J1.1 An Evaluation of Impacts of Grid Resolution on the Verification of Aviation Weather Forecasts

Michael B. Chapman*, Barbara G. Brown and Agnes Takacs
National Center for Atmospheric Research
Boulder, Colorado

1. Introduction

Past evaluations of both in-flight icing and turbulence algorithms are completed using a set of standard forecast verification techniques defined for this purpose. These techniques treat both the forecasts and observations as YES/NO values even though the forecast field is continuous as opposed to binary. Over the past several years the resolution of many operational weather models has increased substantially. For example, the Rapid Update Cycle (RUC) (Benjamin et al. 1999) has increased from 60km resolution from 1994-1998, to 40km from 1998-2001, to 20km from 2001-2005 and will be increased to 13km resolution some time in 2005. Many of the aviation weather algorithms produced at the National Center for Atmospheric Research (NCAR) have also changed in resolution because of their dependence on models such as the RUC.

This study investigates the effect of increased resolution of the RUC model on the standard verification statistics, for a variety of aviation weather algorithms. The Current Icing Potential (CIP; (McDonough and Bernstein, 1999; Bernstein et al. 2004) and the Graphical Turbulence Guidance (GTG; Sharman et al., 2004) are algorithms generated at NCAR and are evaluated for this study.

Four different methods for matching the forecast value to an observation value are investigated. In the past, for icing and turbulence algorithm evaluations, the forecast value was determined by taking the maximum value of the four grid points surrounding an observation. The max value is used along with the average and two different types of interpolation in order to ascertain whether or not changing the grid resolution has any effect on the technique used to infer a forecast value. The analysis includes several continuous months of forecast and observations for each algorithm.

2. Data

2.1 Current Icing Potential (CIP)

CIP is an operational in-flight icing algorithm that diagnoses icing by combining satellite, surface, radar, lightning and PIREP observations with fields from the 20-km Rapid Update Cycle (RUC) numerical weather prediction model (Benjamin et al. 2001). The output of CIP is an “icing potential” with floating point values from zero (no potential for icing) to 1.0 (icing very likely) (Fig. 1). The evaluation of the CIP icing product includes data from 01 January 2003 – 31 March 2004. All layers (0-42kft) are evaluated together using four different techniques (maximum, average, 1/distance interpolation, and 1/distance^2

*Author Contact Information –
Michael Chapman NCAR, PO Box 3000, Boulder, CO 80307-3000
E-Mail:mchapman@ucar.edu

interpolation) to infer the diagnostic value.

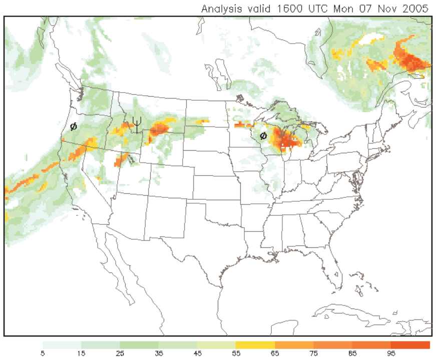

Fig. 1 An example of CIP Icing Potential from the Aviation Digital Data Service (ADDS) website. (http://adds.aviationweather.noaa.gov)

2.1 Graphical Turbulence Guidance (GTG)

GTG is a turbulence forecasting algorithm that uses a combination of several individual turbulence diagnostics generated form the 20km RUC which are weighted by comparisons to turbulence pilot reports. The evaluation of the GTG2 algorithm includes data from 01 January – 31 March 2004. Two altitude ranges, 10 – 20kft and 20 – 46 kft, are evaluated separately using the maximum forecast turbulence value to infer the forecast value. Fig 2 is a plot of the GTG product as displayed in the Aviation Digital Data Service (http://adds.aviationweather.noaa.gov) web page.

