

Quality Assessment Report: Integrated Icing Diagnostic Algorithm (IIDA)

**Barbara G. Brown¹, Jennifer L. Mahoney², Randy Bullock¹, Tressa L. Fowler¹,
Judy Henderson², and Andy Loughe³**

**Quality Assessment Group
Aviation Forecast and Quality Assessment (AFQA) Product Development Team**

31 July 2001

¹ Research Applications Program, National Center for Atmospheric Research (NCAR), Boulder, CO

² Forecast Systems Laboratory, National Oceanic and Atmospheric Administration (NOAA), Boulder, CO

³ Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, CO

Abstract

This report summarizes verification results for the Integrated Icing Diagnostic Algorithm (IIDA), which is designed to diagnose the existence of in-flight icing conditions aloft. The report was prepared by the Quality Assessment Group of the Aviation Weather Research Program's Aviation Forecast and Quality Assessment Product Development Team. The purpose of the report is to summarize the quality of IIDA diagnoses, in anticipation of IIDA's transition through the D4/D5 decision point of the Aviation Weather Technology Transfer process.

The report includes (a) a brief summary of previous evaluations of IIDA and other in-flight icing forecasts and algorithms; (b) results from an intensive evaluation of IIDA diagnoses from winter 2000; and (c) results from ongoing operational verification of IIDA and other icing forecasts by the Forecast System's Laboratory's Real-time Verification System (RTVS). In all cases the IIDA diagnoses are evaluated using *Yes* and *No* pilot reports (PIREPs) of icing conditions. In most cases, the analyses focus on *Yes* reports of moderate-or-greater (MOG) icing severity. While the in-depth analyses considered forecasts and diagnoses during daylight hours only, the RTVS results included night-time hours as well.

The IIDA diagnoses are compared to operational icing forecasts (AIRMETs) and to forecasts from two algorithms – NNICE and VVICE – which were developed at the Aviation Weather Center. In addition, some comparisons consider Liquid Water Content (LWC) forecasts from the Rapid Update Cycle numerical weather prediction system. Although the AIRMETs are quite different from IIDA in many ways (e.g., they are limited to a volume that can be defined in a textual message, and they are intended to depict icing conditions over a 6-h period), they are included in this evaluation as the operational standard that is available to users. IIDA produces both a General Icing and a Supercooled Large Droplet (SLD) field. Both fields consist of values of icing “potential” on a scale from 0 to 1 at each grid point. Most of the analyses focus on the General Icing field, with a few results presented for the SLD field.

IIDA has been evaluated over a number of years (e.g., Brown et al. 1999). However, some previous evaluations were based on earlier versions of the algorithm. In general, it appears that the quality of the IIDA diagnoses has improved somewhat; in particular, differences between the verification statistics for IIDA and other forecasts are somewhat greater in recent years than was the case for earlier versions of the algorithm. A regional evaluation of IIDA suggests that IIDA is best at detecting icing conditions in the Great Lakes region and the Northeast, and somewhat less capable in the South; similar characteristics were noted for the AIRMETs (Kane et al. 2000).

Results of the in-depth evaluation of IIDA for winter 2000 and the RTVS evaluations for winters 2000 and 2001 indicate that:

- IIDA is relatively efficient at detecting icing conditions, with a probability of detection for MOG *Yes* reports [POD_y(MOG)] of 0.75-0.85, and a corresponding probability of detection of *No* reports (POD_n) between 0.6 and 0.7, depending on the IIDA threshold used, and a relatively small percentage (5-8%) of the total airspace impacted by the forecasts. AIRMETs captured similar proportions of *Yes* reports, while capturing a somewhat smaller proportion of *No* reports. AIRMETs were somewhat more efficient in terms of the area covered by the forecasts and somewhat less efficient in terms of the

volume of airspace covered; however, this result is at least partially related to the constraints on the form of the AIRMETS.

- Overall, RTVS analyses – which considered icing diagnoses in both the day-time and the night – indicate that IIDA performance is somewhat better than NNICE and VVICE, and the AIRMETS, in comparisons of PODy, % Volume, and PODn.
- IIDA diagnoses are skillful, as measured by their ability to discriminate between *Yes* and *No* icing situations.
- PODy and PODn values for all of the forecasts and algorithms are somewhat variable from time-to-time. For example, for IIDA with a threshold of 0.25, the PODy(MOG) values for individual diagnoses range from about 0.2 to 1.0, with the middle 50% of values between 0.6 and 0.9. However, the volume covered by the IIDA diagnoses is quite consistent from time-to-time. RTVS evaluations of variations in the statistics from week-to-week suggest that verification statistics for the three algorithms and the AIRMETS exhibit similar variations. PODy(MOG) values for the AIRMETS and VVICE decreased somewhat toward the end of the winter 2001 season, while the PODy(MOG) values for IIDA and NNICE remained somewhat more consistent throughout the season.
- IIDA diagnoses perform fairly well as persistence forecasts out to about three hours. After that period, they are generally out-performed by the AIRMETS.
- RTVS results stratified by altitude suggest that IIDA performs best at lower altitudes (15,000 ft and below) and that IIDA is better than the other algorithms and the AIRMETS at capturing No-icing conditions at all altitudes, while still maintaining a good PODy(MOG) value.
- The SLD field is very efficient at capturing PIREPs reporting severe icing conditions. Although the PODy for severe reports is about 0.3, the diagnoses cover a very small volume, in comparison to the IIDA General Icing field, the LWC forecasts, and the AIRMETS.

In summary, IIDA is skillful at diagnosing General Icing conditions, and the SLD algorithm is efficient at detecting severe icing situations. The algorithm is quite capable of discriminating between *Yes* and *No* icing conditions, and is efficient in limiting the airspace warned. These verification analyses represent a comprehensive evaluation of IIDA over two icing seasons, and thus the results should be representative of most icing situations of concern.

1. Introduction

The Integrated Icing Diagnostic Algorithm (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). The purpose of this report is to document the quantitative evaluations of IIDA that have been undertaken to verify the quality of IIDA diagnoses, in anticipation of the transition of this product through the D4/D5 decision point of the Aviation Weather Technology Transfer (AWTT) process.

2. Approach

IIDA has undergone extensive evaluation since its initial development began, and some of these analyses have been summarized in conference papers and reports. In these studies, the quality of IIDA diagnoses has been compared to the quality of diagnoses and forecasts produced by a number of other icing algorithms as well as the operational icing 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. 1997), along with two other automated in-flight icing algorithms and the icing AIRMETs. Finally, over the last year, IIDA diagnoses from winter 2000 have been evaluated extensively by the Verification Group of NCAR's Research Applications Program (RAP).

Thus, the quality of IIDA diagnoses is considered from three different vantage points in this report:

- (a) A summary of previous evaluations of IIDA's performance;
- (b) A summary of the in-depth analyses undertaken at NCAR.;
- (c) A summary of verification results from the RTVS, including comparisons to the performance of other algorithms and forecasts.

As noted in the Quality Assessment Plan for IIDA prepared by the AWRP's Quality Assessment Group (Brown and Mahoney 2000), the following items represent important aspects of the quality assessment of IIDA:

- (a) IIDA should be evaluated over at least one icing season.
- (b) The quality of IIDA should be compared to the quality of other relevant products [e.g., icing AIRMETs; earlier automated algorithms such as the RAP algorithm (Thompson et al. 1997)].
- (c) A representative set of relevant issue times should be included in the evaluation.
- (d) Each IIDA forecast should be evaluated as a set of *Yes/No* forecasts, by applying a variety of thresholds to the IIDA values.
- (e) Pilot reports (PIREPs) of icing should be used as the verification data.

- (f) Appropriate verification methods should be utilized, to take into account known characteristics of PIREPs.
- (g) Day-to-day variations in the verification statistics should be examined.

These items were taken into account in the assessment presented here. For example, two seasons were used for the RTVS evaluation, and PIREPs were used as the verification data. The verification approach, described in Section 4, which is based on methods that have been developed over a number of years, is documented in several reports and papers.

As noted in item (b) above, it is important to compare the quality of IIDA diagnoses to the quality of one or more standards of reference. Thus, the quality of the IIDA diagnoses is compared to the quality of several other automated forecasting algorithms (e.g., NNICE, VVICE; see Section 3), as well as to the quality of the operational forecasts (i.e., AIRMETs). However, it is important to emphasize that the algorithm forecasts (e.g. IIDA, NNICE, and VVICE) and the AIRMETs are very different types of forecasts, with different objectives. The IIDA algorithm, for instance, is a diagnostic algorithm with hourly updates, which assimilates various datasets to obtain a snapshot of the potential for icing conditions. The AIRMETs, on the other hand, are valid over a 6-h period and are designed to capture the icing conditions as they move through the AIRMET area over the 6-h forecast period. Due to the differences between these forecasts, it is difficult to clearly compare the performance of forecasting algorithms and the AIRMETs, since the two approaches are focused on somewhat different attributes of the icing conditions. However, in order to understand the quality of the IIDA algorithm, it is necessary for comparisons between various forecasts to be made, and for IIDA diagnoses to be compared to the operational standard. These comparisons are made in such a way as to be as fair as possible to both the AIRMETs and IIDA, as described in Section 4, while still obtaining the information needed. Nevertheless, users of these statistics should keep these assumptions in mind when evaluating the strengths and weaknesses of each type of forecast.

3. Data

3.1. Algorithms and forecasts

IIDA and some of the forecast products that are compared to IIDA are briefly described here. Some of these products have been included in previous evaluations of IIDA. Most (e.g., AIRMETs, VVICE) also are included in the ongoing RTVS evaluations.

Integrated Icing Diagnostic Algorithm: IIDA was developed by the IFIPDT, with funding provided by the FAA's AWRP. Every hour, IIDA generates diagnoses of icing conditions. These diagnoses are based on an intelligent combination of observations (satellite, surface, and radar) with 3-h temperature and 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” field and a “Supercooled Large Droplet” (SLD) icing field. Both of these components of IIDA are considered in this evaluation, although most

attention is focused on the General Icing field. The algorithm output for General Icing is a three-dimensional icing “potential” field, with values ranging between 0 and 1 (sometimes re-scaled from 0 to 100) assigned to each RUC grid point. The likelihood of icing is expected to increase with increasing values of icing potential. However, the values are not calibrated to a probability scale. The SLD field also is a three-dimensional field of potential values, ranging between 0 and 1. However, in some cases where the existence of SLD is difficult to ascertain, the algorithm assigns “unknown” to a grid point. Because IIDA produces values across a continuous range, users may select their own threshold for decision-making.

AIRMETS: AIRMETS are the operational forecasts of in-flight icing conditions that are generated by AWC forecasters. The forecasts are produced every six hours and are valid for up to six hours (NWS 1991). AIRMETS may be amended as needed between the standard issue times. However, amended AIRMETS are not considered in this evaluation. The forecasts are in a textual form that can be decoded into latitude and longitude vertices, with tops and bottoms of the icing regions defined in terms of altitude. Unfortunately, some other more descriptive elements of the AIRMETS cannot be decoded and thus are not considered. For comparison with the IIDA diagnoses, and forecasts from other algorithms, the AIRMETS are evaluated over the same time window as the model-based algorithms.

RAP algorithm: The RAP icing algorithm identifies conditions leading to several different icing environments, and it incorporates these structures into four icing forecast components: General, Unstable, Stratiform, and Freezing Rain (Thompson et al. 1997). These four components provide indications of where icing is likely to exist, and they identify the physical structure of the atmosphere (as depicted by the model) leading to the icing prediction. The RAP algorithm is only considered in the studies described in Section 6 (e.g., Brown et al. 1999).

NNICE: The NNICE algorithm was developed by Don McCann at the AWC. The algorithm is based on recognition by a neural network of complex patterns of conditions required for significant icing (e.g., based on T, RH, and slight convective potential). In particular, NNICE uses T, RH, and computed convective potential from the RUC to make its predictions.

VVICE: The Vertical Velocity Icing (VVICE) algorithm also was developed by Don McCann of the AWC. This tool bases aircraft icing forecasts on estimates of ice accumulation potential and the subsequent degradation of aircraft performance. VVICE estimates the Percent Power Increase (PPI) required to overcome the additional drag associated with an accumulation of ice so the aircraft can continue at a steady speed and altitude. More information about VVICE is available on the world-wide-web at <http://www.awc-kc.noaa.gov/awc/help/vviceinfo.html>.

RUC Liquid Water Content (LWC) forecasts: In the current operational version of RUC, LWC predictions are based on the explicit microphysics scheme developed by Reisner et al. (1998). The LWC forecasts used in the comparisons presented here were based on the following criteria: (a) total LWC (cloud water plus rain water) greater than zero; (b) temperature less than 0°C.

3.2. *Observations*

PIREPs indicating the existence or lack of icing conditions are used as the verifying observations for the IIDA assessment. Both explicit *Yes* and *No* PIREPs are considered. In addition, PIREPs that indicate cloud-free skies overhead are used in some cases as an indicator of no-icing conditions, as in Brown et al. (1997); these reports are designated as “Clear-Above” (CA) reports.

