Assessment of the HiRes Current Icing Product (CIP) and Forecast Icing Product (FIP)

As Provided to the Technical Review Panel

Prepared by

Quality Assessment Product Development Team
NOAA/ESRL/GSD/Forecast Impact and Quality Assessment Section

Authors:

Matthew S. Wandishin[2], Brian Etherton[1], Joan Hart[2], Geary Layne[2], and Melissa A. Petty[3]

15 November 2013

Affiliations:

1 – National Oceanic and Atmospheric Administration, Earth System Research Laboratory, Global Systems Division (NOAA/ESRL/GSD)

2- Cooperative Institute for Research in Environmental Sciences (CIRES) and NOAA/ESRL/GSD

3 – Cooperative Institute for Research in the Atmosphere (CIRA) and NOAA/ESRL/GSD

Corresponding Author:

M.A. Petty
NOAA/ESRL/GSD, 325 Broadway, Boulder, CO 80305
Melissa.A.Petty@noaa.gov

EXECUTIVE SUMMARY

The QA PDT was tasked to assess the quality of the High Resolution (HiRes) Current Icing Potential (CIP) and Forecast Icing Potential (FIP) algorithms developed by the National Center for Atmospheric Research. These HiRes products are to replace the current WRF Rapid Refresh (RAP)-based CIP and FIP algorithms currently being used for operational aviation icing decisions. The HiRes CIP and FIP products have undergone a number of modifications, including 1) an increase in horizontal and vertical resolution (from 20km horizontal and 1000 feet vertical to 13km horizontal and 500 feet vertical), 2) an increase in forecast leads from 12 to 18 hours, 3) an extension of the 'scenario' approach to the probability and SLD fields in CIP, 4) upgrades to the cloud top height algorithm, and 5) engineering upgrades and bug fixes.

The assessment has six main areas of investigation and incorporates output from the operational CIP/FIP (RAP) algorithms, the CIP/FIP HiRes, and the NWS-produced G-AIRMETs (Graphical Airmen's Meteorological Advisories), as well as METARs, PIREPs, and satellite observations, to establish a performance baseline. Primary findings include:

- HiRes field distributions are very similar to those from the RAP version
    - There is a small but consistent shift in the HiRes toward higher severity and higher probability in FIP.
    - Severity and probability appear to be strongly correlated
- Characteristics of CIP distributions are different than those for FIP: the largest difference is a strong diurnal signal in CIP in severity coverage in the high layer that is not present in FIP.
- HiRes performs slightly better than RAP but only when using a neighborhood in similar areal extent, i.e., not at the resolution of the product.
- When verifying against METARs, treating SLD unknowns as 'unknown' (rather than as 'no' as is implicitly done on the ADDS display) substantially improves performance.
- Overall, FIP outperforms CIP, except for SLD.
- FIP achieves higher POD and better PSS values than G-AIRMETs with a much smaller volume.
- CIP has a lower POD than G-AIRMETs but is more skillful.
- HiRes FIP captures nearly 80% of the MOG icing inside G-AIRMETs while excluding nearly 80% of the non-MOG icing reports.
- HiRes FIP captures nearly half of the MOG icing reports located outside of a G-AIRMET.
- The difference between CIP and FIP is much greater than between FIPs from successive issuances.

Table of Contents

List of Figures

List of Tables

1 Introduction

This document reports the QA PDT assessment of the High Resolution (HiRes) Current Icing Potential (CIP) and Forecast Icing Potential (FIP) algorithms developed by the National Center for Atmospheric Research. These HiRes products are to replace the current WRF Rapid Refresh (RAP)-based CIP and FIP algorithms currently being used for operational aviation icing decisions. The HiRes CIP and FIP products have undergone a number of modifications, including: 1) an increase in horizontal and vertical resolution (from 20km horizontal and 1000 feet vertical to 13km horizontal and 500 feet vertical), 2) an increase in forecast leads (FIP) from 12 to 18 hours, 3) an extension of the 'scenario' approach in CIP to the probability and super-cooled large drops (SLD) fields, 4) upgrades to the cloud top height algorithm, and 5) engineering upgrades and bug fixes.

The assessment incorporates output from the operational CIP/FIP (RAP) algorithms, the CIP/FIP HiRes, and the NWS-produced G-AIRMETS (Graphical Airmen's Meteorological Advisories), as well as METARs, PIREPs, and satellite observations, to establish a performance baseline, and has six main areas of investigation:

1. Characteristics of the product fields
2. Overall performance and meteorological accuracy of the HiRes CIP/FIP as compared to the current operational version, the RAP CIP/FIP
3. Performance of the HiRes CIP/FIP relative to the G-AIRMETs
4. Performance of HiRes CIP/FIP as a supplement to the G-AIRMET forecasts
5. Consistency of CIP/FIP HiRes forecasts between the various forecast issue and leads

The results and conclusions obtained from the QA PDT assessment will be provided to a Technical Review Panel as input to the decision on whether the HiRes CIP/FIP algorithms are ready for transition to operations at the National Weather Service (NWS).

2 Data

This section describes the forecast and observation data that were included in the assessment, along with the principal stratifications that were used. The time period for this study was January through March 2013.

2.1 Forecasts

2.1.1 CIP/FIP

The output from the grid-based CIP/FIP algorithms include: calibrated icing probability, icing severity, and potential for SLD (including freezing drizzle and freezing rain). The methodology used for producing CIP can be found in Bernstein et al., 2005. The spatial and temporal attributes of the CIP/FIP RAP and HiRes versions are outlined below.

Table 2.1: Attributes of the CIP/FIP RAP and HiRES.

	CIP/FIP RAP	CIP/FIP HiRes
Issues	Every hour	Every hour
Leads	CIP: 0	CIP: 0
	FIP: 1, 2, 3, 6, 9, 12	FIP: 1-12, 15, 18
Horizontal Resolution	20km	13km
Altitudes	1,000-30,000ft, 1,000ft increments	500-30,500ft, 500ft increments (500ft omitted – as it was not available throughout the period of study)

2.1.2 G-AIRMET

The Graphical Airmen's Meteorological Advisory (G-AIRMET) is a BUFR formatted time-series depiction of aviation hazards occurring with occasional or greater frequency throughout the conterminous U.S. and adjacent coastal waters (Murphy, 2010), and is a forecast for moderate or greater icing covering an area of at least 3000 mi^2. The G-AIRMET is issued 4 times per day (0300, 0900, 1500, and 2100 UTC) with forecast leads every 3 hours out to 12 h and from altitudes at the surface to 45,000 ft.

Note: the CIP and FIP are gridded products whereas the G-AIRMETs are human-generated polygons. The mechanics and approaches will account for these forecast differences. Additionally, G-AIRMETs include amendments and corrections. Amendments to the G-AIRMETs were excluded from this evaluation.

2.2 Observations

2.2.1 Voice Pilot Reports (PIREPs)

PIREPs are reported irregularly at the pilot's discretion and include a subjective assessment of many meteorological variables including the existence/absence of icing and a subjective measure of the icing intensity. Included in the icing reports are the location, altitude or range of altitudes, type of aircraft, air temperature, intensity, and type of icing (NWS 2007). The full range of intensity values were used (listed below), as forecasts of 'moderate or greater (MOG)' imply the need for the full range. The 'clear' icing type is used to indicate the possibility of SLD.

Icing intensity

1. Trace: Ice becomes perceptible. The rate of accumulation is slightly greater than sublimation. Deicing/anti-icing equipment is not utilized unless encountered for an extended period of time (over 1 hour).
2. Light: The rate of accumulation may create a problem if flight is prolonged in this environment (over 1 hour). Occasional use of deicing/anti-icing equipment removes/prevents accumulation. It does not present a problem if deicing/anti-icing is used.
3. Moderate: The rate of accumulation is such that even short encounters become potentially hazardous, and use of deicing/anti-icing equipment or diversion is necessary.
4. Severe: The rate of accumulation is such that deicing/anti-icing equipment fails to reduce or control the hazard. Immediate diversion is necessary.

Icing types

1. Rime: Rough, milky, opaque ice formed by the instantaneous freezing of small super-cooled water droplets.
2. Clear: A glossy, clear or translucent ice formed by the relatively slow freezing of large super-cooled water droplets.
3. Mixed: This is a combination of rime and clear.

