

7.B3 VERIFICATION OF AVIATION ICING ALGORITHMS FROM THE SECOND ALLIANCE ICING RESEARCH STUDY

Michael Chapman*, Anne Holmes, and Cory Wolff
National Center for Atmospheric Research
Boulder, Colorado USA

1. INTRODUCTION

The Second Alliance Icing Research Study (AIRSII) was conducted over southeastern Canada from November 2003 to March 2004. One of the main objectives of the project was to develop and evaluate systems to diagnose and forecast in-flight icing conditions over short time periods. Several operational in-flight icing products were available during the field project for evaluation. They include Current Icing Potential (CIP; Bernstein et al. 2005), Forecast Icing Potential (FIP), System of Icing Geographic identification in Meteorology for Aviation (SIGMA; Lebot 2004), the GOES-derived Cloud Products (GDCP; Minnis et al. 1995, 2001), the Global Environmental Multi-scale (GEM) model and the Penn State University/National Center for Atmospheric Research (PSU/NCAR) Mesoscale Model (MM5). Also participating in the field project were several research aircraft including the NASA Glenn Twin Otter, the University of North Dakota Citation, and the National Research Council Convair-580. Each of these aircraft was equipped with standard cloud microphysical probes including a CSIRO hot wire liquid water probe (King probe) and a Rosemount icing detection probe.

For this particular study a statistical verification of several of the in-flight icing nowcasting algorithms (CIP, SIGMA, and GDCP) was performed using King liquid water content (King-LWC) and Rosemount probe heating cycles (ROSE) from the NASA Glenn Twin Otter and the NRC Convair-580 research aircraft as verifying observations. A verification study based on pilot reports (PIREPs) of icing conditions was also performed over the continental U.S. (CONUS) for the time period bracketing the AIRS-II field campaign (01 October 2003 - 31 March 2004). CIP and SIGMA 1500 and 2100 UTC runs and 1445 and 2045 UTC GDCP products were verified against the research aircraft and PIREP datasets.

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

2. DATA

2.1 CIP icing potential field

CIP is a physically-based, situational 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).

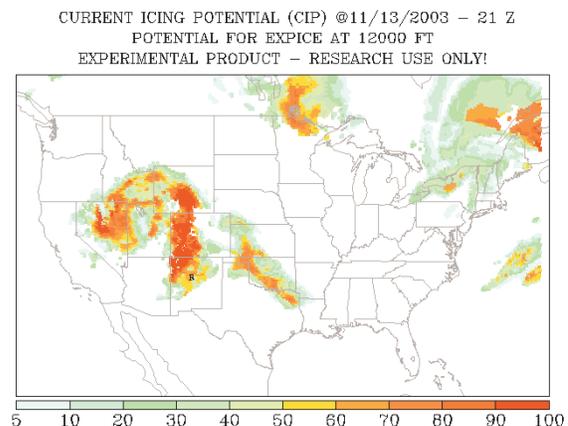


Fig. 1 Example of CIP Icing Potential for 13 Nov 2003 at 21Z.

2.2 SIGMA index field

SIGMA is an operational, diagnostic, in-flight icing algorithm developed at Meteo-France. It is normally run over Europe and utilizes a combination of several different data sources. These data include relative humidity (RH), temperature (T), and vertical velocity (VV) from the French numerical weather prediction model ARPEGE, infrared satellite imagery from METEOSAT, and radar imagery from ARAMIS. For the evaluations in this paper, SIGMA was modified to run over the CONUS and the AIRSII field project area. In particular, for this evaluation SIGMA used forecasts of RH, T, and VV from the 20-km RUC as well as the cloud mask from CIP to generate its icing index (Fig. 2). This index, much like CIP, is defined as a potential of icing with values

ranging from 0.0 (No Icing likely) to 10.0 (Icing likely).

