Quality Assessment Report
Forecast Icing Product – Icing Probability

Michael B. Chapman and Jennifer L. Mahoney
National Oceanic and Atmospheric Administration
Earth System Research Laboratory
Global Systems Division
Boulder, Colorado

Quality Assessment Product Development Team

05 April 2007

1. INTRODUCTION

This report presents the results of a statistical analysis of the Forecast Icing Product (FIP) Icing Probability Field (FIP_Prob). The purpose of this report is to document the performance of the FIP_Prob as part of the Aviation Weather Technology Transfer (AWTT) process. The FIP_Prob was generated through the calibration of the FIP Icing Potential (FIP_IP) field by using a method developed by Brown and Bernstein (2006). This method was originally designed to calibrate the Current Icing Product (CIP) Icing Potential into an Icing Probability Field. To objectively assess the quality of the FIP_Prob, several analyses were performed including a study of the reliability of the calibration of the product and an analysis of the skill of the algorithm. Pilot Reports (PIREPs) of YES/NO icing were collected and compared with the FIP_Prob over the 01 January – 31 March 2005 (Winter-2005) time period.

This report is organized in the following way. Section 2 describes the algorithm and observational data sets. A description of the methods and techniques used in this study is outlined in section 3. The results of the study are presented in section 4. The conclusions and discussion are presented in section 5.

2. DATA SETS

For this study, the FIP_Prob was evaluated over the Winter-2005 time period using PIREPs as observations of YES/NO icing which were located over the CONUS between 0-30 kft. The initial calibration of the FIP_Prob was accomplished by using every third day over Winter-2005 (Kucera et al. 2007). This was consistent with Chapman et al. (2006) where the same one-month period was set aside for FIP Severity development purposes. The remaining two-thirds of the FIP_Prob dataset was independently evaluated for this report.

2.1 FIP_IP

The FIP_Prob field is generated through the calibration (Brown and Bernstein 2006) of the FIP_IP field (McDonough et al. 2004). The FIP algorithm is a physically-based situational forecast icing product which combines 20-km Rapid Update Cycle (RUC) (Benjamin et al. 2004) model fields (e.g. Temperature, relative humidity, precipitation, vertical velocity, and super-cooled liquid water) with fuzzy logic to generate potential, severity, and super-cooled large drop (SLD) fields. Forecasts with 3-h, 6-h, 9-h, and 12-h lead times were available for evaluation. Fig. 1 is a plot of the maximum Icing Probability from the CIP as displayed on the Aviation Digital Data Service (ADDS) webpage (http://adds.aviationweather.gov). Figure 2 is a plot of the Icing Probability Field from the CIP thresholded at 25% with maximum icing severity overlaid as displayed on the ADDS webpage. The ADDS site also displays a hybrid Icing Probability/Severity field similar to Fig. 2 but at a 50% threshold. While Figs 1 and 2 are plots of the CIP product, the future FIP_Prob plots will probably be very similar.