2.2.2 METAR observations

Routine surface report (METAR) data are used to provide observations of icing conditions at the surface and to infer SLD events between the surface and the cloud ceiling. For instance, when freezing rain or freezing drizzle is recorded in the METAR, an SLD event is then inferred to exist between the surface and the cloud base (lowest cloud layer of at least "broken" coverage) (Madine 2008). This information is used to assess the quality of the CIP/FIP SLD parameter.

2.2.3 Satellite Data

Data products from the satellites CloudSat and CALIPSO combined with a temperature field provides a way to measure the boundaries for which icing conditions exist. CloudSat and CALIPSO are polar-orbiting satellites within the A-Train constellation. Each flies in a sun-synchronous orbit that is 705 km above the earth's surface. The ground track repeats every 233 orbital revolutions, or every 16 days (Stephens et al. 2002).

CloudSat carries a Cloud Profiling Radar (CPR), which sends out a series of short pulses at a 94 GHz frequency, providing a very detailed view of clouds from space. The CPR on CloudSat offers desirable sensitivity to cloud particles and provides data with an along-track resolution of ~1.7 km, a cross-track resolution of ~1.4 km, and a vertical resolution of 480 m with over-sampling every 240 m from the surface up to 30 km. While CloudSat has a history of use in verification studies (Kay et al. 2009), the satellite has recently had technical issues, and at present, only operates during daylight hours.

Among the primary objectives of the CALIPSO mission is the quantitative evaluation of clouds and cloud processes in the global atmosphere, and the evaluation of the relationship between vertical profiles of liquid water and ice. Instruments on this satellite utilize three receiver channels to measure the intensity of a returned lidar signal as a function of distance from the device. Information on the size and type of particles is computed from a ratio of measurements taken at two different wavelengths. The lidar signal is more prone to attenuation (signal loss) than the microwave signal sent from CloudSat (Platt 2011). The vertical resolution of the lidar ranges from 30 m near the surface to 300 m at higher altitudes, with a horizontal resolution of 333 m at the surface (NASA 2010).

2.3 Stratifications

Performance results were stratified spatially, temporally, and according to certain icing intensity thresholds.

Altitude bins

Results are aggregated into the following altitude ranges:

Stratification	CIP/FIP RAP	CIP/FIP HiRes
low	1000 – 10,000 ft	1000 – 10,000 ft
middle	11,000 – 20,000 ft	10,500 – 20,000 ft
high	21,000 – 30,000 ft	20,500 – 30,000 ft

Temporal Stratification

Forecast performance is stratified by forecast issue and lead times. The issues and leads included in each component of investigation depend upon the forecasts involved, and are typically an intersection of standard issue and lead times for the products included in the evaluation.

Intensity Stratification

The majority of the focus of the evaluation of icing intensity is on the Moderate-or-Greater level, but all CIP/FIP categories are assessed. However, PIREPs, CIP/FIP, and G-AIRMETs all use different measures of icing severity. Table 2 shows how intensity values are related between the three data sets.

Table 2.2: Mappings of Icing Severity Categories

(ADDS) PIREP	CIP/FIP category	G-AIRMET
Neg	None	N
Neg-Clr		
Trace	Trace	N
Trace-Light	Light	N
Light		
Light-Mod	Moderate	Y
Mod		
Mod-Severe	Heavy	Y
Heavy		
Severe		

Icing Probability Stratifications

Consistent with information provided by the Aviation Digital Data Service (ADDS), CIP and FIP icing severity are masked using three probability thresholds: > 5%, ≥ 25%, and ≥50%.

SLD Stratifications

Consistent with information provided by the Aviation Digital Data Service (ADDS), values of SLD potential are masked using three thresholds: < 0% (unknown), between 0% and 5%(no SLD), and ≥ 5% (SLD present).

3 Approach

One potential enhancement to FAA procedures includes the flexibility for aviation decision makers to consider input from a variety of forecast weather products for aviation icing decision guidance. Since the FIP and CIP products may be used in conjunction with other operational icing forecasts or individually to support flight planning, six main areas of forecast performance were investigated. These include

1. Characteristics of the product fields
2. Overall performance and meteorological accuracy of the HiRes CIP/FIP, including the extensions in lead-time and vertical coverage, as compared to the current operational version, the RAP CIP/FIP
3. Performance of the HiRes CIP/FIP relative to the G-AIRMETs
4. Performance of HiRes CIP/FIP as a supplement to the G-AIRMET forecasts to determine if CIP/FIP, when used in conjunction with the G-AIRMETs, adds meteorological detail for aviation operational planning
5. Consistency of CIP/FIP HiRes forecasts between the various forecast issue and leads

The mechanics of the assessment include: 1) a PIREPs-based technique for verification of both severity and SLD using a neighborhood-based approach for comparing forecasts to observations, 2) a satellite-based technique for icing occurrence that will provide an upper and lower bound for areas in which icing is possible, and 3) a METARs-based technique for SLD.

4 Methods

A variety of verification approaches are employed in this assessment. They are described in the following subsections.

4.1 CIP/FIP Field Characteristics

The makeup of the CIP/FIP fields is first evaluated using value-based distributions. Distributions were generated for each field: bins for CIP/FIP severity are generated per severity category, and the probability and SLD values are binned from 0 to 1.0 using a bin size of 0.01. Distributions for G-AIRMETs are not computed given that they are binary fields.

4.2 Forecast-Observation Pairing Techniques

In order to enable forecast comparisons and evaluation of quality, forecasts and observations are matched spatially and temporally using the following mechanics.

4.2.1 PIREP-based

4.2.1.1 CIP/FIP (Probability and Severity) to PIREP (Severity)

As in previous evaluations, PIREPs are matched to the CIP/FIP grid using a neighborhood of grid points surrounding the PIREP. The CIP/FIP intensity in the neighborhood that best matches the PIREP intensity is taken as the associated forecast value. The neighborhoods are defined as follows:

- For RAP, the nearest CIP/FIP flight level and the levels above and below are included, resulting in 3 vertical levels. A 2x2 horizontal neighborhood is used at each flight level, resulting in an overall neighborhood of 12 grid points.
- For HiRes, two neighborhoods are used:
    - The CIP/FIP flight level closest to the PIREP flight level is included, along with the two levels above and below, resulting in 5 vertical levels around the PIREP. A 3x3 horizontal neighborhood of grid points around the PIREP are included at each flight level, resulting in an overall neighborhood of 45 grid points. This provides a neighborhood of roughly the same volume as the 12-point RAP neighborhood.
    - The 12-point neighborhood as described for the RAP, but at the HiRes's 13km horizontal and 500 ft vertical resolution. The neighborhood includes the nearest CIP/FIP flight level and the levels above and below, with a 2x2 horizontal neighborhood at each level, resulting in the same number of grid points used as for the RAP.

On the boundaries of the grid, the subset of points available in the neighborhood is used for a best match. Gridpoints located below the model surface elevation are also excluded. In the 'best match' approach, if there is not a perfect match (a CIP/FIP intensity directly matching the PIREP intensity)

the closest match is determined by first searching all higher intensities for the closest higher, then searching all lower intensities.

For temporal matching, all PIREPs within a time window of [-30, 30) minutes around the forecast valid time are used to verify FIP. Because PIREPs prior to the analysis time are incorporated in CIP, only a time window of [0, 30) minutes around the analysis time were used to verify CIP.

4.2.1.2 G-AIRMET to PIREP (Severity)

PIREPs are matched to a G-AIRMET if they fall within the G-AIRMET's boundaries and within a time window of [-30, 30) minutes of the G-AIRMET valid time.

4.2.2 METAR-based

METARs are also included as an additional observation set for verification of SLD, using icing event data (FZRA, FZDZ) together with a reported cloud layer of at least "broken". The ceiling value is used to estimate the depth (also the top, with the bottom being at ground level) of the observed SLD layer.

For METARs that indicate SLD, SLD is assumed present from the ground to cloud base. For METARs that indicate no-SLD, the observation is assumed valid from the ground to either cloud base (if the METAR indicates snow) or to 30,000 feet (if the METAR indicates clear skies).