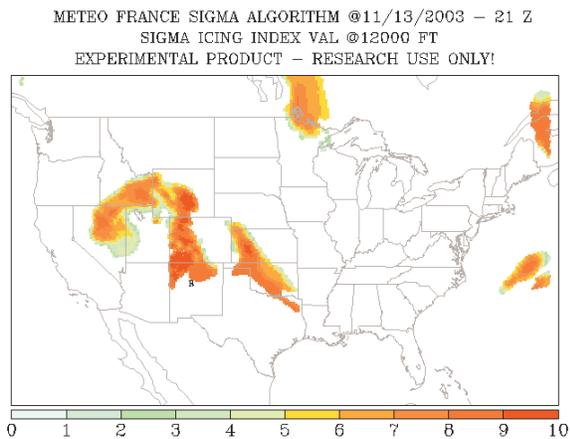


Fig. 2 Example of SIGMA from the same time period as Fig 1.

2.3 NASA Langley Research Center (LaRC) GDCPs

The NASA LaRC GDCPs are generated by combining the GOES-10 and GOES-12 satellite data and using the Visible Infrared Solar-Infrared Split-window Technique (VISST) during daylight hours (Minnis et al. 1995, 1998). A complex cloud identification method (Trepte et al. 1999) is initially used to identify whether a specific pixel is to be classified as cloudy or clear. When a pixel is classified as cloudy, the VISST is used to ascertain characteristics such as cloud phase, liquid or ice water path, effective temperature, effective height, optical depth, and particle size. For this study the cloud phase (PHASE) will be evaluated (Fig. 3).

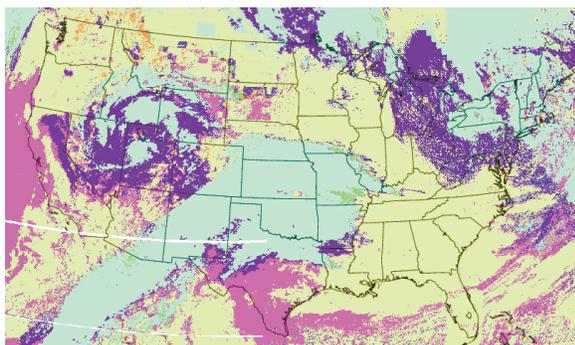


Fig. 3 An example of the NASA LaRC GDCP cloud phase product from the same date as in Figs 1 and 2 but valid for 2045Z.

In order to accurately evaluate PHASE, the data were mapped from a 5km grid over the RUC domain to a 20 km grid. This re-mapping resulted in

16 satellite pixels being mapped to each 20-km grid box and allowed for a direct comparison with the 20-km CIP and SIGMA output. The 20-km PHASE product was generated simply by counting the number of pixels that were characterized as representing SLW. If 3 pixels out of a possible 16 were listed as SLW, the 20-km point would have a “3” recorded for the SLW category.

2.4 Research Aircraft data

The data from the NASA Glenn Twin Otter and the NRC Convair-580 were available for the evaluations. In order to directly compare the research aircraft data to the 20-km gridded algorithms, the data were smoothed over 20-km segments by calculating the amount of time it took for the aircraft to traverse 20km using the average airspeed. Upon completion of a 20km segment, the individual latitudes and longitudes were averaged. Fig. 4 is a map of the research aircraft locations. Also recorded for each segment were median temperature; average, minimum, and maximum altitude; average airspeed; average, median, minimum, and maximum King-LWC; and maximum ROSE. A total of 680 20-km segments were generated by the research aircraft.

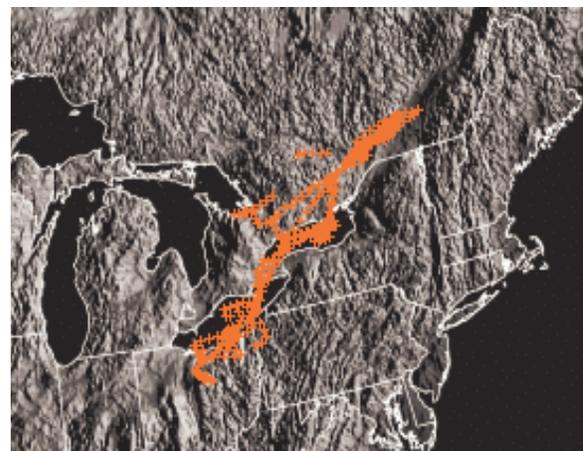


Fig. 4 Locations of 20km research aircraft segments for the AIRSII field project.

