9.5 VERIFICATION OF THE IN-FLIGHT ICING DIAGNOSTIC ALGORITHM (IIDA)

Barbara G. Brown[1], Jennifer L. Mahoney[2], and Tressa L. Fowler[3]

1. INTRODUCTION

The In-flight Icing Diagnostic Algorithm (IIDA) was recently approved to become an operational tool for use in diagnosing the existence of icing conditions aloft. Objective verification was an important component of the development of IIDA. In addition, an in-depth quality assessment was a critical aspect in the decision process to transition IIDA from an experimental to an operational product. Some results of this quality assessment are presented and discussed in this paper.

2. BACKGROUND

IIDA is an automated system to diagnose locations of icing conditions aloft. This system was developed by the In-Flight Icing Product Development Team (IFIPDT) of the FAA's Aviation Weather Research Program (AWRP; Sankey et al. 1997). Every hour, IIDA generates diagnoses of icing conditions. These diagnoses are based on an intelligent combination of observations (satellite, surface, radar) with 3-h temperature and relative humidity forecasts from the Rapid Update Cycle (RUC) numerical weather prediction system (Benjamin et al. 1999). The concepts underlying the development of IIDA are described in McDonough and Bernstein (1999). IIDA produces both a "general icing" and a "supercooled large droplet" (SLD) icing field, indicating icing potential on a scale from 0 to 1. This paper will concentrate on results for the general icing field; results for the SLD diagnoses will be described only briefly here, but are considered in greater detail in Brown et al. (2001).

IIDA has been extensively evaluated over the last several years and the quality of IIDA diagnoses has been compared to the quality of forecasts and diagnoses produced by a number of other icing algorithms as well as the operational forecasts (AIRMETs) issued by the Aviation Weather Center (AWC; Brown et al. 1999). In addition, the algorithm has been evaluated in near-real-time since April 1998 by the Real-Time Verification System (RTVS) at NOAA's Forecast Systems Laboratory (Mahoney et al. 2002), along with two other automated in-flight icing algorithms and the AIRMETs (results are available at *http://www-ad.fsl.noaa.gov/afra/rtvs/RTVS-project_des.html*). This paper focuses on recent evaluations of IIDA diagnoses for winter 2000.

3. APPROACH

As in previous icing verification studies, the verification data are Yes and No pilot reports (PIREPs) of icing conditions. The basic verification approach is described in Brown et al. (1997) and extended in Brown et al. (1999). Thus, the primary verification statistics are Probability of Detection (POD), Impacted Area, and Impacted Volume. POD is computed for both Yes and No PIREPs, with the resulting statistics denoted PODy and PODn, respectively. Due to certain characteristics of PIREPs, it is inappropriate to compute the False Alarm Ratio (FAR) and a number of other common verification scores (e.g., the critical success index; Brown and Young 2000).

PODy can be interpreted as the proportion of Yes PIREPs that are correctly diagnosed to be in regions with icing; PODn is the proportion of No icing PIREPs that are correctly diagnosed to be in regions with no icing. A subset of Yes PIREPs with moderate-or-greater (MOG) reported icing severity is considered in these analyses, since MOG icing is of most concern operationally and is the focus of the operational outlooks (i.e., AIRMETs; NWS 1991). PODn is computed separately using explicit no-icing PIREPs ("clear-above" PIREPs, from which no-icing conditions can be inferred, are not considered here). Impacted Area and Volume generally are reported as percentages of the total area/volume possible that is impacted by a Yes diagnosis/forecast (i.e., as % Area and % Volume, respectively).