Yes PIREPs are grouped according to reported icing severity, including “All” (all severities); “MOG” (moderate or greater severity); and “Severe.” PIREPs included in the RTVS analyses were filtered using lightning observations to remove PIREPs that might be associated with convection.

4. **Mechanics**

Basic procedures used to evaluate the forecasts and diagnoses are briefly described in this section.

4.1. *Forecast/observation matching procedures*

Each PIREP is either matched or interpolated to the four closest grid points at a particular model level, for all model levels within the range of altitudes identified by the PIREP. The NCAR verification system uses a four-gridpoint matching procedure, in which the most extreme forecast value at the surrounding gridpoints is matched to a PIREP; RTVS uses an interpolation method to estimate the forecast value at the location of the PIREP, using forecasts at the four closest grid points. Previous comparisons of these approaches have indicated that the verification results are robust to this difference in matching approaches (Brown et al. 2000). The icing forecasts are evaluated over the entire continental U.S. domain considered by the AIRMETs, which includes some coastal waters.

4.2. *Time window*

The current version of IIDA incorporates information from PIREPs in the hour prior to the forecast valid time. Thus, starting in winter 2000, IIDA verification analyses only use PIREPs in a time window of one hour following the forecast valid time (previous evaluations used a time window including both one hour prior to and one hour after the valid time). The same one-hour period is used to evaluate the AIRMETs and other icing forecasts, so that the same PIREPs are used for all types of forecasts and the results are directly comparable. All of the NCAR evaluations are limited to daylight hours (1200, 1500, 1800, 2100, 0000, and 0300 UTC), whereas the RTVS results include all IIDA issue times, including night-time hours.

5. Verification methods

The verification methods used for the IIDA evaluations are based on standard verification concepts that recognize the underlying framework for forecast verification and the associated high dimensionality of the verification problem (e.g., Murphy and Winkler 1987). The methods described here were developed by the QAG and by members of the In-flight Icing PDT. They are described in greater detail in Brown (1996) and Brown et al. (1997). The methods are also outlined in the Quality Assessment Plan for IIDA (Brown and Mahoney 2000).

The icing forecast verification methodology outlined by Brown et al. (1997) treats the icing forecasts and observations as *Yes/No* values. This method can be extended to forecasts with values on a continuous scale using the approach outlined in Brown et al. (1999). In particular, the icing forecasts for IIDA and other algorithms with continuous output can be converted to a set of *Yes/No* forecasts by application of a variety of thresholds. For instance, application of a threshold of 0.50 to IIDA forecasts would lead to a *Yes* forecast for all grid points with an IIDA value greater than or equal to 0.50, and a *No* forecast for all grid points with an IIDA value less than 0.50. The verification methods are based on the *Yes/No* two-by-two contingency table (Table 1), where the rows represent the forecasts, and the columns represent the observations. Each cell in this table contains a count of the number of times a particular forecast/PIREP pair was observed. Counts can represent an individual forecast field, or can be accumulated across days, weeks, months, and so on. Note that for icing forecasts, the counts in the verification table are PIREP-based (i.e., the sum of the counts is the total number of *Yes* and *No* PIREPs that were included in the analysis), and not all forecast grid points are represented.

Table 1. Basic contingency table for evaluation of dichotomous (e.g., *Yes/No*) forecasts. Elements in the cells are the counts of forecast-observation pairs.

<i>Forecast</i>	<i>Observation</i>		<i>Total</i>
	<i>Yes</i>	<i>No</i>	
<i>Yes</i>	YY	YN	YY+YN
<i>No</i>	NY	NN	NY+NN
<i>Total</i>	YY+NY	YN+NN	YY+YN+NY+NN

Table 2 lists the verification statistics that are included in the IIDA evaluations. Due to characteristics of the PIREPs, certain restrictions must be placed on the verification statistics that can be computed from Table 1 for evaluation of icing forecasts. In particular, some measures and statistics cannot be appropriately estimated (Brown and Young 2000). Thus, Table 2 does not include measures such as the False Alarm Ratio, Critical Success Index, and Bias, which might be considered in the evaluation of other types of dichotomous forecasts; these measures cannot be computed for verification of icing forecasts. Hence, PODY and PODn are the primary verification statistics that are included in the evaluation.

PODY and PODn are estimates of the proportions of *Yes* and *No* observations that were correctly forecasted, respectively. Together, PODY and PODn measure the ability of the forecasts to discriminate between *Yes* and *No* icing observations. The True Skill Statistic (TSS)

(Doswell et al. 1990), also known as Hanssen-Kuipers discrimination statistic (Wilks 1995), summarizes this discrimination ability. Note, however, that it is possible to obtain the same value of TSS for a variety of combinations of POD_y and POD_n. Thus, it always is important to consider both POD_y and POD_n along with TSS.

The % Area is the percent of the total possible area (i.e., the continental U.S.) that has a *Yes* forecast at some model level above, and % Volume is the percentage of the total volume of air space that has a *Yes* forecast. These measures indicate the spatial extent of the forecasts. Volume Efficiency is the ratio of POD_y to % Volume; this efficiency statistic represents the POD_y per unit % Volume. Although Volume Efficiency is a convenient way to summarize the combination of POD_y and forecast extent, comparisons of this statistic for different forecasts can be misleading unless both POD_y and % Volume are also considered.

Table 2. Verification statistics to be used in the evaluation of IIDA.

<i>Statistic</i>	<i>Definition</i>	<i>Description</i>
POD_y	$YY/(YY+NY)$	Probability of Detection of “Yes” observations
POD_n	$NN/(YN+NN)$	Probability of Detection of “No” observations
TSS	$POD_y + POD_n - 1$	True Skill Statistic
% Area	$(\text{Forecast Area}) / (\text{Total Area}) \times 100$	% of the area of the continental U.S. where icing is forecast to occur at some level
% Volume	$(\text{Forecast Volume}) / (\text{Total Volume}) \times 100$	% of the three-dimensional airspace over the continental U.S. where icing is forecast to occur
Volume Efficiency	$(POD_y \times 100) / \% \text{ Volume}$	POD _y (x 100) per unit % Volume
Curve Area	Area under the curve relating POD _y and 1-POD _n	Overall skill, based on Signal Detection Theory (Relative Operating Characteristic curve)

The relationship between POD_y and 1-POD_n for different algorithm thresholds is the basis for the verification approach known as “Signal Detection Theory” (SDT). This relationship can be represented for a given algorithm by the curve joining the (1-POD_n, POD_y) points for different algorithm thresholds. The resulting curve is known as the “Relative Operating Characteristic” (ROC) curve in SDT. The goal is for the ROC curve to lie close to the upper left corner of the diagram. The area under this curve is a measure of overall forecast skill (e.g., Mason 1982), and provides another measure that can be compared among the algorithms, which is not dependent on the threshold used. A forecast with no skill would have an ROC area of 0.5 or less.

The curves relating PODy to % Volume and % Area, with points representing each algorithm threshold, also are of interest. These curves represent the trade-off between large PODy values and the spatial extent of the forecast. As with the ROC, curves for better forecasting systems lie toward the upper left corner of the diagram.

6. Results of previous studies

6.1. *Brown et al. 1997*

This paper, which is attached as Appendix 1, provides an earlier comparison of a number of icing forecasting techniques, including the RAP algorithm and the AIRMETs, based on forecasts and observations collected in winter 1994. Most importantly, the paper defined the underlying methodology for verification of icing forecasts and diagnoses that is utilized for the IIDA verification studies presented in this report. Overall results for the RAP algorithm, 0-h forecasts, were $PODy(MOG) = 0.74$, $PODn = 0.52$, % Area = 43%, and % Volume = 6.5%.

6.2. *Brown et al. 1999*

This paper provided the first published evaluation of IIDA, and included comparisons of IIDA to a number of other algorithms and forecasts, including NNICE and the AIRMETs, using data collected from December 1997 through March 1998. Results of this study indicated that IIDA performed better than algorithms that had been developed previously (e.g., the RAP algorithm). In particular, IIDA forecasts were able to capture the same proportion of *Yes* PIREPs while covering a smaller volume. IIDA also performed better than NNICE in this respect. These results also suggested that IIDA was somewhat more efficient than the AIRMETs in terms of volume coverage, but that the AIRMETs were more efficient in terms of areal coverage. Verification statistics for IIDA with a threshold of 0.20 were the following: $PODy(MOG) = 0.74$; $PODn = 0.56$; % Area = 42%; % Volume = 8%. In comparison, the statistics for AIRMETs were: $PODy(MOG) = 0.73$; $PODn = 0.60$; % Area = 35%; and % Volume = 10%. This paper is included as Appendix 2.

6.3. *Kane et al. 2000*

Kane et al. (2000) considered regional variations in the performance of IIDA and the AIRMETs, using data collected during winter 2000. The results indicate that IIDA performance does vary somewhat from region to region. In particular, for a given IIDA threshold (e.g., 0.25), the PODy value is largest in the Great Lakes region and the Northeast, and smallest in the South. This paper is included as Appendix 3.

7. Results for winter 2000

A special verification effort was undertaken using the RAP verification system, to provide an in-depth evaluation of IIDA diagnoses for winter 2000. Specifically, IIDA output and PIREPs were archived and analyzed for the period 20 January through 21 March 2000. Because the data were archived in real time, the verification results represent the performance of IIDA in an operational setting. The IIDA verification statistics are compared to verification statistics for the AIRMETs and LWC forecasts from the RUC, and are limited to valid times of 1200, 1500, 1800, 2100, 0000, and 0300 UTC. Overall statistics are presented, as well as variations in the statistics from day-to-day and by valid time. In addition, the use of IIDA as a persistence forecast is considered, to provide a more meaningful comparison to the AIRMETs. All of the analyses concern the General Icing component of IIDA, except for the set of results for the SLD component that are presented in Section 7.7.

7.1. Overall results

Overall verification results for IIDA for all valid times combined are shown in Fig. 1, with the results for AIRMETs included for comparison. This figure shows the relationships between PODy and % Area, % Volume, and 1-PODn, for MOG PIREPs. As shown by the plots in this figure, the curve for the IIDA diagnoses is lower than the AIRMET point for PODy(MOG) vs. % Area (Fig. 1a). This result suggests that the AIRMETs cover a smaller areal extent, in general, for a given PODy(MOG) value. Most likely, this result is due to the fact that IIDA values may exceed a threshold in relatively narrow layers, which will contribute significantly to the areal coverage, but only a little to the volume coverage. In contrast, the AIRMETs have a coherent vertical structure.

The curve for PODy(MOG) vs. % Volume (Fig. 1b) indicates that IIDA is able to attain a relatively large PODy(MOG) value with a relatively small volumetric coverage, relative to the AIRMETs. At least in part, this result is also due to the fact that the AIRMETs are restricted to a cylindrical structure, whereas the three-dimensional icing field defined by IIDA may have an uneven top and/or bottom, and it may even have holes in the middle.

The curve relating PODy to 1-PODn (Fig. 1c) shows that IIDA diagnoses have positive skill, and they are able to successfully discriminate between *Yes* and *No* observations of icing. In particular, the IIDA curve lies above the 45° line, which is the “no-skill” line in a ROC diagram. The area under the curve is 0.76. For comparison, as noted in Section 5, the curve area for no-skill forecasts is 0.5. The diagram also indicates that the AIRMETs have positive skill; however, because the AIRMET skill is represented by a single point, it is not meaningful to compute the ROC area for these forecasts.

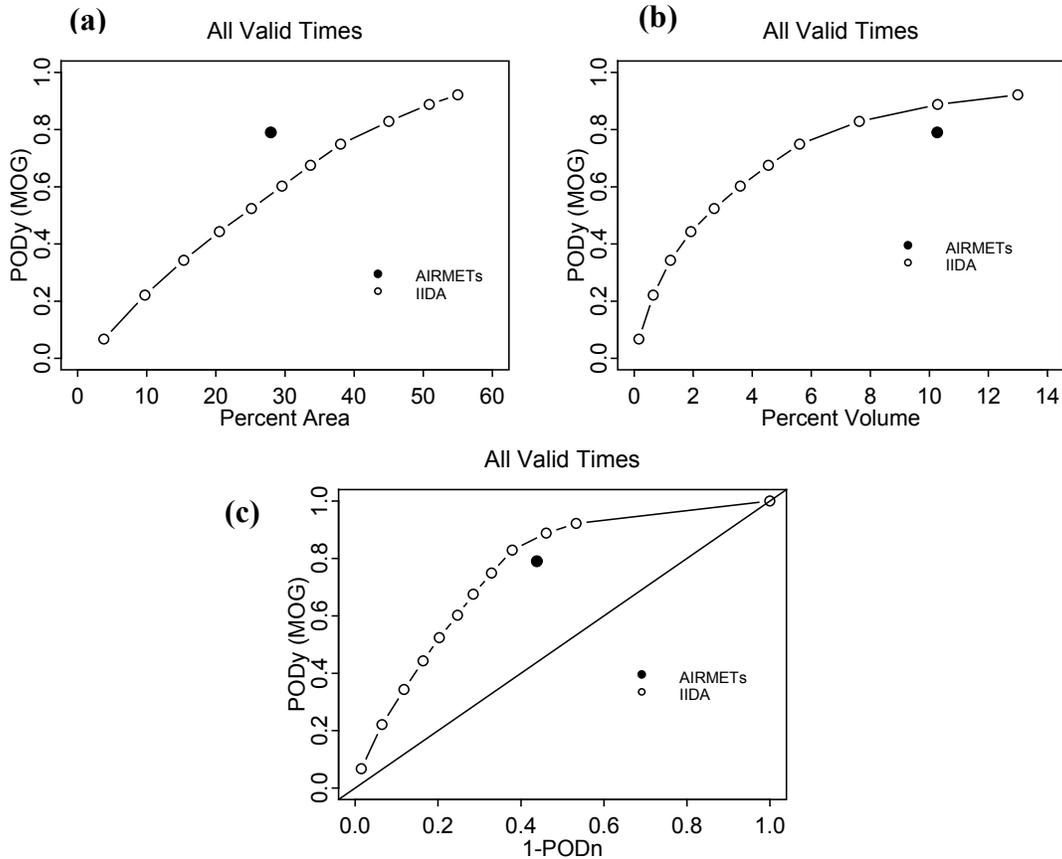


Figure 1. Overall verification statistics for IIDA and AIRMETs for winter 2000, for all valid times combined, based on MOG PIREPs: (a) PODY vs. % Area; (b) PODY vs. % Volume; and (c) PODY vs. 1-PODn. Each point on the IIDA curves represents a different threshold used to define *Yes/No* icing forecasts. The thresholds (starting in the upper right corner) are 0.0, 0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, and 0.95. The (1,1) point is also included in Fig. 1c to complete the curve.

