1.4 STATISTICAL MODELS OF AIRCRAFT ICING

Matthew Pocernich, Cory Wolff and Tressa Fowler *

National Center for Atmospheric Research
Boulder, Colorado

1. INTRODUCTION

A goal of the National Center for Atmospheric Research (NCAR) Research Application Program (RAP) icing project is to improve the forecasts of icing conditions that threaten aircraft. The Current Icing Potential (CIP) is an expert system model that uses observations and numerical weather prediction (NWP) output to infer information about cloud physics and behavior. Currently, the presence and intensity of icing can only be verified using aircraft observations. Historically, this has been done using pilot reports (PIREPs). A less subjective source of icing conditions is data collected by a specially equipped research aircraft. Instrumentation on the aircraft collects information on the occurrence of icing, temperature and liquid water content. This article discusses an effort to statistically post-process CIP diagnoses to improve model performance. Statistical models were created using both PIREPs and data from the research aircraft. Models were compared using relative operating characteristic (ROCs) curves and by calculating the areas beneath these curves. Information about variable importance and model weights are discussed.

This article is structured as follows. First, sources of data are explained. A description of the CIP output, PIREPs and data collected by the icing research aircraft are presented in Section 2. The suite of statistical models were used to post-process the CIP data are described in Section 3. These models include logistic regression, neural networks and random forests. Predictions made by these models as well as the direct output from the CIP are combined using an algorithm called stacked generalization. As detailed in Section 4, this technique calculates weights for each model by means of a cross validation algorithm.

Section 5 presents results that compare the performance of the CIP diagnoses with that of the combination of statistical models and the CIP diagnoses. PIREP and CIP data used for these evaluations were collected during the winter of 2003. These two methods are compared using the ROCs and the empirically calculated areas beneath these curves. Information about the weight given to each model and some information about variable selection is also gathered. An alternative model was constructed using data collected by the research aircraft during the winter of 2002. A statistical model made with this data is compared with the icing potential produced by the CIP. Included in these results is a presentation of variables collected by the experimental icing plane and data available from the CIP algorithm. Conclusions and discussions follow in Section 6.

*Corresponding author address: Matthew Pocernich, National Center for Atmospheric Research, Research Applications Program, Boulder, CO 80307-3000; email: pocernic@ucar.edu

2. DATA

Three sources of data are used in this report. They are PIREP data that are reported by selected aircraft, CIP diagnoses and the accompanying NWP information and icing data from the 2002 flights of the NASA Glenn Research Center Twin Otter research aircraft. These data sources are discussed in the following sections.

2.1 Current Icing Potential (CIP)

The CIP is a physically based expert system used to create an hourly, three-dimensional diagnosis of icing and super cooled large drop (SLD) potential. It combines surface, satellite, radar observations, PIREPs, and numerical weather model data using decision tree and fuzzy logic methods to assign a potential for icing and/or SLD conditions on a 0 to 1 scale. The CIP uses the Rapid Update Cycle 2 (RUC-II) for its environmental information. A complete description of the CIP algorithm can be found in Bernstein et al. (2004). Because the icing and SLD potentials are typically verified using PIREPs, it is very difficult to calibrate them, so they cannot be thought of as a probability for icing or SLD conditions to exist at a certain location.

2.2 Pilot Reports

Because PIREPs are the most widely available observations of icing conditions, they are commonly used to verify icing forecasts. However, PIREPs have known deficiencies as verification data (Kane et al., 1998). In particular, they are not systematic and they are biased in both time and space. Most PIREPs are made during daylight hours when most aircraft are flying. Additionally, certain days of the week that have greater air traffic also tend to have greater numbers of PIREPs. Finally, more

PIREPs are made during the winter season that the other, warmer seasons.

The number of PIREPs received in an area is a function of both the presence of icing and the air traffic. Those regions with the greatest icing frequency in conjunction with heavy air traffic naturally have the most PIREPs. These regions include the Great Lakes, Ohio-Mississippi Valley, Pacific Northwest and Great Plains. Many locations in the western US have few or no PIREPs. Desolate parts of Utah and Nevada are good examples.

