

# **Forecast Icing Potential (FIP): Quality Assessment Report**

**B.G. Brown<sup>1,2</sup>, T.L. Fowler<sup>1</sup>, J.L. Mahoney<sup>2,3</sup>, M. Chapman<sup>1</sup>, R. Bullock<sup>1</sup>,  
and J. Henderson<sup>3</sup>**

**Quality Assessment Product Development Team  
Aviation Weather Research Program**

**24 April 2003**

---

<sup>1</sup> Research Applications Program, National Center for Atmospheric Research, PO Box 3000, Boulder CO 80307-3000.

<sup>2</sup> Contact information: [bgb@ucar.edu](mailto:bgb@ucar.edu) (Barbara Brown) or [jennifer.mahoney@fsl.noaa.gov](mailto:jennifer.mahoney@fsl.noaa.gov) (Jennifer Mahoney).

<sup>3</sup> Forecast Systems Laboratory, NOAA, 325 Broadway, Boulder CO 80303.

# Contents

<b>Section</b>	<b>Page</b>
<b>Summary</b> .....	iii
<b>1. Introduction</b> .....	1
<b>2. Approach</b> .....	1
<b>3. Algorithms and forecasts</b> .....	3
<b>4. Data: Model output and PIREPs</b> .....	5
<b>5. Methods</b> .....	6
5.1 Matching methods .....	6
5.2 Statistical verification methods .....	6
5.3 Stratifications .....	9
5.4 Time periods.....	10
<b>6. Results</b> .....	10
6.1 FIP version comparisons .....	11
6.2 Model version comparisons .....	11
6.3 Overall results and results by lead time .....	17
6.4 Variability in verification statistics .....	21
6.5 Comparisons by altitude .....	23
6.6 Monthly time series .....	28
6.7 PIREP evaluation .....	32
<b>7. Conclusions and discussion</b> .....	32
<b>Acknowledgments</b> .....	33
<b>References</b> .....	34

# **Forecast Icing Potential (FIP): Quality Assessment Report**

**B.G. Brown, T.L. Fowler, J.L. Mahoney, M. Chapman, R. Bullock,  
and J. Henderson**

**Quality Assessment Product Development Team  
Aviation Weather Research Program**

**April 2003**

## **Summary**

This report summarizes assessments of the quality of forecasts of icing conditions produced by the Forecast Icing Potential (FIP) algorithm. FIP was developed by the Inflight Icing Product Development Team of the Federal Aviation Administration's Aviation Weather Research Program (FAA/AWRP), and is currently being considered for transition to an operational product through the Aviation Weather Technology Transfer (AWTT) process.

The performance of FIP has been evaluated over several seasons by the AWRP Quality Assessment Group (now, the Quality Assessment Product Development Team). Ongoing real-time and long-term evaluations are available on the Real-Time Verification System (RTVS; <http://www-ad.fsl.noaa.gov/fvb/rtvs/icing/index.html>), developed by the National Oceanic and Atmospheric Administration's Forecast Systems Laboratory (NOAA/FSL). In addition, in-depth analyses of the algorithm's performance have been undertaken at the Research Applications Program at the National Center for Atmospheric Research (NCAR/RAP). Both the real-time and post-analysis evaluations have involved meteorological/statistical verification of the icing forecasts.

This report concentrates on results of objective evaluations of FIP during fall 2001, winter/spring 2002, and fall 2002. The evaluation considers a recent enhancement to the algorithm, as well as the impacts of a new version of the numerical weather prediction model that is the basis of the FIP forecasts. Trends and seasonal variations in the verification statistics since April 2001 are also considered using results from RTVS.

The forecasts were verified using Yes and No icing observations from pilot reports (PIREPs) indicating either "moderate or greater" icing severity or "no icing." FIP predictions were evaluated as Yes/No icing forecasts by applying a threshold to convert the algorithm output to a Yes or No value. A variety of thresholds were applied to the algorithm output, in order to examine the full range of FIP performance characteristics. The verification analyses were primarily based on the algorithms' ability to discriminate between Yes and No observations, as well as the extent of their coverage. In addition,

forecasts based on Airmens' Meteorological Advisories (AIRMETs), the operational forecasts issued by the National Weather Service Aviation Weather Center (AWC), were evaluated to provide a standard of comparison. Several thousand individual FIP forecasts were considered in this evaluation. The number of Yes (No) PIREPs considered in the evaluation ranged from 1,000 to 11,000 (2,500 to 16,000) depending on the forecast lead time and season.

Results of the evaluation indicate that FIP is skillful at discriminating between Yes and No icing conditions. FIP also provides relatively efficient forecasts, covering comparatively small volumes for a given icing detection rate. Using a threshold of 0.05, FIP correctly classifies 84% of the Yes PIREPs and 69% of the No PIREPs, while covering approximately 12% of the airspace volume over the CONUS. The forecast quality is relatively insensitive to lead time, and is maintained up to about 24,000 ft. Detection rates and volumes covered vary from day-to-day, with volume coverage somewhat more consistent from day to day than the other statistics. Trends in FIP performance over the last two years indicate that FIP maintains its forecasting capability through the summer months somewhat better than the AIRMETs; this result is likely due to the fact that in the summer operational icing advisories are often incorporated into the Convective Significant Meteorological Advisory (c-SIGMETs), also issued by the AWC.

The operational numerical weather prediction model that is used by FIP [i.e., the Rapid Update Cycle (RUC)] evolved from a 40-km horizontal resolution to a 20-km resolution in mid-April 2002. Since the 20-km version is the new operational standard, it was important to evaluate changes in FIP performance associated with the new model resolution. Results of a comparison of verification results based on both versions of the model indicate only a small variation in the verification statistics that can be associated with the change to the finer-resolution model. In many respects, algorithm performance apparently improved with implementation of the new version of the model. In particular, the FIP forecasts based on the 20-km RUC are better able to discriminate between Yes and No PIREPs.

In summary, evaluations of FIP demonstrate that it is a skillful forecasting algorithm that is generally able to discriminate between Yes and No icing PIREPs, with relatively efficient forecasts. The quality of FIP forecasts is relatively insensitive to variations in the PIREPs used for the analyses and does not degrade with altitude below 24,000 ft.

## 1. Introduction

This report summarizes basic results of an evaluation of the forecasting capability of the Forecast Icing Potential (FIP) algorithm. This algorithm is under consideration for transition from experimental to operational through the Aviation Weather Technology Transfer (AWTT) process. FIP was designed to predict in-flight icing conditions over the continental U.S. (CONUS). It has been evaluated over several seasons by the Quality Assessment Product Development Team [QAPDT; formerly the Quality Assessment Group (QAG)] of the Federal Aviation Administration's Aviation Weather Research Program (FAA/AWRP). Long-term and real-time verification statistics on the performance of FIP are available on the Real-Time Verification System (RTVS) developed by the National Oceanic and Atmospheric Administration's Forecast Systems Laboratory (NOAA/FSL) (Mahoney et al. 1997, 2002). In addition to the real-time analyses, FIP forecasts were evaluated in-depth in post-analysis. The analyses presented in this report focus primarily on in-depth evaluations of forecasts from fall 2001, winter/spring 2002, and fall 2002. Long-term performance trends from RTVS are also considered.

During its development, FIP was known as the Integrated Icing Forecast Algorithm (IIFA). Performance of IIFA forecasts was initially considered in a report prepared for the transition of FIP/IIFA to the experimental stage of the AWTT process (Brown et al. 2001a).

In most of the analyses included in this report, FIP performance is compared to the forecasting performance of the operational icing forecasts. These forecasts are issued by the NWS National Centers for Environmental Prediction Aviation Weather Center (NWS/NCEP/AWC). It is important to emphasize, however, that the goal of this report is **not** to provide a comprehensive evaluation of the AIRMETs (see Section 2).

The report is organized as follows. The study approach is presented in Section 2. Section 3 briefly describes the algorithms and forecasts that were included in the evaluation, and the data that were utilized are discussed in Section 4. The verification methods are described in Section 5, and results of the study are presented in Section 6. Finally, Section 7 includes the conclusions and discussion.

## 2. Approach

The FIP predictions are based on data from the RUC (Rapid Update Cycle) model (Benjamin et al. 1998), with model output obtained from the NCEP Environmental Modeling Center (NCEP/EMC). Model forecasts with lead times of 3, 6, 9, and 12 h, and valid times between 1200 and 0300 UTC, were included in the verification study. In addition, the icing Airmens' Meteorological Advisories (AIRMETs),

which are the operational icing forecasts issued by forecasters at the AWC, were included for comparison purposes (i.e., as noted earlier, this report is not intended as an evaluation of icing AIRMETs).

In mid-April 2002 a new version of the RUC model became operational. Two of the major differences between the old and new versions of the model are (i) increased horizontal resolution (from 40 to 20 km) and (ii) implementation of a new microphysics parameterization scheme. Because the 20-km version of the model is the new standard, which will be employed by the operational version of FIP, it is important to understand the sensitivity of FIP performance to this change. This report includes some comparisons of FIP performance on the 20-km vs. 40-km versions of RUC for a short period in early April 2002. In addition, results for the fall of 2002 (when the new version of RUC was the operational standard) are compared to results for fall of 2001 (when the old version of RUC was the standard). Although a direct comparison of the results for these two periods is not appropriate due to differences in the weather in the two periods, it is valuable to obtain a sense of whether major differences can be observed. To help sort out whether any observed differences are simply due to grid resolution changes, as opposed to fundamental changes in model performance, the results for both periods are also compared to results based on a 40-km degraded version of the 20-km model.

The verification approach applied in this study is identical to the approach taken in previous studies (e.g., Brown et al. 2001b, 2002). In particular, the algorithm forecasts and AIRMETs were verified using Yes and No PIREPs of icing. The algorithm forecasts were transformed into Yes/No icing forecasts by determining if the algorithm output at each model grid point exceeded or was less than a pre-specified threshold; a variety of different thresholds was utilized to examine the full range of performance of the algorithm. The Yes/No forecasts were evaluated using standard verification techniques available for Yes/No forecasts, where observations are based on PIREPs. In addition, the amount of airspace impacted by the forecasts was considered. For most analyses, only PIREPs reporting moderate or greater (MOG) icing severity were included as Yes reports.

In evaluating an algorithm or forecast, it is important to compare the quality of the forecasts to the quality of one or more standards of reference. Thus, the quality of the FIP forecasts is compared to the quality of the operational forecasts (i.e., AIRMETs). These forecasts represent the operational forecast information that is currently available to a user/decision-maker. Although both types of information could be used by the user/decision-maker, it is important to emphasize that the FIP forecasts and the AIRMETs are very different types of forecasts, with different objectives. FIP forecasts generally are understood to be valid at a particular time. The AIRMETs, on the other hand, are valid over a 6-h period and are designed to capture icing conditions as they move through the AIRMET area over the period. Due to the differences between these forecasts, it is difficult to clearly compare their performance. However, in order to understand the quality of FIP, it is necessary for FIP forecasts to be compared to the operational standard, especially since both types of information will be available to users. The comparisons are made in such a way as to be as fair as possible to both the

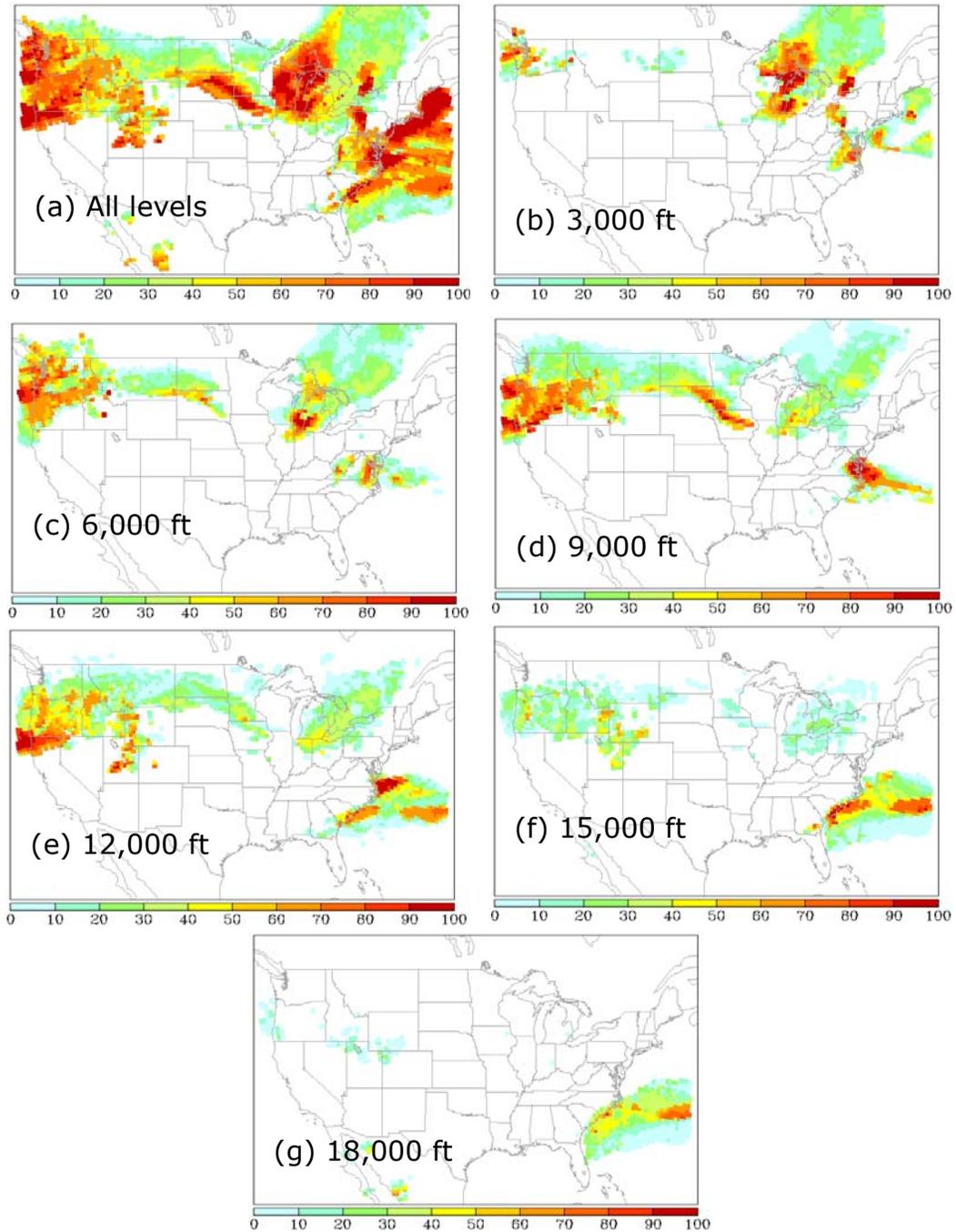
AIRMETs and FIP, as described in Section 4, without degrading the information available from FIP. Nevertheless, users of these statistics should keep these assumptions in mind when evaluating the strengths and weaknesses of each type of forecast.