Table 3 presents a summary of some of the overall verification statistics for IIDA and the AIRMETs, for two IIDA thresholds. The results in this table suggest that IIDA – for these particular thresholds – is able to capture a relatively large number of PIREPs while covering a relatively small volume of airspace. In addition, for a comparable value of PODY, IIDA was able to correctly classify a large percentage of the negative icing reports. For these thresholds, IIDA correctly classified about 65% of the *No* PIREPs, 70–80% of the *Yes* PIREPs, and about 90% of the clear-above (CA) PIREPs, while forecasting icing over 5–8% of the airspace volume.

The results in Table 3 also indicate that the POD values depend somewhat on the type of PIREP that is considered, and these variations are consistent for both IIDA and the AIRMETs. In particular, PODY values increase somewhat as the reported severity increases. Moreover, PODn values are much larger for the inferred no-icing conditions (i.e., CA reports) than for the explicit *No* reports. These results are consistent with those reported by Brown et al. (1997).

Table 3. Overall verification statistics for IIDA and the AIRMETs, for the period 20 January through 21 March 2000, for all valid times combined.

Forecast	POD _y			POD _n		TSS	% Area	% Vol	Vol Eff.
	All	MOG	SVR	No	CA				
IIDA-0.15	0.80	0.83	0.87	0.62	0.90	0.45	45.0	7.6	10.9
IIDA-0.25	0.70	0.75	0.76	0.67	0.93	0.42	38.1	5.6	13.4
AIRMETs	0.76	0.79	0.90	0.56	0.91	0.35	28.0	10.3	7.7

In considering the statistics in Table 3, it is important to recognize that the measures presented have uncertainty associated with them, due to sampling and other errors. Table 4 provides 95% confidence intervals for POD_y(MOG) and POD_n(No), computed using methods appropriate for application to PIREP-based statistics, which are described in Kane and Brown (2000). These intervals have a range of about 0.05 for IIDA and 0.06-0.08 for the AIRMETs. The intervals for IIDA-0.25 and the AIRMETs overlap somewhat. However, the intervals for IIDA-0.15 are distinct from the intervals for both IIDA-0.25 and the AIRMETs, indicating that the IIDA-0.15 measures are significantly different from the IIDA-0.25 and AIRMET measures.

Table 4. 95% confidence intervals for POD_y(MOG) and POD_n(No) values shown in Table 3.

Forecast	Statistic	
	POD _y (MOG)	POD _n (No)
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

7.2. Variations with valid time

Figure 2 shows the same plots as in Fig. 1, with a separate line for each valid time. In general, the lines for the various valid times are quite consistent with each other, and it is difficult to identify a specific pattern of variation among valid times. The results in Table 5 make this comparison a bit more specific, for the same IIDA thresholds considered in Table 3. The results in Table 5 suggest that there is a slight tendency for POD_y values to be a bit larger for early-day periods (1200 – 2100 UTC), but this tendency is quite small. Overall statistics, such as TSS and Volume Efficiency, indicate that there is no particular trend with valid time. The most notable variation in Table 5 is associated with the POD_y values for severe PIREPs. However, the large variation in these statistics is most likely related to the very limited number of severe reports in the dataset, as shown in Table 6, which lists the numbers of each type of PIREP that were included in the analysis, by valid time. For individual valid times, the number of severe reports ranges from 21 to 44.

Characteristics of the variations of the statistics with valid time can be summarized using the ROC area statistics presented in Table 7. As shown in this table, the ROC Curve Area values vary only a small amount among the valid times. The largest value (0.80) is associated with the

1800 UTC valid time, while the smallest value (0.74) is associated with the 1200 UTC valid time.

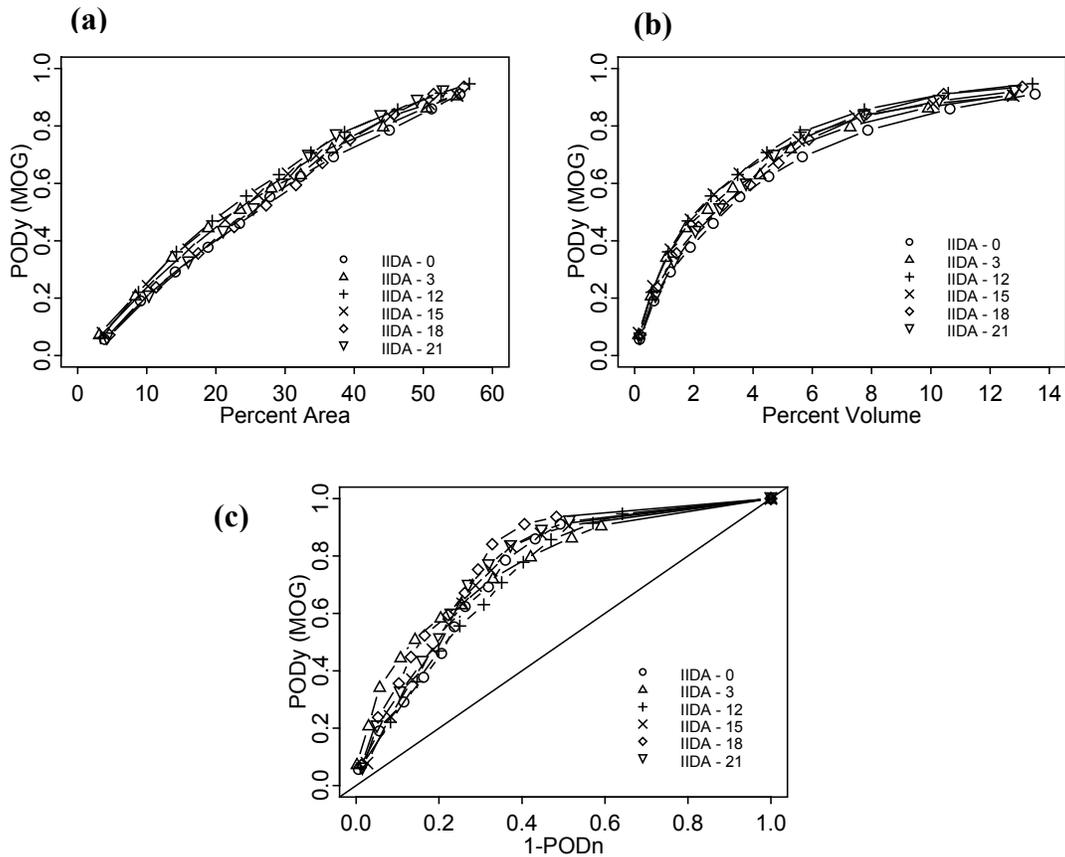


Figure 2. As in Figure 1, showing overall verification results for IIDA by valid time.

Since IIDA verification statistics seem to be quite stable among the valid times, the remainder of the results focuses on statistics for all valid times combined. This approach has the advantage of providing a larger sample of PIREPs to use in estimating the verification statistics, which should lead to more stable results (e.g., especially for statistics where the number of observations is quite limited, such as for severe conditions).

7.3. Day-to-day variations

It is important to also consider the variations in the verification statistics from day-to-day, to consider how much variability can be expected for different forecasts, and what the expected range of values is for various statistics. This variability is most conveniently displayed using box plots, which provide an easy way to compare the distributions of values. Figure 3 shows box plots of various verification statistics, organized according to the value of the IIDA threshold that was used to define the *Yes/No* forecasts, for all valid times combined.

Table 5. Verification statistics for IIDA by valid time, for two IIDA thresholds, for the period 20 January through 21 March 2000.

Valid time (UTC)	PODy			PODn		TSS	% Area	% Vol	Vol Eff.
	All	MOG	SVR	No	CA				
<i>Threshold = 0.15</i>									
1200	0.81	0.85	0.90	0.53	0.89	0.38	46.3	7.8	10.9
1500	0.80	0.84	0.79	0.62	0.91	0.46	45.2	7.5	11.2
1800	0.81	0.84	0.81	0.68	0.92	0.52	45.8	7.8	10.8
2100	0.79	0.83	1.00	0.63	0.91	0.46	43.8	7.7	10.8
0000	0.78	0.79	0.91	0.63	0.87	0.42	45.1	7.9	10.0
0300	0.77	0.79	0.78	0.59	0.91	0.38	44.2	7.3	10.8
All	0.80	0.83	0.87	0.62	0.90	0.45	45.0	7.6	10.9
<i>Threshold = 0.25</i>									
1200	0.72	0.77	0.87	0.60	0.91	0.37	38.6	5.6	13.8
1500	0.70	0.76	0.53	0.67	0.93	0.43	38.7	5.5	13.8
1800	0.72	0.76	0.81	0.71	0.94	0.47	39.4	5.9	12.9
2100	0.72	0.76	0.89	0.68	0.93	0.44	37.4	5.7	13.3
0000	0.68	0.70	0.86	0.67	0.90	0.37	37.0	5.7	12.3
0300	0.66	0.71	0.62	0.68	0.93	0.39	36.8	5.3	13.4
All	0.70	0.75	0.76	0.67	0.93	0.42	38.1	5.6	13.4

Table 6. Counts of PIREP observations included in the analyses, by PIREP type and valid time. Note that a single PIREP may be assigned to multiple levels; each level is counted.

PIREP type	Valid time (UTC)						
	1200	1500	1800	2100	0000	0300	All
Yes – All	1,987	3,254	2,650	3,008	1,299	1,289	13,487
Yes – MOG	779	1,114	1,076	1,254	598	619	5,440
Yes – SVR	38	34	21	44	22	32	191
No	1,238	2,117	1,796	1,479	494	467	7,591
CA	10,965	16,041	13,096	9,855	3,541	3,099	56,597

Table 7. ROC Curve Areas by valid time.

Valid Time (UTC):	1200	1500	1800	2100	0000	0300	All
Curve area:	0.74	0.76	0.80	0.77	0.75	0.77	0.76

Figures 3a-c show the distributions of PODy and PODn values. One interesting aspect of these figures is the larger variability associated with the PODy(MOG) values, in comparison to the PODy(All) distributions. In particular, the boxes – which contain the middle 50% of the distribution – are quite a bit larger for PODy(MOG) than for PODy(All). This result may be at least partially related to the smaller numbers of MOG PIREPs available to compute PODy(MOG), relative to the number of All PIREPs. For most IIDA thresholds, PODy values as large as 1 and as small as 0 were observed. However, in general, the central part of the distributions of PODy values are a decreasing function of threshold, while the central part of the PODn distributions is an increasing function of threshold. For an IIDA threshold of 0.25, the middle half of the observations have (a) a PODy(All) value between about 0.55 and 0.85; (b) a PODy(MOG) value between about 0.6 and 0.9; and (c) a PODn value between about 0.6 and 0.85.

Distributions of % Area and % Volume also exhibit a fair amount of variability from day-to-day, and with threshold. In particular, Figs. 3d-e show that the distributions of both % Area and % Volume are decreasing functions of threshold. However, these distributions exhibit less day-to-day variability than the PODy distributions. That is, they appear to be “tighter,” perhaps due to the fact that they are not impacted by the availability and distribution of pilot reports. For an IIDA threshold of 0.25, the middle half of IIDA forecasts have a % Area value between about 30 and 45%, and a % Volume value between 4 and 7%.