In this report, the PIREPS have been coded to indicate the presence or absence of icing. This is due to the lack of consistency in the PIREP severity field. PIREP severities greater than trace are considered positive. Positive PIREPS are only considered if they are from twin engine commuter aircraft that carry about 20 to 70 passengers. Such aircraft tend to fly in icing conditions relatively frequently and have good visual indicators of icing.

2.3 NASA Twin Otter

Engineers and scientists at the NASA-Glenn Research Center have been studying the formation of icing and its properties in an effort to improve aircraft safety (see Miller et al. (1998)). As part of this research, an instrumented aircraft is flown into known icing conditions to take measurements of the environments associated with icing and their effects on aircraft performance. The aircraft is a DHC-6 Twin Otter, which is a twin engine, propeller driven plane that has been modified for use as a research platform. It has instruments to detect icing and measure cloud properties such as temperature and liquid water content.

Data from these instruments were used in this study to help verify and improve the CIP diagnosis. The data were averaged over five-minute periods to be applied to a grid point in the algorithm. To ensure that the data were consistent, they were only used if the plane's altitude changed by less than 1000 ft and if its speed remained above 90 kt. Also, the five-minute average temperature had to stay below 0^o C so that icing was possible. The main instrument used for this study was the Rosemount icing detector. This instrument is a metal shaft that protrudes from the aircraft and oscillates at a certain frequency. When ice accretes on the instrument, the frequency drops until it reaches a threshold. Then voltage must be applied to shed the ice and return it to the base frequency. For the purposes of this study, if the cycling process occurs more than once in five minutes then icing will be inferred for that time period and location.

3. STATISTICAL MODELS

Three statistical models are used to model the presence and absence of icing: logistic regression, neural networks and random forests. These approaches were selected because of their past use in modeling icing and similar phenomena and are described below.

3.1 Logistic Regression (or GLM)

Logistic regression is a specific member of the generalized linear model (GLM) family and is suitable for binary responses (Chambers and Hastie, 1992). General linear models are described by their link function and their variance function. The link function describes the relation between the mean and the linear predictors. In the case of linear regression the link function is $\log\{\frac{\mu}{(1-\mu)}\}$ where μ is the mean. A principle benefit of GLM models is that the predictand is not transformed to address assumptions about the error. Additionally, μ is constrained to the interval [0, 1].

3.2 Neural Network

Feed-forward neural networks are a way to generalize linear regressions. Neural networks consist of a hidden layer of nodes that provide an intermediate level for the input information. Predictions are made from this input layer using an "activation threshold" function (Venables and Ripley, 1994). In this project, there are twelve input nodes, three hidden nodes and two output nodes. The number of hidden nodes was selected based on past experience.

3.3 Random Forest

The random forest is an extension of the bagging algorithm proposed by Breiman (2001). Bagging is a classification and regression tree algorithm that independently grows trees on a bootstrapped sample of the original data. In bagging, at each node a split is made using the best of all variables. In random forests, a split is made on a randomly selected subset of all regressors. This makes the process significantly faster and allows more trees to be grown than in bagging. In the end, a prediction is made based on a vote amongst all the trees.

In random forests, several measures of variable importance are available. This paper considers the following method. After a tree is constructed, the hold-out data are used to estimate the effect of each variable. For each variable, the values are randomly permuted and the increase in the mean squared error of each prediction is recorded. These data are aggregated amongst all trees, providing some information about what variables are most influential. It is important to note that if two

variables are highly dependent, their importance may be underrepresented. Random forests are explained in Liaw and Wiener (2002). For this paper, 100 trees were constructed at each iteration and at each node four variables were randomly selected.

4. COMBINATION OF STATISTICAL MODELS

In order to determine the optimal weights to assign to each statistical and CIP diagnosis a stacking procedure was followed, as described by Smyth and Wolpert (1999). The procedure consists of the following steps. The data are randomly partitioned into groups. In this study, five groups were used. In turn, each group is withheld and a model is constructed with the remaining data. This model is used on the hold-out data. Weights that minimize the sum of the squared errors for all models are estimated using a non-linear optimization routine. The error is the difference between the forecast and the observation of the model. The weight assigned to each model is constrained to be kept non-negative. In this study, the weights were not constrained to add up to one. This allows the combination algorithm to make adjustments for biases. To provide information about the usefulness of each statistical model, the values of these weights are noted for each time step.