In December 2002, an error was discovered in the decoder used to translate textual PIREPs into a digital form. The error was associated with determining the location of a small subset of the PIREPs. While we expected this error to have relatively minor impacts on the verification results, we concluded that it would be best to limit most of the analyses presented in this report to include PIREPs that have been correctly decoded. Thus, all of the analyses, except some long-term statistics, are based on the corrected decoder. Impacts of the PIREP decoder error are considered briefly in Section 6, and will be described more fully in a subsequent report.

### **3. Algorithms and forecasts**

*FIP*: The FIP forecasting technique uses a physically-based, situational, fuzzy logic technique to integrate various model variables (cloud microphysical values, temperature, humidity, and vertical velocity). The algorithm is modeled after the Current Icing Potential (CIP; Bernstein et al. 2001), which integrates various observations with model output to create a diagnosis of icing conditions; however, because FIP produces a forecast rather than a diagnosis, observations are not available and must be inferred from the model output. The FIP values are a measure of the “potential” for icing occurrence; that is, they are not calibrated probabilities but larger values indicate a greater chance of icing. An example of a FIP forecast is presented in Fig. 1. In this figure, the maximum FIP values for particular layers are shown, as well as the composite values based on the FIP values in the whole column. An improved version of FIP was implemented in the fall of 2002; performance of this version of the algorithm is compared to the previous version in Section 6, and the improved version is considered in most of the subsequent analyses. Further information about FIP and its development can be found in McDonough et al. (2003).

*AIRMETs*: AIRMETs are the operational forecasts of icing conditions. These forecasts are produced by AWC forecasters every six hours and are valid for up to six hours (NWS 1991). AIRMETs may be amended as needed between the standard issue times. 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 in AIRMET verification analyses. For comparison with the forecasts from FIP, the AIRMETs are evaluated over the same time window as the model-based algorithms.



**Figure 1. Example FIP grid for 7 March 2003, 6-h forecast valid at 2100 UTC. Maximum column value is shown (a) as well as values at several flight levels.**

#### 4. Data: Model output and PIREPs

Model output was obtained from the RUC model, which is run operationally at NWS/NCEP/EMC (Benjamin et al. 1998). The model vertical coordinate system is based on a hybrid isentropic-sigma vertical coordinate. The RUC assimilates data from commercial aircraft, wind profilers, rawinsondes and dropsondes, surface reporting stations, and numerous other data sources. The model produces forecasts on an hourly basis; however, as mentioned in Section 2, only a subset of forecast and lead-time combinations was used in this study. Data for the 20-km version of the RUC model (which became operational on April 17, 2002) were obtained from the FSL mass store system. Although the RUC domain extends further in all directions, the verification analyses were limited to the domain covered by the AIRMETs, which is shown in Fig. 2.



**Figure 2. Total domain of the AIRMETs, used for the FIP verification analyses.**

The FIP algorithm was applied to the model output files to create algorithm output files. This part of the process was undertaken by the FIP algorithm developers. As part of this process, the algorithm output data were interpolated to flight levels (i.e., every 1,000 ft) so the algorithm could be verified on flight levels rather than the raw model levels. The AIRMETs were decoded to extract the relevant location, altitude range, and other information.

All available Yes and No icing PIREPs were included in the study. These reports include information about the severity of icing encountered, which was used to categorize the reports. In particular, reports of moderate to extreme icing were included in the “Moderate-or-Greater” (MOG) category; most of the analyses are based on this category of icing conditions. As in the verification studies for CIP (Brown et al. 2001b, 2002), an additional form of No-icing reports was included in some of the analyses. These PIREPs are the “Clear Above” (CA) reports, in which a pilot remarks that the sky is clear above a particular flight level; the lack of icing conditions at higher levels can be inferred from these reports. The CA reports represent a different type of negative information, and thus are treated separately from the explicit “No” reports.

## **5. Methods**

This section summarizes methods that were used to match forecasts and observations, as well as the various verification statistics that were computed.

### **5.1 Matching methods**

The methods used to connect PIREPs to the forecasts are the same as have been used in previous evaluations of CIP and other in-flight icing algorithms (e.g., Brown et al. 1997, 2001a,b, 2002). In particular, both the post-analysis and RTVS systems connect each PIREP to the forecasts at the nearest 8 grid points (four surrounding grid points; two levels vertically). However, the RTVS uses bi-linear interpolation to compute the appropriate forecast value, whereas the post-analysis system matches the PIREP to the largest forecast value among the eight surrounding gridpoints. As in the evaluations of CIP, a time window of +1 hour after the model valid time was used to evaluate both the FIP forecasts and the AIRMETS.

### **5.2 Statistical verification methods**

The statistical verification methods used to evaluate the results for this study are the same as the methods used in previous studies and are consistent with the approach described by Brown et al. (1997). These methods are briefly described here.

Icing forecasts and observations are treated here as dichotomous (i.e., Yes/No) values. AIRMETS essentially are dichotomous by definition (i.e., a location is either inside or outside the defined AIRMET region). The algorithm forecasts are converted to a variety of Yes/No forecasts by application of various thresholds for the occurrence of icing. The thresholds used for FIP are  $2 \times 10^{-5}$ , 0.05, 0.10, 0.15, ..., and 0.95. RTVS includes results for thresholds of 0.02, 0.15, 0.25, 0.45, 0.65, and 0.85. Once the forecasts are converted to Yes/No forecasts, the basic verification approach makes use of the two-by-two contingency table (Table 1). In this table, the forecasts are represented by the rows, and the columns represent the observations. The entries in the table represent the joint distribution of forecasts and observations.

Table 2 lists the verification statistics used in this evaluation. As shown in this table, PODy and PODn are the primary verification statistics based on the 2x2 verification table. Together, PODy and PODn measure the ability of the forecasts to discriminate between (or correctly categorize) Yes and No icing observations. This discrimination ability is summarized by the True Skill Statistic (TSS), which frequently is called the Hanssen-Kuipers discrimination statistic (Wilks 1995). Note that it is possible to obtain the same value of TSS for a variety of combinations of PODy and PODn. Thus, it always is important to consider both PODy and PODn, as well as TSS.

**Table 1: Contingency table for evaluation of dichotomous (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

It will be noted that Table 2 does not include the False Alarm Ratio (FAR), a statistic that is commonly computed from the 2x2 table. Due to the non-systematic nature of PIREPs, it is not appropriate to compute FAR using these observations. This conclusion, which also applies to statistics such as the Critical Success Index and Bias, is documented analytically and by example in Brown and Young (2000). In addition, due to characteristics of PIREPs and their limited numbers, other verification statistics (e.g., PODy and PODn) should not be interpreted in an absolute sense, but can be used in a comparative sense, for comparisons among algorithms and forecasts. Moreover, PODy and PODn should not be interpreted as probabilities, but rather as *proportions of PIREPs that are correctly forecast*.

As shown in Table 2, three other statistics are utilized for verification of the icing forecasts: % Area, % Volume and Volume Efficiency (VE). The % Volume statistic is the percent of the total possible airspace volume<sup>4</sup> that has a Yes forecast. The % Area indicates the proportion of the surface area of the forecast domain that is associated with a Yes icing forecast at some flight level above. VE considers PODy relative to the volume covered by the forecast, and can be thought of as the POD per unit volume. *The VE statistic must be used with some caution, however, and should not be used by*

---

<sup>4</sup> The total possible area (limiting coverage to the area of the continental United States that can be included in AIRMETS) is 9.5 million km<sup>2</sup>. The total possible volume thus is about 120 million km<sup>3</sup>.

*itself as a measure of forecast quality.* For example, sometimes it is easy to obtain a large VE value when PODy is very small. An appropriate use of VE is to compare the efficiencies of forecasting systems that have nearly equivalent values of PODy. In fact, *none of the statistics should be considered in isolation* – all should be examined in combination with the others to obtain a complete picture of forecast quality.

**Table 2: Verification statistics used in this study.**

<b>Statistic</b>	<b>Definition</b>	<b>Description</b>	<b>Interpretation</b>	<b>Range</b>
<b>PODy</b>	$YY/(YY+NY)$	Probability of Detection of Yes observations	Proportion of Yes observations that were correctly forecasted	0-1 Best: 1 Worst: 0
<b>PODn</b>	$NN/(YN+NN)$	Probability of Detection of No observations	Proportion of No observations that were correctly forecasted	0-1 Best: 1 Worst: 0
<b>TSS</b>	$PODy + PODn - 1$	True Skill Statistic; Hanssen-Kuipers discrimination	Level of discrimination between Yes and No observations	-1 to 1 Best: 1 No skill: 0
<b>ROC Curve Area</b>	Area under the curve relating PODy and 1-PODn	Area under the curve relating PODy and 1-PODn (i.e., the ROC curve)	Overall skill (related to discrimination between Yes and No observations)	0 to 1 Best: 1 No skill: 0.5
<b>% Area</b>	$[(\text{Forecast Area}) / (\text{Total Area})] \times 100$	% of the total area (e.g., CONUS) that has a Yes forecast at some level above	% of the area that is impacted by a Yes forecast at one or more flight levels above	0-100 Smaller is better
<b>% Volume</b>	$[(\text{Forecast Vol}) / (\text{Total Vol})] \times 100$	% of the total air space volume that is impacted by the forecast	% of the total air space volume that is impacted by the forecast	0-100 Smaller is better
<b>Volume Efficiency (VE)</b>	$(PODy \times 100) / \% \text{ Volume}$	PODy (x 100) per unit % Volume	PODy relative to airspace coverage	0-infinity Larger is better

As in previous icing forecast verification analyses, emphasis in this report will be placed on PODy, PODn, and % Volume. Use of this combination of statistics implies that the underlying goal of the algorithm development is to include most Yes PIREPs in the forecast “Yes icing” region, and most No PIREPs in the forecast “No icing” region (i.e., to increase PODy and PODn), while minimizing the extent of the forecast region, as represented by % Volume.

The relationship between PODy and 1-PODn for different FIP algorithm thresholds is the basis for the verification approach known as “Signal Detection Theory” (SDT). For a given algorithm or forecasting system, this relationship can be represented by the curve joining the (1-PODn, PODy) points for different algorithm thresholds. The resulting curve is known as the “Relative Operating Characteristic” (ROC) curve in SDT. ROC curves measure the skill of a set of forecasts at discriminating between Yes and No observations. The area under an ROC curve is a measure of overall forecast skill (e.g., Mason 1982), another measure that can be compared among the forecasts.

% Volume and % Area plots provide two additional overall views of forecast performance. These plots show the relationships between (i) PODy and % Volume, and (ii) PODy and % Area for various thresholds. For all three plots (ROC, % Area, and % Volume), curves for better forecasts are located closer to the upper lefthand corner of the diagram (e.g., see Fig. 7). It is important to understand the differences among these three plots: as noted previously, the ROC measures the forecasts’ ability to discriminate between Yes and No observations of icing. In contrast, the % Area and % Volume plots measure the trade-offs between increasing PODy and increasing the amount of airspace that is impacted by the forecasts. However, while the % Volume plots measure the trade-offs with actual three-dimensional airspace, the % Area plots measure the trade-offs with that volume projected to the surface.

Quantification of the uncertainty in verification statistics is an important aspect of forecast verification that often is ignored. Confidence intervals provide a useful way of approaching this quantification. However, most standard confidence interval approaches require various distributional and independence assumptions, which generally are not satisfied by forecast verification data. As a result, the QAPDT has developed an alternative confidence interval method based on re-sampling statistics, which is appropriate for icing forecast verification data (Kane and Brown 2000). This approach is applied to some of the statistics considered in this report.

### **5.3 Stratifications**

The verification results are stratified in a variety of different ways. The time periods are all considered separately (see Section 5.4), and the results are also stratified by lead time. In addition, variations in FIP performance at different flight levels are considered by stratifying the data according to altitude.

## 5.4 Time periods

A variety of different time periods were included in this analysis: (a) 15 September to 31 October 2001, including some sub-periods; (b) 1 January to 20 April 2002, including some sub-periods; and (c) 1 October to 15 November 2002. In addition, a long-term evaluation of FIP on RTVS includes forecasts from the period 15 April 2001 through 28 February 2003. The periods, sub-periods, and types of comparisons each period was used for are summarized in Table 3.