Day-to-day variability in verification statistics is common to all types of forecasts, not just IIDA. In particular, the AIRMET verification statistics exhibit similar variability, as shown in Fig. 4. These box plots show the distributions of verification statistics for all valid times combined, for the AIRMETs and for IIDA with thresholds of 0.15 and 0.25. For PODy(MOG), the AIRMET and IIDA-0.15 statistics have similar distributions, whereas the distribution for IIDA-0.25 values is located somewhat below the other two distributions. The PODy(MOG) values for IIDA-0.25 also appear to be somewhat more variable, but 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 (which they cannot exceed). Distributions of PODn for IIDA (Fig. 4b) are located somewhat higher than the distribution for the AIRMETs; all three distributions exhibit similar variability.

As expected, the % Area distributions are higher for IIDA than for the AIRMETs (Fig. 4c), and the % Volume distribution for the AIRMETs is higher than the corresponding distributions for IIDA (Fig. 4d). One notable feature of Fig. 4d is the narrow range of % Volume values associated with IIDA – the distributions of % Volume are very tight, as noted earlier with respect to Fig. 3e. 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. It also is possible that the AIRMET volumes are constrained to be influenced to some extent by non-meteorological factors, which impact the volume warned.

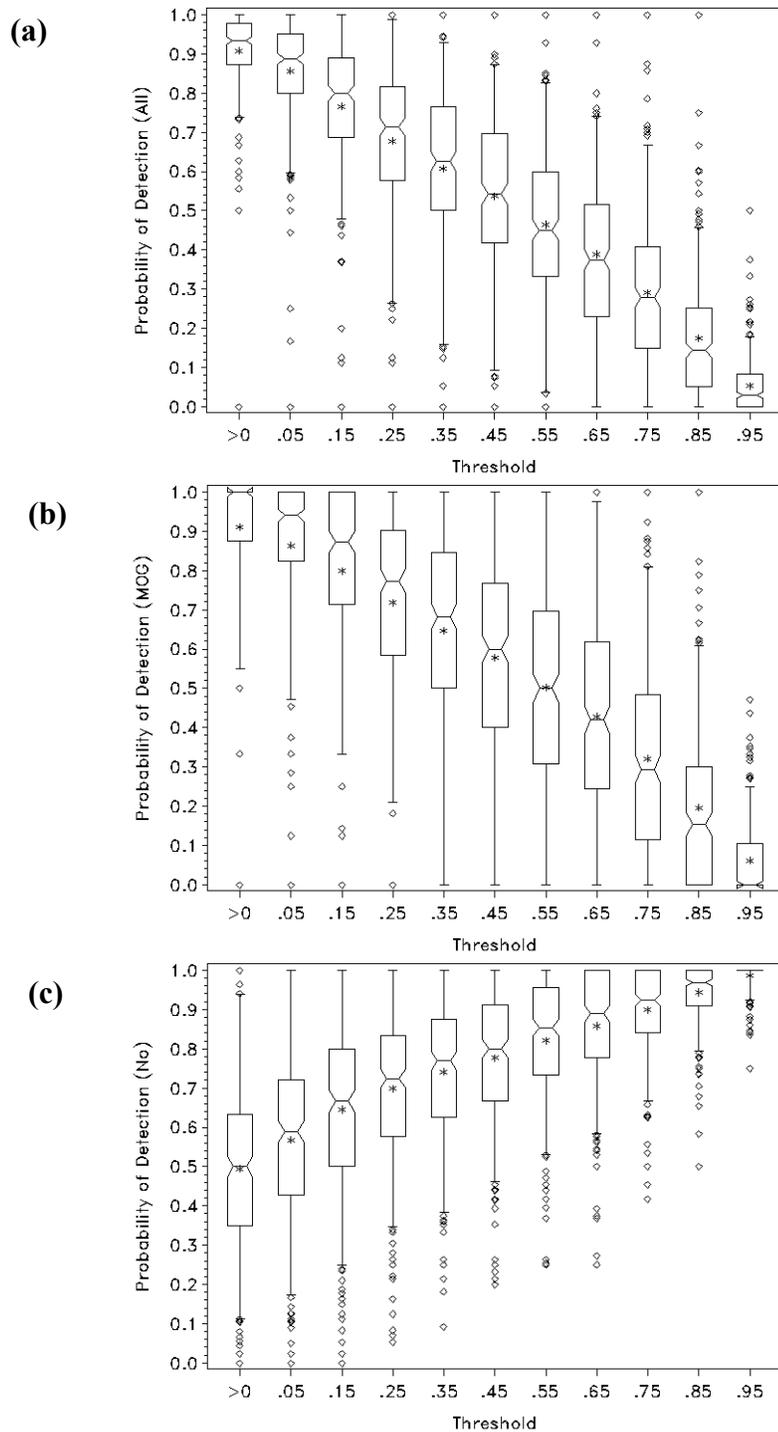


Figure 3. Box plots showing distributions of daily values of verification statistics for IIDA, by threshold: (a) $POD_y(All)$ (b) $POD_y(MOG)$; (c) POD_n ; (d) % Area; and (e) % Volume. Line inside each box represents median value; bottom and top of boxes are 0.25th and 0.75th quantile values, respectively; 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, respectively.

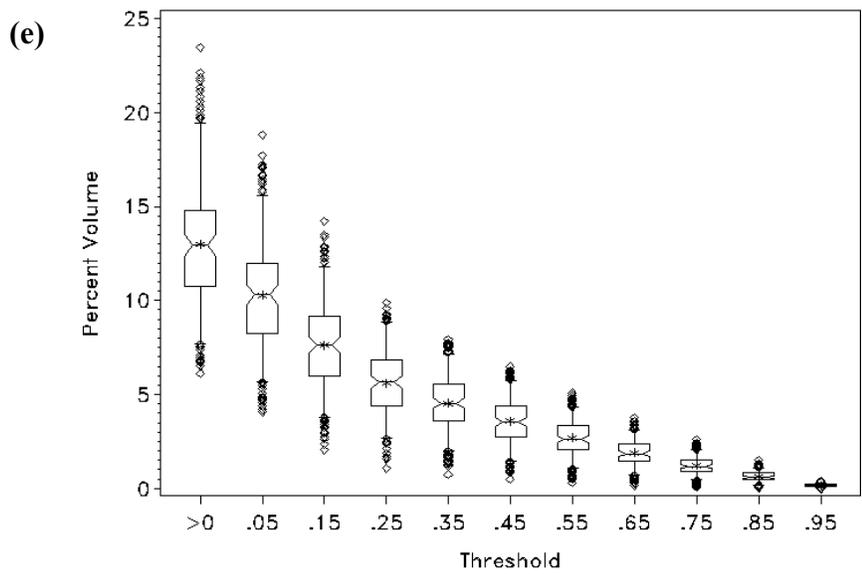
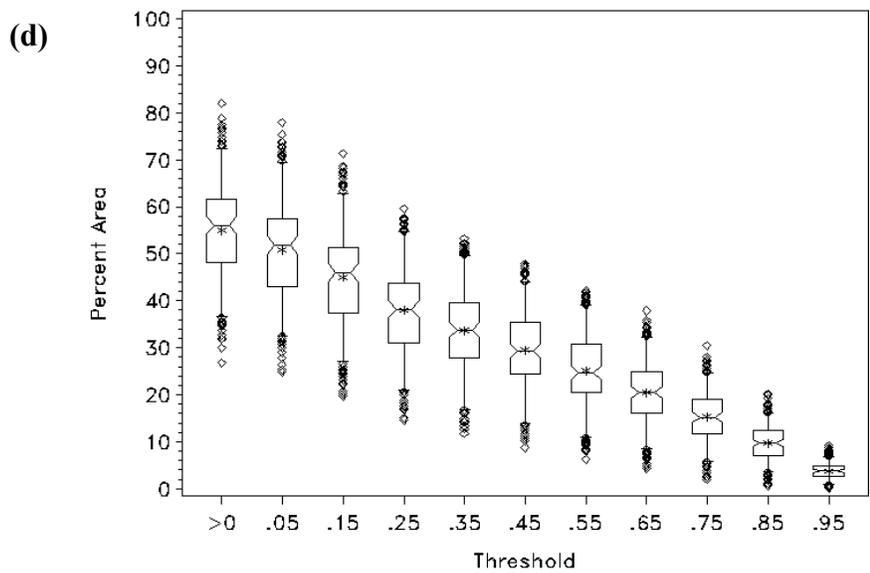


Figure 3, cont.

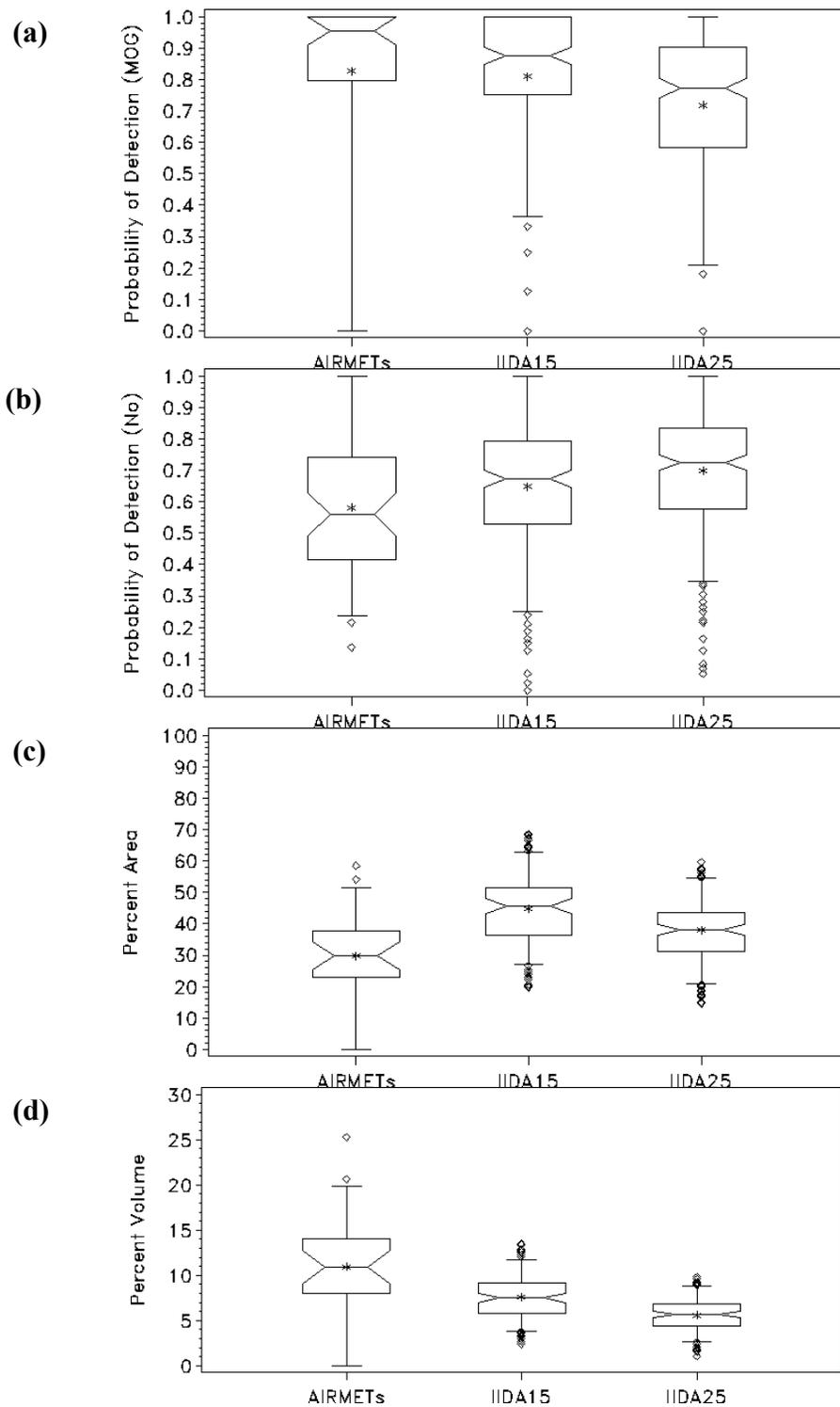


Figure 4. Distributions of verification statistics for IIDA and AIRMETs, as in Fig. 3, for all valid times combined: (a) $POD_y(MOG)$; (b) POD_n ; (c) % Area; and (d) % Volume.

7.4. IIDA persistence

As discussed in Section 2, the AIRMETs represent a very different kind of forecast from the IIDA diagnoses. One of the most important differences between the two systems is that the AIRMETs attempt to identify regions that will experience icing conditions over the subsequent *six hours*, whereas the IIDA diagnosis essentially is a *nowcast*, representing conditions at the time the diagnosis is updated. This distinction has not been directly taken into account in the preceding analyses.

Although IIDA is not intended to be a forecast, it is of interest to examine its capabilities as a forecast, by essentially treating the IIDA diagnostic as a persistence forecast. This treatment allows a more appropriate comparison to the AIRMETs, and provides information to users regarding the length of time IIDA diagnoses might be assumed to be meaningful.

Fig. 5 shows verification results for IIDA diagnoses, as well as IIDA treated as a 3-h and a 6-h forecast, and the AIRMETs. Results in this figure suggest that the IIDA persistence forecasts perform fairly well for the first three hours. However, the performance deteriorates quite a bit by the end of the 6-h period. In particular, if the AIRMET point is used as a point of reference, the 3-h forecast curve lies just above the point of reference, while the 6-h forecast curve lies below the point. These results are consistent across valid times. They suggest that the AIRMETs generally provide a better forecast of icing conditions in the latter part of the 6-h period than is provided by a persisted IIDA diagnosis.