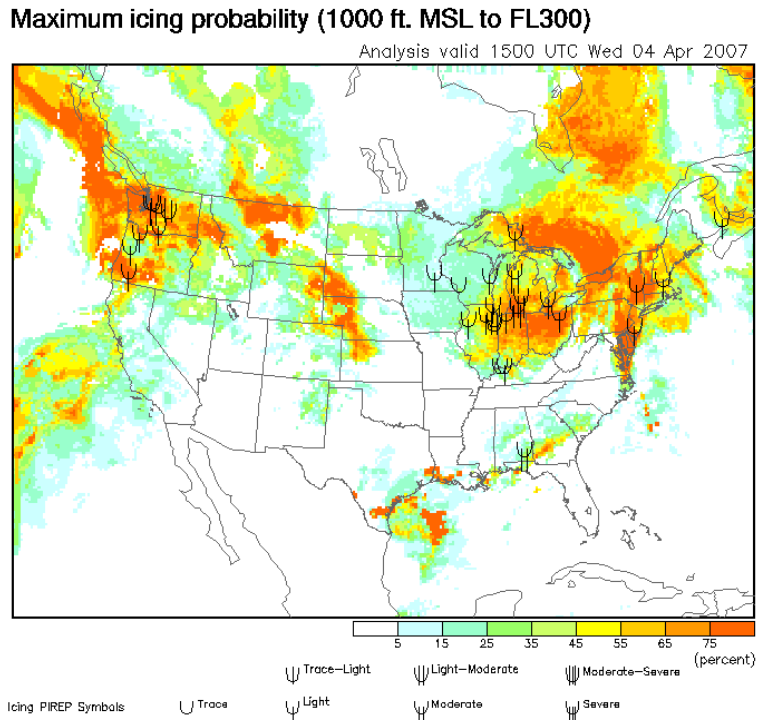

Figure 1. Plot of the CIP Maximum Icing Probability forecast (colors show intensity) with available PIREP observations plotted as symbols. See key for intensities. Image courtesy of ADDS (http://adds.aviationweather.gov)

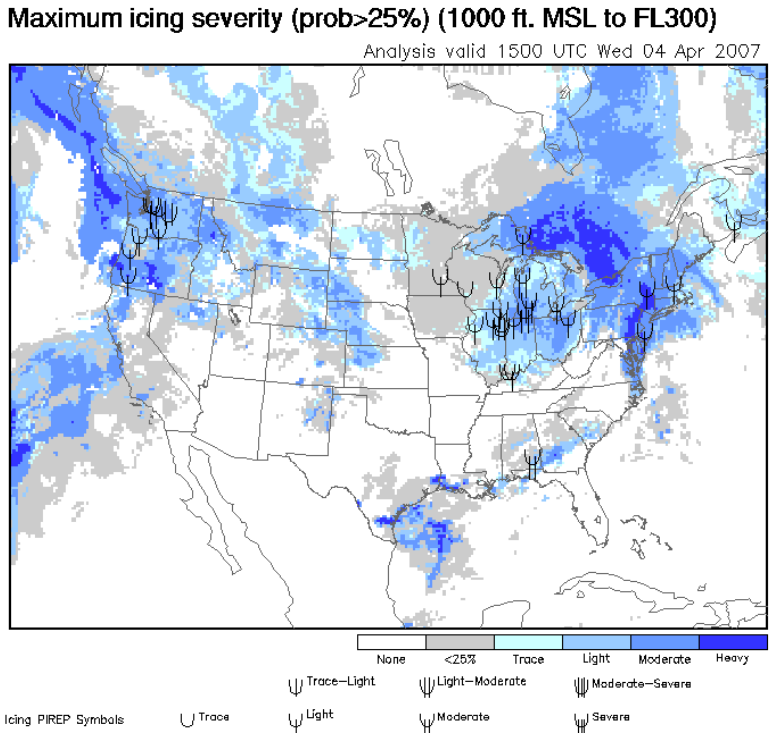

Figure 2. Plot of the CIP Maximum Icing Severity field overlaid on Icing Probability at a threshold of 25%. Available PIREPs are also displayed (symbols). Image courtesy of

ADDS (http://adds.aviationweather.gov)

2.2 PIREPs

PIREPs are categorical reports of icing (or lack thereof) conditions as experienced by pilots. While subjective in nature, PIREPs are considered "truth" for verification purposes in this study due to the lack of any available objective observations of in-flight icing. PIREPs are also non-systematic and biased both spatially and temporally (Kane et al. 1998). For example, PIREPs are generally located in high air- traffic areas during high air-traffic times (e.g. around airports and during daylight hours). PIREPs are also subject to the experience of the pilot as well as the type of aircraft being flown. Figures 1 and 2 show an example of the available hourly icing PIREPs overlaid with the CIP Icing Probability product for a specific time.