**Table 3. Periods and sub-periods used in the FIP verification analyses.**

Period	Sub-periods	Comparisons	# of forecasts
1 – 31 Oct 2001 (Fall 2001)	1 – 31 Oct 2001	<ul style="list-style-type: none"> <li>FIP versions</li> <li>RUC model versions</li> </ul>	728
1 Jan – 20 Apr 2002 (Winter/spring 2002)	1 Jan – 20 Apr 2002	<ul style="list-style-type: none"> <li>FIP versions</li> <li>Overall FIP performance (40-km)</li> </ul>	2,576
	1-20 Apr 2002	<ul style="list-style-type: none"> <li>RUC model versions</li> </ul>	451
1 Oct – 15 Nov 2002 (Fall 2002)	1 Oct – 15 Nov 2002	<ul style="list-style-type: none"> <li>Overall FIP performance (20-km)</li> </ul>	1,110
	1-31 Oct 2002	<ul style="list-style-type: none"> <li>RUC model versions</li> </ul>	728
15 April 2001 – 28 Feb 2003		<ul style="list-style-type: none"> <li>Long-term variations</li> </ul>	

## 6. Results

Basic results of the FIP evaluations are described in this section. The post-analysis verification analyses were limited to dates and times when algorithm output for both FIP and the AIRMETs were available. The results are organized in sub-sections as follows: (i) comparison of FIP versions; (ii) model version comparisons; (iii) overall results and results by lead time; (iv) day-to-day variations in results; (v) variations with altitude; and (vi) seasonal variations. Finally, some basic results of the PIREP comparisons, associated with the PIREP decoder error, are described. Many of the results are based on the combination of verification counts (i.e., the counts from Table 1) across a large number of forecasts to compute statistics that represent the performance of the whole set of forecasts.

## 6.1 FIP version comparisons

In anticipation of operational implementation of FIP, the FIP developers determined that a few upgrades to the algorithm would be valuable. Thus, the version of FIP to be implemented is slightly different from the version of the algorithm that has been running experimentally and evaluated by RTVS<sup>5</sup>. Hence, a comparison of the performance of the new and old versions of the algorithm was required. For this evaluation, two time periods, fall 2001 and winter/spring 2002 were examined. For both periods, the operational model was the 40-km version of RUC.

Figures 3 and 4 show % Volume plots and ROC diagrams for the two versions of the algorithm for the two periods, for 6-h forecasts. As shown in these figures, the performance of the algorithm appears to have been slightly improved in the new version, although the differences in performance are not significant. In particular, the curves for Version 1 are located slightly further toward the upper left in most of the diagrams. These results are consistent for all lead times. Thus, all further evaluations in this report are based on the new version of the algorithm.

## 6.2 Model version comparisons

The change to the new version of RUC involved a change in the model microphysics as well as an increase in model grid resolution. To evaluate the impacts of these two changes in the model on FIP performance, two types of comparisons were undertaken. First, results based on the old version of the model (on a 40-km scale) were compared to results based on the new model on a 20-km scale. Second, results of this evaluation were compared to results based on the new model with grid resolution degraded from 20 km to 40 km. Two comparisons were included: (i) results based on the 20-km version of the model for the period 1-31 Oct 2002 were compared to results based on the 40-km version of the model for the period 1-31 Oct 2001; and (ii) results based on both versions of the model were compared for the period 1-20 April 2002, just prior to the time when the 20-km RUC became operational. The first of these comparisons has the disadvantage of using data from two different years, so the results are not really directly comparable. The second period has the disadvantage of having a more limited number of forecasts, which makes the statistics somewhat less stable and robust. Overall results for these comparisons are shown in Figures 5 and 6, for 6-h forecasts.

The plots in Fig. 5 suggest a small difference in results between fall 2001 and fall 2002. In particular, the FIP forecasts for 2001 achieved a somewhat larger POD<sub>y</sub> value than the FIP forecasts for fall 2002, for the same values of % Volume and POD<sub>n</sub>. However, these differences are not statistically significant. It is not clear from these diagrams how much of these differences is related to changes in the model, and how

---

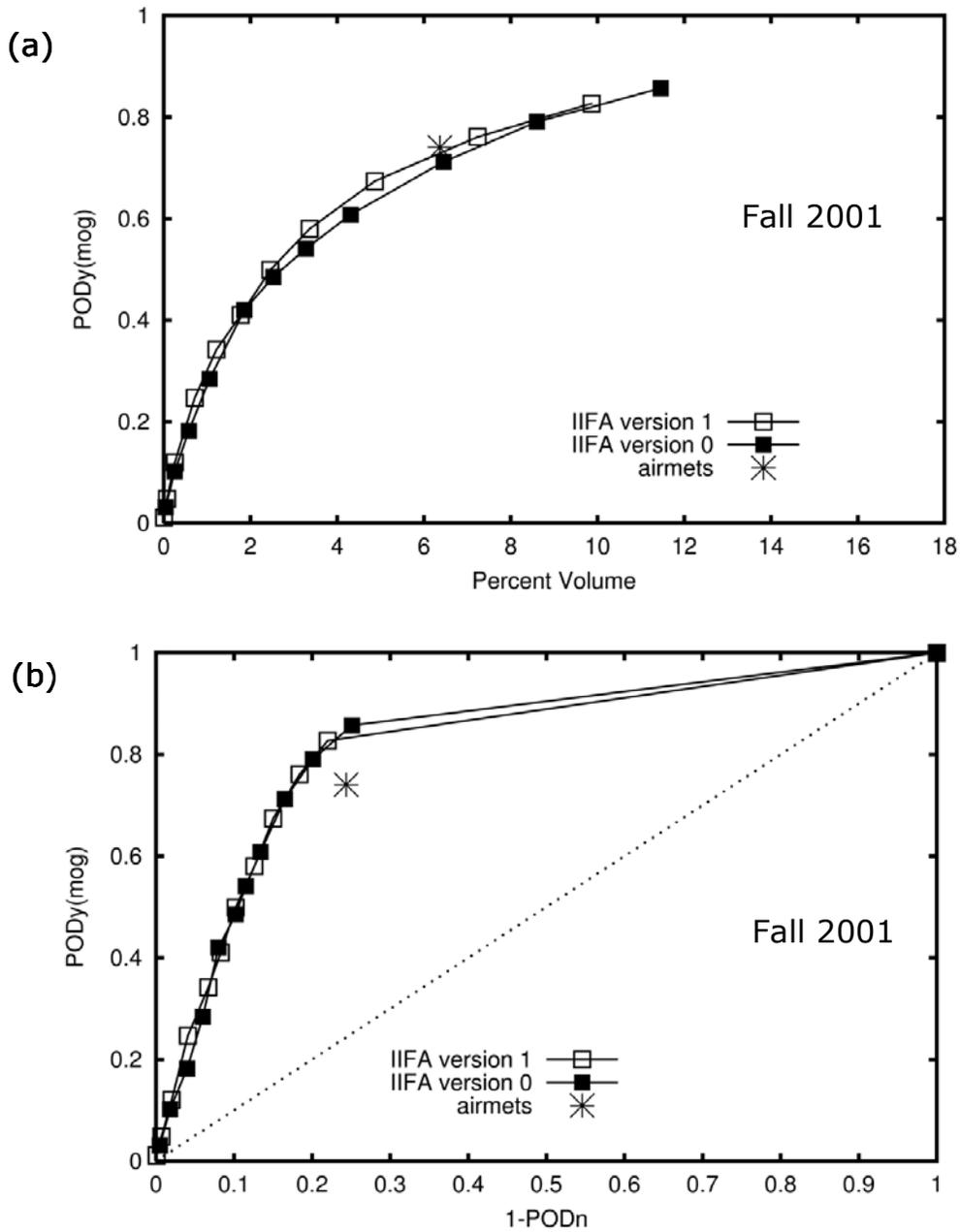
<sup>5</sup> The new version of FIP was implemented on RTVS in December 2002 and is the current version of the algorithm being evaluated on that system.

much is simply the effect of weather variations between the two seasons. It is interesting to note that the AIRMET verification statistics are quite different for these two periods, which suggests the weather variations may have had an important impact on the FIP verification statistics. The curves associated with Fall 2002 also indicate that there is relatively little difference between the performance of FIP on the 20-km RUC vs. the degraded 20-km RUC for this period.

The results in Fig. 6, for 1-20 April 2002, suggest a few differences between the verification statistics for the two versions of the model. In particular, the ROC (POD<sub>y</sub> vs. 1-POD<sub>n</sub>) plots (Fig. 6b) suggest that FIP is most successful at discriminating between Yes and No PIREPs when the 20-km version of the model is used. The degraded 20-km version of the model also produced somewhat better verification results – in terms of the ROC – than the old 40-km version of the model. These results are consistent for all lead times. This result suggests that at least some of the improvement in verification statistics is due to fundamental changes in the model microphysical parameterization that were implemented in the new version of the model. Results for POD<sub>y</sub> vs. % Volume (Fig. 6a) indicate much smaller variations in performance among the model versions, with a slight suggestion that the old version of the model performs somewhat better than the new version; this result also is evident for the 3- and 12-h forecasts, but not for the 9-h forecasts (not shown). Variability in these results may be due to the relatively small number of forecasts that were available during the period when all of the model versions were available.

Comparison of the individual points in Fig. 6 also suggest that the FIP calibration is somewhat different for the new version of the RUC. In particular, the threshold value required to achieve a particular POD<sub>y</sub> value is somewhat smaller for the 20-km RUC. Although this result implies that the FIP calibration is somewhat different for the two model versions, it does not have an impact on overall performance of the algorithm.

The results in Figs. 5 and 6 indicate that – overall – FIP performance was not degraded by the change to the new 20-km version of the RUC. The remainder of the results considered in this report are based on two time periods with both versions of the model: (a) 1 January through 20 April 2002 (40-km RUC); and (b) 1 October through 15 November 2002 (20-km RUC).



**Figure 3. (a) % Volume and (b) ROC plots for Fall 2001, showing performance of 6-h forecasts from old version of FIP (IIFA version 0) vs. performance of new version of FIP (IIFA version 1), for Fall 2001. From right to left, FIP thresholds are  $2 \times 10^{-5}$ , 0.05, 0.15, 0.25, 0.35, 0.45, ..., 0.95.**