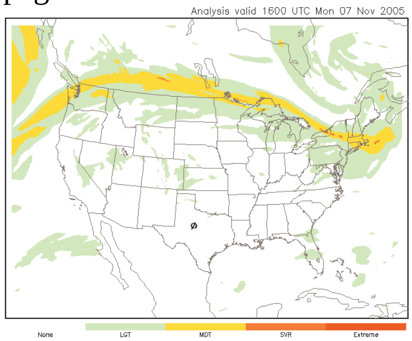

Fig 2 An example of the GTG algorithm taken from the ADDS website.

2.2 Pilot Reports (PIREPs)

PIREPs, which signify an observation of icing/turbulence or lack thereof, are vital because they are the primary “ground truth” observations available to verify the presence or absence of icing/turbulence at a specific location and time. However, standard PIREPs have several drawbacks. They are subjective in nature, typically do not provide high resolution information, and underreport the absence of icing and turbulence.

3. Verification Technique

The methods utilized in the evaluation of CIP/GTG are based on standard techniques of forecast verification. They are described in greater detail in Brown et al. (1997). The icing forecast verification methodology treats icing/turbulence forecasts and observations (PIREPs) as Yes/No values. Brown et al. (1999) outlines how this method is able to be extended to verify continuous, rather than binary fields. Icing/turbulence diagnoses produced by CIP/GTG can be converted into a set of Yes/No values by applying a variety of thresholds. For example, applying a threshold of 0.35 to CIP diagnoses would lead to a Yes value for all grid points with an icing potential greater than or equal to 0.35 while each grid point with a value less than 0.35 would be assigned a No value. The verification methods are based on a two-by-two contingency table (Table 1). Each cell in this table contains a count of the number of times a particular forecast/observation pair was observed.

Table 1. Contingency table for YES/NO forecasts. Elements in cells are forecast-observation pairs.

Forecast	*Observation*		*Total*
	YES	*NO*	
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 verification statistics that are included in this evaluation. They are estimates of the proportions of Yes and No observations that are correctly diagnosed. Together, PODy and PODn measure the ability of the forecasts to discriminate between Yes and No icing observations. Percent Volume is another statistic that is used to assess the efficiency of an algorithm. Table 2 gives the definition and description of these statistics.

Table 2. Verification Statistics used for the evaluation of CIP.

Statistic	*Definition*	*Description*
PODy	YY/(YY+NY)	Probability of detection of YES observations
PODn	NN/(NN+YN)	Probability of detection of NO observations
%Vol	[(Forecast Volume)/ (Total Volume)] x 100	Percent of total airspace that is impacted by the forecast

The relationship between PODy and 1-PODn for different thresholds is the basis for the verification approach known as “Signal Detection Theory” (SDT). This relationship can be represented for a given algorithm with the curve joining the (1-PODN, PODy) points for different thresholds. The resulting curve is known as the “Relative Operating Characteristic” (ROC) curve in SDT. When PODy is plotted on the y-axis, the closer a given curve comes to the upper left corner, the better the forecast. The area under the curve is a measure of overall forecast skill and provides another measure that can be compared among forecast products. This measure is not dependent on the threshold used. A forecast with zero skill would have an ROC area of 0.5.

4.1 Results

4.1 GTG2

The analysis of GTG2 involves an evaluation using the maximum turbulence value at the closest grid point to the PIREP location, the four surrounding grid points, and the nine surrounding grid points. Changing the number of grid points is a simple way of simulating the verification at different scales. Fig 3 shows the results of the evaluation at mid-levels (10-20kft) for the 6-h forecast. The ROC curves in Fig 3a show very similar results while the PODy increases with the number of grid points for a given percent volume in Fig 3b. The evaluation represented by the plots in Fig 4 is similar to Fig 1 but for higher levels (20 – 46kft). The results are similar for Figs 3 and 4 as well.