3. METHODS

This section summarizes the verification methods used to match the forecast to the observations as well as the statistics generated for analysis after specific verification techniques (as described in section 3.2) were applied to the matched forecast/observation pairs.

3.1 Matching Methods

The methods used to match the PIREPs to the FIP_Prob forecast are similar to methods used in past studies (Chapman et al. 2006 and Fowler et al. 2006) of the CIP and other icing products. The PIREP is matched to the closest four surrounding grid points as well as the four grid points above and below the initial flight level. The maximum FIP_Prob value of the 12 grid points is then compared with the icing intensity of the PIREP. PIREPs that were available between the lead time and the lead time+1-h were used for the evaluation.

3.2 Verification Methods

The analysis performed for this report is organized in two specific sections. Section 3.2.1 describes the verification method that has been commonly used in past quality assessment studies in order to assess skill. The methods used to assess the reliability of the calibration of the FIP_Prob product, similar to a study done by Fowler et al. (2006), are found in Section 3.2.2.

3.2.1 Skill Assessment

The verification methods used for this section are common techniques that treat the FIP forecasts and icing observations from PIREPs as dichotomous YES/NO values (Brown et al. 1997). This is achieved by applying a set of pre-determined thresholds to the FIP forecast that separate the forecast into either the YES or NO category. The PIREPs are thresholded in several ways to decide whether or not to record them as YES

or NO observations. For example, if the analysis requires the positive icing observations to be of Moderate Or Greater (MOG) icing values, then the reports that are categorized with intensities from light-to-moderate icing to severe icing are classified as YES observations. In this case, PIREPs with an explicit report of "No icing" are classified as NO observations. If the observations are thresholded for "light" icing, all positive icing PIREPs are recorded as YES observations. This means that PIREPs with any positive icing categories are classified as YES observations, and the PIREPs with an icing intensity of zero are classified as NO observations. A 2X2 contingency table is then generated by summing up the counts of the matched YES/NO forecast/observation icing pairs over each threshold (*Table 1*).

Table 1. 2X2 Contingency Table for YES/NO Forecast/Observation Pairs.

	Observation		
Forecast	**YES**	**NO**	**Total**
YES	YY	YN	YY+YN
NO	NY	NN	NY+NN
Total	YY+NY	YN+NN	YY+YN+ NN+NY

PODy and PODn are the primary statistics that are computed from *Table 1*. They are estimates of the percentage of YES/NO observations that are correctly captured by the forecast (Brown et al. 1999). When used together, these two statistics measure the algorithm's ability to discriminate between the YES/NO observations. Percent Volume (%Vol) is a measure of the percent of total airspace volume that may be encompassed by a positive forecast of icing. The %Vol and PODy can be analyzed together to assess the efficiency of the forecast. *Table 2* provides a definition and description of the three statistics discussed above.

Table 2. Definitions and Descriptions of Statistics Computed from 2X2 Contingency Table

Statistic	**Definition**	**Description**
PODy	YY/(YY+NY)	Probability of Detection of "Yes" observations
PODn	NN/(YN+NN)	Probability of Detection of "No" observations
% Vol	100 x (Forecast Volume) / (Total Volume)	% of the 3-dimensional forecast domain where icing is expected to occur

The relationship between PODy and 1-PODn for different thresholds is borrowed from the verification approach known as "Signal Detection Theory" (SDT) (Reference – Wilks, 2006). The curve that joins the points X=1-PODn and Y=PODy for the different thresholds, represents this SDT relationship and is referred to as the Relative Operating Characteristics (ROC) curve. Analysis of the ROC curve provides insight into the "skill"

the algorithm. For example, the closer the curve is to the upper left-hand corner of the plot the better the forecaster skill.