Because IIDA values can cover a continuous range between 0 and 1, the verification analyses are based on applying several different thresholds to IIDA to create Yes/No icing diagnoses. That is, a Yes diagnosis is inferred at a grid point if the IIDA value equals or exceeds the threshold; a No diagnosis is inferred if the value at a grid point is less than the threshold. The relationship between the values of PODy and % Volume for different thresholds is of interest because it measures the trade-off between increased PODy values and increased coverage by the forecast/diagnosis. The relationship between PODy and 1-PODn for different algorithm thresholds also is of interest and is the basis for the verification approach known as "Signal Detection Theory" (SDT; e.g., Mason 1982). The curve joining the (1-PODn, PODy) points for different thresholds is known as the "Relative Operating Characteristic" (ROC) curve; the area under this curve is a measure of skill (with a value of 0.5 indicating no skill). Ideally, the ROC curve will lie above the diagonal no-skill line, toward the upper left corner of the diagram.

In the verification analyses, PIREPs are either matched or interpolated to the four closest grid points at a particular model level. The NCAR verification system uses a four-grid-point matching procedure, in which the most extreme forecast value at the surrounding grid points is matched to a PIREP. RTVS uses an interpolation approach to estimate the algorithm value at a PIREP location. Previous comparisons of these approaches have indicated that the verification results are relatively insensitive to this difference in approach.

The current version of IIDA incorporates information from PIREPs in the hour prior to the forecast valid

[1]*Corresponding author address*: Barbara G. Brown, National Center for Atmospheric Research (NCAR), PO Box 3000, Boulder CO 80307-3000; e-mail: bgb@rap.ucar.edu

[2]Forecast Systems Laboratory, NOAA/DOC, Boulder CO

[3]NCAR, Boulder, CO

time. Thus, the verification analyses only use PIREPs in a time window of one hour following the forecast valid time. The NCAR evaluations are limited to every three hours from 1200-0300 UTC (when the most PIREPs are available), whereas the RTVS results include all IIDA valid times (i.e., every hour).

In some cases, verification results for IIDA are compared to verification results for the AIRMETs, because AIRMETs are the current operational standard for icing information. AIRMETs have a somewhat different form from IIDA. They are produced every six hours and are valid for up to six hours, although they also can be amended or canceled during that period. The forecasts are in a textual form and are decoded into latitude and longitude vertices, with tops and bottoms of the icing regions defined in terms of altitude. For comparison purposes, AIRMETs are evaluated over the same time window as IIDA. In essence, we are assuming that a user has available two types of information about icing at a particular time, AIRMETs and IIDA. The question of interest is: What are the differences between the forecasts/ diagnoses for that time? Unfortunately, this approach may not represent the entire set of information available from an AIRMET (much of which cannot be digitally decoded because it is in a non-standard format). However, it is meaningful from the perspective of a user who is considering the two products simultaneously. Moreover, the comparison is very important to compare IIDA to the operational standard for in-flight icing, which is the icing AIRMET.

4. RESULTS

Results of recent evaluations of IIDA and the AIRMETs are described in this section. Overall results and variations with time and threshold are presented first, followed by some RTVS results showing variations with altitude and time. Only results for winter 2000 are presented; results for winter 2001 are quite similar and are included in Brown et al. (2001).

Overall verification results for IIDA and the AIRMETs for winter 2000 are shown in Figure 1. This figure shows the relationship between PODy (MOG PIREPs) and 1-PODn, PODy and % Area, and PODy and % Volume for various IIDA thresholds. Note that the AIRMETs are represented by a single point in each graph, since they are by nature a Yes/No forecast. The ROC curve (Fig. 1a) indicates that IIDA is skilful in discriminating between Yes and No PIREPs; in fact, the area under this curve is 0.76, much larger than the no-skill value of 0.50.

The PODy vs. % Area curve (Fig. 1b) suggests that the AIRMETs are more successful than IIDA at limiting the areal coverage of the forecasts. This result is not surprising since positive IIDA diagnoses at any level contribute to the areal coverage (even if they only occur at a single level). With respect to volume coverage, IIDA performs somewhat better than the AIRMETs (Fig. 1c), since the AIRMET point is located below the IIDA curve. This result also is not surprising, since the AIRMETs are restricted to a “cake”-like shape (with a solid bottom and top across the entire AIRMET region), whereas the IIDA values are free to vary across the domain, so some regions have narrower vertical coverage than others.