2.5 Pilot reports (PIREPs)

PIREPs, which signify an observation of icing or lack thereof, are vital because they are the primary “ground truth” observations available to verify the presence or absence of icing 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. Table 1 lists the

counts of the numbers of PIREPs available over the evaluation period for each icing intensity. Intensities equal to 5 or more are considered moderate or greater (MOG) PIREPs.

Table 1. Numbers of PIREPs available over the CONUS from 01 October 2003 – 31 March 2004 for the 15Z and 21Z time periods.

<i>Intensity</i>	<i>Description</i>	<i>Counts</i>
8	Severe	33
7	Heavy	5
6	Moderate-Heavy	187
5	Moderate	1596
4	Light-Moderate	3952
3	Light	4538
2	Trace-Light	3112
1	Trace	628
0	No Icing	8574

3. VERIFICATION METHODS

3.1 PIREP evaluations

Evaluations were accomplished by comparing the CIP algorithm icing potential, SIGMA icing index, and PHASE product to PIREPs of positive and negative icing. Because the PHASE product is two dimensional [i.e., the individual pixels are only valid at or near cloud top height (CTH), a field included in the GDCP suite], the simultaneous evaluation of the three products required that the algorithms only be evaluated with observations that were located within 305m (1,000ft) or 915m (3,000ft) below the median CTH. The purpose of the “buffer” below the median CTH was to enable the PHASE product to be evaluated in three dimensions and over the same volume as CIP and SIGMA. The average CTH was calculated either by averaging the CTH values of pixels that were considered to represent clouds or by averaging the CTH of pixels that were only considered to have supercooled liquid water (SLW), depending on which cloud phase was being evaluated. In order to better compare the PHASE product to the CIP and SIGMA icing products, the SLW product is used throughout these evaluations. The algorithm values at the four grid boxes surrounding the PIREP were then examined. Since CIP incorporates information from PIREPs in the hour prior to the analysis time, only PIREPs from the hour following the valid time were included in the analysis. Statistics were then computed and analyzed.

A separate evaluation of CIP and SIGMA was also performed using all of the available PIREPs and all available levels.

3.2 Research aircraft evaluations

Evaluations were accomplished by using an approach much like that used for the PIREP evaluations. The two types of research aircraft observations used were the 20-km average King LWC at $T < 0^{\circ}\text{C}$ and the number of Rosemount heating cycles over 20 km. They were used separately in evaluation of the algorithms, except in cases where the King LWC was in the 0.001-0.025 g/m^3 range. When this occurred, the Rosemount probe data was used to determine the presence of icing.

3.3 Verification Technique

The methods utilized in the evaluation of the icing algorithms are based on standard techniques of forecast verification, as described by Brown et al. (1997). The icing forecast verification methodology treats icing forecasts and observations as Yes/No values. Brown et al. (1999) outlines how this method can be extended to verify continuous, rather than binary fields. Icing diagnoses produced by CIP/SIGMA can be converted into a set of Yes/No values by applying a variety of thresholds. For example, applying a threshold of 0.20/2.0 to CIP/SIGMA diagnoses would lead to a Yes value for all grid boxes with an icing potential greater than or equal to 0.20/2.0 while each grid box with a value less than 0.20/2.0 would be assigned a No value. The CIP and SIGMA values are chosen as the maximum of those available from the four 20-km grid boxes surrounding the aircraft location.

For the PHASE product, the four surrounding 20-km boxes contain 64 satellite pixels. These data can be thresholded at 8 pixel intervals. For example, one threshold would test whether more than 8 pixels indicating SLW were present. If this condition was met, it would be counted as a "Yes" icing diagnosis.

The verification methods are based on a two-by-two contingency table (Table 2). Each cell in this table contains a count of the number of times a particular forecast/observation pair was observed at a specific threshold.

Table 2. Contingency table for YES/NO forecasts. Elements in cells are 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+ NN+NY

Table 3 presents a list of the thresholds used in the evaluation of the three algorithms. The PHASE thresholds are in numbers of SLW pixels contained in the four surrounding 20-km grid spaces surrounding an observation.

Table 3. Thresholds used for verification of CIP, SIGMA, and PHASE.