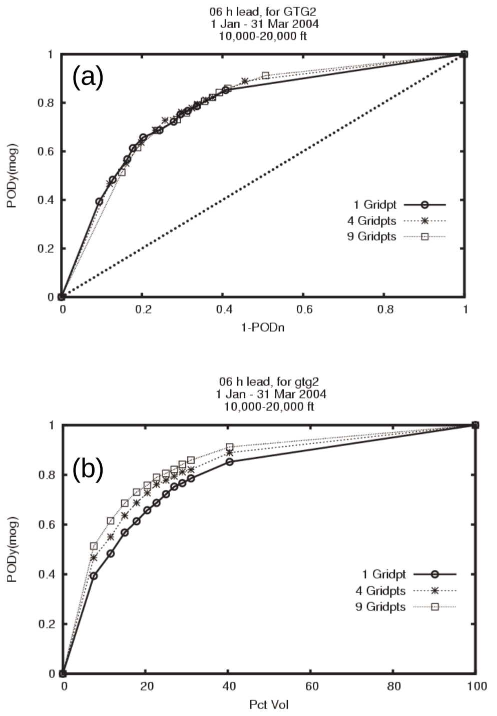

Fig 3 (a) ROC curves and (b) % Volume plots for comparing GTG2 performance for different numbers of grid points at mid-levels.

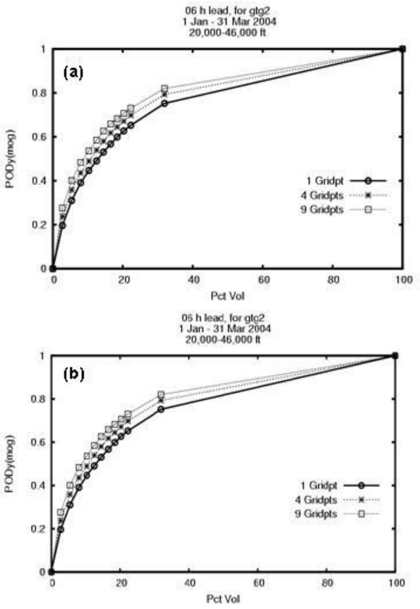

Figure 4 (a) ROC curves and (b) % Volume plots for comparing GTG2 performance for different numbers of grid points at upper levels (20-46kft).

The similar ROC curves in Figures 3a and 4a show that the results are insensitive to the number of grid points used for matching. This insensitivity to the scale of the evaluation shows that the verification technique is relatively robust when comparing PODy to 1-PODn statistics. Conversely, it is important to note that even though the curves do not change, the actual statistics do change. In fact, both the PODy and 1-PODn statistics increase as the number of grid points increases. The increase in PODy is evident in Figs 3a and 4a with the curve shifting to the left as the grid point number increases.

4.2 CIP

CIP is evaluated using several methods in order to assess how changes in scale might affect the verification results. Evaluations similar to those for GTG2 show similar results. As the number of grid points used to infer the maximum icing potential increases, both the PODy and 1-PODn statistics increase. Also included in this assessment are four different methods used to infer the icing potential value at all grid points to be evaluated. Table 3 is a list of the four methods used to infer the icing potential.

Table 3. Methods used to infer icing potential values

Method	*Description*
Maximum	Max icing potential from four surrounding grid points
Average	Average icing potential from four surrounding grid points

Interpolation (1/D)	Icing potential inferred by interpolation using 1/Distance as the weight
Interpolation ($1/D^2$)	Icing potential inferred by interpolation using 1/Distance Squared as the weight

Figs 5-7 are ROC plots of the four methods for: four surrounding grid points (Fig 5), nine surrounding grid points (Fig 6), and sixteen surrounding grid points (Fig 7). For each of these plots the points are virtually indistinguishable as they fall along the same curve. This is important because it shows how robust this technique is at assessing an algorithms ability to distinguish between YES and NO reports. In the past, the maximum icing potential was utilized for verification. These results show that any one of the four methods suffice when assessing the skill of an algorithm.