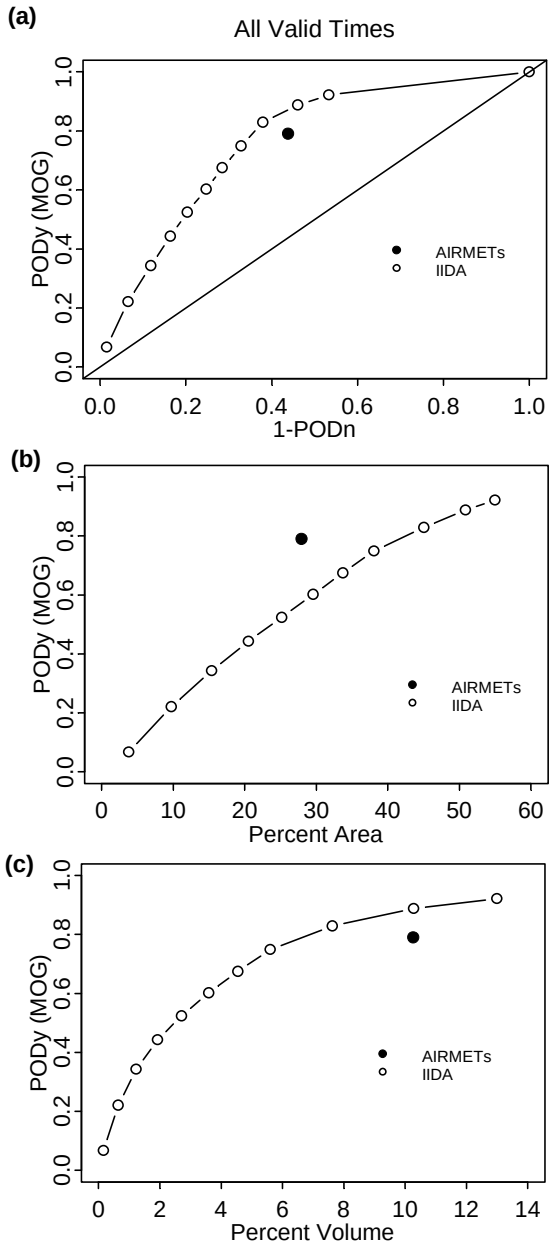

FIGURE 1. Relationship between PODy (MOG PIREPs) and (a) 1-PODn, (b) % Area, and (c) % Volume, for IIDA and AIRMETs, winter 2000, all valid times combined. Each point on the IIDA curves represents a different threshold used to define Yes/No icing forecasts. The thresholds used (starting in the upper right corner) are 0.0, 0.05, 0.15, 0.25, 0.35, ..., 0.95.

The results in Figure 1 are based on a combination of results for all valid times that were included in the NCAR analysis (i.e., every third hour between 1200 and 0300 UTC). Though not shown here, results for the individual valid times are consistent with those shown in Figure 1, except for a slight tendency for PODy values to be larger in earlier daytime periods (1200 - 2100 UTC). Overall, these results indicate little trend in the verification statistics with valid time.

It is of interest to know how much variability is associated with the verification statistics. One approach to obtaining this information is through confidence intervals, as shown in Table 1. This table shows 95% confidence intervals for PODy and PODn for two IIDA thresholds and for the AIRMETs, based on approaches that are appropriate for these types of data (Kane and Brown 2000). Results in Table 1 indicate the confidence interval for PODy (PODn) has a width of 0.04 - 0.08 (0.03 - 0.06), with the widest intervals associated with the AIRMETs (i.e., indicating greater underlying variability in the estimates for AIRMETs).

TABLE 1. 95% confidence intervals for PODy (MOG) and PODn for two IIDA thresholds and the AIRMETs, for verification statistics computed for winter 2000.