3.2.2 Reliability Assessment

The goal of the original calibration of the FIP_Prob into the Icing Probability field (Kucera et al. 2007) is to accurately represent the probability of in-flight icing around a particular FIP grid point. Reliability diagrams (Wilks 2006) were used to assess the quality of this calibration, and to determine how the newly calibrated FIP_Prob product relates to the estimated relative frequency of icing. In accurate calibrations, icing occurs approximately *(100 x p)%* of the time when *p* is forecast (Fowler et al. 2006). Because of the non-systematic nature of the PIREPs, the climatological probability of icing is not known. Therefore, an estimated value of 4.64% for the icing climatology was used to provide consistency with past evaluations involving the CIP (Brown and Bernstein 2006, Fowler et al. 2006 and Kucera at al. 2007). This approximation of the climatological probability of icing is calculated and described in the initial calibration study for the CIP (Brown and Bernstein 2006)

4. RESULTS

Sections 4.1 and 4.2 document the results of the verification technique described in section 3. The reliability of the calibration will be discussed first in section 4.1, followed by the skill discussion in section in 4.2.

4.1 Reliability

Figure 3 is a reliability diagram of the calibrated FIP_Prob versus the approximated icing event probability for the four lead times (3-h, 6-h, 9-h, and 12-h). The results show that the 3-h, 6-h, and 9-h lead times are reasonably calibrated up to 35% probability with a slight underestimation as the lead times increase. The 3-h and 6-h lead times continue to be accurately calibrated up to 45% probability. The results show between the 15-35% probabilities the 12-h lead-time is underestimated. As the event probability increases to 45%, the results indicate an underestimate of the icing probability for the 9-h forecast. Similarly, as the event probability reaches 55%, the 3-h and 6-h icing probability forecasts are underestimated by nearly 10%. Because this plot only shows the average over the entire Winter-2005 time period, an analysis of the variability on a day-to-day timescale was performed to better assess the calibration and is shown in Figs 4-7.

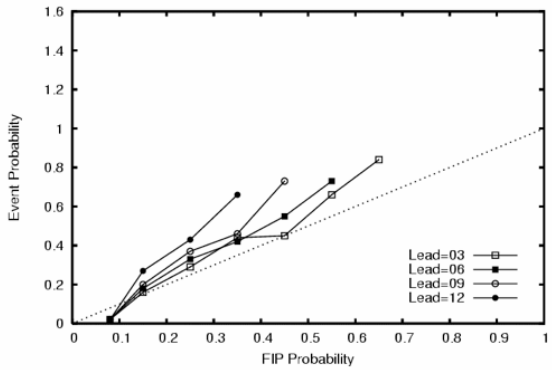

Figure 3. Reliability diagram for FIP_Prob calibration for all lead times.

Figure 4 is a reliability diagram for the 03-h lead time. To assess the variability of these data at the different forecast probability thresholds, boxplots were generated from an analysis of the daily calibration data. The results in Fig. 4 show that the FIP_Prob field for this lead time is calibrated reasonably well as shown by the overlap of the X=Y line and the boxplot median probability values for forecast probability thresholds less than 55%. A slight underestimation is indicated above an event probability of 45%.

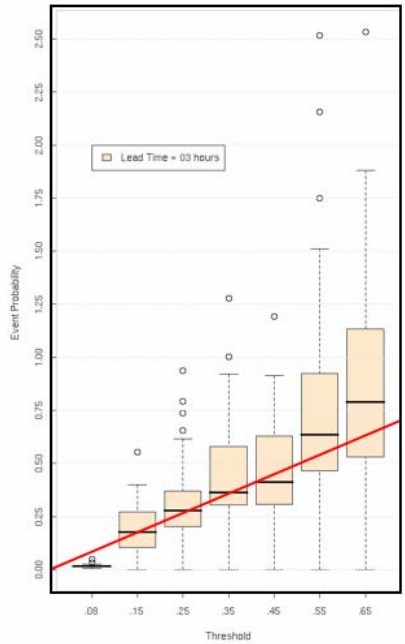

Figure 4. Reliability diagram for FIP_Prob calibration at lead=3 with X=Y line (red).