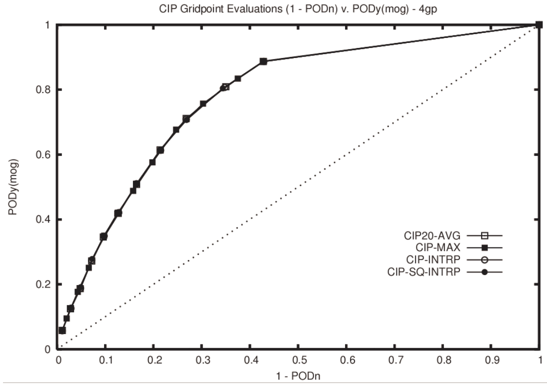

Fig 5. ROC curve for CIP evaluation at four surrounding grid points for four differing methods of inferring the icing potential value

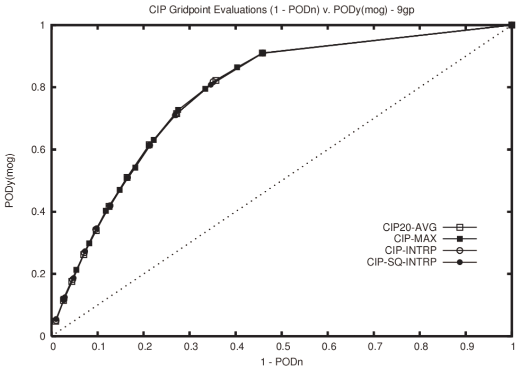

Fig 6. ROC curve for CIP evaluation at nine surrounding grid points for four differing methods of inferring the icing potential value

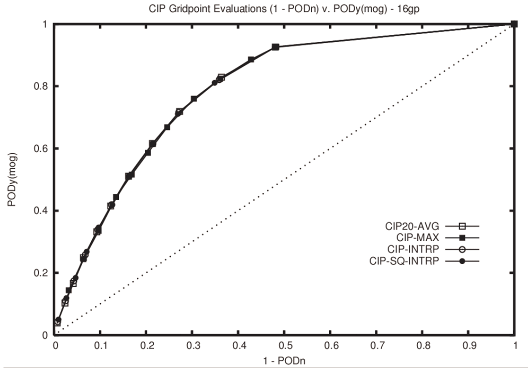

Fig 7. ROC curve for CIP evaluation at sixteen surrounding grid points for four differing methods of inferring the icing potential value

Figs 8-10 show plots of the PODy versus percent volume. These plots indicate that the PODy statistics for the two interpolation method and the averaging method are almost identical, while the PODy statistics for the maximum icing potential value increases as the number of grid points increases. This makes sense because, as the number of grid points increases, the PODy for the maximum inference method should increase as well since there is more of a chance of picking up higher valued grid points. Since the percent volume of positive icing potential does not change, the curve will move towards the right as grid resolution increases.

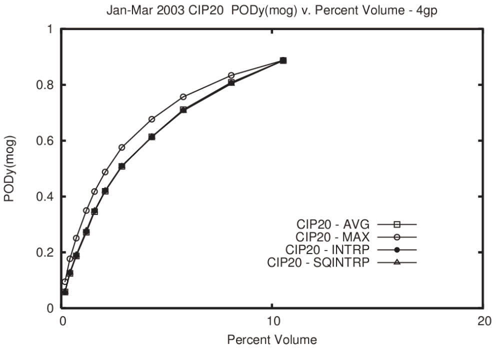

Fig 8. PODy v. %Volume curve for CIP evaluation at four surrounding grid points for four differing methods of inferring the icing potential value

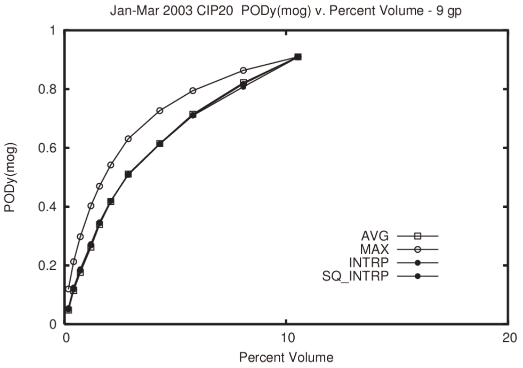

Fig 9. PODy v. %Volume curve for CIP evaluation at nine surrounding grid points for four differing methods of inferring the icing potential value