Thresholds		
<i>CIP</i>	<i>SIGMA</i>	<i>PHASE</i>
>0	>0	>0
0.05	0.5	8
0.15	1.5	16
0.25	2.5	24
0.35	3.5	32
0.45	4.5	40
0.55	5.5	48
0.65	6.5	56
0.75	7.5	64
0.85	8.5	-
0.95	9.5	-

POD_y and POD_n 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, POD_y and POD_n measure the ability of the forecasts to discriminate between Yes and No icing observations. Other common verification scores (e.g., false alarm ratio, critical success index) cannot be computed due to the nature of the verification data (Brown and Young 2000). Table 4 gives the definition and description of these statistics.

Table 4. Verification Statistics used for the evaluation of CIP, SIGMA, and PHASE.

<i>Statistic</i>	<i>Definition</i>	<i>Description</i>
<i>POD_y</i>	YY/(YY+NY)	Probability of detection of YES observations
<i>POD_n</i>	NN/(NN+YN)	Probability of detection of NO observations
<i>TSS</i>	POD _y +POD _n -1	True Skill Statistic
<i>Area under ROC curve (AUC)</i>	Area under the curve relating POD _y and 1-POD _n	Area under curve relating POD _y and 1-POD _n (ROC curve)

The relationship between POD_y and 1-POD_n 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-POD_n, POD_y) points for different thresholds. The

resulting curve is known as the “Relative Operating Characteristic” (ROC) curve in SDT. When POD_y 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 (AUC) is a measure of overall forecast skill and provides another measure that can be compared among products. This measure is not dependent on the threshold used. A forecast with zero skill would have an ROC area of 0.5.

4. RESULTS

Because GDCP is essentially two-dimensional, it is not considered in all evaluations. In the following subsections, full volume comparisons of CIP and SIGMA are considered first for each type of evaluation, followed by limited-volume comparisons of all three algorithms. PIREP-based results are presented first, followed by comparisons with research aircraft data.

4.1 PIREP evaluation results (CIP & SIGMA)

The time period for this evaluation was 01 October 2003 to 31 March 2004. The 1500 UTC and 2100 UTC valid times were considered. Fig. 5 shows ROC curves for this evaluation, which indicate the capabilities of both algorithms at capturing MOG and negative icing PIREPs. Both CIP and SIGMA have approximately the same scores for the higher thresholds (0.35-0.95 for CIP and 3.5-9.5 for SIGMA) which are evident by the near overlap of the bottom part of both curves. CIP continues to detect the YES PIREPs at lower thresholds (>0-0.25) while SIGMA only captures a few at its lower thresholds (>0 – 2.5). This is evident by the separation of the curves at the top of the plot. Because of this separation, CIP (AUC=0.75) showed slightly more skill than SIGMA (AUC=0.69) for this evaluation.

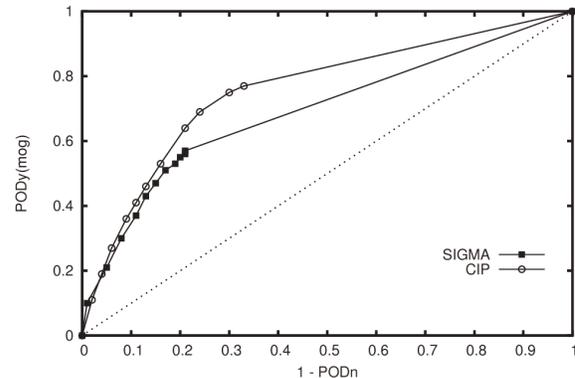


Fig. 5 ROC plot for CIP v. SIGMA PIREP evaluation for 01 Oct 2003 – 31 March 2004.

4.2 PIREP evaluation results (CIP, SIGMA, and PHASE)

The time period evaluated in this part of the study is the same as in section 4.1 (i.e., 1 October 2003 to 31 March 2004). PHASE was included in the comparison, along with SIGMA and CIP. In order to directly compare the three products, observations and diagnostic levels from CIP and SIGMA were limited to an area 305m (1,000ft) below the average GDCP's CTH for the pixels diagnosed as containing SLW in the PHASE product. The ROC curves in Fig 6 show similar results for CIP and SIGMA as those shown in Fig. 5, with a slight reduction in the PODn statistics. The skill of the PHASE product is positive (AUC = 0.56) but less than the skill of CIP (AUC = 0.71) and SIGMA (AUC = 0.66).