Forecast	Statistic	
	PODy(MOG)	**PODn**
IIDA (0.15)	0.81, 0.85	0.60, 0.64
IIDA (0.25)	0.72, 0.77	0.65, 0.69
AIRMETs	0.75, 0.83	0.53, 0.59

The values in Table 1 can also be used to evaluate significant differences among the statistics. Because their confidence intervals do not overlap, the results in Table 1 indicate that the PODy and PODn values for IIDA (0.15) and IIDA (0.25) are significantly different from each other. In addition, the PODn value for the AIRMETs is significantly smaller than the PODn values associated with both IIDA thresholds. However, the PODy value for the AIRMETs is not significantly different from either of the other PODy values, due to the width of the interval for the AIRMET PODy.

The variability in the statistics also can be examined through depictions of the distributions, as in Figure 2. This figure shows box plots of the distributions of PODy(MOG) and % Volume for the AIRMETs and for IIDA with thresholds of 0.15 and 0.25. For PODy(MOG), the AIRMETs and IIDA-0.15 statistics have similar distributions, whereas the distribution for the IIDA-0.25 values is somewhat below the other two distributions. The PODy(MOG) values for IIDA-0.25 also appear to be somewhat more variable, as measured by the sizes of the boxes and the distances between the ends of the upper and lower whiskers. However, this result may be related to the fact that PODy is bounded at 1.0 and the AIRMET and IIDA-0.15 PODy values are more frequently close to that upper bound.

The % Volume distribution for AIRMETs is higher than the corresponding distributions for IIDA (Fig. 2b). One notable feature of Fig. 2b is the narrow range of % Volume values associated with IIDA - the distributions of % Volume are very tight. This result is consistent for all IIDA thresholds (not shown here). Thus, although the detection rates associated with IIDA diagnoses are fairly variable from time to time, the extent of the regions covered is quite consistent from time to time.

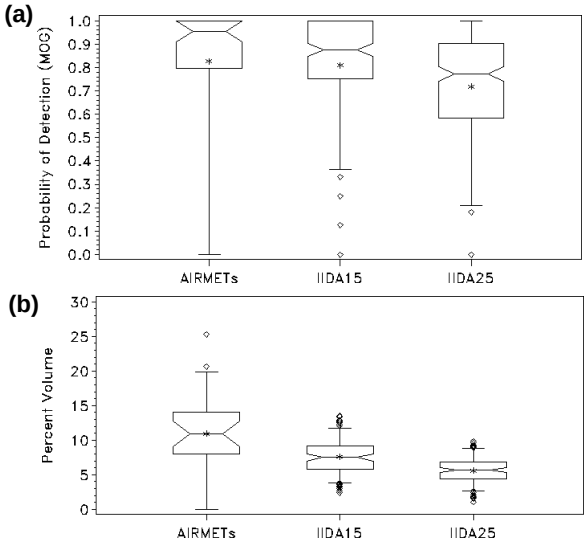

FIGURE 2. Box plots showing distributions of individual values of verification statistics for AIRMETS and IIDA (two thresholds - 0.15 and 0.25): (a) PODy(MOG); (b) % Volume. Line inside each box represents median value; bottom and top of boxes are 0.25th and 0.75th quantile values, respectively; ends of bottom and top whiskers are 0.05th and 0.95th quantile values; and points extending below and above whiskers are in lower and upper 5% of distribution.

Fig. 3 shows weekly time series plots of PODy and PODn for IIDA (threshold of 0.15) and the AIRMETs, obtained from the RTVS. These plots show how the statistics vary from week-to-week through the winter period, and they demonstrate that the variations are relatively consistent for IIDA and the AIRMETs. The PODn time series also seems to have a slight increasing trend, with somewhat larger values in the later part of the winter.