5. RESULTS

5.1 Models based on PIREPs

For nine weeks from December 2002 until March 2003 statistical models were constructed using two weeks of CIP output and NWP data to predict the presence or absence of icing as reported on PIREPs. The CIP and statistical models are verified with data from the subsequent week. The results of these models were compared with PIREPs collected during the following week. This process was repeated for each week in the nine week period.

Based on the larger areas beneath each ROC curve, the combination of statistical models provides a slight improvement in skill over the directed CIP diagnoses (Table 1). On a weekly basis, this improvement ranges from 0.3 to 4.8 percent. Compositing the results the entire trial period, a 3.1 percent improvement is noted (Table 1 and Figure 2).

The random forest consistently received the greatest proportion of weight (see Table 2). These weights are consistently greater than that placed on any of the remaining models. Neural networks did not contribute significantly to any weekly prediction.

Variable importance information was gathered for each random forest model during the creation step. Since

Table 1: *Empirically calculated areas under ROC curves by week.*

Week	CIP	Stats Mod	Diff.	% Diff.
1	0.77	0.81	0.03	4.1
2	0.82	0.84	0.02	2.5
3	0.80	0.80	0.00	0.3
4	0.75	0.78	0.03	3.4
5	0.81	0.83	0.03	3.2
6	0.79	0.81	0.01	1.5
7	0.80	0.83	0.03	4.2
8	0.77	0.81	0.04	4.8
9	0.84	0.88	0.04	4.1
Avg.	0.80	0.82	0.03	3.1

Table 2: *Weights placed on statistical and CIP output - by week*

Week	CIP	glm	tree	nnet
1	0.09	0.24	0.56	0.07
2	0.07	0.15	0.75	0.03
3	0.10	0.00	0.89	0.01
4	0.16	0.04	0.81	0.01
5	0.00	0.10	0.88	0.02
6	0.14	0.10	0.69	0.04
7	0.00	0.13	0.85	0.02
8	0.08	0.00	0.94	0.01
9	0.00	0.11	0.92	0.00
Average	0.07	0.10	0.81	0.02