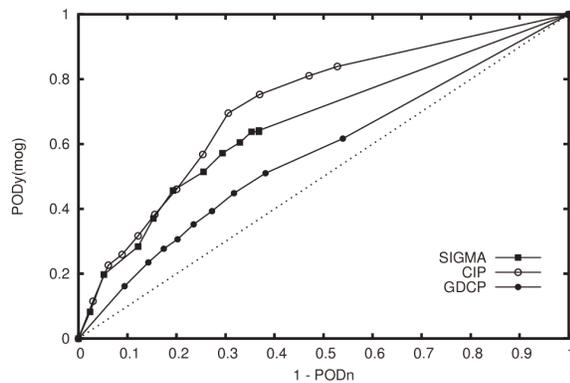


Fig. 6 ROC plot for CIP, SIGMA, and PHASE PIREP evaluation for 01 Oct 2003 – 31 March 2004.

4.3 Research aircraft evaluation (CIP & SIGMA)

For this part of the evaluation the two algorithms were evaluated over the AIRSII time period wherever Twin Otter or Convair-580 research aircraft data were available and could be expanded to 20-km segments. For Figure 7, icing conditions were defined by KingLWC > 0 g/m³ and Temp < 0 °C). The ROC curve for CIP in Fig. 7 is similar to the CIP curves in Figs. 5 and 6 (AUC = 0.71). An increase in both algorithms' ability to detect YES reports at lower thresholds is apparent, but the lower PODn values result in overall similar skill to that found in previous test. SIGMA's AUC fell slightly, though, to 0.61.

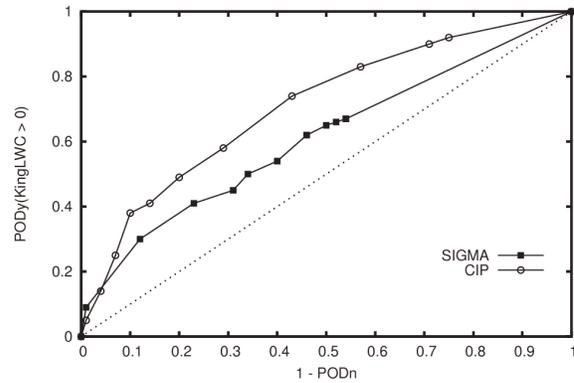


Fig. 7 ROC plot for CIP v. SIGMA AIRSII Research Aircraft evaluation using KingLWC as the observation for 01 Oct 2003 – 31 March 2004.

Figure 8 shows verification results when icing conditions are defined by ROSE>0. Using this test, CIP and SIGMA performance is virtually equivalent at higher thresholds, with the ability of SIGMA to detect YES reports at lower thresholds again, not as strong as CIP's. This equated to larger skill for CIP (AUC=0.66) than with SIGMA (AUC=0.61).

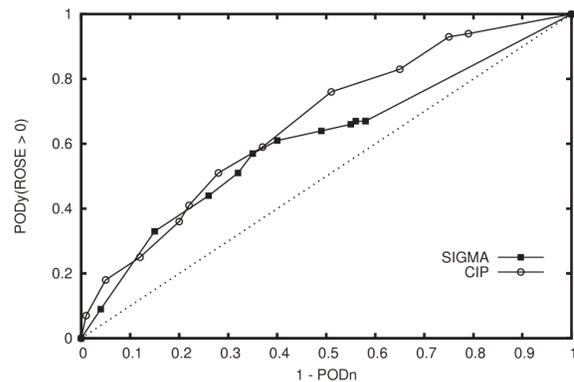


Fig. 8 ROC plot for CIP v. SIGMA AIRSII Research Aircraft evaluation using Rosemount heating cycles as the observations for 01 Oct 2003 – 31 March 2004

4.4 Research aircraft evaluation (CIP, SIGMA and PHASE)

This evaluation was accomplished much like the evaluation in section 4.2, with King LWC measurements from the research aircraft used as observations as opposed to PIREPs. Due to the lack of matching observation/diagnoses pairs, the ROC curves that were used in the previous results sections could not be reliably generated. Table 5 is a list of PODy and PODn statistics for each algorithm at a single threshold (CIP=0.35, SIGMA=3.5, and PHASE=24); these particular thresholds were selected to produce similar values of PODy. Thus, for each algorithm, the detection of King LWC > 0

g/m^3 was relatively close [POD_y = 0.76 (CIP), 0.74 (SIGMA), and 0.76 (PHASE)] while the ability to detect NO observations was different [POD_n = 0.67 (CIP), 0.59 (SIGMA), and 0.33 (PHASE)]. When comparing the POD_y and POD_n results together, CIP (TSS = 0.43) and SIGMA (TSS=0.33) seem to have more skill than PHASE (TSS=0.09).