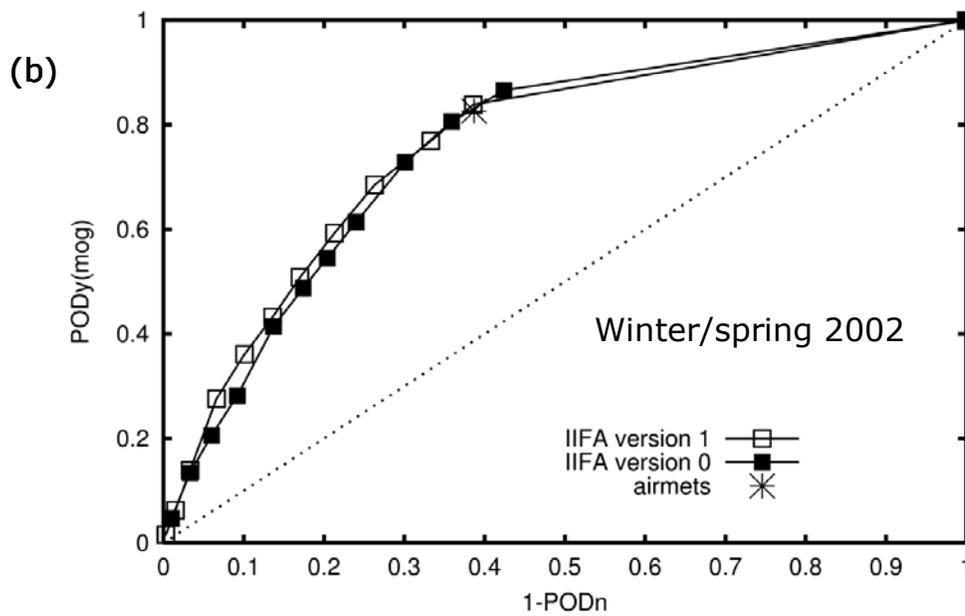
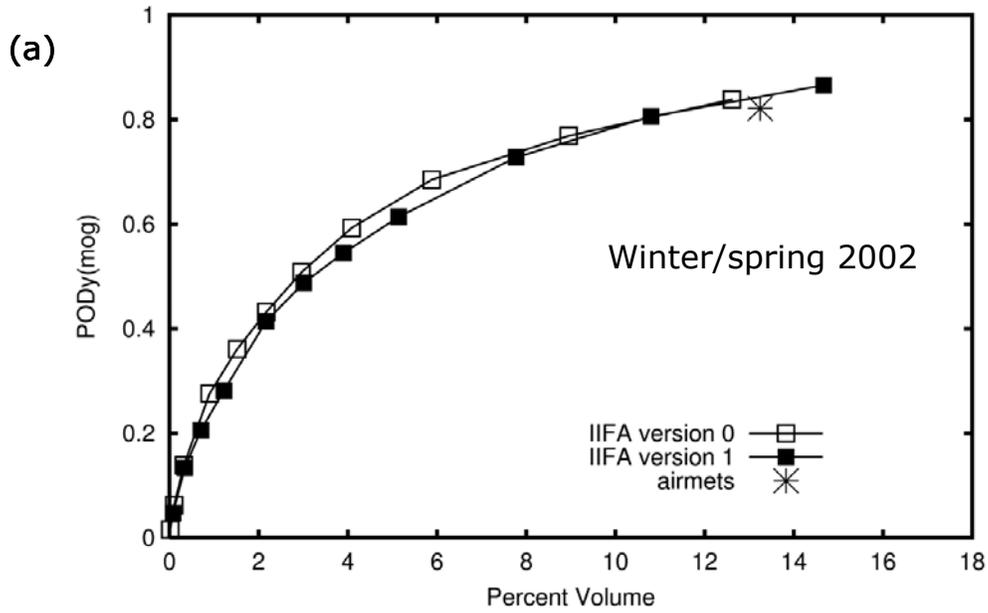
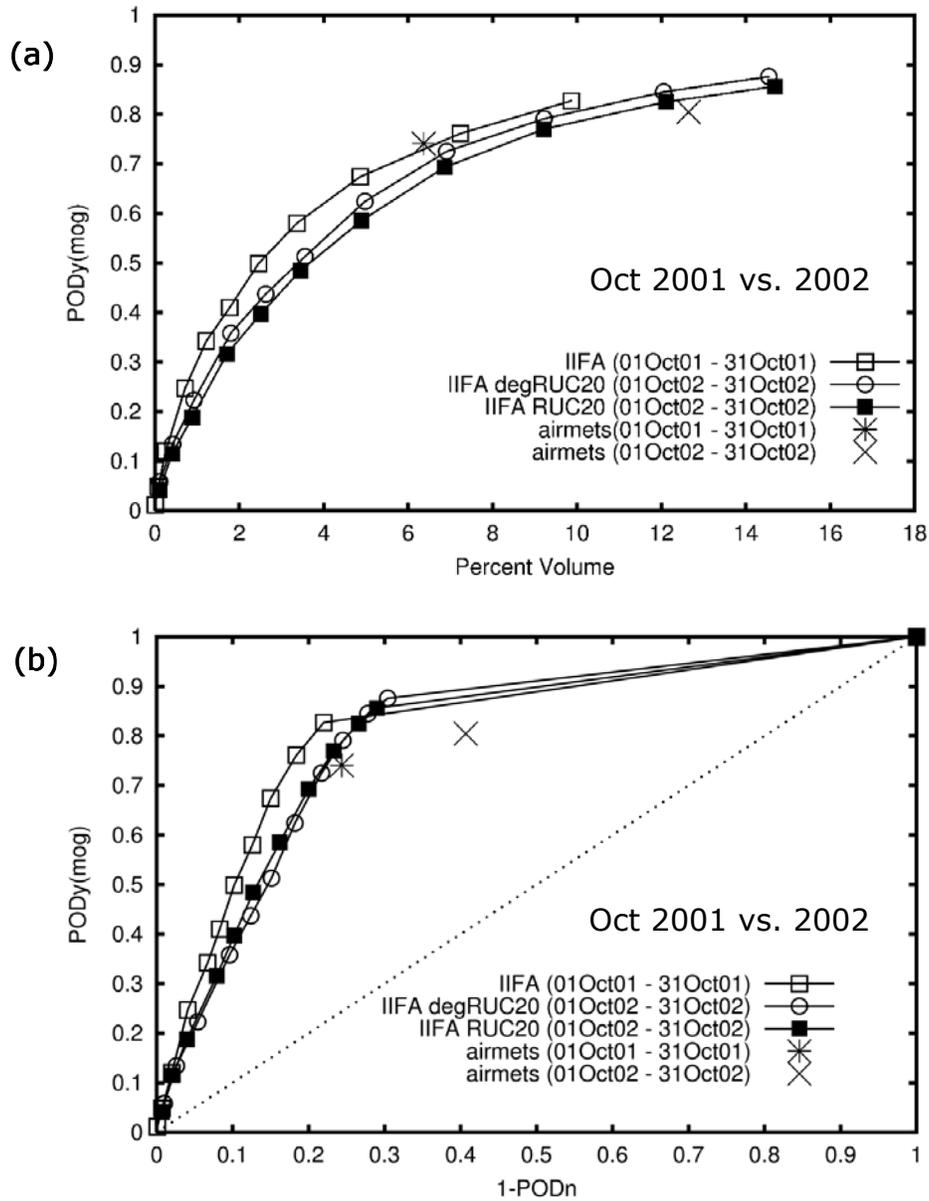
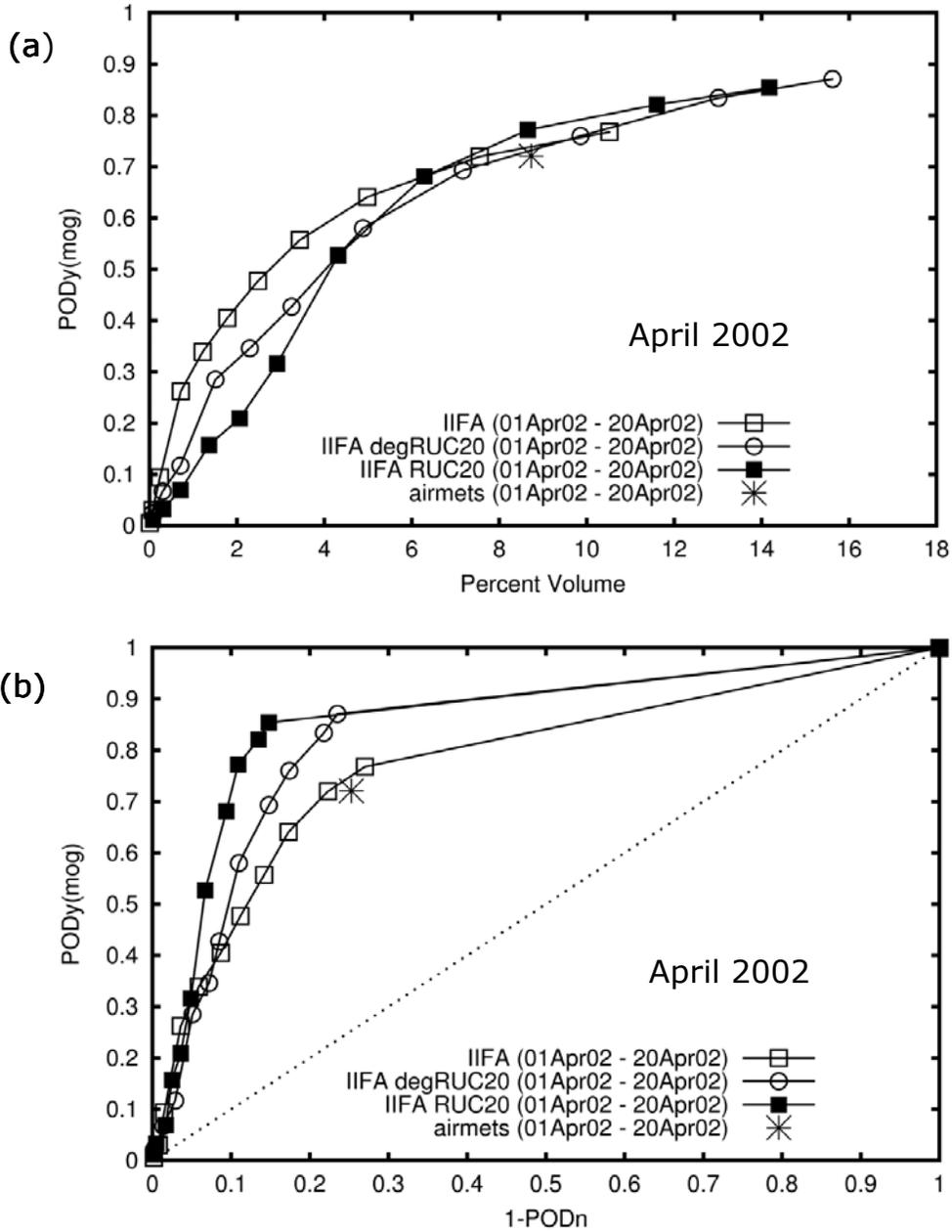


Figure 4. As in Figure 3, for Winter/Spring 2002.



**Figure 5. (a) % Volume and (b) ROC plots for Oct 2001 and 2002, showing performance of 6-h FIP/IIFA forecasts based on the old 40-km version of RUC [open squares; “IIFA...”] vs. FIP/IIFA based on the new 20-km RUC [closed squares; “IIFA RUC20...”] and FIP/IIFA based on the 20-km RUC degraded to 40 km [open circles; “IIFA degRUC20...”]. From right to left, FIP thresholds are  $2 \times 10^{-5}$ , 0.05, 0.15, 0.25, 0.35, 0.45, ..., 0.95. Performance of AIRMETs for the two periods is also shown.**



**Figure 6. (a) % Volume and (b) ROC plots for 1-20 April 2002, showing performance of 6-h FIP/IIFA forecasts based on the old 40-km version of RUC [open squares; "IIFA .."] vs. FIP/IIFA based on the new 20-km RUC [closed squares; "IIFA RUC20..."] and FIP/IIFA based on the 20-km RUC degraded to 40 km [open circles; "IIFA degRUC20..."]. From right to left, FIP thresholds are  $2 \times 10^{-5}$ , 0.05, 0.15, 0.25, 0.35, 0.45, ..., 0.95.**

### 6.3 Overall results and results by lead time

Overall verification results for FIP, by lead time, are shown in Figures 7 and 8, for winter/spring (1 January – 20 April) and fall (1 October – 15 November) 2002, respectively. Results for CIP (IIDA) are also shown in Fig. 7. These figures suggest that the differences in results by lead-time are very small and not statistically significant. In fact, for the winter/spring season (Fig. 7), no differences are evident in the plots. For fall 2002 (Fig. 8), very small differences are visible, with performance for the 3-h (12-h) forecasts slightly better (worse) than the performance for other lead times. Fig. 7 also indicates that the performance of CIP (IIDA) is slightly better than the performance of FIP. The algorithm developers suggest this difference is not as large as might be expected because FIP utilizes more advanced interest maps that have not yet been implemented into CIP.

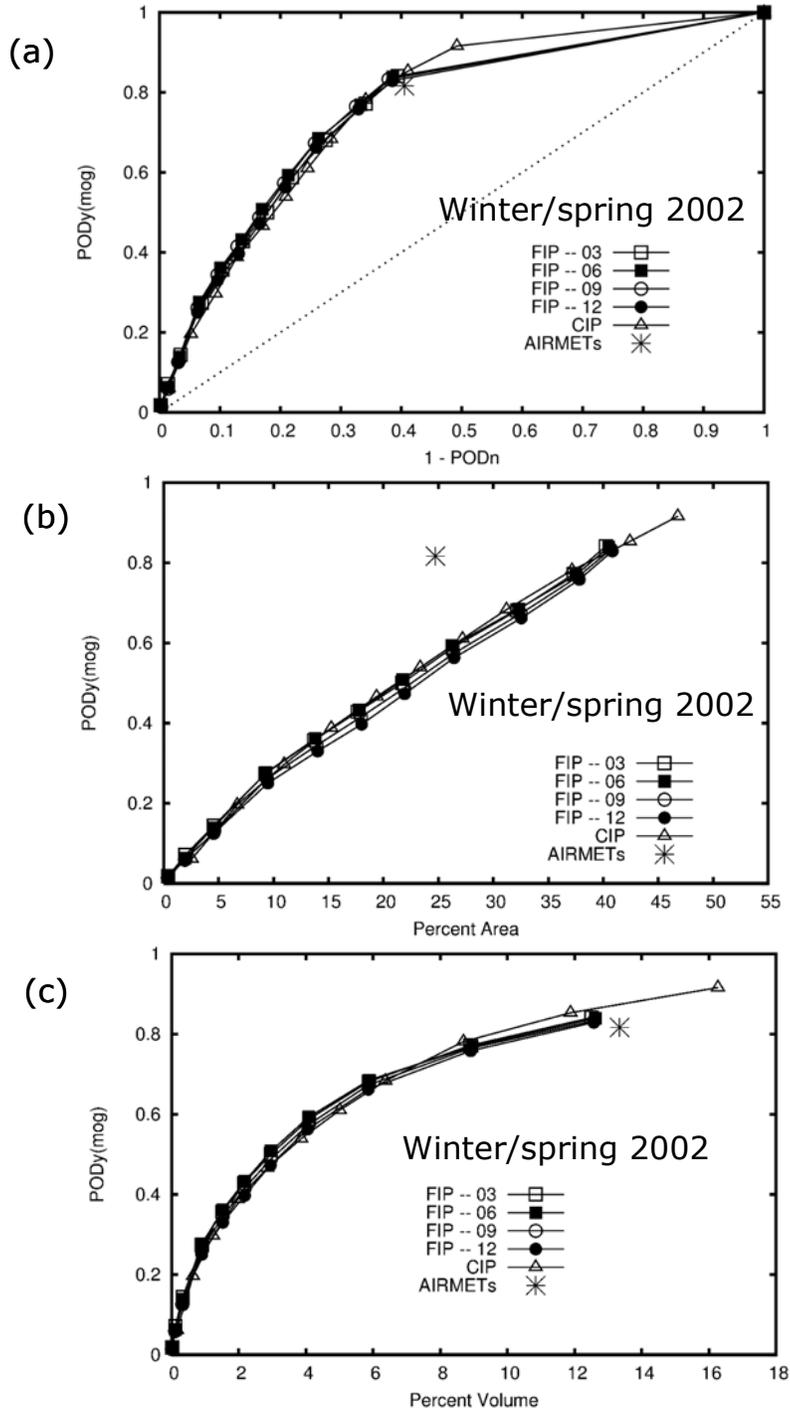
Comparisons of the % Area and % Volume plots in Figs 7 and 8 indicate that FIP is relatively efficient in terms of trade-offs between PODy and volume impacted (i.e., % Volume), but less efficient in terms of trade-offs with areal coverage (i.e., % Area). As noted in Brown et al. (1997), this result is likely due to two factors. First, while the AIRMETs are restricted to a “cakelike” definition of the icing volume (i.e., with solid top, bottom, and interior), FIP (and other automated algorithms) can identify specific smaller volumes, which can lead to greater volume efficiency. However, FIP also can identify thin layers and small regions, which contribute a great deal to the % Area, but very little to the % Volume. In general, the % Volume plot provides a more meaningful evaluation of the impacted airspace trade-offs associated with increasing PODy.