The CIP/FIP neighborhood that contains the METAR location, from the lowest CIP/FIP level up to the chosen top level, is compared to the METAR report. For METARs that indicate SLD, at least one of the vertical levels in the column of CIP/FIP grid boxes above the METAR site is expected to contain SLD. For the METARs that indicate no SLD, it is expected that all grid boxes above the site, up to the chosen top, will not contain SLD.

4.2.3 Satellite-based

An estimate of the upper and lower boundaries of where icing has the potential to form were derived each from CloudSat and CALIPSO using the cloud classification field from each product in conjunction with the temperature field from the RAP analysis. Only satellite swaths intersecting the RAP domain are considered. The temperature analysis from the RAP is mapped to the satellite swath using a nearest-neighbor interpolation, resulting in a RAP temperature analysis with the same resolution and profile as the CloudSat and CALIPSO data. Using the cloud classification and temperature thresholds conducive to icing, the temperature and satellite analyses are combined to identify areas of potential icing conditions. Table 3 shows the combinations of cloud type and RAP temperature that result in possible icing.

Table 4.1: Combinations of cloud type from CALIPSO and CloudSat with temperatures from RAP that produce possible icing conditions. The cloud types are as follows: Sc = stratocumulus, St = stratus, Ac = altocumulus, As = altostratus, Cu = cumulus, Ns = nimbostratus. Ci = cirrus, Cs = cirrostratus, and Cc = cirrocumulus.

Cloud Types	Temperature range for icing
Sc, St	0°C to -10°C
Ac, As	0°C to -20°C
Cu, Ns, deep convective	0°C to -25°C

Ci, Cs, Cc	no icing

Two different approaches are used for combining the cloud type and temperature information:

- Strict : cloud type and temperature range by type (Note: attenuated regions of CALIPSO not included)
- Relaxed: any cloud type and full temperature range (0 to -25C) (Note: attenuated regions of CALIPSO are included)

The strict and relaxed approaches will each be applied to CloudSat and CALIPSO. Because CloudSat and CALIPSO are on different grid representations, each of the four derived fields (Strict and Relaxed from CloudSat, Strict and Relaxed from CALIPSO) are projected to a common satellite grid that is between the resolution of the CloudSat and CALIPSO grids. The common satellite grid has a resolution significantly finer than that of the CIP/FIP.

The intersection of the Strict fields from CloudSat and CALIPSO defines the minimum area of icing potential that should be identified by the CIP/FIP products. The union of the Relaxed fields from CloudSat and CALIPSO identifies the maximum area where CIP/FIP should identify icing potential, i.e., CIP/FIP should not identify icing outside this area. CIP and FIP grids are matched to the Strict intersection field and Relaxed union fields using the common satellite grid.

Temporal matching is achieved by matching the CIP and FIP to the satellite data that occurs within +/- 30 minutes of the CIP/FIP valid time.

4.3 Evaluations

Terminology and score definitions are first provided for reference in the subsequent sections:

MOG:	Moderate- or-Greater Icing
LTMod :	Less-than-Moderate Icing
POD (= PODy):	proportion of all observed events that are correctly forecast to occur, in this case, of detecting icing at a specific threshold
PODn:	proportion of all observed non-events that are correctly forecast to occur.
POFD (= 1 – PODn):	proportion of all observed non-events that are mistakenly forecast to be events, in this case, detecting icing less than the specified threshold
CSI:	proportion of all forecast and observed events that were forecast correctly
PSS:	POD – POFD (Peirce Skill Score, aka True Skill Score, TSS)

4.3.1 CIP/FIP evaluation

4.3.1.1 CIP/FIP Severity

4.3.1.1.1 PIREPs-based

PIREPs are used to determine product skill in detecting CIP/FIP MOG severity as well as SLD. To be consistent with ADDS displays, the severity field is masked using probability values of 0.05, 0.25, and 0.50.

Due to the non-systematic nature of the verification data set (PIREPs), the "yes" observations and "no" observations must be treated separately (Carriere et al. 1997). As a result, it becomes inappropriate to compute several common statistics that would otherwise be computed and analyzed (e.g. Critical Success Index, Bias, and False Alarm Ratio). The rationale for this is well documented by Brown and Young (2000) and Carriere et al. (1997).

The association of the CIP/FIP product to PIREPs as described in Section 4.2.1 yields the following contingency table:

Hit:	forecast = yes; obs = yes
False alarm:	forecast = yes; obs = no
Miss:	forecast = no; obs = yes
Correct no:	forecast = no; obs = no

where 'yes' signifies that the forecast or observation equals or exceeds a given threshold, and 'no' signifies that the forecast or observed value is less than the threshold. POD, POFD, and PSS are computed from the contingency table.

4.3.1.1.2 Satellite-based

The Strict intersection and Relaxed union fields identifying potential areas of icing as described in Section 4.2 provide a lower and upper bound, respectively, for areas in which icing is possible. The Strict area defines the minimal regions where CIP/FIP should identify icing potential; the Relaxed area defines the maximal regions where CIP/FIP should identify icing potential (CIP/FIP should not identify icing outside of this area). Statistics are computed using CIP/FIP intensity and probabilities consistent with the ADDS display (5%, 25%, 50%). POD is computed for the Strict field, which represents the fraction of all grid cells in the Strict area for which CIP/FIP (correctly) identifies icing potential. POFD is computed for the Relaxed field, which represents the fraction of grid cells outside the Relaxed area for which CIP/FIP (incorrectly) identifies icing potential.

4.3.1.2 CIP/FIP SLD

For evaluation of CIP and FIP predictions of SLD, the SLD forecast is compared with PIREP reports of SLD, freezing rain, freezing drizzle, or intensity level of Severe with a "clear" icing type. We acknowledge that there are likely to be few PIREPs with SLD (as this is an area to avoid). METARs are also included as an additional observation using icing event data (FZRA, FZDZ) together with a reported cloud layer of at least "broken".

For the ADDS display, SLD potential is converted to a yes/no, where all SLD potential ≥ 0.05 is defined as a 'yes' forecast of SLD. As a result, in the display grid points with values of 'unknown' are treated as 'no' forecasts. For this evaluation, however, the performance of the CIP/FIP SLD forecasts is assessed considering each of the three possible treatments of the 'unknown' points. POD, POFD, and PSS scores will be calculated separately for each case, considering the 'unknown' points as 'yes' forecasts, as 'no' forecasts, and leaving them as 'unknown', or essentially removing those points from the verification.

4.3.2 CIP/FIP compared to G-AIRMET

When comparing CIP/FIP and G-AIRMET fields, CIP/FIP forecasts are matched to the G-AIRMET forecasts of the same issue and lead times. The same caveats for PIREPs (due to their non-systematic nature) listed in Section 4.2.1 also hold for G-AIRMET comparisons. The values of POD, POFD, and forecast volume from CIP/FIP and G-AIRMET are compared. In addition to the MOG CIP/FIP thresholds, thresholds at other intensities are also included in the comparison to assess how other CIP/FIP intensities perform in comparison to the G-AIRMET.

G-AIRMETs are, by definition, forecasts of MOG icing. Therefore, the contingency table is defined as:

Hit: MOG PIREP inside a G-AIRMET

False alarm: LTMod inside a G-AIRMET

Miss: MOG PIREP outside a G-AIRMET

Correct no: LTMod PIREP outside a G-AIRMET

The G-AIRMET contingency-table statistics POD, POFD, and PSS are then compared to the CIP/FIP contingency-table statistics as determined above.

4.3.3 CIP/FIP as supplement to G-AIRMET

In this study we provide a complementary view of CIP and FIP performance by considering their contribution as a supplement to G-AIRMETs. Inside a G-AIRMET, where MOG icing is predicted, CIP/FIP disagreement can potentially lower false alarm rates by reducing forecast volume. Outside a G-AIRMET, where MOG icing is not predicted, CIP/FIP disagreement can potentially reduce the likelihood of encountering an icing event without drastically increasing forecast volume.