Figure 5 is similar to Fig. 4 but for the 6-h lead time. It also shows that the forecast probability field is roughly calibrated with a slight underestimation of the probabilities at thresholds greater than 35 %. The large variability and underestimate of icing probability at the 55% threshold is most likely due to the algorithms under-forecasting of rare icing events.

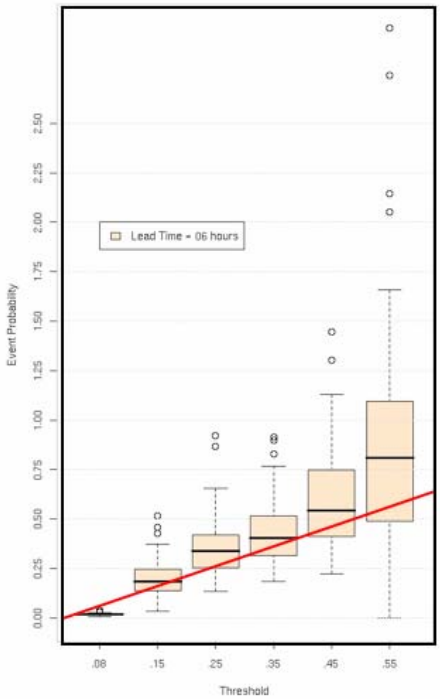

Figure 5. Reliability diagram for FIP_Prob calibration at lead=6 with X=Y line (red).

The results illustrated in Fig. 6 are similar to the results in the last two figures (Figs 4 and 5) with the FIP_IP for the 9-h lead time being roughly calibrated up to 15% probability. However, at forecast thresholds greater than 15%, the underestimate is much larger at 9-h than for the 3 or 6-h lead times.

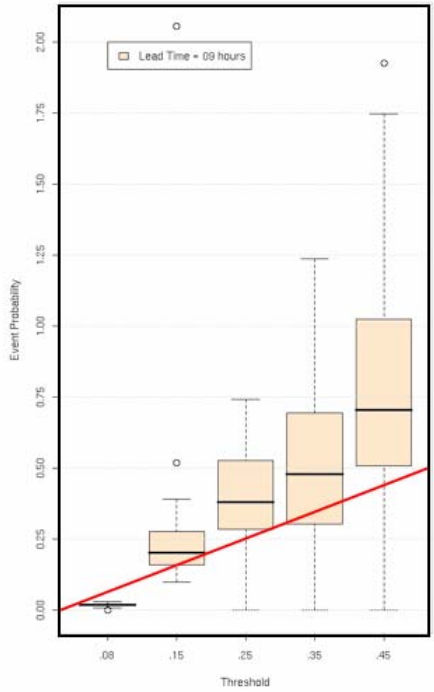

Figure 6. Reliability diagram for FIP_Prob calibration at lead=09 with X=Y line (red).

Figure 7 illustrates similar results for the FIP_Prob at the 12-h lead time. The results show that the FIP_Prob is roughly calibrated up to 15% probability, but is significantly underestimated at thresholds greater than 15%. Part of this underestimation of the event probability could be the result of the variability in the PIREP data set or the difference between the types and/or frequency of icing events used for the one-month initial calibration period versus the two-month period that was evaluated in this study.

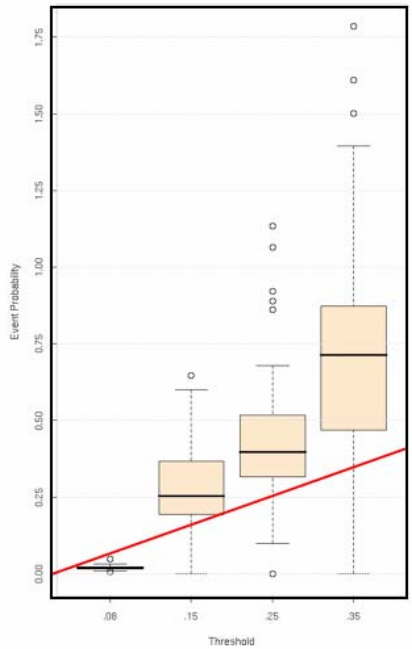

Figure 7. Reliability diagram for FIP_Prob calibration at lead=12 with X=Y line (red)