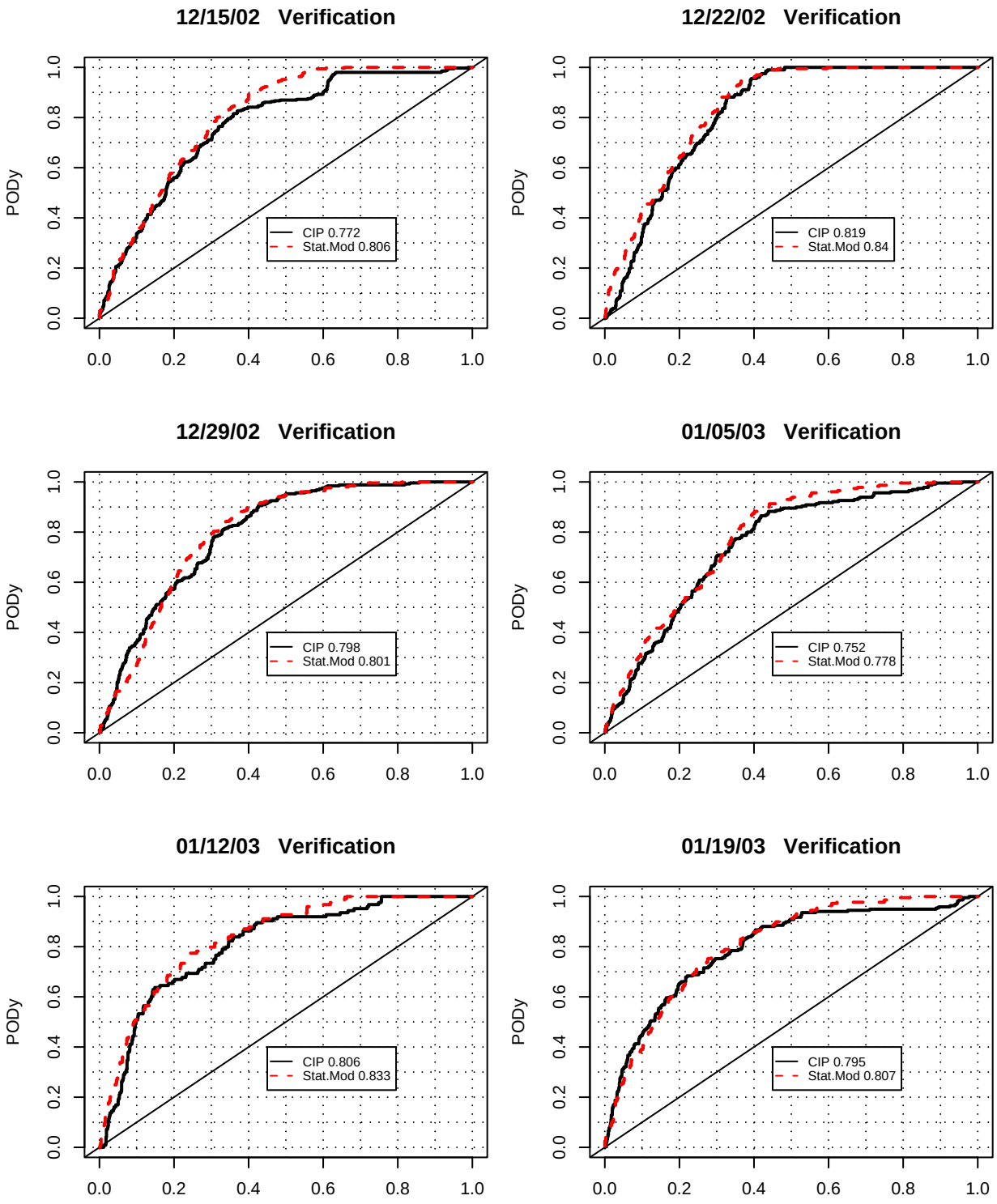

Figure 1: *Examples of ROC curves generated for the week forecast starting on the date listed. The statistical models were generated using the previous two week's data. Values in text box are the empirically calculated areas under the curves.*

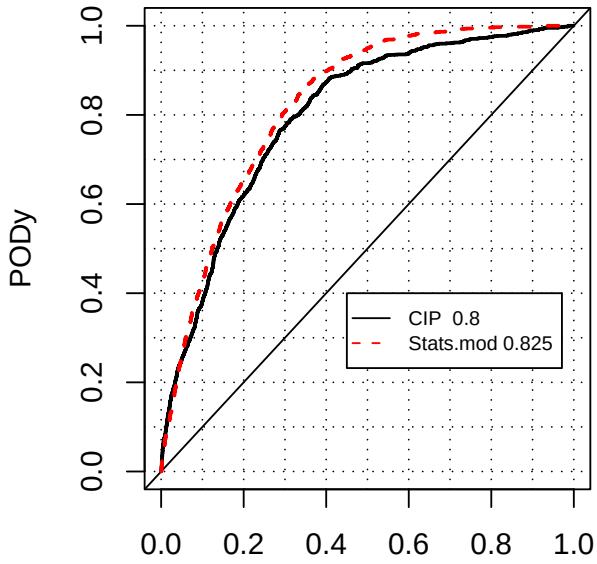

Figure 2: *ROC plot with verification results from all verification periods grouped together. 14,869 records.*

the random forest algorithm begins by bootstrapping the data, a portion of the data is not used. Using this holdout data, in turn, the value for each variable is randomly permuted and the increase in the mean squared error on the holdout data is noted. The more important the variable in the model prediction, the greater the increase in error when it is randomized. In Table 3, one notes that the CIP diagnosis is frequently the most important variable. Temperature and relative humidity are the most important NWP variables. During weeks 2 through 4 temperature is the most important variable.

5.2 NASA Twin Otter Data

The routes of the NASA Twin Otter in the winter of 2002 originated in Cleveland with the the intent of measuring cloudy areas with high potential for icing (Figure 3). Figure 4 illustrates the relation between variables. This figure is presented as a tool used for exploratory data analysis and intended to show the relationship between variables. The histograms of the respective variables are found along the diagonal. The first variable, labeled “Cycles”, is the number of heating cycles of the icing probe per five minutes. The greater the number of cycles, the higher the accumulation rate of ice on the aircraft. The flights are coded by color. The triangles and circles indicate the presence of icing and no icing using the criteria described in Section 2.3. Of particular interest here is the strong relationship between the number of icing cycles and the average liquid water content (ALWC). This is as expected since by definition the greater the liquid water content, the more the icing. Since ALWC is measured on the plane, hopefully, we would find a variable or set of variables produced or used by the CIP, NWP or observation that would show a similarly strong relationship with the icing accumulation rate. Unfortunately, such a variable is not immediately obvious.