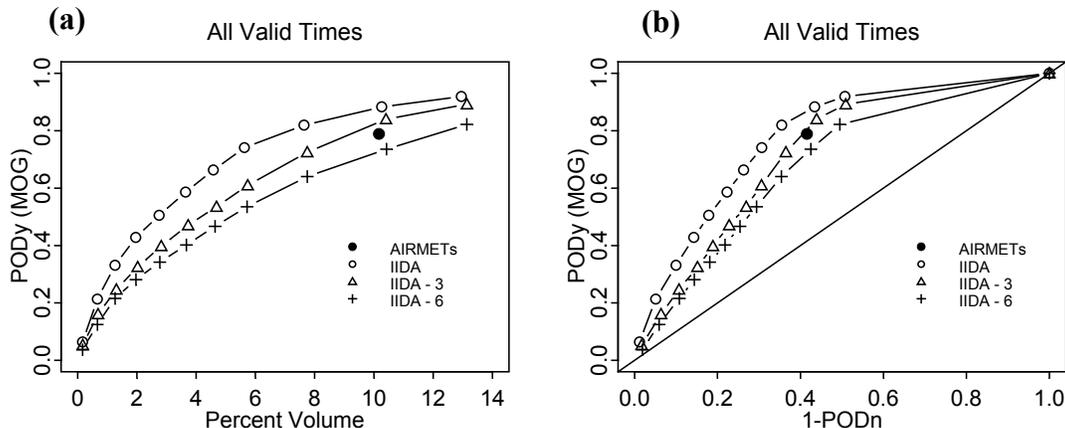


Figure 5. Verification curves for IIDA, 3-h IIDA persistence, 6-h IIDA persistence, and AIRMETs, all valid times combined: (a) PODy(MOG) vs. % Volume; and (b) PODy(MOG) vs. 1-PODn.

7.5. Comparisons to RUC liquid water forecasts

The explicit liquid water content (LWC) forecasts produced by the RUC represent another possible approach to identifying locations of icing conditions in the atmosphere. A representative sample of verification results for the RUC LWC forecasts (3-h forecasts valid at 1800 UTC) is shown in Fig. 6, along with comparable results for IIDA and the AIRMETs. As noted in Section 3, the LWC forecasts used for this comparison were based on the following criteria: (a) LWC > 0; and (b) temperature less than 0°C. The results in Fig. 6 indicate that the RUC LWC forecasts

have some skill; in fact the points for the RUC LWC statistics lie along the curves for IIDA. Although the LWC forecasts only covered a very small volume, they also only correctly classified about 20% of the MOG PIREPs.

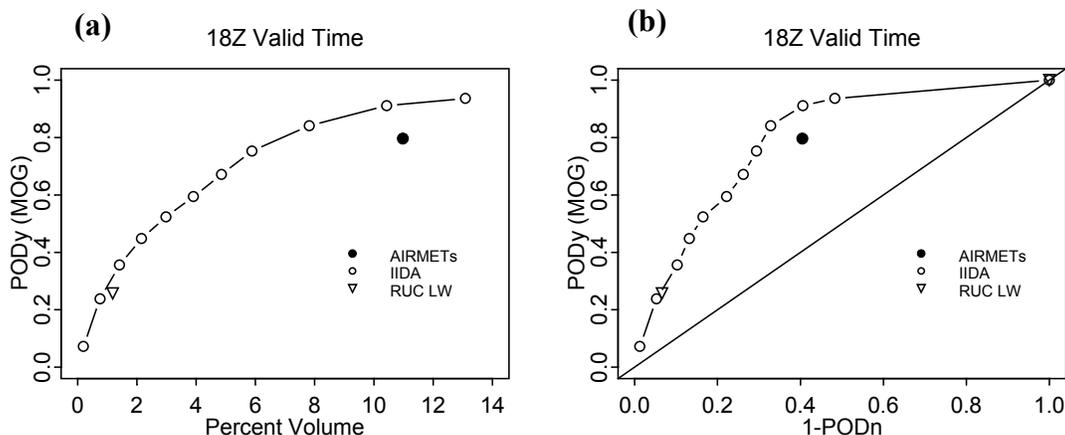


Figure 6. Verification curves for IIDA, with points representing statistics for 3-h RUC LWC forecasts and AIRMETs, all valid at 1800 UTC: (a) PODy(MOG) vs. % Volume; and (b) PODy(MOG) vs. 1-PODn.

7.6. Sensitivity to missing data

Computation of the IIDA diagnoses requires a variety of different types of data, including satellite, surface observations, and radar (McDonough and Bernstein 1999). The evaluation described so far has been based on an “operational” version of IIDA, which was run in real time during the operational period. In some cases, not all of the required datasets were available. Thus, the reported performance represents the level of quality that can be expected when IIDA is run operationally, and is occasionally subjected to missing data.

It was of interest to determine if the reported performance was adversely affected by occasional missing datasets. Thus, IIDA was run in a post-processing mode, with complete datasets for all cases, and the resulting verification statistics were compared to the statistics that have been presented here. The results (not shown) indicate very little change in overall performance associated with use of the complete datasets.

7.7. Verification of SLD field

As mentioned in Section 3, IIDA provides an indication of the potential for SLD icing, as well as the potential for General Icing conditions. Unfortunately, verification data are quite limited for SLD conditions. Brown et al. (1999) used the combination of MOG conditions and clear or mixed icing type reported by PIREPs as an indicator of SLD conditions. Here, severe PIREPs are used for most of the comparisons.

Table 8 shows comparisons of verification statistics for the IIDA SLD field, 3-h RUC LWC forecasts, and the IIDA General Icing field. The choice of threshold for the IIDA General Icing algorithm in Table 8 was based on the PODy for severe (SVR) PIREPs. In particular, the threshold was selected so that the PODy values for severe PIREPs are quite similar for the four categories of forecasts. In particular, the PODy for severe PIREPs is approximately 0.3 in all cases, with a threshold of 0.85 used for the IIDA General Icing case.

The results in Table 8 indicate that while the PODy values for severe PIREPs are similar among the different algorithms, the values of PODn, % Area, and % Volume are quite different. In particular, the SLD algorithm covers a much smaller area and volume of air-space than the other algorithms. This result is reflected in the Volume Efficiency statistics, which are much larger for the SLD algorithm than for the other algorithms. The SLD thresholds used in this analysis are quite small, with associated small PODy values for All and MOG PIREPs. Thus, the SLD field is not a good indicator of less severe icing, even though it is quite efficient at identifying severe conditions.

For comparison, consider the results presented in Table 3. While the AIRMETs and the IIDA General Icing algorithm capture a larger proportion of the severe icing reports, they do so at the expense of covering a much larger proportion of the area and volume. The SLD algorithm captures about 30% of the severe reports, but it does so very efficiently.

Table 8. Verification statistics for IIDA SLD field (with thresholds of 0.05 and 0.15), IIDA General Icing field (with a threshold of 0.85) and 3-h RUC LWC forecasts, for the period 20 January through 21 March 2000, for all valid times combined. TSS and Volume Efficiency are based on PODy(SVR).

Algorithm	PODy			PODn		TSS	% Area	% Vol	Vol Eff.
	All	MOG	SVR	No	CA				
SLD-0.05	0.12	0.14	0.32	0.96	0.99	0.28	2.6	0.3	106.7
SLD-0.15	0.08	0.10	0.27	0.98	1.00	0.25	1.8	0.2	135.0
IIDA-0.85	0.19	0.22	0.33	0.94	0.99	0.27	9.7	0.6	55.0
RUC-LW	0.22	0.26	0.30	0.94	1.00	0.24	13.0	1.1	27.3

8. RTVS comparisons

The IIDA diagnostic algorithm, NNICE, VVICE, and the AIRMETs, were evaluated from 1 January – 31 March 2000 and 2001 by the RTVS, with statistical output provided through the web-based interface (<http://www-ad.fsl.noaa.gov/afra/rtvs>; link to icing). Only a selection of the available results is presented here. The PIREPs used to evaluate the forecasts were located outside of convective regions. Only *Yes* PIREPs reporting MOG severity were included, and Clear Above PIREPs were not considered in the analyses. As in the results presented in Section 7, a verification window of 1 h after the issue time was applied for verification purposes. For

VVICE and NNICE, results for the 0-h lead-time are shown. All valid times, including both day-time and night-time hours, were combined to compute the statistics.

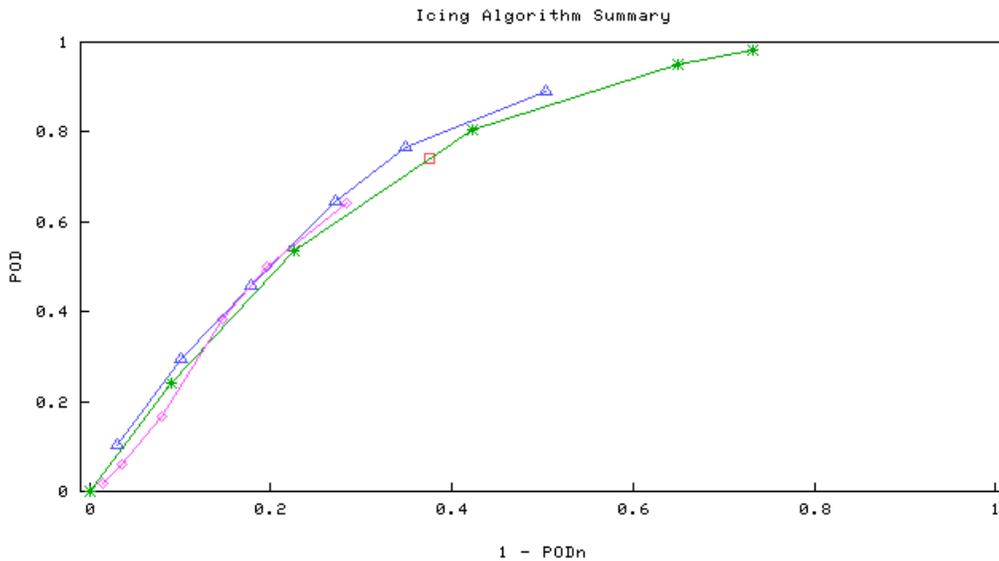
8.1. Overall results

Algorithm performance for IIDA, NNICE, VVICE, and the AIRMETs for the period from 1 January – 30 March 2000 and 2001 (hereafter denoted as Eval2000 and Eval2001) is summarized in Figs. 7-10. In these figures, each line on the plots represents one of the three algorithms, and the AIRMETs are represented by a single point. Each symbol on a line indicates a unique threshold used to define the icing potential. The values of POD_y vs. 1-POD_n are shown in Fig. 7 for Eval2000 and Fig. 8 for Eval2001. The plots of POD_y vs. % Volume are shown in Fig. 9 for Eval2000 and Fig. 10 for Eval2001. In all cases, the best statistical scores should approach the values in the upper left hand corner of the plot, where POD_y approaches 1.0, 1-POD_n approaches 0, and % Volume is minimized.

Statistically, the overall differences between the forecasts are small. However, the IIDA algorithm does indicate an improvement over the other forecasts, which is particularly evident in Eval2001, as indicated by higher POD_y(MOG) values at all thresholds for a given 1-POD_n value (Fig. 8). When the % Volume is considered (Figs. 9 and 10), IIDA and NNICE remain the top performers. For instance, for all values of % Volume, the POD_y(MOG) values are larger for IIDA than for the other forecasts, particularly between volumes of 5 - 15%. It should be noted that the AIRMET volumes are designed to contain moving areas of icing over a 6-h period. Therefore, the actual volume intended to be valid for the AIRMETs at any particular time may be smaller than the value indicated in this evaluation. However, that information is difficult to precisely obtain from the text-based AIRMET message.

8.2. Weekly results

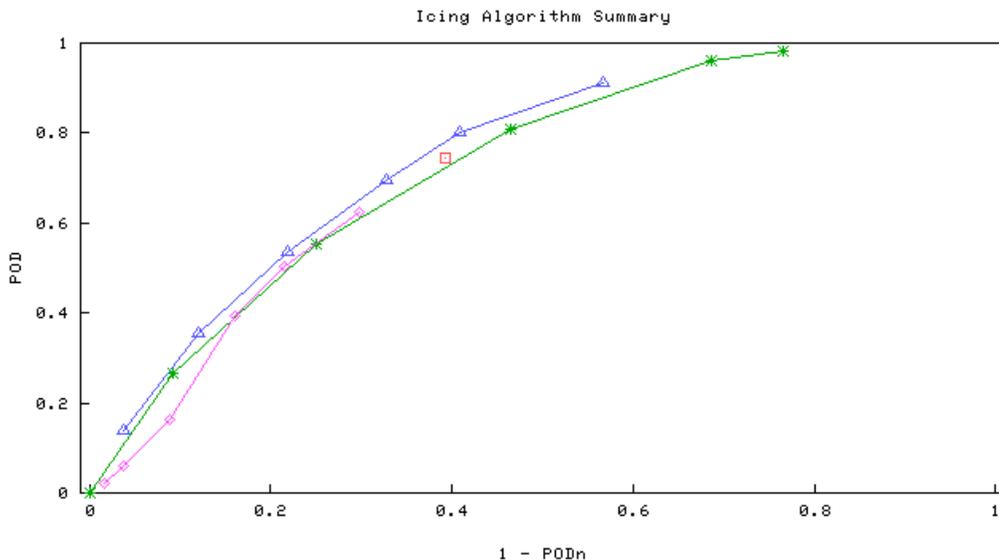
Time series plots of POD_y(MOG), POD_n, and % Volume for Eval2000 and Eval2001 are shown in Figs. 11-16. Each line on the plots represents a statistic for one of the forecasts at a specific threshold. With the exception of VVICE, the thresholds chosen for display were selected so that the POD_y(MOG) values for all the forecasts were similar. Further testing is needed to determine a more representative threshold for VVICE. Each point on the lines represents a statistic computed over a 7-day period during Eval2000 and Eval2001.