Inside the G-AIRMET, where MOG icing is forecast to occur, the goal is to reduce the forecast volume (or the number of false alarms) without missing too many of the MOG observations captured by the G-AIRMET. Outside the G-AIRMET, where MOG icing is not forecast to occur, the focus is reversed: the goal is to capture as many of the missed MOG observations as possible without unduly increasing the number of false alarms.

As mentioned in section 4.1, when making comparisons to PIREPs, the neighborhood approach is used for the CIP/FIP algorithms, but in comparing G-AIRMET to PIREPs, the 'in or out' metric described above is used.

4.3.4 Consistency

In this area of investigation, the consistency of the CIP and FIP is assessed. Auto-lag correlation skills are computed, on forecasts of adjacent lead-times, to identify if there are any sudden changes in the correlation between forecasts at a particular lead-time. For example,

FIP 3-hour forecast is compared to the FIP 2-hour forecast, valid at the same time.

FIP 2-hour forecast is compared to the FIP 1-hour forecast, valid at the same time.

FIP 1-hour forecast is compared to the CIP analysis, valid at the same time.

From this type of comparison, gradual forecast improvement as the lead-time decreases is expected.

This comparison demonstrates hour-by-hour consistency in the FIP forecasts. An additional comparison is performed between the FIP forecast at each lead time and the CIP analysis corresponding to the valid time of the forecast, to determine general consistency of the FIP product with the CIP.

5 Results

5.1 CIP/FIP Field Characteristics

Before looking at the verification scores, it is useful to examine characteristics of the fields themselves, specifically distributions of the forecast values. Starting with the severity fields, Fig. 5.1 shows the distribution of the FIP severity categories in the lower (1-10 kft) layer masked by the 5%, 25%, and 50% probability fields (similar results obtain for CIP, not shown) for both the HiRes

and RAP versions. As lower probability areas are excluded the distributions shift to the right, favoring higher severities. This demonstrates a positive correlation between probability and severity: lower probabilities are generally associated with lower severity and higher probabilities are generally associated with higher probability.

In addition, note that while the pattern holds for both the RAP and HiRes versions, there is a modest but steady increase in the HiRes/RAP ratio (green bars) as the severity category increases. One possible explanation for the difference is the smoothing associated with interpolation from the native 13-km grid to the 20-km RAP grid.

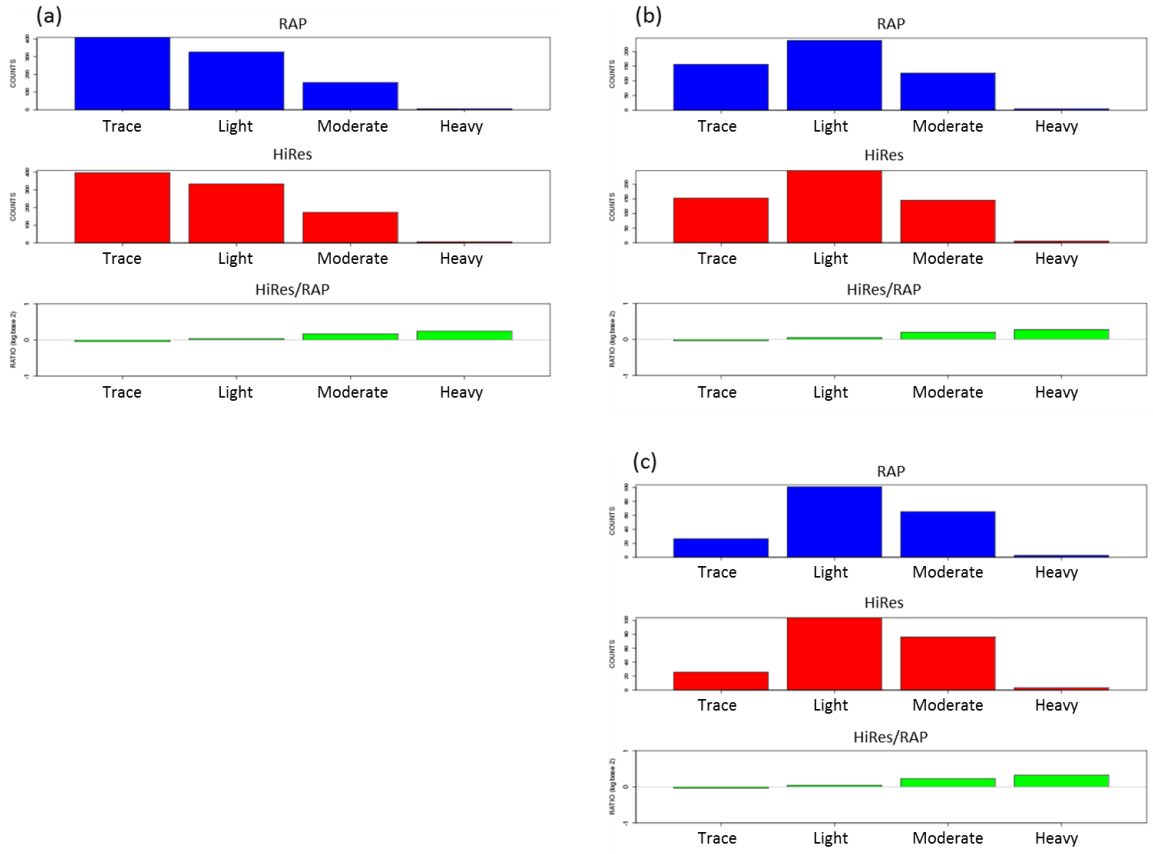

Figure 5.1: Distributions of severity for FIP in the 1000-10,000 ft layer for the RAP (blue) and HiRes (red) versions, along with the ratio between the two (green), using a log base 2 scale. Distributions are plots for the probability masks 5% (a), 25% (b), and 50% (c).

Comparing now between altitude levels (Fig. 5.2) for CIP with the 25% probability mask shows a similar picture, with a shift toward higher severity categories as elevation increases. The pattern of an increasing HiRes/RAP ratio with higher severity is present and even somewhat stronger than when comparing the distributions across probability masks. Again, the pattern for FIP (not shown) is very similar, but with a somewhat weaker trend in the HiRes/RAP ratio across severity categories.