Variations of the verification statistics with altitude can be examined through the height series plots available on RTVS. Fig. 4 shows variations in PODy and PODn with altitude for IIDA (threshold of 0.15) and the AIRMETs. These plots indicate that the the PODy values generally improve with altitude through lower and middle layers, and then quickly decrease at upper levels. The IIDA PODy values are largest at lower levels, whereas the AIRMET PODy values are largest at middle levels. The IIDA PODn values are relatively large at most levels.

5. SUMMARY AND CONCLUSIONS

This paper has briefly summarized some of the verification analyses that were undertaken for IIDA as part of its development and in preparation for its transition to an operational product. These statistics indicate that IIDA provides skilful diagnoses of icing conditions. The algorithm is able to successfully discriminate between Yes and No icing conditions, and trade-offs between the improved detection rates and the extent of the forecast region are smaller than for the AIRMETs. IIDA performs best at lower altitudes, with performance dropping off at higher levels. Some results that were not shown here, but that are documented in Brown et al. (2001) include the following: (a) the SLD algorithm is

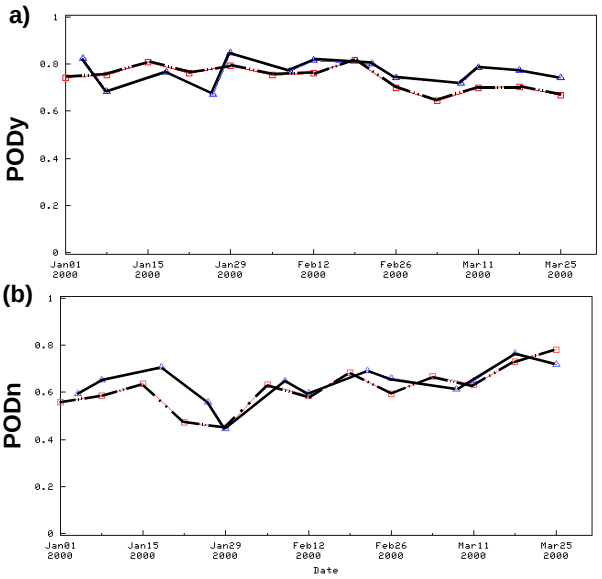

FIGURE 3. Time series plots of weekly PODy(MOG) for IIDA with a threshold of 0.15 (solid) and icing AIRMETs (dashed), for January 1 - 31 March 2000: (a) PODy(MOG) and (b) PODn.

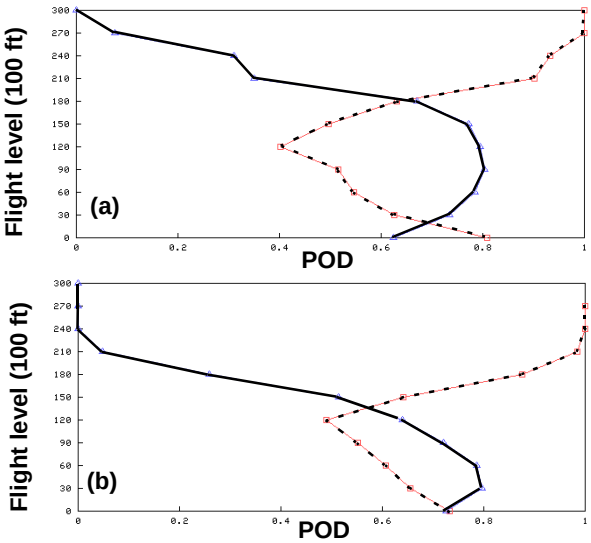

FIGURE 4. Height series plots from RTVS for (a) AIRMETS and (b) IIDA (threshold = 0.15), showing variation of PODy(MOG) (solid) and PODn (dashed) with height.

quite efficient at capturing conditions associated with severe icing reports (i.e., the algorithm captures a relatively large number of severe reports while covering very small regions); and (b) the IIDA diagnoses perform quite well as persistence forecasts, out to at least three hours.