The ROC plots in Figs. 7a and 8a can be summarized using the areas under the ROC curves. These area values are shown in Table 4. The results in this table suggest that the forecast performance was slightly better in the fall, with the 20-km RUC, and that the fall statistics degrade very slightly with lead-time.

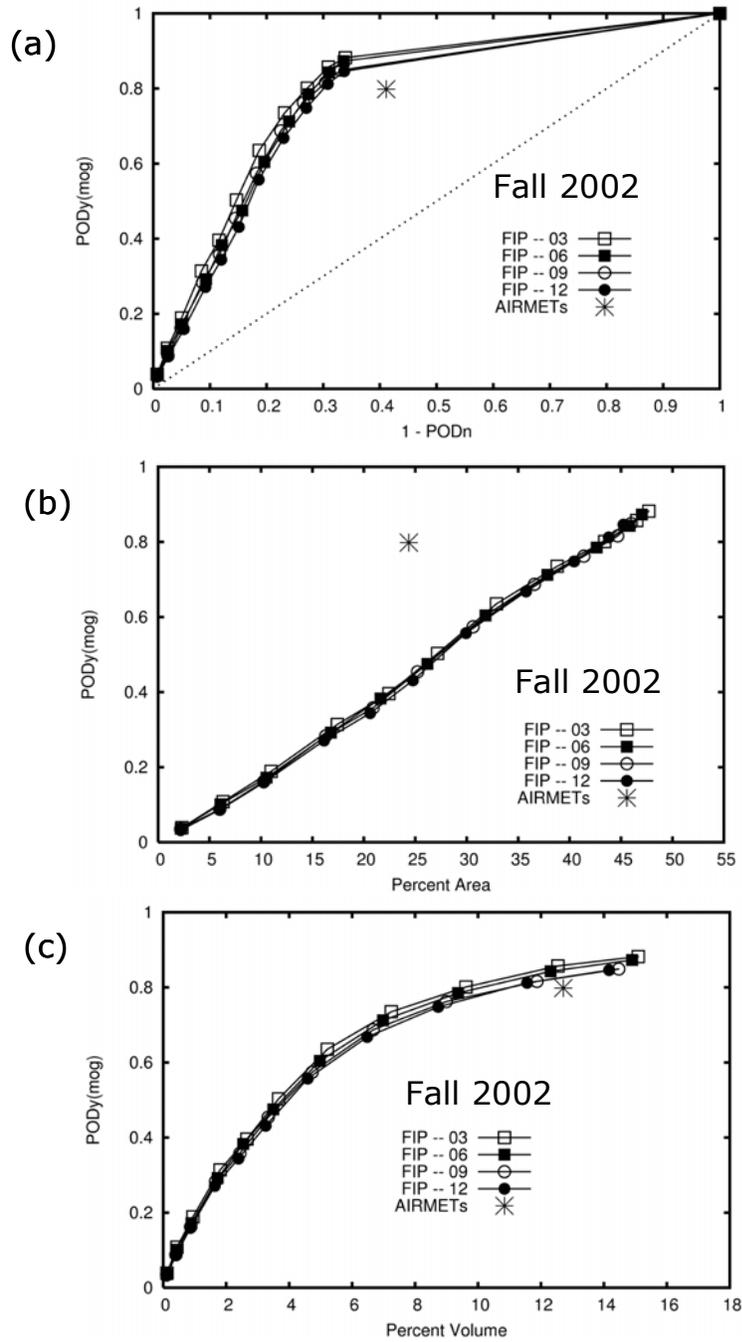
The overall results also can be examined in greater depth by selecting appropriate, comparable thresholds for FIP and comparing the individual statistics. As in previous studies, the rationale used for this process is to select thresholds that lead to PODy values that are approximately the same as the value attained by the AIRMETs. In this case, the thresholds selected<sup>6</sup> are  $2 \times 10^{-5}$  and 0.05. Tables 5 and 6 show the results of this exercise for the 6-h forecasts, for the Winter/spring and Fall 2002 periods, respectively. These tables include a variety of statistics associated with the specified thresholds.

---

<sup>6</sup> Note that these thresholds are different from those used in the evaluation of IIDA (i.e., CIP; Brown et al. 2001b). This difference relates to differences in calibration of the two algorithms. In the case of IIDA, thresholds of 0.15 and 0.25 led to PODy values close to the AIRMET value, whereas for FIP the appropriate thresholds are  $2 \times 10^{-5}$  and 0.05.



**Figure 7. (a) ROC, (b) % Area and (c) % Volume plots for FIP and CIP for winter/spring (1 January – 20 April 2002), by lead time. From right to left, FIP and CIP thresholds are  $2 \times 10^{-5}$ , 0.05, 0.15, 0.25, 0.35, 0.45, ..., 0.95.**



**Figure 8. (a) ROC, (b) % Area, and (c) % Volume plots for FIP for Fall (1 October – 30 November) 2002, by lead time. From right to left, FIP thresholds are  $2 \times 10^{-5}$ , 0.05, 0.15, 0.25, 0.35, 0.45, ..., 0.95.**

**Table 4. ROC areas for winter/spring 2002 and fall 2002 FIP performance, by lead time.**

Lead time (h)	ROC area	
	Winter/Spring 2002 (40-km RUC)	Fall 2002 (20-km RUC)
3	0.76	0.80
6	0.76	0.79
9	0.76	0.78
12	0.76	0.77

**Table 5: Verification statistics for all 6-h forecasts (all issue times combined), for thresholds with PODy (MOG PIREPs) about the same as the PODy(MOG) for AIRMETs, for winter/spring 2002.**

Forecast	Threshold	PODy (All)	PODy (MOG)	PODn	PODn (CA)	TSS	Ave % Area	Ave % Vol	VE
AIRMETs	--	0.78	0.82	0.60	0.89	0.42	24.7	13.3	6.17
FIP	$2 \times 10^{-5}$	0.82	0.84	0.61	0.87	0.45	40.5	12.6	6.67
	0.05	0.75	0.77	0.67	0.91	0.44	37.5	8.9	8.65

**Table 6: As in Table 5, for fall 2002.**

Forecast	Threshold	PODy (All)	PODy (MOG)	PODn	PODn (CA)	TSS	Ave % Area	Ave % Vol	VE
AIRMETs	--	0.77	0.80	0.59	0.87	0.39	24.4	12.7	6.30
FIP	$2 \times 10^{-5}$	0.83	0.87	0.66	0.85	0.53	47.0	14.9	5.84
	0.05	0.80	0.84	0.69	0.88	0.53	45.8	12.3	6.83

Two values of POD<sub>y</sub> are included in Tables 5 and 6 – one for All severities and one for MOG severities. In almost all cases, POD<sub>y</sub> (MOG) is a bit larger than POD<sub>y</sub> (All). This result, which is consistent with previous results, suggests that the MOG PIREPs are somewhat easier for the forecasts to capture than are PIREPs associated with less severe conditions. The POD<sub>n</sub> values for explicit “No” PIREPs also are smaller than the POD<sub>n</sub> values for CA PIREPs, which suggests that the CA PIREPs also are somewhat easier to discriminate from the other types of PIREPs.

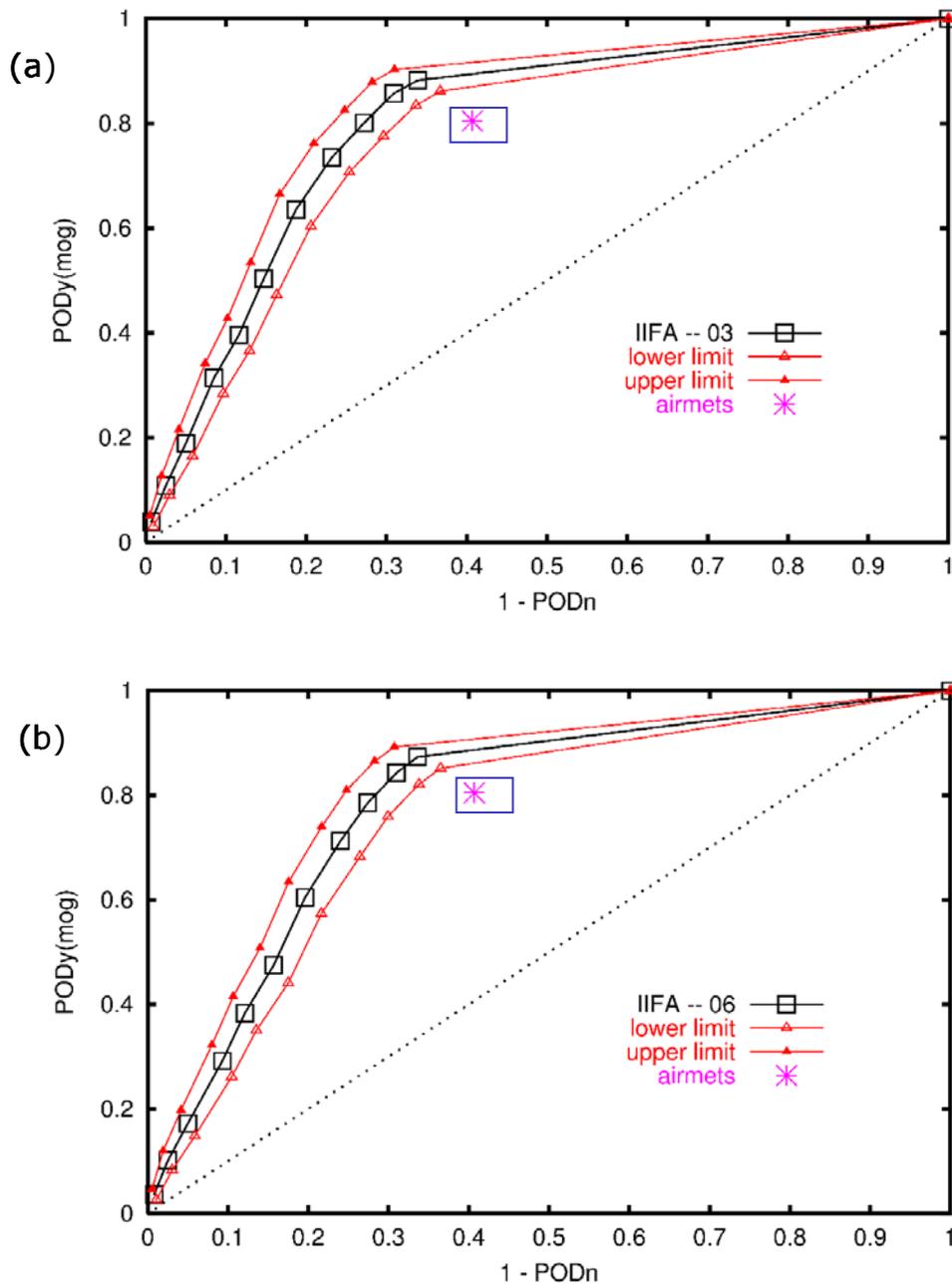
The TSS values in Tables 5 and 6 again indicate that the FIP is skillful at discriminating between Yes and No PIREPs, as are the AIRMETs. With respect to % Area, as expected, the smallest values are associated with the AIRMETs. In terms of the % Volume values in Tables 5 and 6, the smallest value is achieved by FIP in the winter/spring; in the fall, % Volume values for FIP and the AIRMETs are quite similar. Because % Volume is strongly related to POD<sub>y</sub>, the small variations in POD<sub>y</sub> in Tables 5 and 6 may have had some impact on these results. Thus, in some cases it is more appropriate to consider the Volume Efficiency (VE) values. As shown in Tables 5 and 6, FIP and the AIRMETs both have fairly large (and comparable) VE values.

#### **6.4 Variability in verification statistics**

The verification statistics discussed in the previous section vary from forecast to forecast. In addition, some sampling variability is expected to be associated with the statistics. This variability is examined in this section using (a) confidence intervals for the verification measures and (b) box plots (i.e., depictions of the distributions) of the verification statistics for individual forecasts.

Figure 9 shows the ROC (i.e., POD<sub>y</sub> vs. 1-POD<sub>n</sub>) plots for FIP with 3- and 6-h lead times, for Fall (1 October through 15 November) 2002. Additionally, approximate 95% confidence intervals on the ROC curve have been added to these graphs. The lines in this plot suggest that the width of the confidence intervals is at most about +/- 0.07 for both POD<sub>y</sub> and 1-POD<sub>n</sub>. The confidence intervals are narrowest at the extremes (i.e., when POD<sub>y</sub> and 1-POD<sub>n</sub> are close to 0 or 1). The AIRMET confidence intervals fall outside the confidence intervals for the FIP statistics for both lead times; thus the AIRMET forecasts are significantly different from the IIFA forecasts with respect to POD<sub>y</sub> and POD<sub>n</sub>. Results for other lead times are similar.