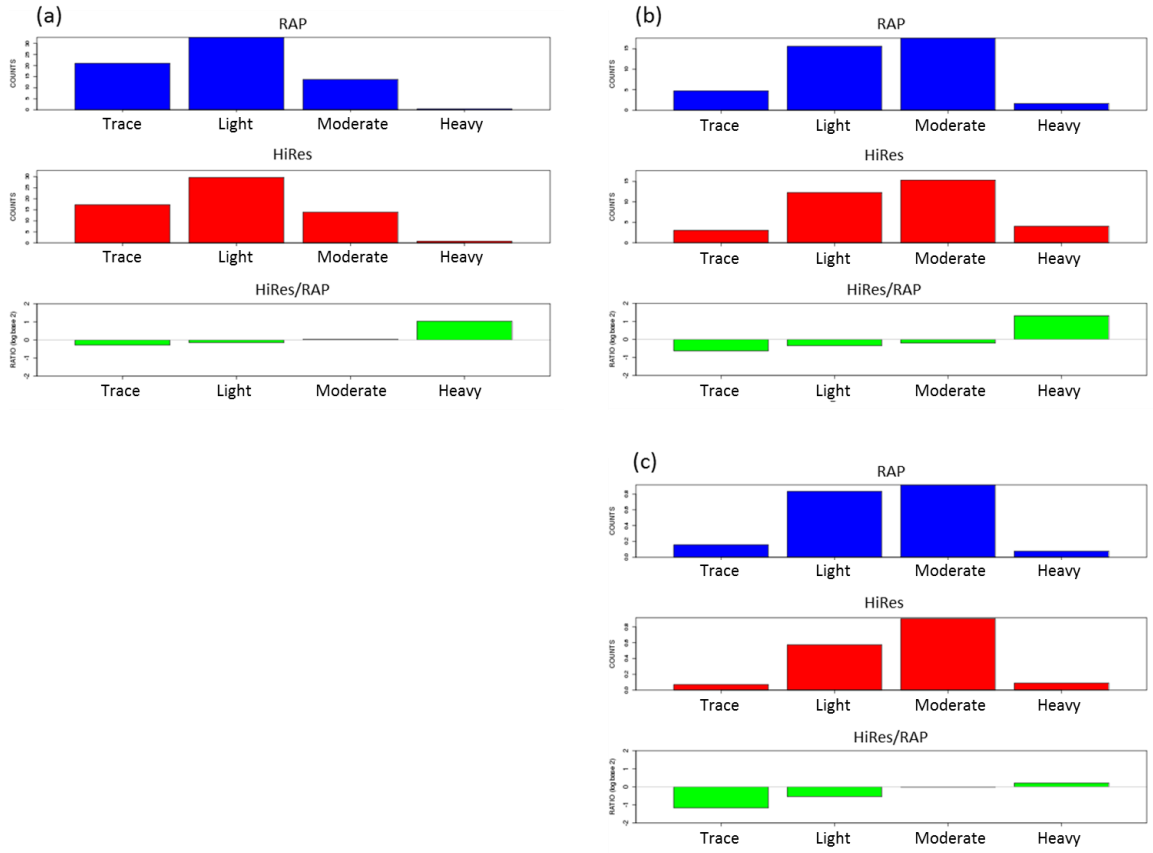

Figure 5.2: As in Fig. 5.1, but for CIP with the 25% probability mask for the low (a), middle (b), and high layers (c).

Shifting the focus to the probability fields, CIP in the low layer favors probabilities below 10%, but then remains nearly flat out to 70% whereupon the counts once again decline (Fig. 5.3a). Note the spike at 85%. There is a suggestion of a slight shift toward lower probabilities in the HiRes, relative to the RAP, but the two distributions are nearly identical. The FIP low layer probability distribution also favors lower probabilities (Fig. 5.3b), but the decline is nearly linear in FIP, out to an apparent cap at 80% but without the spike that is present in the CIP distribution. Again, the distribution remains unchanged moving from RAP to HiRes. For the middle layer (Fig. 5.3c), the distribution resembles a combination of the FIP and CIP low layer probability distributions: There is a linear decline in counts out to about 30% at which point the distribution levels out until dropping off above 60%.

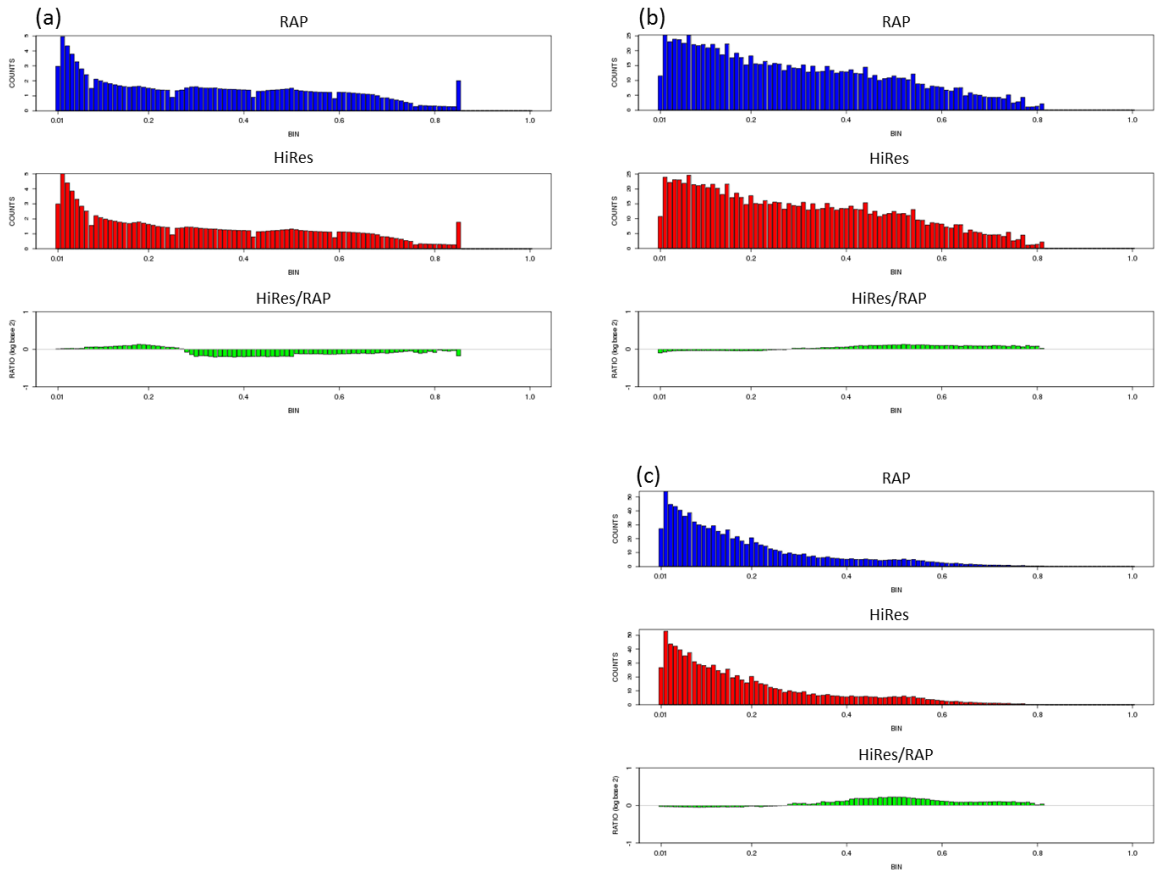

Figure 5.3: Distributions of probability values for CIP low layer (a), FIP low layer (b), and FIP middle layer (c).

The FIP probability distributions in Fig. 5.3 combine not only all issuances but all lead times masking the sensitivity of the distributions to the lead hour. Figure 5.4 shows the middle layer FIP probability distributions for 1-h, 6-h, and 12-h leads. Two features stand out: the steady movement of the probability cap toward lower values with longer lead times and a secondary peak in the tail of the distribution that appears to move with the probability cap.