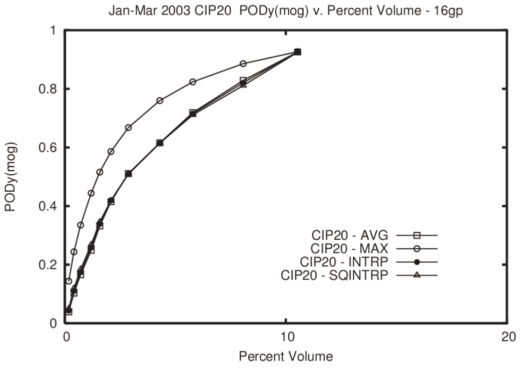

Fig 8. PODy v. %Volume curve for CIP evaluation at sixteen surrounding grid points for four differing methods of inferring the icing potential value

The results from the CIP evaluations show that the verification method used to evaluate the skill of CIP is very robust. Like in the GTG2 evaluation, the number of grid points used in inferring the forecast/diagnoses value does not change the ROC curve. The PODy and 1-PODn statistics are simply re-calibrated along the skill line. It also shows that the skill results are relatively insensitive to the method (maximum, average, two interpolation techniques) used to infer the forecast/diagnoses value. However, the results show that one must use care in assessing the volume efficiency of CIP because the results vary if the maximum value (at 4, 9, and 16 surrounding grid points) is used as opposed to the three other methods.

5. Conclusions

This study shows similar results for both icing and turbulence evaluations. In both cases (*Section 4.1 and 4.2*) the ROC curves are not affected by model resolution. In the case (*Section 4.2*) where the forecast inference method differs, the resulting ROC curves are almost identical. These results are very promising in that they show how robust this verification technique can be as the grid resolution increases. These results also lend confidence to results of past evaluations of CIP and GTG when the ROC curve is used as a measurement of skill.

Evaluations of the PODy vs. Percent Volume curves show results that are similar for CIP and GTG. For the GTG evaluation, as the resolution of the evaluation decreases the curves shift to the left (greater volume efficiency). This can be attributed to an increase in the PODy statistics as the resolution broadens. This makes sense because the maximum forecast value is used for the GTG evaluation which allows for contributions from a greater number of grid points as the grid resolution decreases while the percent of positive volume forecast stays the same.

For the CIP evaluation, as the resolution of the evaluation decreases, three resulting curves remain relatively close (average, 1/Distance interpolation, and 1/D2 interpolation) while the curve representing the maximum inference method shifts to the left (greater volume efficiency). These results make evident the fact that even though the method utilized for the measurement of skill (ROC curves) does not seem to be affected by changing grid resolution or forecast inference method, care must be taken when choosing the method for assessing the volume efficiency of either GTG or CIP.

6. 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.

7. 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*

Benjamin, S.G., G. A. Grell, S. S. Weygandt, T. L. Smith, T. G. Smirnova, B. E. Schwartz, D. Kim, D. Devenyi, K. J. Brundage, 2001: The 20-km version of the RUC. *Preprints, 18th Conference on Weather Analysis and Forecasting and the 14th Conference on Numerical Weather Prediction,* Fort Lauderdale, Fl., 30 July -02 August, 2001, American Meteorological Society (Boston).

Bernstein, B.C., F. McDonough, M.K. Politovich, and B.G. Brown, 2000: A research aircraft verification of the Integrated Icing Diagnostic Algorithm (IIDA). *Preprints, 9th Conf. on Aviation, Range and Aerospace Meteorology,* 11-15 Sept., Orlando, FL, 280-285.

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 Conf on Aviation, Range and Aerospace Meteorology,* Dallas, TX, 10-15 January, American Meteorological Society (Boston), 48-52.

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 Conf. on Aviation, Range and Aerospace Meteorology, 10-15 Jan., Dallas TX. 467-471.*

Sharman, R., J. Wolff, G. Wiener, and C. Tebaldi, 2004: Technical Description Document for the Graphical Turbulence Guidance Product 2 (GTG2). Report, submitted to the Federal Aviation Administration Aviation Weather Research Program (FAA/AWRP); available from R. Sharman (sharman@ucar.edu).