Generated on 13 Jul 2001 by NOAA/FSL/RTVS

- 1 - PODn vs. MOG PODy for airmets verified by ncpireps, National Scale, FH=0h (2000-01-01)
- ▲— 1 - PODn vs. MOG PODy for IIDA verified by ncpireps, National Scale, FH=0h (2000-01-01)
- *— 1 - PODn vs. MOG PODy for NNICE verified by ncpireps, National Scale, FH=0h (2000-01-01)
- ◇— 1 - PODn vs. MOG PODy for VVICE verified by ncpireps, National Scale, FH=0h (2000-01-01)

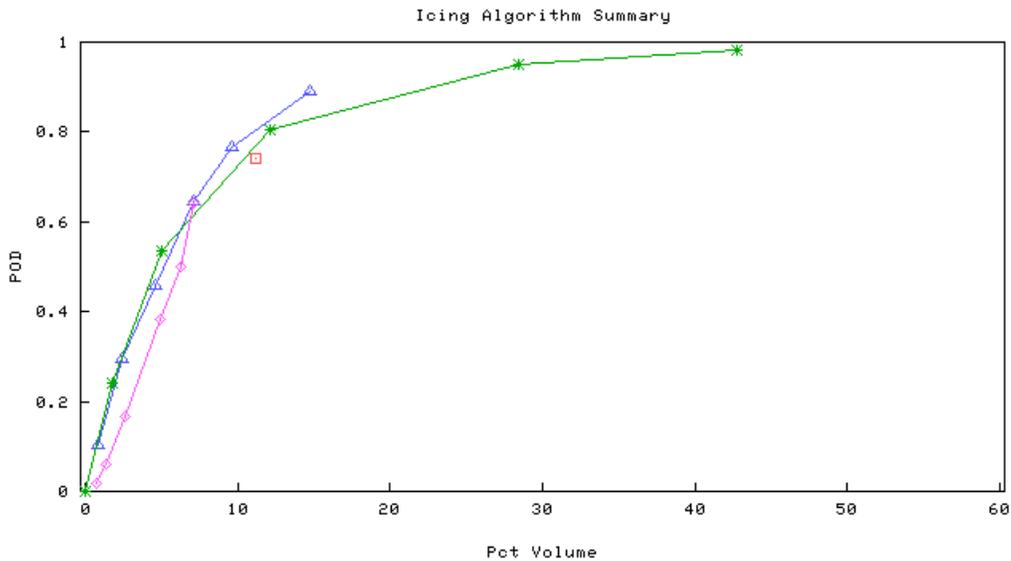
Figure 7. Algorithm summary plots of PODy(MOG) vs. 1-PODn for IIDA (triangle), NNICE (*), VVICE (diamond), Icing AIRMETS (square), from 1 January – 31 March 2000 (Eval2000). Each dot on a line represents a unique threshold. Thresholds are 0.02, 0.15, 0.25, 0.45, 0.65, and 0.85 for IIDA; 0.5, 0.9, 2.5, 3.5, 4.0, and 5.0 for NNICE; and 0.01, 0.05, 0.08, 0.12, 0.16, and 0.20 for VVICE.



Generated on 13 Jul 2001 by NOAA/FSL/RTVS

- 1 - PODn vs. MOG PODy for airmets verified by ncpireps, National Scale, FH=0h (2001-01-01)
- ▲— 1 - PODn vs. MOG PODy for IIDA verified by ncpireps, National Scale, FH=0h (2001-01-01)
- *— 1 - PODn vs. MOG PODy for NNICE verified by ncpireps, National Scale, FH=0h (2001-01-01)
- ◇— 1 - PODn vs. MOG PODy for VVICE verified by ncpireps, National Scale, FH=0h (2001-01-01)

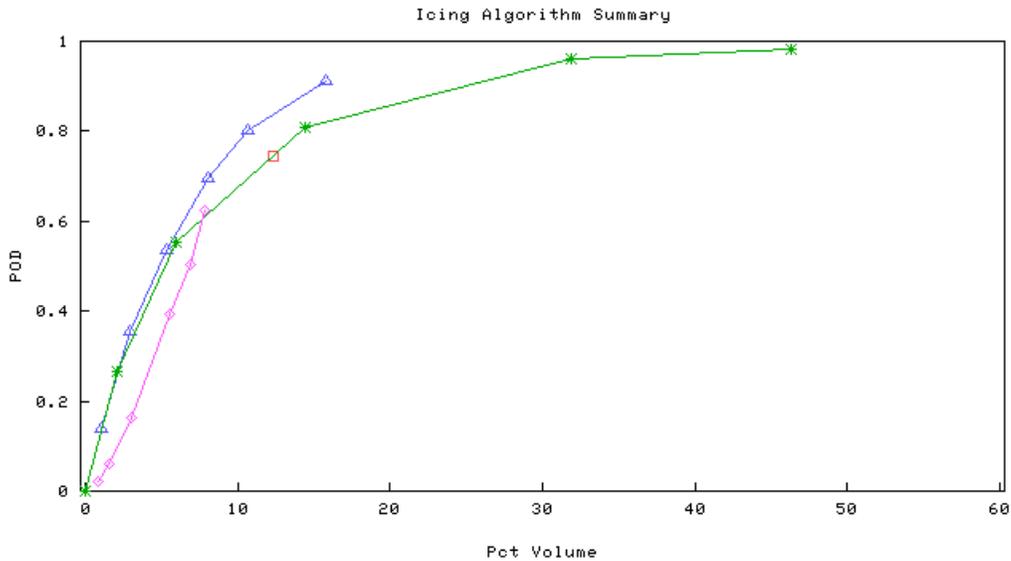
Figure 8. As in Fig. 7, for the period 1 January – 31 March 2001 (Eval2001).



Generated on 13 Jul 2001 by NOAA/FSL/RTVS

- Pct Volume vs. MOG PODy for airmets verified by ncpireps, National Scale, FH=0h (2000-01-01)
- △— Pct Volume vs. MOG PODy for IIDA verified by ncpireps, National Scale, FH=0h (2000-01-01)
- *— Pct Volume vs. MOG PODy for NNICE verified by ncpireps, National Scale, FH=0h (2000-01-01)
- ◇— Pct Volume vs. MOG PODy for VVICE verified by ncpireps, National Scale, FH=0h (2000-01-01)

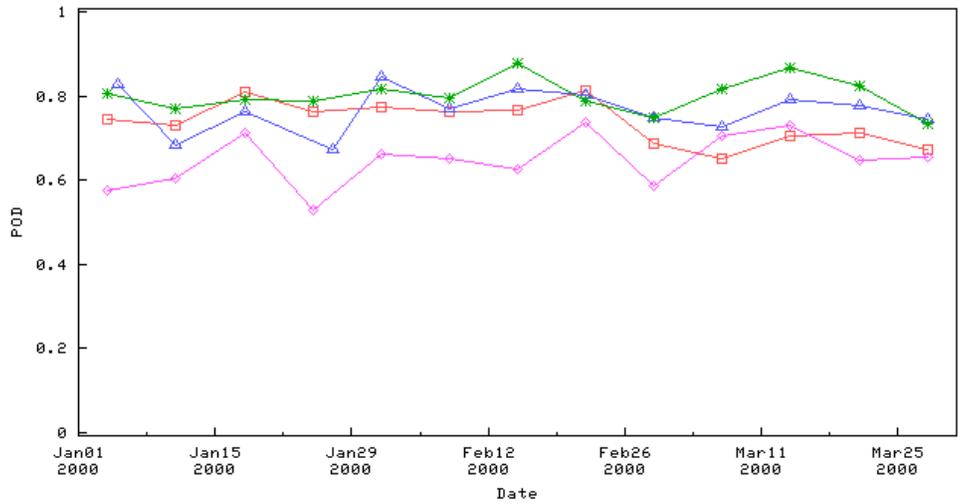
Figure 9. As in Fig. 7, for PODy(MOG) vs. % Volume, for Eval2000.



Generated on 13 Jul 2001 by NOAA/FSL/RTVS

- Pct Volume vs. MOG PODy for airmets verified by ncpireps, National Scale, FH=0h (2001-01-01)
- △— Pct Volume vs. MOG PODy for IIDA verified by ncpireps, National Scale, FH=0h (2001-01-01)
- *— Pct Volume vs. MOG PODy for NNICE verified by ncpireps, National Scale, FH=0h (2001-01-01)
- ◇— Pct Volume vs. MOG PODy for VVICE verified by ncpireps, National Scale, FH=0h (2001-01-01)

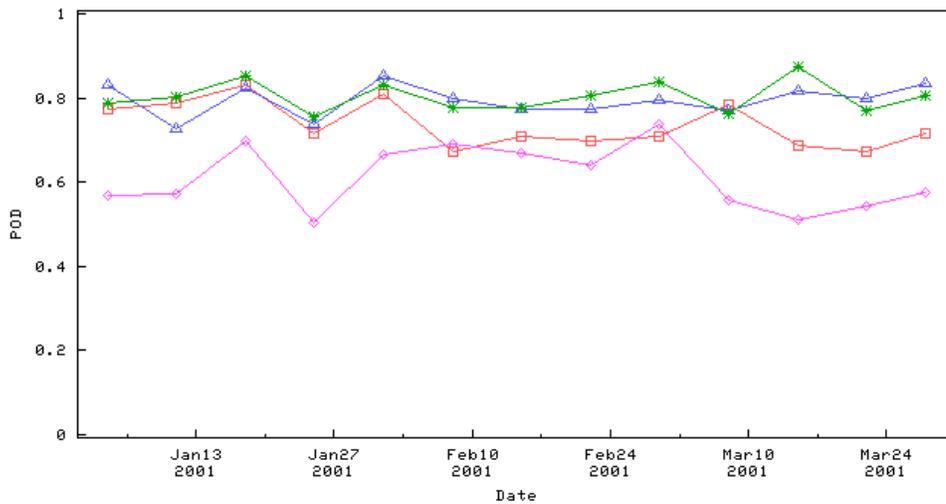
Figure 10. As in Fig. 7, for PODy(MOG) vs. % Volume, for Eval2001.



Generated on 13 Jul 2001 by NOAA/FSL/RTVS

- MOG PODy for airmets verified by ncpireps, National Scale, FH=0h (2000-01-01)
- △- MOG PODy for IIDA (0.15) verified by ncpireps, National Scale, FH=0h (2000-01-01)
- * MOG PODy for NNICE (2.5) verified by ncpireps, National Scale, FH=0h (2000-01-01)
- ◇- MOG PODy for VVICE (0.01) verified by ncpireps, National Scale, FH=0h (2000-01-01)

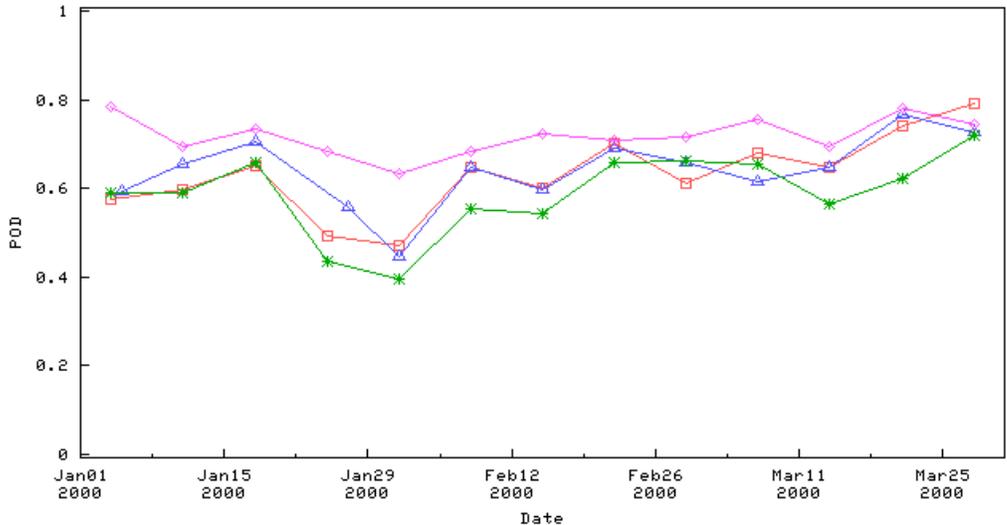
Figure 11. Time series plots of weekly PODy(MOG) for IIDA (triangle), NNICE (*), VVICE (diamond), Icing AIRMETS (square), from 1 January – 31 March 2000 (Eval2000). Each line represents a unique threshold for IIDA (0.15), NNICE (2.5), VVICE (0.01), and AIRMETS (no threshold) and each dot is a PODy(MOG) value computed over a 7-day period.



