guidance PODs for both thunderstorms and snow were in the single digits. The forecasters also showed consistent improvement over model guidance as measured by the FARs. However, the margin of improvement wass not as pronounced as with PODs.

The CSIs revealed a paradox, since the data showed -- with respect to both thunderstorms and snow -- the model guidance outperformed the TAF more often than the TAF outperformed the model guidance. This was in contrst to the PODs and FARs that indicated the forecasters improved upon the model guidance. The answer to this paradox lies in the number of false alarm hours, which the forecasts issued far more often than the model guidance.

Analysis also showed TEMPO groups hurt the TAF more often than they improved it, indicating a routine over-use of TEMPO groups. The verification scores for PROB30 groups were consistently well below the 30% threshold prescribed by NWSI 10-813, indicating a similar over-use of the PROB30 groups,.

Acknowledgements

The author thanks Henry Steigerwaldt (Science and Operations Officer) and Jerry Orchanian (Warning Coordination Meteorologist), both of WFO Nashville, for their reviews, comments, and suggestions regarding this manuscript.

REFERENCE

Schaefer, J. T, 1990: The Critical Success Index as an indicator of warning skill. *Wea. and Forecasting*, 5, 570-575.