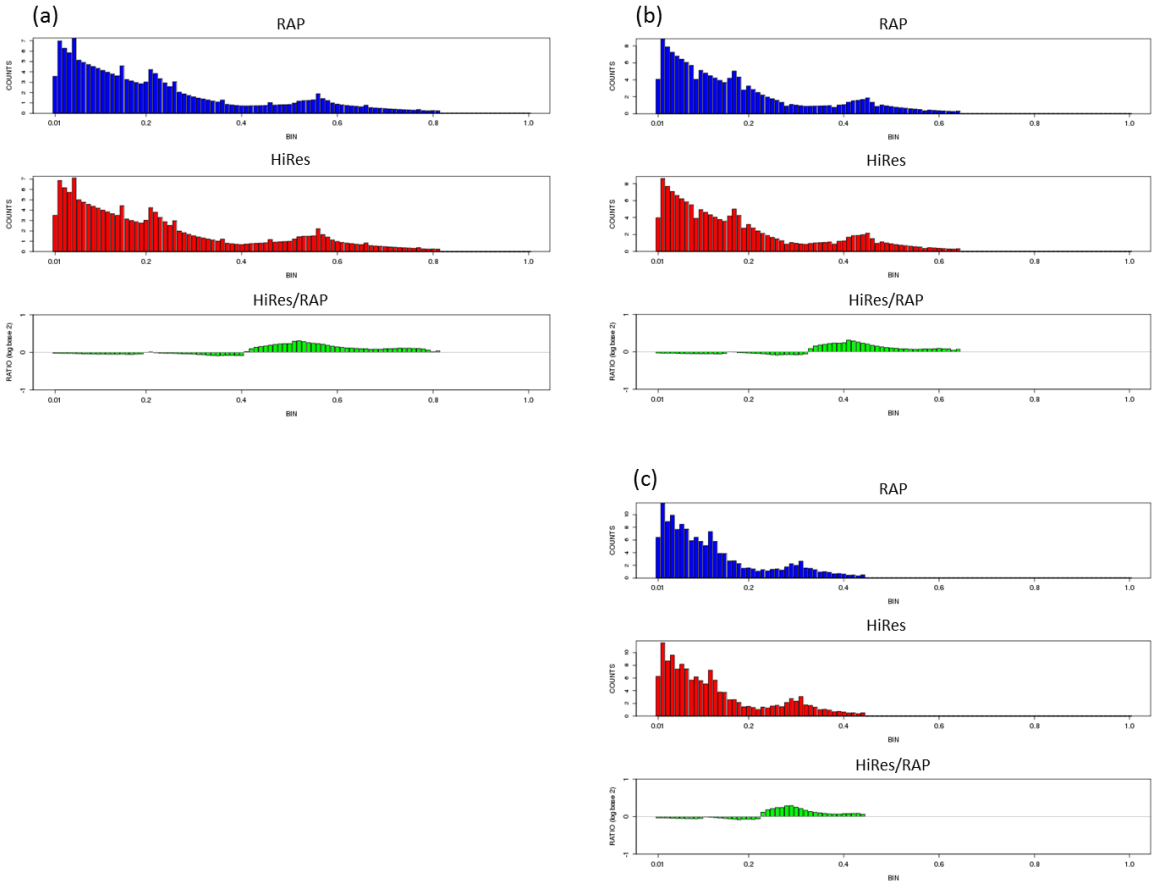

Figure 5.4: Distributions of probability values for CIP low layer (a), FIP low layer (b), and FIP middle layer (c)., but for the FIP middle layer at 1-h (a), 6-h (b), and 12-h (c) leads.

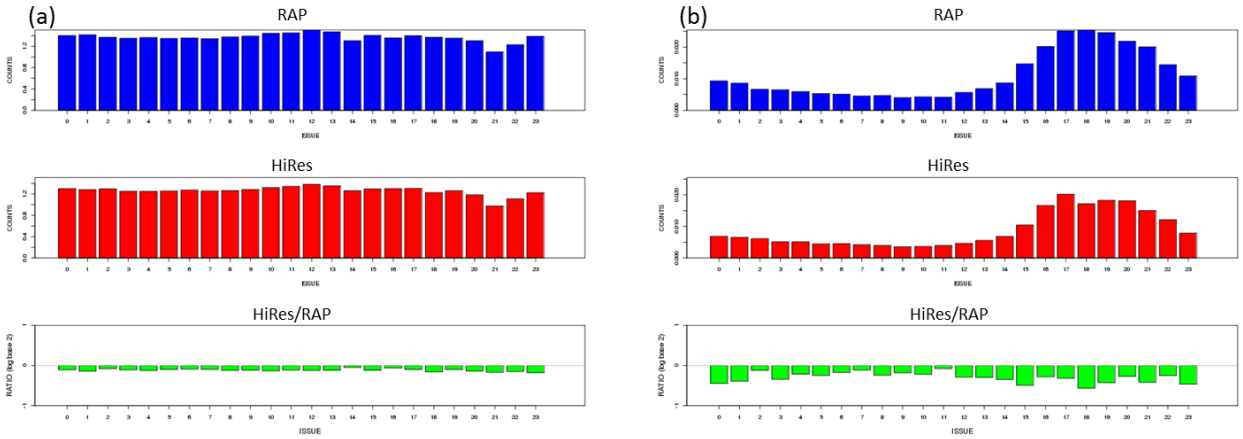

Figure 5.5: As in Fig. 5.1 but for CIP probability as a function of issue hour for the low layer (a) and the high layer (b).

The distribution of probability (all categories) as a function of issue hour for CIP reveals a substantial difference between the low and high altitude layers (Fig. 5.5). The distribution for the low layer is nearly uniform across valid hours, while for the high layer there is a distinct diurnal pattern with a minimum in the early morning hours (0900-1100 UTC) and a maximum around mid-day (1700-1800 UTC). The diurnal signal could be a result of the association of high layer icing to convection. Note that there is no diurnal signal in the corresponding FIP distribution (Fig. 5.6).

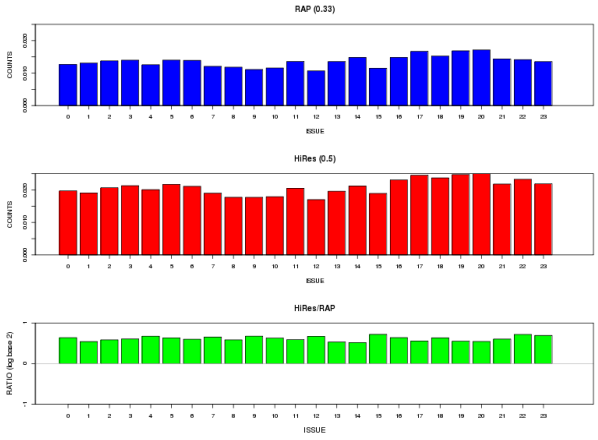

Figure 5.6: As in Fig. 5.5, but for FIP high layer at 1-h lead.

Finally, we consider the distributions of SLD (Fig.5.7). The CIP SLD low-level distributions (Fig. 5.7a) are considerably less smooth than those previously examined, but have a fairly uniform distribution with nearly evenly spaced spikes superimposed upon it. This behavior is more pronounced in the RAP than it is in the HiRes. For nearly 80% of the bins, there is more SLD in the HiRes version than in the RAP, however the other 20% of the bins strongly favor the RAP. As a result, there appears to be almost no overall bias in SLD coverage. The other notable feature of the distributions is the large spike at 1.0. The CIP middle-layer distribution (not shown) is very similar to the low level; the high level (not shown) has too few samples for any definitive remarks. For FIP (Fig.5.7b), the spike is even more pronounced (it accounts for about half of all SLD potential) but is shifted down to 0.67. The middle layer distribution (not shown) has a similar shape but only 1/3 of the SLD coverage of the low layer. It should be noted that on the ADDS display all SLD potentials above 0.05 are included as a binary (yes/no) field and so these features are invisible to users.

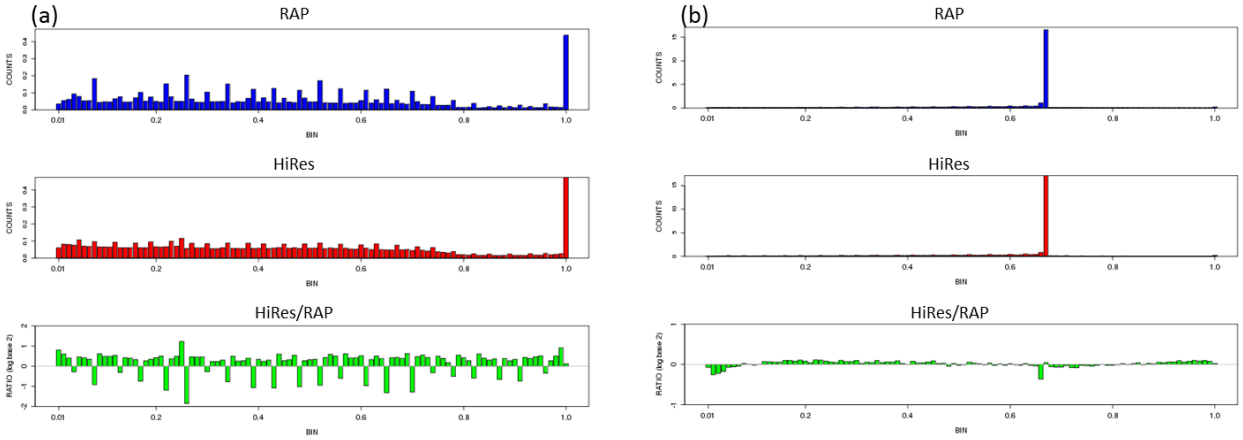

Figure 5.7: Distribution of SLD potential for CIP low layer (a) and FIP low layer (b).

5.2 Overall Performance

5.2.1 Severity

As explained in Section 4.1, CIP and FIP severity are verified against PIREPs, using the neighborhood approach. Figure 5.8 shows the accuracy of CIP and FIP MOG forecasts for the three probability masks aggregated over all vertical levels and all forecast leads. With the 5% probability mask, FIP outperforms CIP for both the HiRes and RAP versions: there is a substantial jump in the POD with only a small accompanying increase in POFD. Using the 25% probability mask, the increase in POD from CIP to FIP is smaller than with the 5% mask, and only slightly more than the increase in POFD. For the 50% probability mask, there is no increase in POFD, but the FIP POD is substantially lower than the CIP POD. Recall from Section 5.1 that FIP employs a cap on the probability values that decreases with increasing lead time. Figure 5.8 includes all FIP forecast leads and so is strongly affected by that cap and the corresponding decrease in forecast coverage.