The confidence interval curves were derived by estimating confidence intervals on both POD<sub>y</sub> and POD<sub>n</sub> via the bootstrap empirical method. The bootstrap is a statistical technique that relies on repeated computer-generated random sampling to estimate distributions, variances, confidence limits, etc. For more information on the method applied here, see Kane and Brown (2000); for more information regarding the bootstrap procedure, see Efron and Tibshirani (1993).



**Figure 9. ROC (POD<sub>y</sub> vs. 1-POD<sub>n</sub>) curves with approximate 95% confidence intervals for (a) 3-h and (b) 6-h FIP (IIFA) forecasts, for Fall (1 October – 30 November) 2002. Boxes around AIRMET points represent 95% confidence intervals for AIRMET POD<sub>y</sub> and 1-POD<sub>n</sub>.**

A convenient way to examine day-to-day variations in the verification statistics is through box plots, which show the distributions of values of the statistics. As an example, Fig. 10 shows box plots of PODy and % Volume associated with individual FIP thresholds, for 6-h FIP forecasts for Fall (1 October through 15 November) 2002. As shown in these plots, the distributions of PODy and % Volume decrease with increasing FIP value. The PODy values are fairly variable (as indicated by the sizes of the boxes), especially for middle threshold values; this result is partly due to the fact that PODy is limited to the range 0-1, so is constrained to be less variable when approaching either 0 or 1. This variability is at least partially due to the small numbers of PIREPs that are available to verify any one forecast. The % Volume values exhibit less variability from day to day; in fact these distributions are quite narrow for any given FIP threshold. Results for other lead times and time periods are consistent with the results shown in Fig. 10.

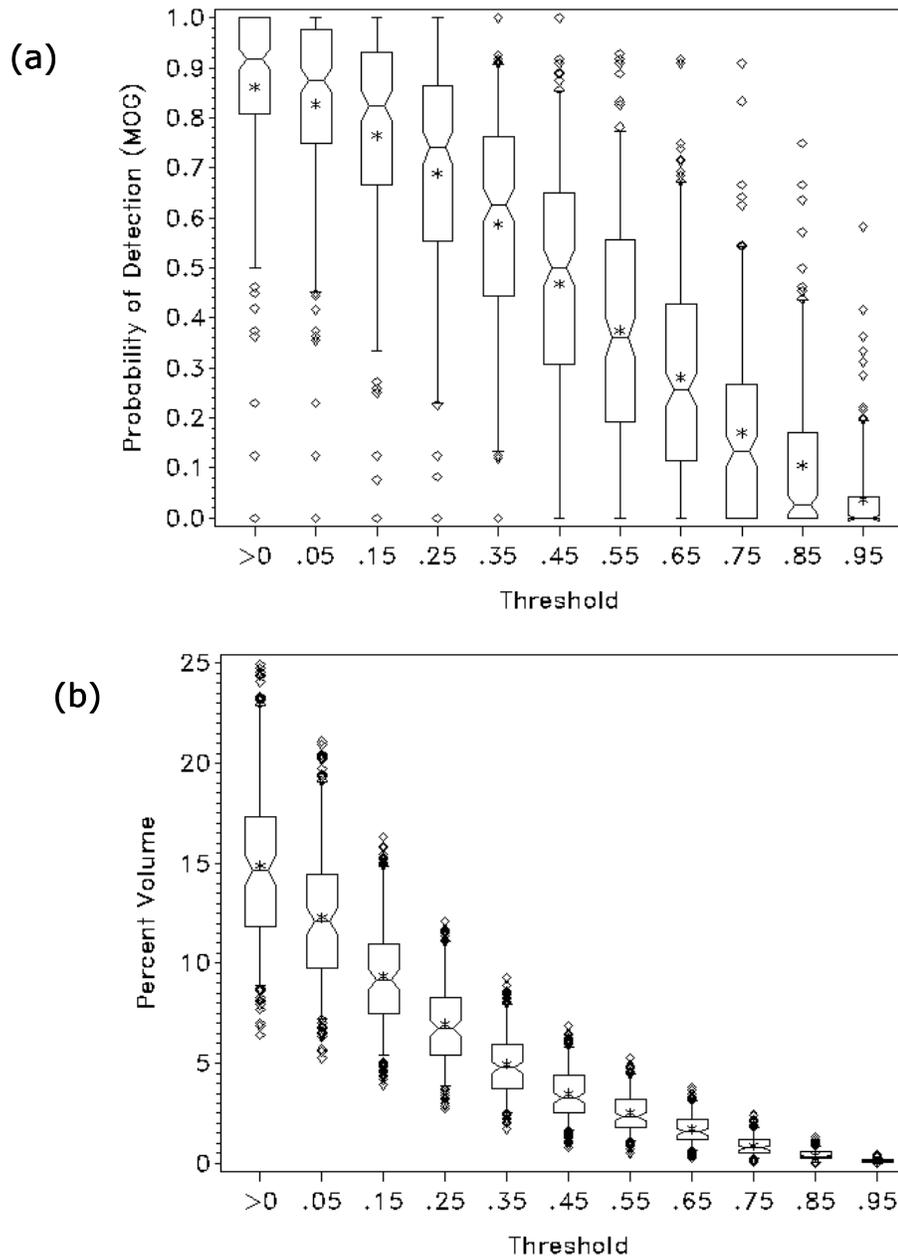
Figure 11 provides a closer look at the day-to-day variations in the statistics for two FIP thresholds and for the AIRMETs, for 6-h FIP forecasts from the fall of 2002. This figure suggests that day-to-day variability in the FIP statistics is somewhat less than the variability in the AIRMET statistics, as represented by the sizes of the boxes. In general, except for % Area, the locations of the distributions are similar. As expected, the % Area distributions for FIP are higher than the corresponding distributions for the AIRMETs.

## 6.5 Comparisons by altitude

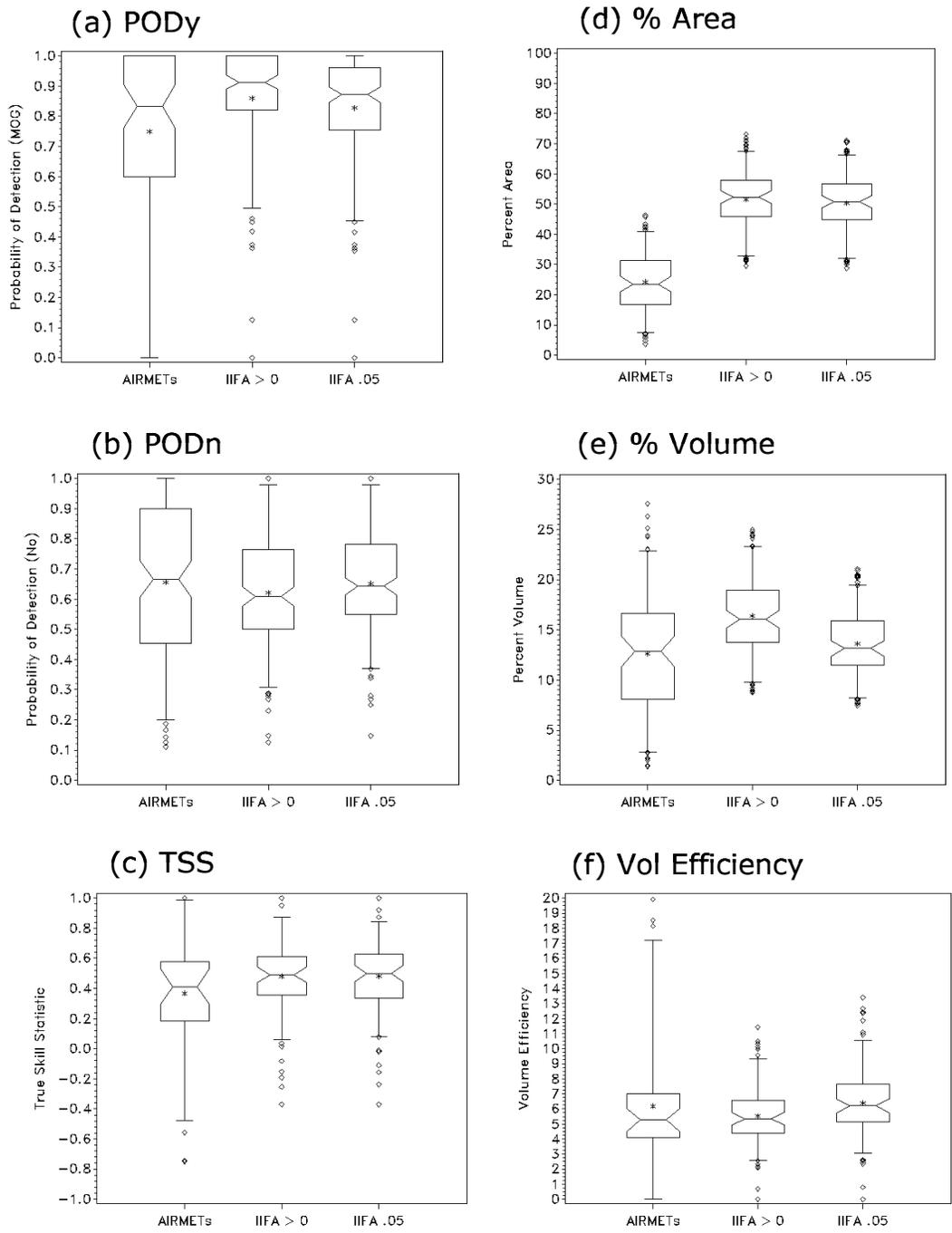
To assess the performance of icing forecasts at different altitudes, verification statistics were computed separately for each 3,000 ft interval from the surface to 30,000 ft. Plots of statistics at all altitudes are presented for FIP with a 6-h forecast lead-time. Plots for other lead times are similar, and have been excluded for brevity.

Figure 12 shows the PODy vs. % Volume graphs for 6-h forecasts in winter/spring (1 January through 20 April) 2002 and fall (1 October through 15 November) 2002, respectively. Each altitude range is represented by a separate curve. For both seasons, the performance is best at lower altitudes and degrades slightly as height increases. For the highest altitude bands, the volumes were extremely small. Since the points for these ranges are concentrated near the origin of the graph, it is somewhat difficult to see these curves.

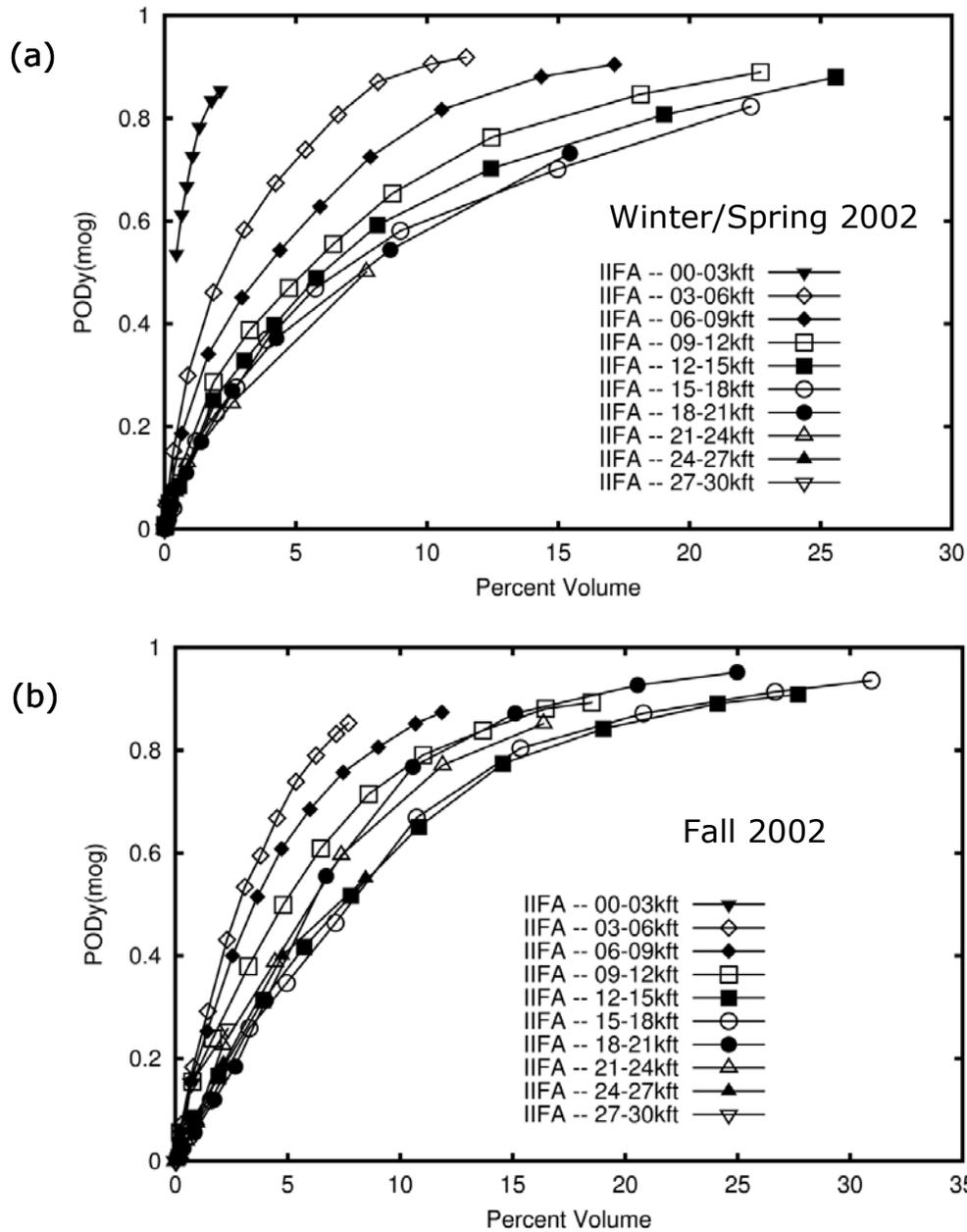
Figure 13 shows the ROC (PODy vs. 1-PODn) plots for the two seasons. As in the previous figures, each altitude range is represented by a separate curve. For winter/spring, FIP performs best at the surface, with performance decreasing gradually as the height increases. However, for the fall time period, FIP performs best near the surface and between 18 and 24,000 ft, and worst between 12 and 18,000 ft. FIP performance in the other altitude ranges falls between these extremes.