Table 5. Statistics for CIP, SIGMA, and PHASE for AIRSII research aircraft evaluation.

Algorithm (Thresh)	POD _y	POD _n	TSS
CIP(0.35)	0.76	0.67	0.43
SIGMA(3.5)	0.74	0.59	0.33
PHASE (24)	0.76	0.33	0.09

5. CONCLUSIONS

5.1 PIREP evaluations

For each PIREP evaluation, CIP and SIGMA showed similar results. The two algorithms had approximately equivalent skill at discriminating between the YES and NO icing reports at higher thresholds (0.35-0.95 for CIP and 3.5-9.5 for SIGMA). For the lower thresholds, SIGMA was unable to match CIP's ability to detect YES reports. A possible reason for this discrepancy is that CIP's fuzzy logic scheme allows for diagnoses of icing at lower values of temperature and relative humidity than SIGMA's scheme does. The algorithmic difference may be partially attributable to the development of SIGMA on a different model (ARPEGE) that may have different moisture characteristics.

PHASE showed positive skill in discriminating between YES and NO PIREPs. Haggerty et al. (2005) showed that the GDCP over-predicts CTH. Bernstein et al (2005) noted that under some circumstances CIP has a similar problem. This problem would have also cascaded into the SIGMA dataset. CTH over-prediction would keep all three products from detecting negative icing conditions between the product estimated CTHs and the actual CTH. The result would be smaller POD_n statistics. In addition, it was assumed for this evaluation that PHASE was predicting positive icing wherever SLW was indicated. This assumption is not always the case and would also result in a smaller POD_n. The SLW-only PHASE test did show similar skill to a prior study (Politovich et al. 2004) where pilot reports that were located both above and below the CTH measured by the algorithm were included in the evaluation, regardless of phase ("all-cloudy" in Fig. 11). In contrast, the skill of PHASE was evaluated using only the SLW pixels for the evaluation done

here. As a result, the algorithm showed the same amount of skill as the all-cloudy (liquid, ice, and SLW) evaluation from Politovich et al. (2004).

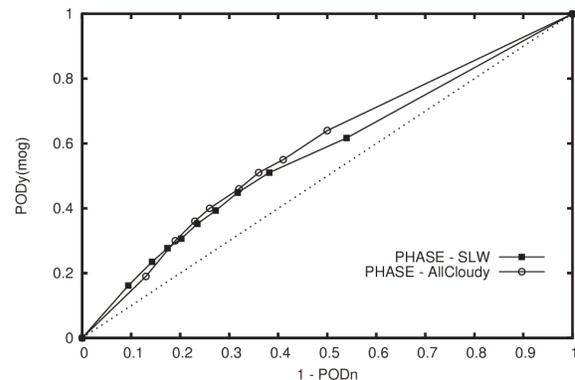


Fig. 11 ROC plot of PHASE [(All-cloudy) +/-3,000ft (915m) CTH] versus PHASE [(SLW) - 1,000ft (305m) CTH].

5.2 Research aircraft evaluations

The results of the CIP and SIGMA evaluations were different for the higher thresholds depending on which observation (KingLWC or ROSE) was used. While the statistics for SIGMA were consistent for the two types of observations, the statistics for CIP were somewhat dependent on the observation type. A possible explanation for this difference may be from ram-air temperature rise on the Rosemount probe, which can inhibit the accretion of ice on the the probe at higher temperatures (e.g. 0 to about -2C for the Twin Otter). This could result in simultaneous observations of positive KingLWC and ROSE=0 when very warm icing conditions are present.