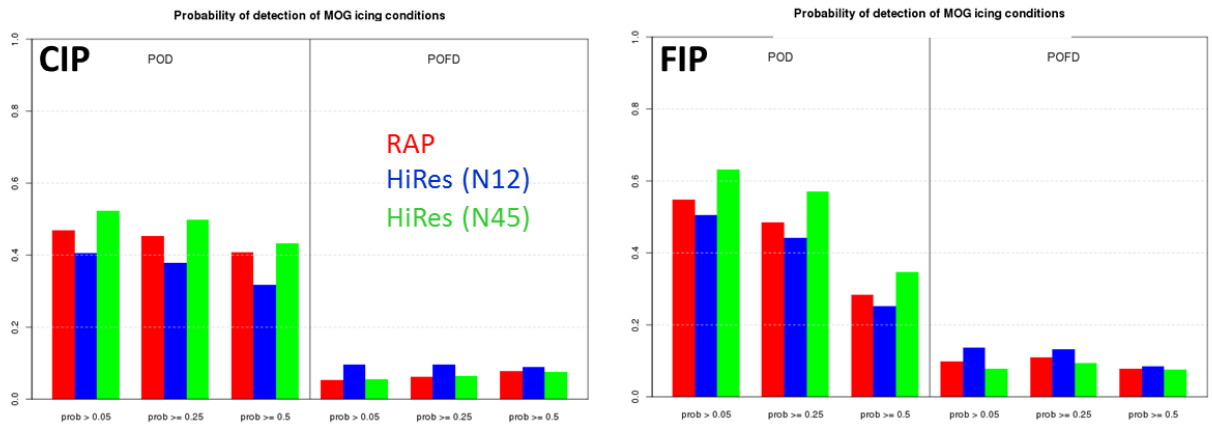

Figure 5.8: Probabilty of dection (POD) and Probability of False Detection (POFD) for CIP (a) and FIP (b) for the three probability masks for RAP (red), HiRes with the 12-point neighborhood (N12; blue), and HiRes with the 45-point neighborhood (N45; green).

Using the 12-point neighborhood for HiRes results in a drop in skill relative to the RAP version. The HiRes (N12) POD is lower and its POFD is slightly higher. By contrast, using the 45-point neighborhood results in an increase in skill for both CIP and FIP and for all probability masks. The volume of the HiRes 45-point neighborhood is nearly identical to the RAP neighborhood while the volume of the HiRes 12-point neighborhood is substantially smaller, thus requiring the HiRes to be more precise in its placement of icing events. The fact that the HiRes skill improves over the RAP for the 45-point neighborhood but decreases for the 12-point neighborhood suggests that the quality of the information at the original grid resolution has improved but the increase in the information resolution does not match the increase in the grid resolution.

Figure 5.8 showed the FIP performance aggregated over all lead times; viewing the performance as a function of lead time reveals very little sensitivity to the lead hour (Fig. 5.9). The notable exception to this is when the 50% mask is used, where the effect of the increasingly severe probability cap is evident. The superiority of the HiRes(N45) over the RAP version holds not just in aggregate, but for all lead times as well. (The RAP version extends out to 12 hours only, and so the 15-h and 18-h HiRes leads are not included.) Similarly, the improvement of the FIP over the CIP holds for all leads (with the exception of the 50% mask curves).

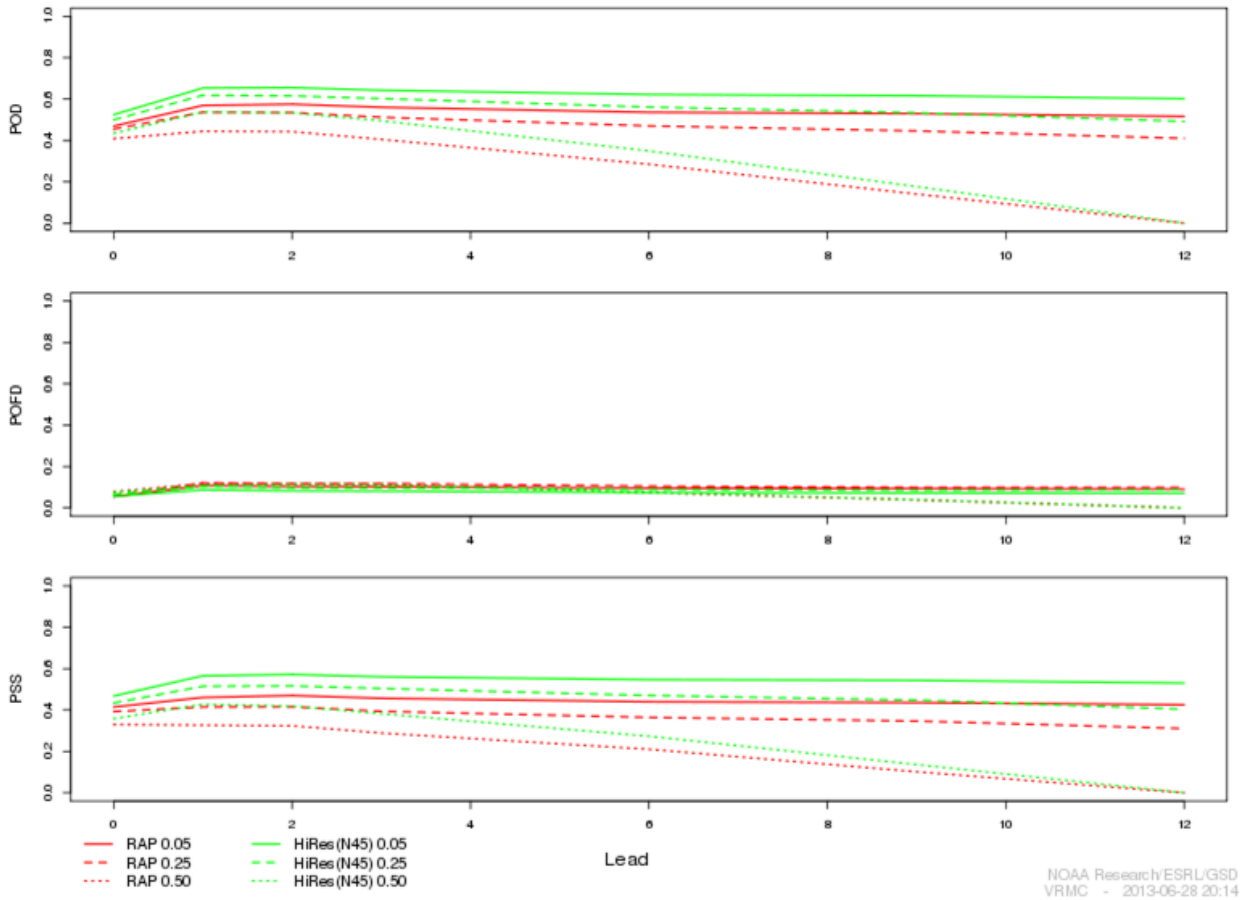

Figure 5.9: POD (top), POFD (middle), and PSS (bottom) as a function of forecast lead time for the RAP (red) and HiRes (N45) (green), for the 5% (solid), 25% (dashed), and 50% (dotted) probability masks.

Comparing the performance for the different altitude layers reveals that the improvement in skill of the HiRes (N45) over the RAP is greatest in the low layer (Fig. 5.10; only the 1-h FIP is shown to avoid the effects of the probability mask). The POD is highest in the middle layer, but at the cost of a large increase in POFD, consistent with the greater coverage of MOG icing forecasts in the middle layer than in the low layer. The POFD drops again for the high layer, but the POD declines substantially as well, reflecting the small number of MOG icing forecasts above 20,000 ft.