Generated on 13 Jul 2001 by NOAA/FSL/RTVS

- MOG PODy for airmets verified by ncpireps, National Scale, FH=0h (2001-01-01)
- △- MOG PODy for IIDA (0.15) verified by ncpireps, National Scale, FH=0h (2001-01-01)
- * MOG PODy for NNICE (2.5) verified by ncpireps, National Scale, FH=0h (2001-01-01)
- ◇- MOG PODy for VVICE (0.01) verified by ncpireps, National Scale, FH=0h (2001-01-01)

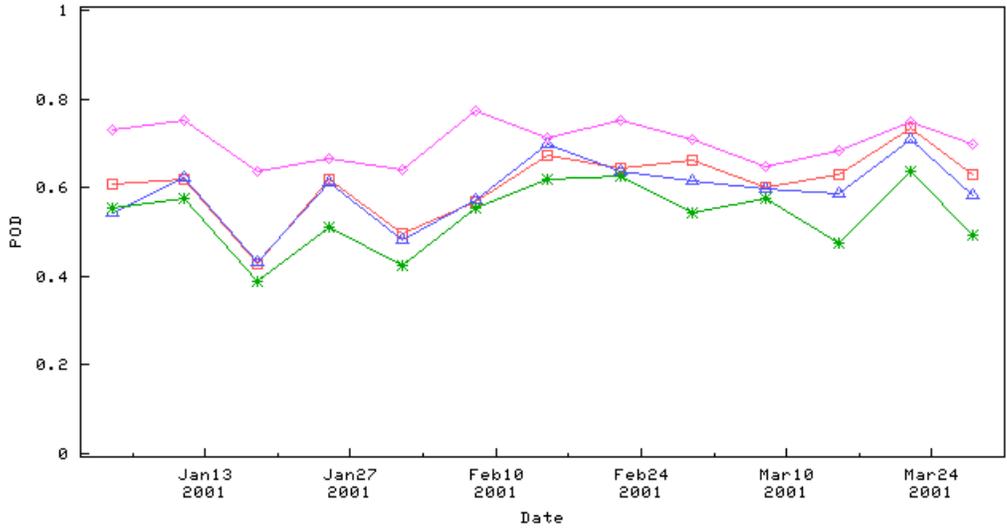
Figure 12. As in Fig. 11, for the period 1 January – 31 March 2001 (Eval2001).



Generated on 13 Jul 2001 by NOAA/FSL/RTVS

- PODn for airmets verified by nopireps, National Scale, FH=0h (2000-01-01)
- △— PODn for IIDA (0.15) verified by nopireps, National Scale, FH=0h (2000-01-01)
- *— PODn for NNICE (2.5) verified by nopireps, National Scale, FH=0h (2000-01-01)
- ◇— PODn for VVICE (0.01) verified by nopireps, National Scale, FH=0h (2000-01-01)

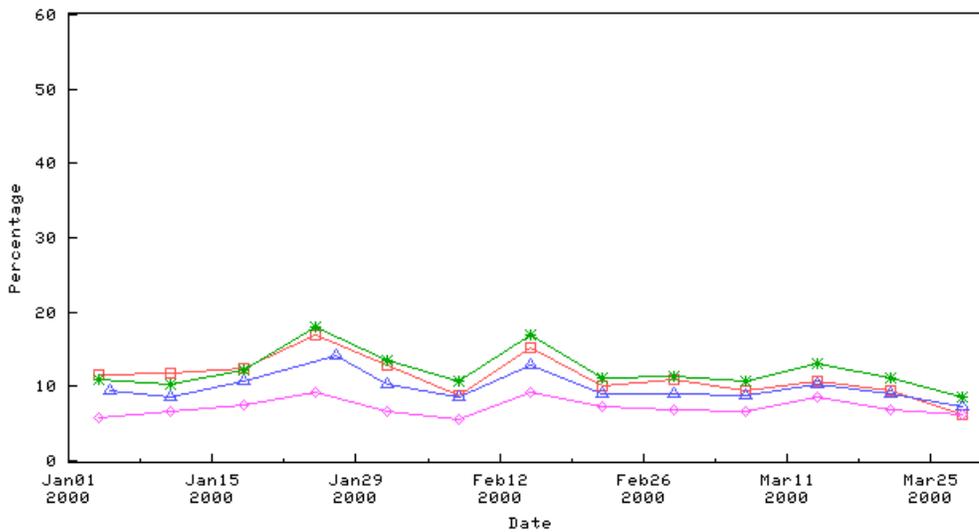
Figure 13. As in Fig. 11, for weekly time series of PODn, for Eval2000.



Generated on 13 Jul 2001 by NOAA/FSL/RTVS

- PODn for airmets verified by nopireps, National Scale, FH=0h (2001-01-01)
- △— PODn for IIDA (0.15) verified by nopireps, National Scale, FH=0h (2001-01-01)
- *— PODn for NNICE (2.5) verified by nopireps, National Scale, FH=0h (2001-01-01)
- ◇— PODn for VVICE (0.01) verified by nopireps, National Scale, FH=0h (2001-01-01)

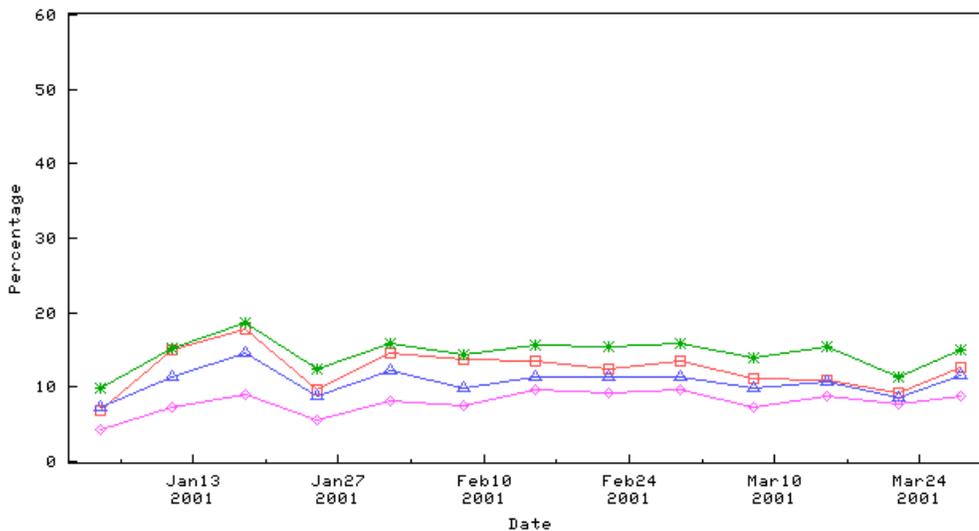
Figure 14. As in Fig. 11, for weekly time series of PODn, for Eval2001.



Generated on 13 Jul 2001 by NOAA/FSL/RTVS

- Pct Volume for airmets verified by nopireps, National Scale, FH=0h (2000-01-01)
- ▲— Pct Volume for IIDA (0.15) verified by nopireps, National Scale, FH=0h (2000-01-01)
- *— Pct Volume for NNICE (2.5) verified by nopireps, National Scale, FH=0h (2000-01-01)
- ◇— Pct Volume for VVICE (0.01) verified by nopireps, National Scale, FH=0h (2000-01-01)

Figure 15. As in Fig. 11, for weekly time series of % Volume during Eval2000.



Generated on 13 Jul 2001 by NOAA/FSL/RTVS

- Pct Volume for airmets verified by nopireps, National Scale, FH=0h (2001-01-01)
- ▲— Pct Volume for IIDA (0.15) verified by nopireps, National Scale, FH=0h (2001-01-01)
- *— Pct Volume for NNICE (2.5) verified by nopireps, National Scale, FH=0h (2001-01-01)
- ◇— Pct Volume for VVICE (0.01) verified by nopireps, National Scale, FH=0h (2001-01-01)

Figure 16. As in Fig. 11, for weekly time series of % Volume during Eval2001.

Examination of the time series of weekly PODy(MOG) values for IIDA, NNICE, VVICE, and the AIRMETs for Eval2000 (Fig. 11) and Eval2001 (Fig. 12) indicates that the best statistics are produced by IIDA, NNICE, and the AIRMETs. The PODy(MOG) values for VVICE are well below the lines for the other forecasts, even though the VVICE threshold is quite small. Comparisons among the forecasts for Eval2001 (Fig. 12) indicate that the PODy(MOG) values for the AIRMETs and VVICE decrease somewhat as March approaches, while the PODy(MOG) values for IIDA and NNICE remain high. When PODn is considered (Fig. 13 for Eval2000; Fig. 14 for Eval2001), the IIDA and AIRMET time series are nearly identical throughout the period, while the line for NNICE is below the others. The % Volume curve (Fig. 15 for Eval2000; Fig. 16 for Eval2001) is lowest for VVICE and IIDA, with larger values of % Volume recorded for NNICE and the AIRMETs in most weeks. In general, the quality of the IIDA diagnoses, as measured by PODy(MOG), PODn, and % Volume, appears to be relatively consistent throughout January and March, while somewhat larger variations in the verification statistics for NNICE, VVICE, and the AIRMETs are noted.

8.3. Results by height

Height series plots of PODy(MOG) and PODn for Eval2000 and Eval2001 are shown in Figs. 17-20. Heights above 30,000 ft. were excluded from this evaluation. Each line on the plots represents the statistic for one of the forecasts at a specific threshold. These thresholds correspond to those presented on the time series plots. Each symbol on a line represents a statistical value computed over the entire period from 1 January – 31 March at a specific height.

As shown in Figs. 17 (Eval2000) and 18 (Eval2001), overall, the largest PODy(MOG) values for the forecasts are associated with altitudes below 15,000 ft, with a steady decrease in PODy(MOG) as the height increases to 30,000 ft. With the exception of the lowest two height levels, IIDA generally performs better than the other forecasts at altitudes below 15,000 ft. However, above 15,000 ft, the PODy(MOG) values for IIDA drop off somewhat more quickly than the values for the other forecasts. The AIRMETs and NNICE perform somewhat better than the other algorithms at heights above 15,000 ft. When PODn is considered, IIDA performance is somewhat better than the performance of the AIRMETs and NNICE at almost all altitudes, as shown by the curves in Figs. 19 and 20 (Eval2000 and Eval2001, respectively). In particular, the PODn line for IIDA is consistently located further to the right of the other lines at most heights. The large PODn values for VVICE are a response to the low PODy(MOG) values shown in the previous diagrams.

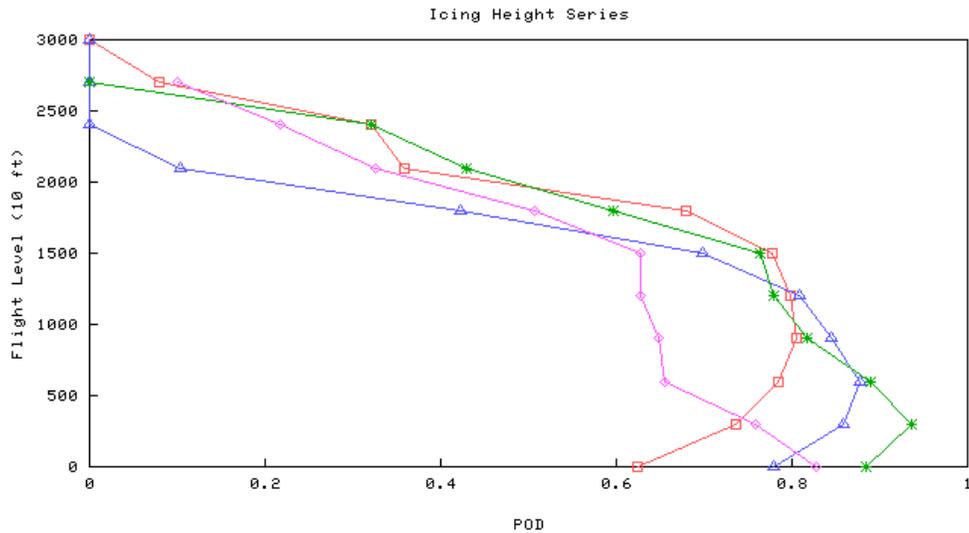


Figure 17. Height series plots of PODy(MOG) for IIDA (triangle), NNICE (“*”), VVICE (diamond), and Icing AIRMETs (square), from 1 January – 31 March 2000 (Eval2000). Each line represents a unique threshold for IIDA (0.15), NNICE (2.5), VVICE (0.01), and AIRMETs (no threshold). The statistical values were computed over the entire 3-month period.