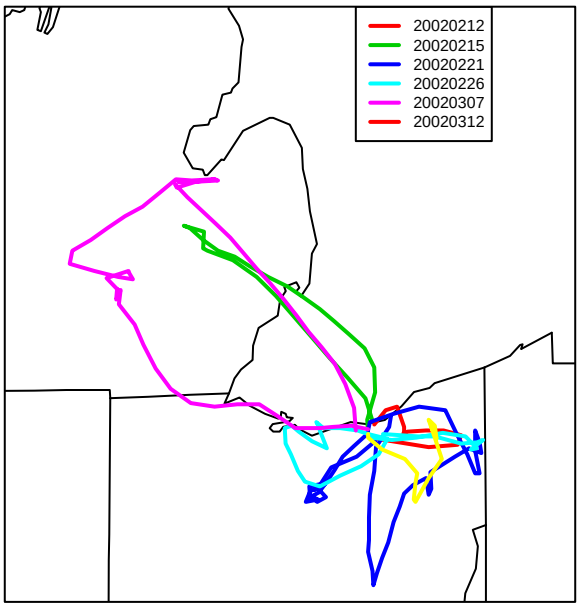

Figure 3: *Routes of NASA Twin Otter plane flown from Cleveland in Winter 2002.*

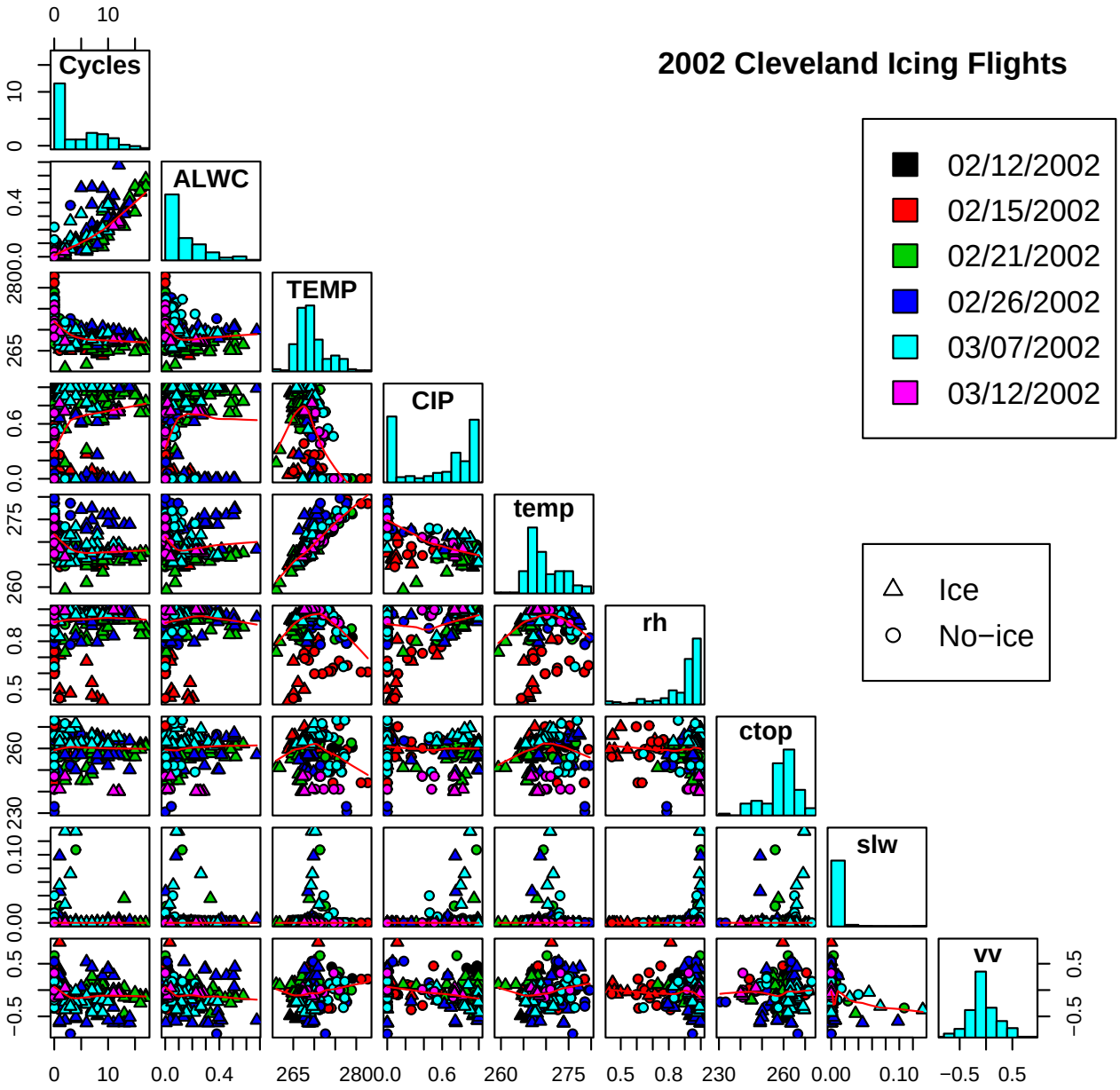

Figure 4: *Q-Q plot of data collected by the NASA Twin Otter and information provided by CIP and NWP model. The red line is a loess curve and is presented as a non-parametric indicator of trend. Data are from winter 2002. The names of variables are found in the Appendix.*

Table 3: *Increase in mean squared error when variable values were randomly permuted, for hold-out Random Forest data. Variable definitions are found in the Appendix.*

Week	CIP	temp	rh	ctopt	anyp	ccnt	slw	vv	sld
1	8.6	5.5	5.5	1.8	0.5	0.4	0.5	1.1	4.0
2	8.2	11.6	6.7	1.5	1.2	0.8	1.0	1.1	5.3
3	8.1	12.9	9.0	2.2	2.1	1.3	1.2	1.9	5.6
4	9.3	10.7	10.1	5.3	2.6	1.0	1.0	3.1	4.6
5	7.6	6.4	7.2	4.9	2.2	0.9	0.7	2.9	2.5
6	7.0	3.4	5.6	2.9	2.9	0.8	0.2	2.1	2.1
7	5.9	3.8	3.8	2.6	1.0	0.3	0.5	2.2	2.4
8	7.1	6.6	5.8	4.4	1.2	0.6	0.8	3.0	2.4
9	9.9	6.4	7.3	4.8	1.1	0.9	1.0	3.0	4.1
10	8.8	5.7	6.1	2.5	0.7	0.6	0.7	2.3	4.0

5.3 Models based on and compared with NASA Twin Otter data