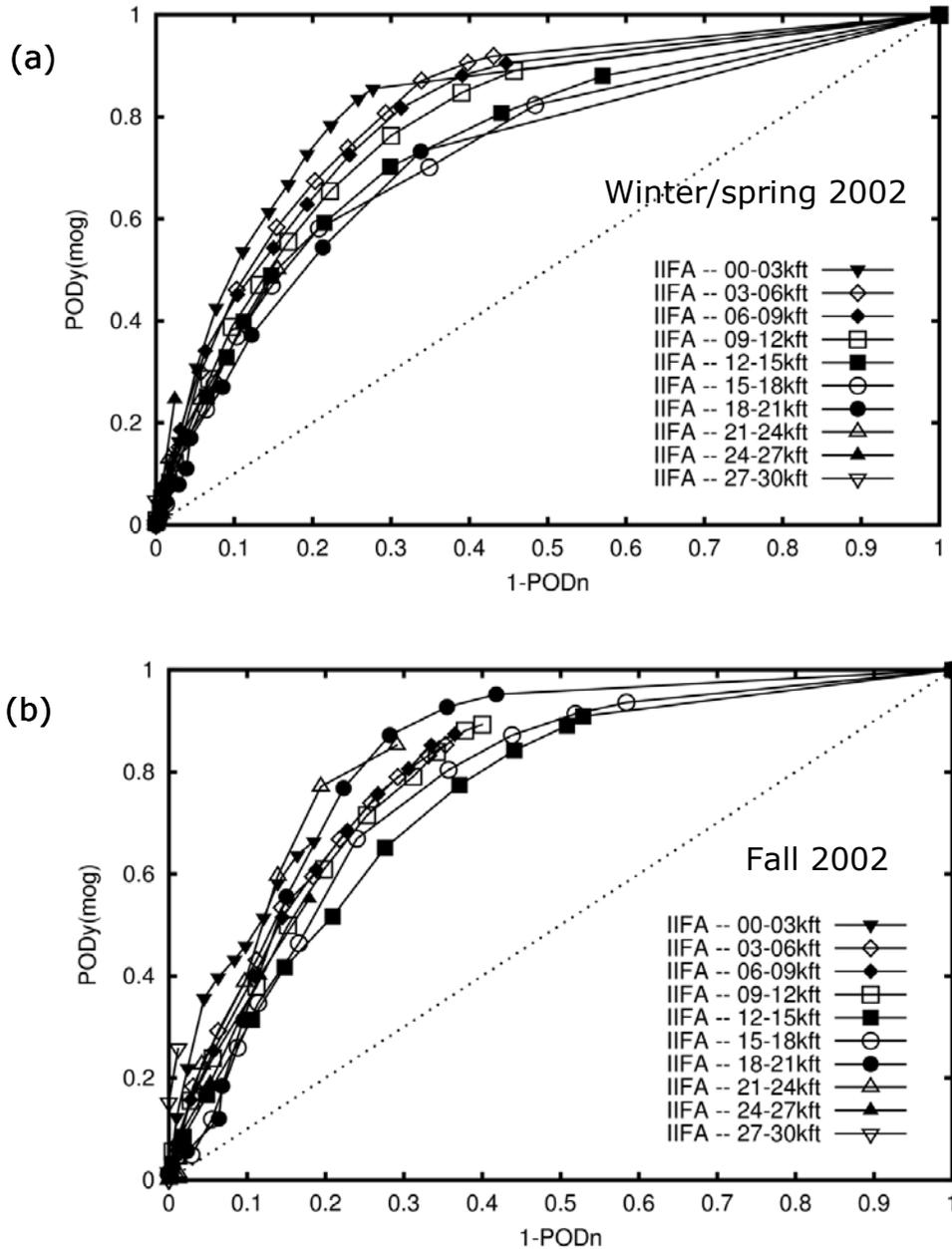
**Figure 10. Box and whisker plots showing distributions of (a) PODy(MOG) and (b) % Volume statistics for individual FIP forecasts, with 6-h lead time for Fall (1 October – 15 November) 2002. The boxes enclose the middle 50% of the distribution (i.e., between the 0.25<sup>th</sup> and 0.75<sup>th</sup> quantile values, with the middle line showing the median value. Ends of the whiskers represent the 0.05<sup>th</sup> and 0.95<sup>th</sup> quantile values of the distributions. Points at top and bottom represent values in the upper and lower 5% of the distribution.**



**Figure 11. Box plots showing distributions of verification statistics for AIRMETs and 6-h FIP forecasts for fall (1 Oct-15 Nov) 2002: (a) PODy(MOG); (b) PODn; (c) TSS; (d) % Area; (e) % Volume; and (f) Volume efficiency. Statistics based on two FIP thresholds [ $2 \times 10^{-5}$  (“IIFA>0”) and 0.05] are shown.**



**Figure 12. PODy vs. % Volume plots for individual altitude ranges for 6-h FIP (IIFA) forecasts for (a) Winter/spring (1 Jan – 20 Apr) 2002; and (b) Fall (1 Oct – 30 Nov) 2002.**



**Figure 13. ROC (POD<sub>y</sub> vs. 1-POD<sub>n</sub>) plots by altitude range for 6-h FIP forecasts, for (a) Winter/spring (1 Jan – 20 Apr) 2002 ; and (b) Fall (1 Oct – 30 Nov) 2002.**

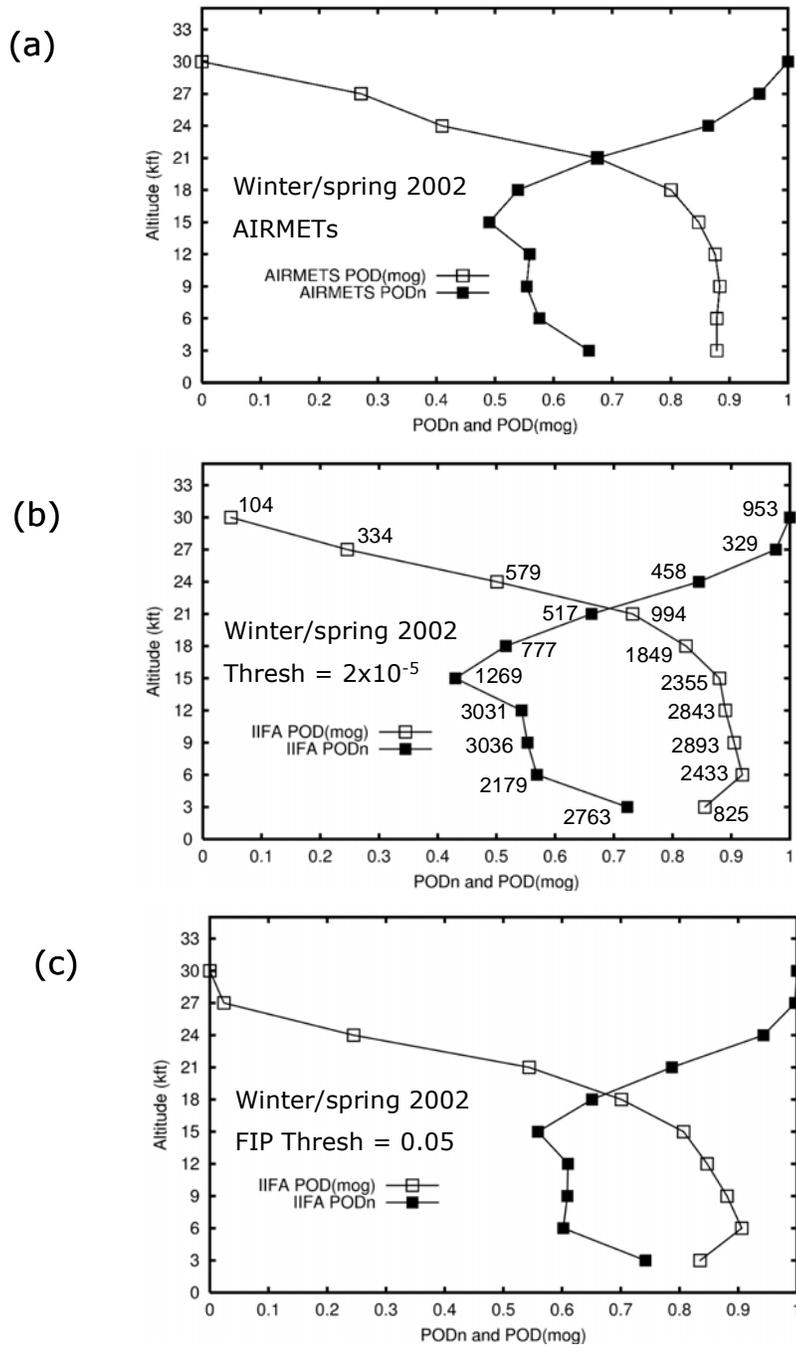
Figure 14 shows height series plots for the winter/spring period (1 January to 20 April 2002). These plots show PODy and PODn as a function of altitude. Separate plots are presented for the AIRMETs and 6-h FIP forecasts with two different thresholds:  $2 \times 10^{-5}$  and 0.05. The PODy and PODn values for the AIRMETs are consistent from the surface to about 18,000 ft. From about 18,000 ft to 30,000 ft, the AIRMET PODn values gradually increase, while the AIRMET PODy values decrease rapidly from 18,000 ft to 30,000 ft. For this same time period, the FIP verification statistics behave somewhat similarly to the AIRMETs. In particular, the PODy increases from the surface to 3,000 ft, decreases slightly between 3 and 18,000 ft, then decreases rapidly from 18 to 30,000 ft. Similarly, the FIP PODn values decrease from the surface to 15,000 ft, then increase rapidly up to 30,000 ft. The FIP forecasts seem to have some skill up to about 24,000 ft. At higher levels, the numbers of PIREPs are relatively small, which may lead to apparent decreases in forecast capability.

Figure 15 presents the height series plots for Fall (1 October to 15 November) 2002. For this period, the AIRMET PODn also is consistent up to about 18,000 ft, then increases toward 1. The PODy decreases slightly up to 18,000 ft, then rapidly up to 30,000 ft. For both FIP thresholds, PODy (PODn) increases (decreases) up to 18,000 ft, then drops (gains) up to 30,000 ft.

## 6.6 Monthly time series

It is instructive to consider long-term trends in the performance of FIP and to examine variations in performance by season. Long-term statistics provided by RTVS are utilized for this analysis. Although most of the results presented in previous sections of this report are based on a slightly different version of FIP than was implemented in RTVS for most of the period considered, the verification results for different versions of the algorithm have been shown to be relatively consistent. In addition, recently discovered discrepancies in the decoded PIREPs have been shown to only have a small impact on the verification statistics (see Section 6.7). Thus, the RTVS long-term statistics should fairly represent seasonal performance of the algorithm. Trends and seasonal variations in the AIRMET performance are also considered simply to provide a baseline for the evaluation.

Monthly time series plots of PODy and PODn for the AIRMETs and for 6-h FIP forecasts based on two thresholds (0.02 and 0.15) are shown in Fig. 16. The AIRMET statistics appear to have a fairly strong seasonal cycle, with decreased PODy and increased PODn in the summer months, and the opposite effects in the winter months. This characteristic of the AIRMET statistics is most likely due to the fact that most icing conditions during the summer are associated with convection, which is accounted for in the Convective SIGMETs issued by the AWC. Thus, fewer icing AIRMETs are issued during the summer months than during other times of the year. The FIP statistics also show some seasonal variations, especially in PODn, but these variations are relatively small. The results for other lead times are consistent with the statistics shown in Fig. 16.