Another difference between the PIREP and research aircraft results was a reduction in the POD_n statistics for both algorithms. This could be an artifact of the research aircraft data. Specifically, when the Twin Otter has a substantial amount of ice build-up, the aircraft generally will exit the icing conditions just above CTH in order to evaluate aircraft performance. Since the aircraft is just above the cloud top, the observation is recorded as a No while the 25-mb vertical resolution of the RUC and CIP's approach of allowing icing to be diagnosed at the first model level above its diagnosed CTH may result in a yes diagnosis there. The lower POD_n for the PHASE product (0.33) in section 4.3 might also be attributed to its overestimation of CTH.

6. FUTURE WORK

Expansion of the previous verification to include forecast (FIP, GEM, and MM5) as well as the liquid water path product from NASA LaRC is in the

process of being completed. Results will be summarized during time of presentation. A more comprehensive verification will also be completed and summarized in a journal paper in the future.

7. ACKNOWLEDGEMENTS

Thanks to Pat Heck and Kirk Ayers from NASA Langley for generating the NASA LaRC GDCPs for all of these evaluations on such short notice. Thanks to Ben Bernstein from NCAR and Christine LeBot from Meteo-France for providing SIGMA for these evaluations. Thanks to Patrick Boylan for his help in generating some of these statistics.

This project is supported by the NASA Applied Sciences Program and the NASA Aviation Safety and Security Program through the NASA Advanced Satellite Aviation-weather Products (ASAP) project. NCAR is sponsored by the National Science Foundation.

8. REFERENCES

- 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, B.G. Brown, T.P. Ratvasky, D.R. Miller, C.A. Wolff, and G. Cunning, 2005: Current Icing Potential (CIP): Algorithm Description and Comparison with Aircraft Observations. *J. Applied Met.*, In press.
- Brown, B.G., G. Thompson, R.T. Brientjes, 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.
- 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, 9th Conference on Aviation, Range, and Aerospace Meteorology*, Orlando, FL, 11-15 Sept., American Meteorological Society (Boston), 393-398.
- Haggerty, J., G. Cunning, B. Bernstein, M. Chapman, D. Johnson, M. Politovich, C. Wolff, P. Minnis, R. Palikonda, 2005: Integration of Advanced Satellite Cloud Products into an Icing Nowcasting System. *WWRP Symposium on Nowcasting and Very Short-range Forecasting*, Toulouse, France, 5-9 September.
- LeBot, Christine, 2004: SIGMA : System of Icing Geographic identification in Meteorology for Aviation. *11th Conference on Aviation, Range, and Aerospace*, Hyannis, Mass. 4-8 October 2004, American Meteorological Society (Boston).
- Minnis, P., D.P. Kratz, J.A. Coakley, Jr., M.D. King, D. Garber, S. Mayor, D.F. Young, and R. Arduini, 1995: Cloud Optical Property Retrieval (Subsystem 4.3), "Clouds and the Earth's Radiant Energy System (CERES) Algorithm Theoretical Basis Document, Volume III: Cloud Analyses and Radiance Inversions (Subsystem 4)", *NASA RP 1376 Vol. 3*, edited by CERES Science Team, pp. 135-176.
- Minnis P. D.P. Garber, D.F. Young, R.F. Arduini, and Y. Takano, 1998: Parameterization of reflectance and effective emittance for satellite remote sensing of cloud properties. *J. Atmos. Sci.*, 55, 3313-3339.
- Minnis P., W.L. Smith, D.F. Young, L. Nguyen, A.D. Rapp, P.W. Heck, S. Sun-Mack, Q. Trepte, and Y. Chen, 2001: A near real-time method for deriving cloud and radiation properties from satellites for weather and climate studies. *Proc. AMS 11th Conf. Satellite Meteorology and Oceanography*, Madison, WI. 15-18 Oct, 477-480.
- Politovich, M.K., P. Minnis, D. B. Johnson, C. A. Wolff, M. Chapman, P. W. Heck, and J. A. Haggerty, 2004: Benchmarking In-Flight Icing Detection Products for Future Upgrades. *11th Conference on Aviation, Range, and Aerospace*,

Hyannis, Mass. 4-8 October 2004, American Meteorological Society (Boston).

Trepte, Q., Y. Chen, S. Sun-Mack, P. Minnis, D.F. Young, B.A. Baum, and P.W. Heck. 1999: Scene identification for the CERES cloud analysis subsystem. Proc. *AMS 10th Conf. Atmos. Rad.*, Madison, WI, June 28-July2, 169-172.