To avoid the deficiencies of the PIREP information, a statistical model was built using the research aircraft measurements as the predictand. A comparable model was not made using PIREPs because relatively few PIREPs were available for this area at these times. To verify this statistical model and compare the model with CIP output a cross validation procedure was used. Each flight was treated as holdout data and a statistical model was formulated with the data from the other five flights. Since these flights occurred over a month time period, there is not the same chronological order present in the PIREP based models. Essentially, the ROC curves from the statistical model and the CIP are quite similar (Figure 5). The ROC area for the statistical model is slightly greater than that for the CIP forecast. ROC values are not calculated for individual flights because with so few records the empirically calculated areas are suspect.

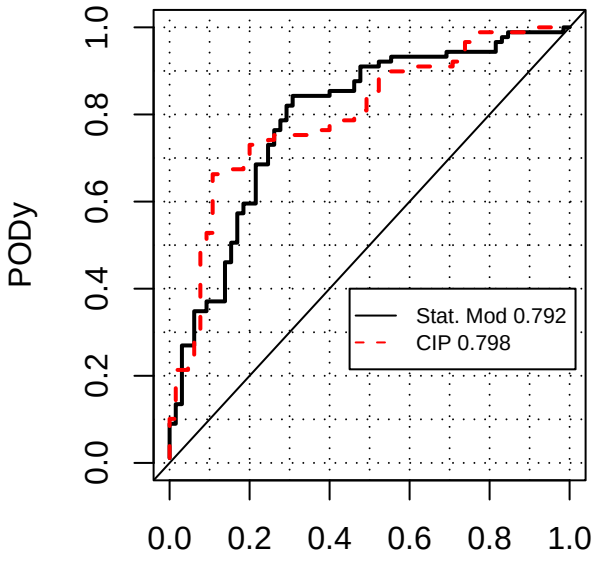

Figure 5: *ROC curve for models created using NASA Twin Otter data instead of PIREPs. There were 154 independent records.*