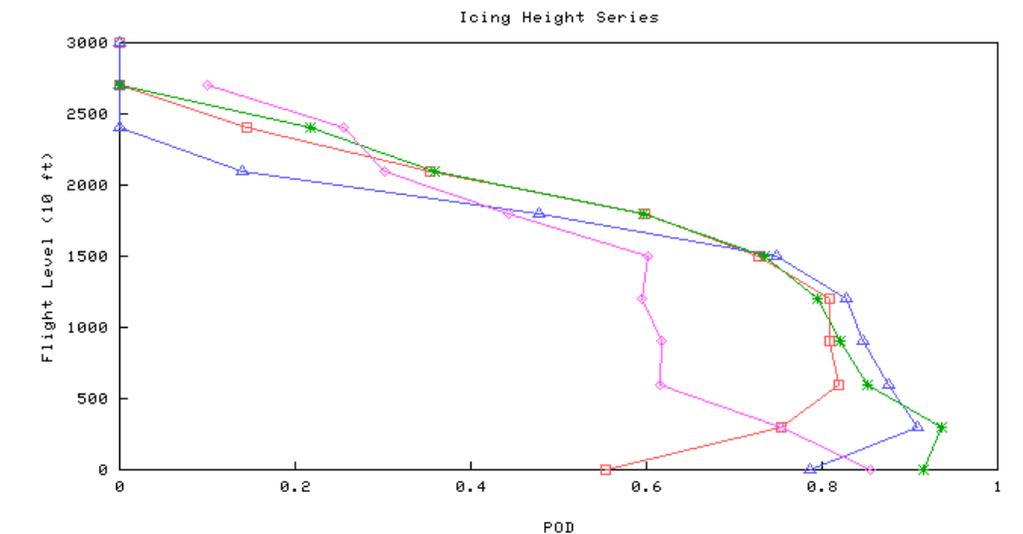


Figure 18. As in Fig. 17, for the period 1 January – 31 March 2001 (Eval2001).

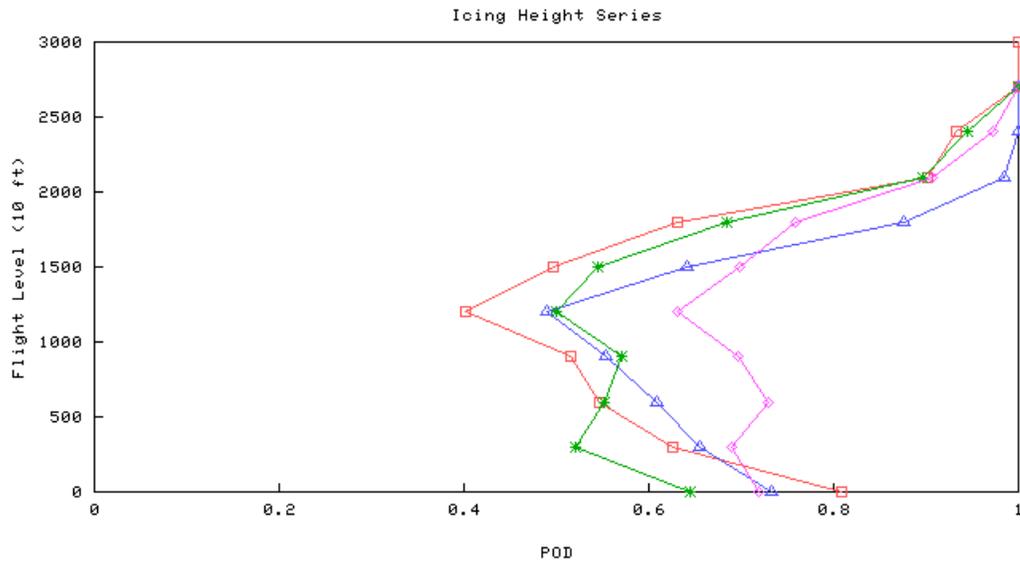


Figure 19. As in Fig. 17, for PODn height series during Eval2000.

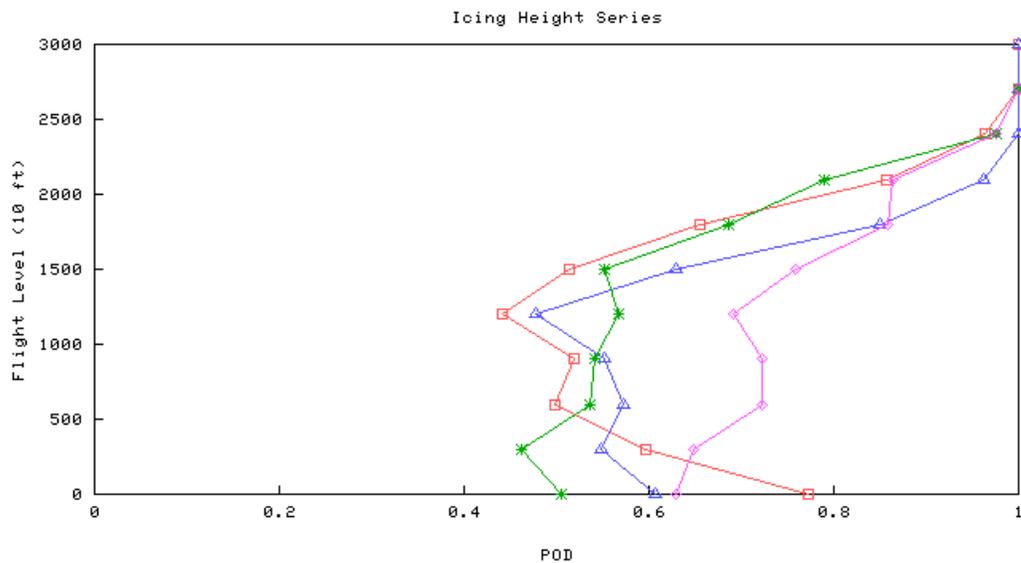


Figure 20. As in Fig. 17, for PODn height series during Eval2001.

9. Summary and conclusions

The analyses presented in this report provide a comprehensive evaluation of IIDA over two seasons. The results thus are representative of IIDA's performance in a variety of conditions. The IIDA diagnoses were compared to operational icing forecasts (AIRMETs) and to forecasts based on the output of two algorithms – NNICE and VVICE – which were developed at the Aviation Weather Center. In addition, some comparisons considered RUC LWC forecasts. Although the AIRMETs are quite different from IIDA in many ways (e.g., they are limited to a volume that can be defined in a textual message, and they are intended to depict icing conditions over a 6-h period), they were included in this evaluation because they represent the operational standard that is currently available to users. Nevertheless, as noted earlier, it is important to use care in evaluating performance differences between the AIRMETs and the automated, gridded algorithms, and in comparing their strengths and weaknesses.

Comparison of previous evaluations of IIDA to more recent evaluations suggests that the quality of the IIDA diagnoses has improved somewhat; in particular, differences between the verification statistics for IIDA and the other forecasts are somewhat greater in recent years than was the case for earlier versions of the algorithm. A regional evaluation of IIDA suggests that IIDA is best at detecting icing conditions in the Great Lakes region and the Northeast, and somewhat less capable in the South; similar characteristics were noted for the AIRMETs (Kane et al. 2000).

Results of the in-depth evaluation of IIDA for winter 2000 and the RTVS evaluations for winters 2000 and 2001 indicate that IIDA is relatively efficient at detecting icing conditions, with a PODy(MOG) of 0.75-0.85, and a corresponding PODn value between 0.6 and 0.7, depending on the IIDA threshold used to define the forecasts. Moreover, a relatively small percentage (5-8%) of the total airspace is impacted by the IIDA diagnoses. AIRMETs captured similar proportions of *Yes* reports, while capturing a somewhat smaller proportion of *No* reports. AIRMETs were somewhat more efficient in terms of the area covered by the forecasts and somewhat less efficient in terms of the volume of airspace covered; however, this result is at least partially related to constraints on the form of the AIRMETs. Overall, RTVS verification analyses – which considered both night-time and day-time icing diagnoses – indicate that IIDA performance is somewhat better than the performance of NNICE and VVICE, and the AIRMETs, in comparisons of PODy, % Volume, and PODn. The results of the ROC analyses indicate that the IIDA diagnoses are skillful, as measured by their ability to discriminate between *Yes* and *No* icing situations.

The results also indicate that PODy and PODn values for all of the forecasts and algorithms are somewhat variable from time to time. For example, for IIDA with a threshold of 0.25, PODy(MOG) for individual diagnoses ranges from about 0.2 to 1.0, with the middle 50% of values between 0.6 and 0.9. However, the volume covered by the IIDA diagnoses is quite consistent from time-to-time. RTVS evaluations of variations in the statistics from week-to-week suggest that verification statistics for the three algorithms and the AIRMETs exhibit similar variations. PODy values for the AIRMETs and VVICE decreased somewhat toward the end of the winter 2001 season, while the PODy values for IIDA and NNICE remained somewhat more consistent throughout the season.

Other results of the study include the following items:

- IIDA performs fairly well as a persistence forecast out to about three hours. After that period, they are generally out-performed by the AIRMETs.
- IIDA performance is best at lower altitudes (15,000 ft and below) and IIDA is better than the other algorithms and AIRMETs at capturing No-icing conditions at all altitudes, while still maintaining a good PODy value.
- The IIDA SLD field is very efficient at capturing PIREPs reporting severe icing conditions. Although the PODy for severe reports is about 0.3, the diagnoses cover a very small volume, in comparison to the IIDA General Icing field, the AIRMETs, and the LWC forecasts.

In summary, IIDA is quite skillful at diagnosing General Icing conditions, and the SLD algorithm is very efficient at detecting severe icing situations. IIDA performs well in discriminating between *Yes* and *No* icing conditions, and is efficient in limiting the airspace warned.

Acknowledgments

This research is in response to requirements and funding by the Federal Aviation Administration. The views expressed are those of the authors and do not necessarily represent the official policy and position of the U.S. Government.

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 RUC-2. *Preprints, 8th Conference on Aviation, Range, and Aerospace Meteorology*, Dallas, TX, 10-15 January, American Meteorological Society (Boston), 486-490.

Brown, B.G., 1996: Verification of in-flight icing forecasts: Methods and issues. *Proceedings, FAA International Conference on Aircraft In-flight Icing*, Report No. DOT/FAA/AR-96/81, II, 319-330.

Brown, B.G., and J.L. Mahoney, 2000: Quality Assessment Plan for the Integrated Icing Diagnostic Algorithm. Prepared by the Quality Assessment Group, Aviation Gridded Forecast System Product Development Team, available from B.G. Brown (bgb@ucar.edu), 7 pp.

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*, Orlando, FL, 11-15 September, American Meteorological Society (Boston), 393-398.

Brown, B.G., G. Thompson, R.T. Brintjes, R. Bullock, and T. Kane, 1997: Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Wea. 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*, Dallas, TX, 10-15 January, American Meteorological Society (Boston), 48-52.

Brown, B.G., J.L. Mahoney, R. Bullock, J. Henderson, and T.L. Fowler, 2000: Turbulence Algorithm Intercomparison: 1998-1999 Initial Results. FAA Turbulence Product Development Team Report to FAA Aviation Weather Research Program. NOAA Technical Memorandum OAR FSL-25, NOAA Department of Commerce, 63 pp.

Doswell, C.A., III, R. Davies-Jones, and D.L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576-585.

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*, Asheville, NC, 8-11 May, American Meteorological Society (Boston), 46-49.

Kane, T.L., B.G. Brown, and B.C. Bernstein, 2000: Regional icing algorithm performance analysis. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 September, American Meteorological Society (Boston), 270-273.

Mahoney, J.L., J.K. Henderson, and P.A. Miller, 1997: A Description of the Forecast Systems Laboratory's Real-Time Verification System (RTVS). *Preprints, 7th Conference on Aviation, Range, and Aerospace Meteorology*, Long Beach, American Meteorological Society (Boston), J26-J31.

Mason, I., 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 aircraft icing diagnosis. *Preprints, 8th Conference on Aviation, Range, and Aerospace Meteorology*, Dallas, TX, 10-15 January, American Meteorological Society (Boston), 467-471.

Murphy, A.H. and R.L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

NWS, 1991: National Weather Service Operations Manual, D-22. National Weather Service. (Available at the web site <http://www.nws.noaa.gov>).

Reisner, J., R.M. Rasmussen, and R.T. Brientjes, 1998: Explicit forecasting of supercooled liquid water in winter storms using the MM5 mesoscale model. *Quarterly Journal of the Royal Meteorological Society*, **124**, 1071-1107.

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*, Long Beach, American Meteorological Society (Boston), 7-10.

Thompson, G., R.T. Brientjes, B.G. Brown, and F. Hage, 1997: Intercomparison of in-flight icing algorithms. Part I: WISP94 Real-time Icing Prediction and Evaluation Program. *Wea. Forecasting*, **12**, 878-889.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

Appendices

The appendices are attached as PDF files

Appendix 1: Brown et al. 1997

Appendix 2: Brown et al. 1999

Appendix 3: Kane et al. 2000