4.2 Skill

This section documents the results of FIP_Prob skill assessment over the Winter-2005 time period. Figure 8 is a ROC plot of the four lead times available for evaluation. This figure shows that the FIP_Prob has positive skill at all lead times. In fact, there is little difference between 3-h and the 12-h lead times. These results are similar to the findings from the recent FIP Severity evaluation (Chapman et al. 2006), which showed little difference between the skill of the four lead times for the FIP Severity algorithm.

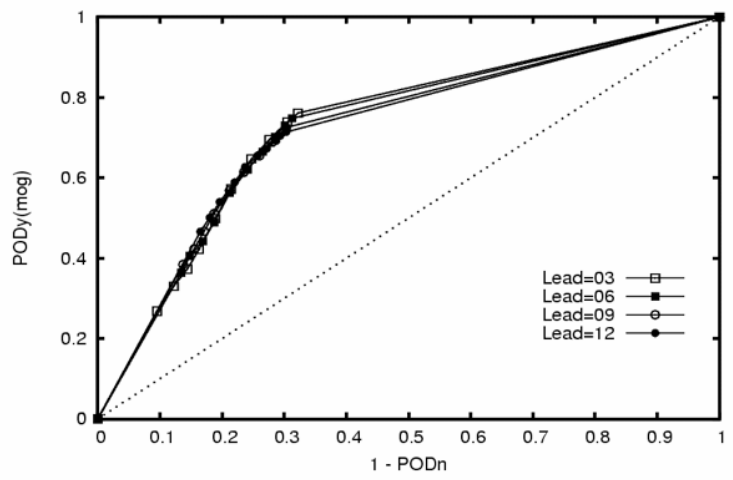

Figure 8. ROC plot for Winter 2005 FIP_Prob product for all lead times.

Figure 9 is a plot of the PODy(MOG) versus the %Vol statistics. These results show a difference slight between the four lead times with 3-h lead time indicating slightly better efficiency than the other three lead times. This result is also similar to Chapman et al. (2006) findings, which indicated that as lead times increased, the forecast was slightly less efficient. These results imply that with an increase in forecast lead time. The FIP_Prob overforecasts icing events.