**Figure 14. Height-series plots showing variations of PODy [POD(mog)] and PODn with altitude for (a) AIRMETS; (b) FIP with a threshold of  $2 \times 10^{-5}$ ; and (c) FIP with a threshold of 0.05, for Winter/spring (1 Jan – 20 Apr) 2002. Numbers on (b) indicate numbers of Yes and No PIREPs used to compute the statistics.**

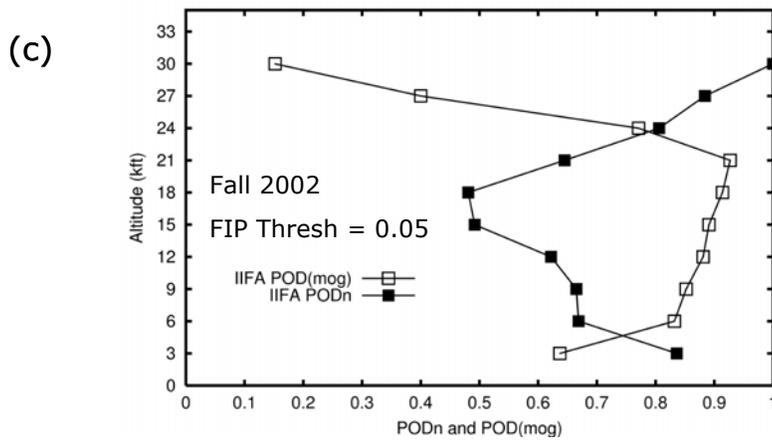
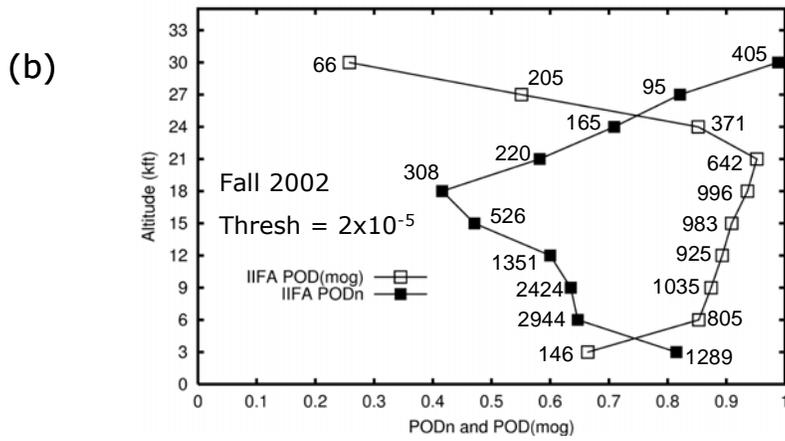
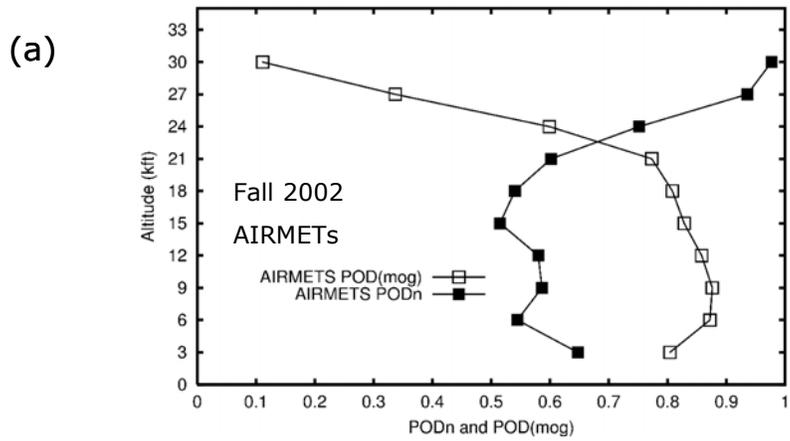
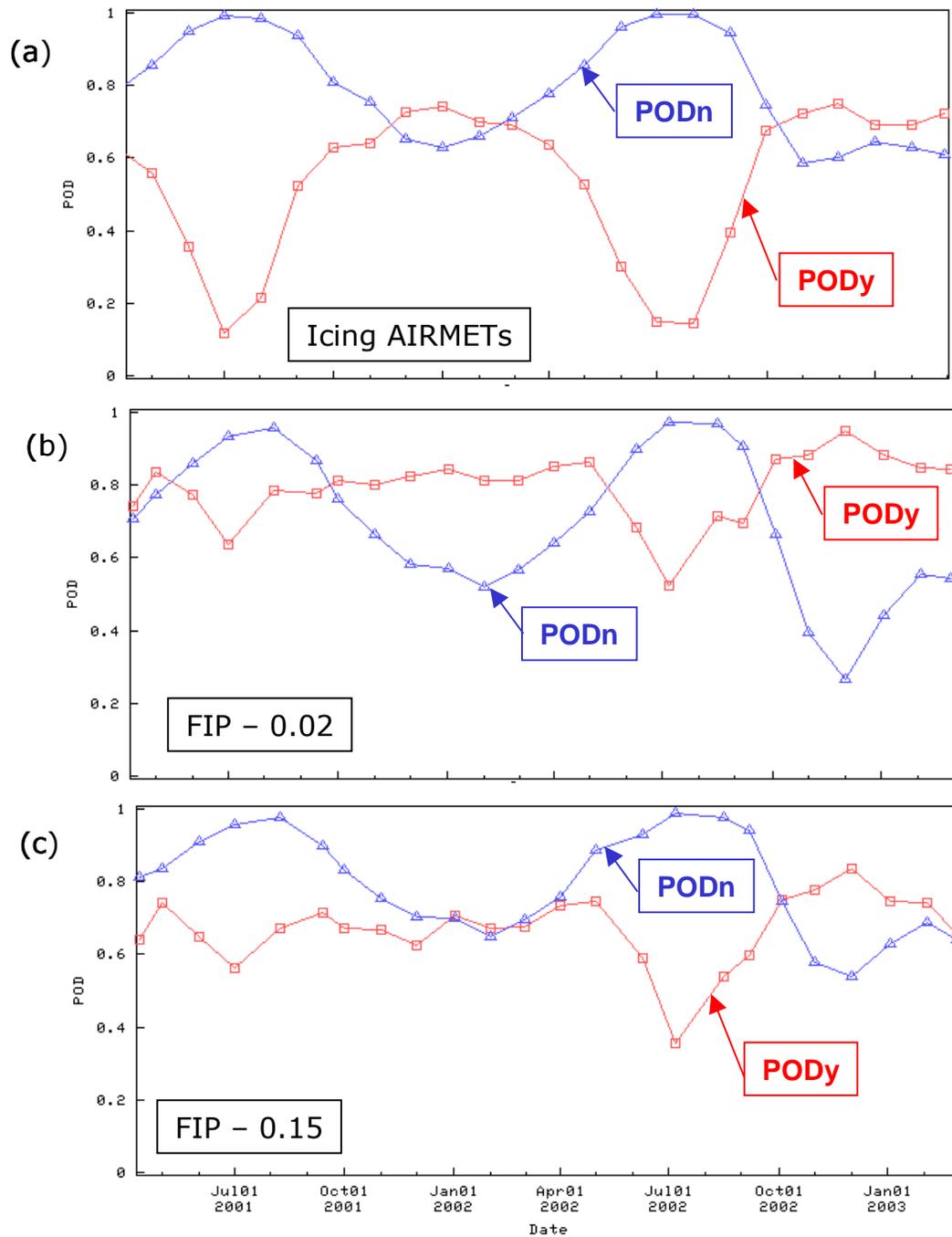


Figure 15. As in Fig. 14, for Fall (1 Oct – 30 Nov) 2002.



**Figure 16. Monthly time series of FIP and AIRMET verification statistics: (a) AIRMETs; (b) 6-h FIP forecasts with a threshold of 0.02; (c) 6-h FIP forecasts with a threshold of 0.15.**

## 6.7 PIREP evaluation

In order to use PIREPs for verification (and a multitude of other applications), they must be decoded from a text message into a digital form. In December 2002, it was discovered that the conversion factor used by the decoder to convert from nautical miles (nmi) to kilometers (km) was incorrect. After the decoder was corrected, the PIREPs were decoded again and compared to those that had been decoded using the incorrect conversion factor. As expected, the error in the conversion factor made very little difference in both the location of the PIREPs and the verification statistics computed using PIREPs. (Nevertheless, the corrected PIREPs were used for all of the verification analyses presented in this report, except the results described in Section 6.6).

In order to assess the effect of the incorrect nmi to km conversion factor, PIREPs from January 2002 were decoded using both the incorrect and corrected versions of the decoder and the resulting digital PIREPs were compared. Of the 39,253 PIREPs decoded for that month, about half (48.9%) of the PIREP locations were unchanged. An additional 24% changed by less than one quarter of a degree in latitude and/or longitude. Only 5% of the PIREPs had changes in latitude and/or longitude of more than three-quarters of a degree.

A comparison of verification statistics for FIP calculated using both the old and corrected PIREPs for the period of January 1<sup>st</sup> to April 20<sup>th</sup> 2002 also reveals little change. As an example, consider verification statistics for FIP with a lead of 3 h and a threshold of 0.05. The calculated POD<sub>y</sub> (MOG) changed from 0.773 to 0.772 when the PIREPs were corrected, while the POD<sub>n</sub> value changed from 0.660 to 0.659. Verification statistics for other lead times and thresholds were similarly small. Clearly, the verification statistics are robust to small errors in PIREP location.

## 7. Conclusions and discussion

This report has summarized evaluations of icing forecasts produced by FIP. This evaluation has followed and built off of several previous evaluations of forecasts and diagnoses produced by FIP and other icing algorithms. The results described here suggest that FIP is a potentially useful icing forecast product. In particular:

- FIP forecasts are skillful, as measured by their ability to discriminate between Yes and No PIREPs of icing.
- FIP forecasts are relatively efficient in terms of the trade-offs between the volume of airspace impacted for a given POD<sub>y</sub> value. They are somewhat less efficient in terms of trade-offs with impacted area (due to the large contributions of thin icing layers to the area computation).

- FIP forecast skill is similar to the skill of the AIRMETS, in terms of the verification approach applied here.
- Day-to-day variations in PODy can be fairly large, partly due to the small numbers of PIREPs available to verify a single forecast. Variations in the volume of airspace covered by FIP are relatively small.
- The skill of FIP forecasts is relatively consistent throughout the year, with relatively small degradations in the summer months.
- FIP forecasts perform best at lower altitudes, but are skillful up to about 21,000 ft or higher.
- FIP skill was not degraded when the algorithm was moved to the 20-km RUC model. In fact some measures of skill were improved when the algorithm was evaluated on the 20-km version of the model. These improvements are likely due to the enhancements to the model microphysics that were incorporated into the new version of the model.
- Errors in the PIREP decoder do not appear to have a major impact on the verification results.

The results described in this report are a small fraction of the verification results that are available. For example, a wide variety of verification information for FIP, other algorithms, and the AIRMETS is available at the RTVS web site (<http://www-ad.fsl.noaa.gov/fvb/rtvs/icing/index.html>).

## **Acknowledgments**

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

We would like to thank the members of the Icing Product Development Team for their support of the independent verification effort over the last several years. We also would like to thank Huming Han of the NOAA/FSL ITS division for obtaining some of the 20-km RUC data that were used in the evaluations. We appreciate the excellent support that Jamie Braid (NCAR/RAP) provided for this study, and also would like to thank Agnes Takacs (NCAR/RAP) for reviewing earlier versions of the report. Finally, we express thanks to the AWRP Leadership Team for their support of independent verification activities.

## References

- Benjamin, S.G., J.M. Brown, K.J. Brundage, B.E. Schwartz, T.G. Smirnova, and T.L. Smith, 1998: The operational RUC-2. *Preprints, 16<sup>th</sup> Conference on Weather Analysis and Forecasting*, Phoenix, AZ, American Meteorological Society (Boston), 249-252.
- Bernstein, B.C., F. McDonough, and M.K. Politovich, 2001: Integrated Icing Diagnosis Algorithm - Technical Description. Report to the FAA Aviation Weather Research Program (Available from M. Politovich, NCAR, P.O. Box 3000, Boulder CO 80307-3000).
- 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. *Weather and Forecasting*, **12**, 890-914.
- Brown, B.G., and G.S. Young, 2000: Verification of icing and icing forecasts: Why some verification statistics can't be computed using PIREPs. *Preprints, 9<sup>th</sup> Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 Sept., American Meteorological Society (Boston), 393-398.
- Brown, B.G., T.L. Fowler, and J.L. Mahoney, 2001a: Verification Results for the Integrated Icing Forecast Algorithm (IIFA). Report to the FAA Aviation Weather Research Program and the FAA Aviation Weather Technology Transfer Board. Available from B.G. Brown ([bgb@ucar.edu](mailto:bgb@ucar.edu)), 4 pp.
- Brown, B.G., J.L. Mahoney, R. Bullock, T. L. Fowler, J. Henderson, and A. Loughe, 2001b: Quality Assessment Report: Integrated Icing Diagnostic Algorithm (IIDA). Report to the FAA Aviation Weather Research Program and the FAA Aviation Weather Technology Transfer Board. Available from B.G. Brown ([bgb@ucar.edu](mailto:bgb@ucar.edu)), 36 pp.
- Brown, B.G., J.L. Mahoney, and T.L. Fowler, 2002: Verification of the in-flight icing diagnostic algorithm (IIDA). *Preprints, 10<sup>th</sup> Conference on Aviation, Range, and Aerospace Meteorology*, Portland, OR, 13-16 May, American Meteorological Society (Boston), 311-314.
- Efron, B. and R. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall, New York.
- Kane, T.L., and B.G. Brown, 2000: Confidence intervals for some verification measures – a survey of several methods. *Preprints, 15<sup>th</sup> Conference on Probability and Statistics in the Atmospheric Sciences*, Asheville, NC, 8-11 May, American Meteorological Society (Boston), 46-49.
- 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, 7<sup>th</sup> Conference*

*on Aviation, Range, and Aerospace Meteorology*, Long Beach, CA, American Meteorological Society (Boston), J26-J31.

Mahoney, J.L., J. K. Henderson, B.G. Brown, J.E. Hart, A. Loughe, C. Fischer, and B. Sigren, 2002: The Real-Time Verification System (RTVS) and its application to aviation weather forecasts. *Preprints, 10<sup>th</sup> Conference on Aviation, Range, and Aerospace Meteorology*, 13-16 May, Portland, OR, American Meteorological Society (Boston), 323-326.

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

McDonough, F., B.C. Bernstein, and M.K. Politovich, 2003: The Forecast Icing Potential (FIP) Technical Description. Report to the FAA Aviation Weather Technology Transfer Board. Available from M.K. Politovich (NCAR, P.O. Box 3000, Boulder, CO 80307-3000), 30 pp.

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

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