6. DISCUSSION AND CONCLUSIONS

The use of a statistical model to post process data produced a slight improvement over the direct CIP output. This improvement may be practically significant. As the CIP is not a calibrated diagnosis, the statistical post-processing may be useful as a way to calibrate such a product. The challenge here is in finding and appropriate verification procedure. An important next step is to quantify the volume of space with positive icing indicated for each event. This procedure will produce ROC–like figures with %Volume plotted on the *x*-axis. From these plots, one can determine whether the performance presented here is achieved by forecasting an excessively large area. Plots of the output need to be examined for

physical consistency and should be compared with CIP itself.

Exploration of the Twin Otter data verified the well-known relationship between liquid water content and icing. Unfortunately, the environmental variables provided by the NWP or observations alone do not provide a reliable prediction of this field due to microphysics limitations. This is evident by the poor correlation between the ALWC and the NWP super-cooled liquid water field. The poor relation between these field may in part be due to spatial variability of the clouds that is not detected by the NWP. Improvements in forecasting of this field may be offered by the MM5 and newer versions of the RUC.

To help overcome the deficiencies in PIREP data, more research aircraft data are becoming available. This includes NASA data from the winter of 2003. Additionally, several planes are scheduled to gather data in the winter of 2004 during AIRS II Alliance Icing Research Study.

APPENDIX

List of predictors used in the prediction of the presence of icing. Superscripts indicate the following: 1 – variables measured by the aircraft, 2 – CIP data, 3 – NWP model (RUC II) data.

Name	Description
Cycles[1]	Rosemount heating cycles per 5 minutes
AWLC[1]	5 minute average cloud liquid water content (g m^{-3})
TEMP[1]	Static air temperature (^{o}K)
CIP[2]	Icing potential
rh[3]	Relative humidity (%)
temp[3]	Temperature (^{o}K)
ctop[3]	cloud top temperature (^{o}K)
anyp[3]	any precipitation $\{0, 1\}$.
vv[3]	vertical velocity (μ bar sec $^{-1}$
slw[3]	combination of cloud and rain mixing ratio (g kg^{-3} * 1000)
ccnt[3]	cloud count

ACKNOWLEDGEMENTS

NCAR is sponsored by the National Science Foundation. This research is partly in response to requirements and funding by the Federal Aviation Administration (FAA). Additional funding was provided from the NCAR Weather and Climate Assessment Initiative. The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA. We would like to thank the NASA Glenn Research Center for providing the Twin Otter research aircraft data. Additional help was provided by Ben Bernstein, Frank McDonough, Linda Mearns and Marcia Politovich.

REFERENCES

Bernstein, B., F. McDonough, M. Politovich, B. Brown, T. Ratvasky, D. Miller, and C. Wolff, 2004: Current icing potential (CIP) Part I: Algorithm description and comparison with aircraft observations. *Submitted to Journal of Applied Meteorology*.

Breiman, L., 2001: Random forests. *Machine Learning*, **45**, 5–32.

Chambers, J. M. and T. J. Hastie, eds., 1992: *Statistical Models in S*. Wasworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

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

Liaw, A. and M. Wiener, 2002: Classification and regression by randomforest. *R News*, **2**, 18–22.
URL `http://CRAN.R-project.org/doc/Rnews/`

Miller, D., T. Ratvasky, B. Bernstein, F. McDonough, and J. Strapp, 1998: NASA/FAA/NCAR supercooled large droplet icing flight research: Summary of winter 96-97 flight operations. *36th Aerospace Science Meeting and Exhibit, Reno, NV, 12-15 January American Institute of Aeronautics and Astronautics.*

Smyth, P. and D. Wolpert, 1999: Linearly combining density estimators via stacking. *Machine Learning*, **36**, 59–83.

Venables, W. and B. Ripley, 1994: *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York.