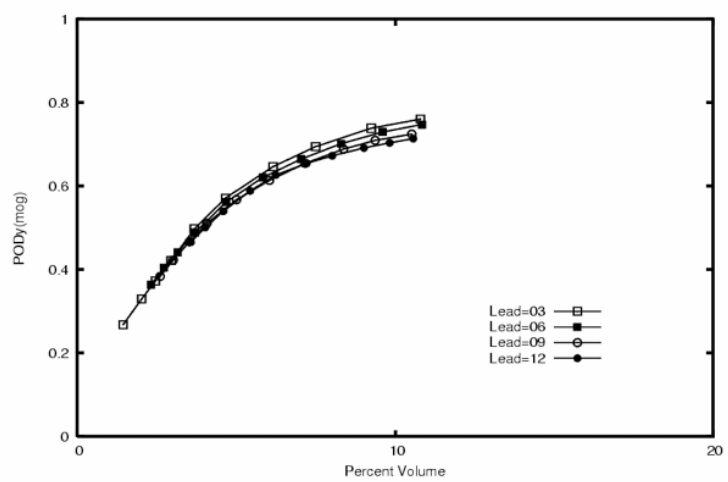

Figure 9. PODy(MOG) vs. %Vol plot for Winter 2005 FIP_Prob algorithm

5. CONCLUSION/DISCUSSION

This study assessed the reliability of the calibration of the FIP_IP field into the FIP_Prob as well as the skill and efficiency of the FIP_Prob after the calibration was performed. The results indicated that the FIP_Prob is reasonably calibrated for the 3-h, 6-h lead times for probabilities less than 35%. The results also indicated that the algorithm is roughly calibrated for the 9 and 12-h lead times at probabilities less than 15%. This could indicate that the algorithm underforecasts the rare events that occur at higher probabilities. Another possibility for this discrepancy might lie in the variability of the PIREP observations or a difference in the type or frequency of icing events between the one-month time period used for the development of the FIP_Prob and the two-month time period used to assess the quality of the FIP_Prob. The results of the skill analysis were consistent with past studies of the FIP_IP and Icing Severity algorithm which showed positive skill over all four lead times with virtually no change in skill as lead time increases. The efficiency analysis was also consistent with past studies and indicated a slight overforecast of positive icing conditions with an increase in lead-time.

In conclusion, more work needs to be accomplished in order to accurately calibrate this field over differing circumstances. For this evaluation, it was assumed that the calibration was homogeneous across the CONUS as well as across all flight levels and was evaluated as such. The results showed that the FIP_Prob was roughly calibrated over the four lead times. In reality, this calibration, as well as the icing climatology value, will most likely be different across geographic regions, seasons, and flight levels. In future evaluations, the calibration of the FIP_Prob will be assessed over a larger time period and stratified over the CONUS at all combined altitudes (similar to this study), different seasons, four different geographic regions, and possibly higher-resolution flight level layers.

6. ACKNOWLEDGEMENTS

The authors would like to thank Paul Kucera and Barb Brown from the National Center for Atmospheric Research for discussions of forecast probability. This evaluation was performed in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are solely opinions of the authors and do not represent the official policy of the FAA.

7. REFERENCES

Benjamin S. G., T. G. Smirnova, K. Brundage, S. S. Weygandt, T. L. Smith, B. Schwartz, D. Dévényi, J. M. Brown, and G. A. Grell, 2004: A 13-km RUC and beyond: Recent developments and future plans. *11th Conf. on Aviation, Range and Aerospace Meteorology,* Hyannis, MA, 3 – 8 Oct, Amer. Meteor. Soc., Boston.

Brown B.G. and B.C. Bernstein, 2006: A calibration approach for CIP based on forecast

performance measures. Report submitted to the FAA Aviation Weather Research Board. Available from B. Brown (bgb@ucar.edu)

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

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

Chapman M.B., M. Pocernich, A. Holmes, P. Boylan, P.A. Kucera, B.G. Brown, J. L. Mahoney, and J.T. Braid, 2006: Quality assessment report: Forecast Icing Product (FIP) – severity. Report submitted to the FAA Aviation Weather Research Board. Available from P. Kucera (pkucera@ucar.edu).

Fowler, T.L., Chapman, M.B., Holmes, A., Brown, B.G., Mahoney, J.L., Braid, J.T., Boylan, P., 2006: Quality Assessment Report- Current Icing Potential (CIP): Probability, Severity, SLD. Report submitted to the FAA Aviation Weather Technology Transfer Board.

Kane, T. L., B.G. Brown, and R. Bruintjes, 1998: Characteristics of pilots reports of icing, *Preprints, 14th Conference on Probability and Statistics*, 11-16 January, Phoenix, p. 90-95.

Kucera, P.A., B.G. Brown, and A. Holmes, 2007: Calibration of the Forecast Icing Potential (FIP) – Icing Potential. Report submitted to the FAA Aviation Weather Research Board. Available from P. Kucera (pkucera@ucar.edu).

McDonough, F., B.C. Bernstein, M.K. Politovich, and C.A. Wolff, 2004: The forecast icing potential (FIP) algorithm. Preprints, *20th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*. Available from the American Meteorological Society.

Wilks, D.S. 2006: *Statistical Methods in the Atmospheric Sciences*. Associated Press, 264 pp