One area of future work that is related to the verification results is calibration of the IIDA "icing potential" values. While these values clearly seem to indicate increasing likelihood of icing with increases in icing potential, the values have not been calibrated to indicate true probabilities. This calibration process would be relatively straightforward if the PIREPs were collected systematically. However, since they are not, the process is much more difficult, and is a subject for further research.

Additional analyses will also include an evaluation of the potential economic value of the icing diagnoses to users, based on the relationship between economic value and attributes of the ROC diagram. These analyses can be accomplished through use of the prototypical cost-loss ratio decision-making model (e.g., Richardson 2000). Although measuring forecast quality is not equivalent to measuring forecast value (Murphy 1993), this approach can provide a meaningful link between the two.

ACKNOWLEDGMENTS

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the author and do not necessarily represent the official policy or position of the U.S. Government. We thank the NCAR and FSL verification teams for their support for this study.

REFERENCES

Benjamin, S.J., J.M. Brown, K.J. Brundage, D. Kim, B. Schwartz, T. Smirnova, and T.L. Smith, 1999: Aviation forecasts from the RIC-2. *Preprints, 8th Conference on Aviation, Range, and Aerospace Meteorology,* 10-15 January, Dallas, TX, American Meteorological Society (Boston), 486-490.

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

Brown, B.G., T.L. Kane, R. Bullock, and M.K. Politovich, 1999: Evidence of improvements in the quality of in-flight icing algorithms. *Preprints, 8th Conference on Aviation, Range, and Aerospace Meteorology*, 10-15 January, Dallas, TX, American Meteorological Society (Boston), 48-52.

Brown, B.G., and G.S. Young, 2000: Verification of icing and turbulence forecasts: Why some verification statistics can't be computed using PIREPs. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology,* 11-15 September, Orlando, FL, American Meteorological Society (Boston), 393-398.

Brown, B.G., J.L. Mahoney, R. Bullock, T.L. Fowler, J. Henderson, and A. Loughe, 2001: Quality assessment report: Integrated Icing Diagnostic Algorithm. Report to the Aviation Weather Research Program, FAA/DOT (available from the corresponding author), 36 pp.

Kane, T.L., and B.G. Brown 2000: Confidence intervals for some verification measures - a survey of several methods. *Preprints, 15th Conference on Probability and Statistics in the Atmospheric Sciences,* 8-11 May, Asheville, NC, American Meteorological Society (Boston), 46-49.

Mahoney, J.L., B.G. Brown, J.K. Henderson, J.E. Hart, A. Loughe, C. Fischer, and B. Sigren, 2002: The Real Time Verification System and its application to aviation weather forecasts. *Preprints, 10th Conference on Aviation, Range, and Aerospace Meteorology,* 13-16 May, Portland, OR, American Meteorological Society (Boston), in press.

Mason, I.B., 1982: A model for assessment of weather forecasts. *Australian Meteorological Magazine*, **30**, 291-303.

McDonough, F., and B.C. Bernstein, 1999: Combining satellite, radar, and surface observations with model data to create a better icing diagnosis. *Preprints, 8th Conference on Aviation, Range, and Aerospace Meteorology,* 10-15 January, Dallas, TX, American Meteorological Society (Boston), 467-471.

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281-293.

NWS, 1991: National Weather Service Operations Manual, D-22. National Weather Service, NOAA (Available at http://tgsv5.nws.noaa.gov/oso/oso1/oso12/wsom/wsomd25.html).

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *QJRMS*, **126**, 649-667.

Sankey, D., K.M. Leonard, W. Fellner, D.J. Pace, and K.L. Van Sickle, 1997: Strategy and direction of the Federal Aviation Administration's Aviation Weather Research Program. *Preprints, 7th Conference on Aviation, Range, and Aerospace Meteorology,* 2-7 Feb, Long Beach, CA, American Meteorological Society (Boston), 7-10.