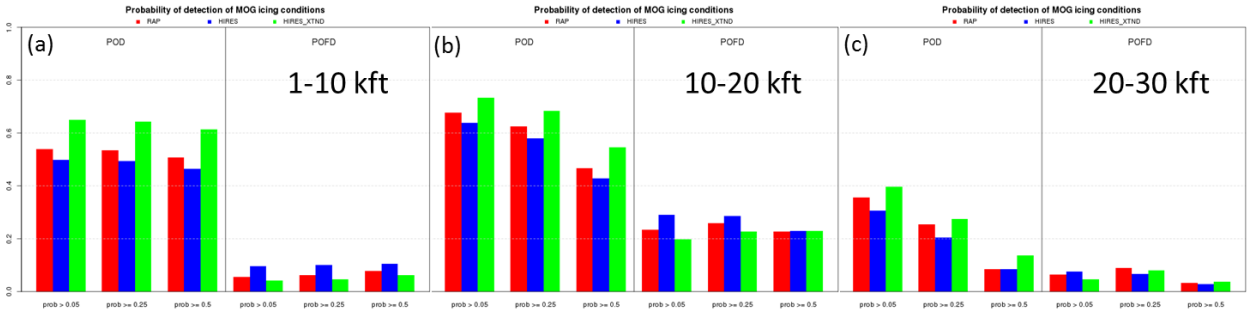

Figure 5.10: As in Fig. 5.8, but for FIP 1-h lead for the low (a), middle (b), and high (c) layers.

As expected, the POD and POFD decrease as the icing event being forecast becomes more severe (Fig. 5.11). The decline accelerates with increasing severity: nearly 90% of all Trace icing observations are captured by the 1-h HiRes (N45) FIP forecast (with the 5% probability mask) (Fig. 5.11a), dropping to around 85% for Light icing (Fig. 5.11b), 65% for Moderate (Fig. 5.11 c), and less than 10% for Heavy (Fig. 5.11 d). Note also, that the significant drop in POD between Light and Moderate is accompanied by a much smaller decline in POFD. In other words, what is seen is not just the expected decline in POD and POFD as the forecast event becomes more rare (as occurs between the Trace and Light categories), rather FIP is also more skillful in forecasting Light icing than in forecasting Moderate icing.

Please note that, due to the consistent nature of the relationship between verification scores using the 12-point and the 45-point neighborhoods, for the rest of the report, the term HiRes will be used to refer to the 45-point neighborhood approach, unless otherwise indicated.

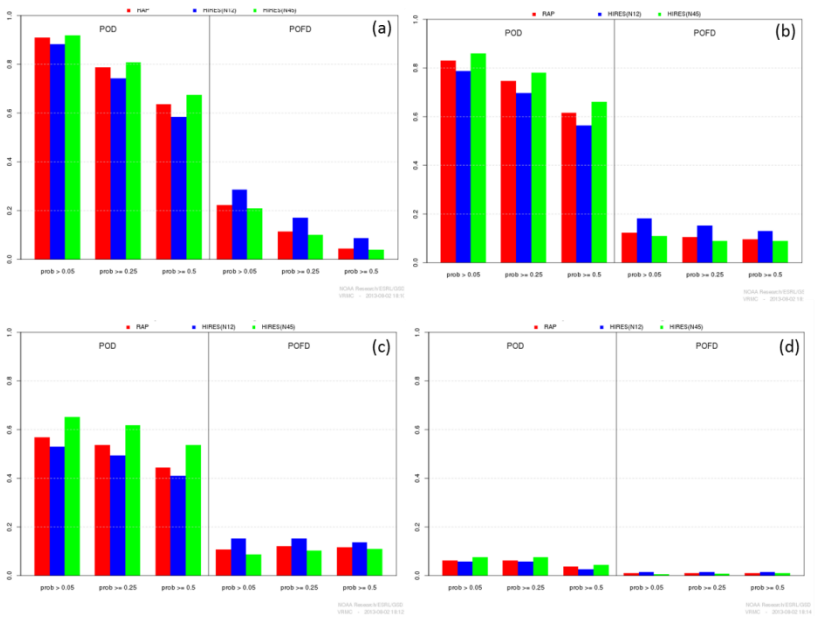

Figure 5.11: As in Fig. 5.8, but for 1-h FIP forecasts of Trace-and-above (a), Light-and-above (b), MOG (c), and Heavy (d) icing.

5.2.2 SLD

CIP and FIP forecasts of Supercooled Large Drops (SLD) consist of a potential field ranging from 0 to 1. This is converted to a binary field by considering all SLD potential ≥ 0.05 to be a positive forecast of SLD and all SLD potential < 0.05 to be a negative forecast of SLD. In addition, however, there is a third category of 'unknown', which ADDS does not display, thereby implicitly treating unknowns as forecasts of no SLD. Furthermore, there is a large difference in the number of 'unknown' forecasts between CIP and FIP: FIP contains about 50% more than CIP, though most of these come at the expense of the 'no' forecast so that on ADDS there is little difference between the two.

It is possible, however, to examine the performance of the SLD forecasts for cases where the 'unknown' forecasts are handled differently. Considering first SLD forecasts verified against PIREPs (Fig. 5.12), treating the unknowns as forecasts of no icing yields almost no false detections; only 10 to 30% of all SLD observations are captured. If the unknowns are instead treated as 'yes' forecasts, the POD jumps to 0.6-0.8, but at the cost of a very large number of false detections such that the skill for most leads is even lower than when the unknowns are treated as 'no' forecasts. Finally, if the unknowns as left as unknown, i.e., only explicit 'yes' or 'no' forecasts are considered, the product captures nearly half of all SLD events while maintaining a very low POFD, yielding the highest skill. Note that whereas the severity forecast consistently shows FIP outperforming CIP for both RAP and HiRes, the opposite is true for the RAP forecasts, where CIP outperforms FIP regardless of how the unknowns are handled. For HiRes, however, the CIP skill has decreased significantly compared to RAP such that FIP outperforms CIP. The decline in CIP skill comes from a large reduction in the POD for the HiRes CIP compared to the RAP CIP.

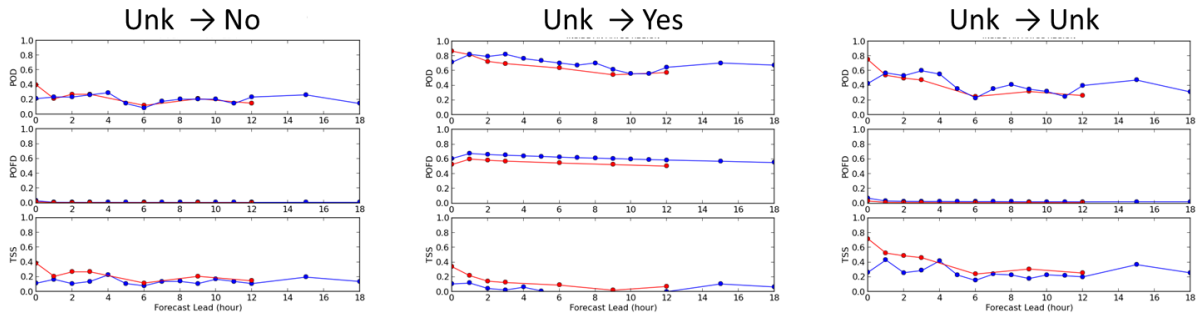

Figure 5.12: Plots of POD, POFD, and PSS for CIP/FIP forecasts of SLD by lead time for HiRes (blue) and RAP (red) where SLD 'unknown' forecasts are treated as 'no' (left), 'yes' (middle), and 'unknown' (right). Forecasts are verified by PIREPs.

When the SLD forecasts are verified against METARs the behavior is somewhat different (Fig. 5.13). SLD, while still rare, is more common in the METAR data than in PIREPs. As a result, even when the unknowns are treated as 'no' forecasts, about 70% of all icing events are correctly forecast (POD is sensitive to the event base rate as the base rate approaches zero), while still maintaining a very low POFD. Once again, both POD and POFD increase if the unknowns are treated as 'yes' forecasts (about 90% of all events are captured for shorter leads), with the skill being marginally lower than when treating the unknowns as 'no'. Leaving the unknowns as 'unknown' results in a very skillful forecast. Unlike when using PIREPs to verify the SLD forecasts, the HiRes CIP mirrors the RAP in outperforming the FIP SLD forecasts.

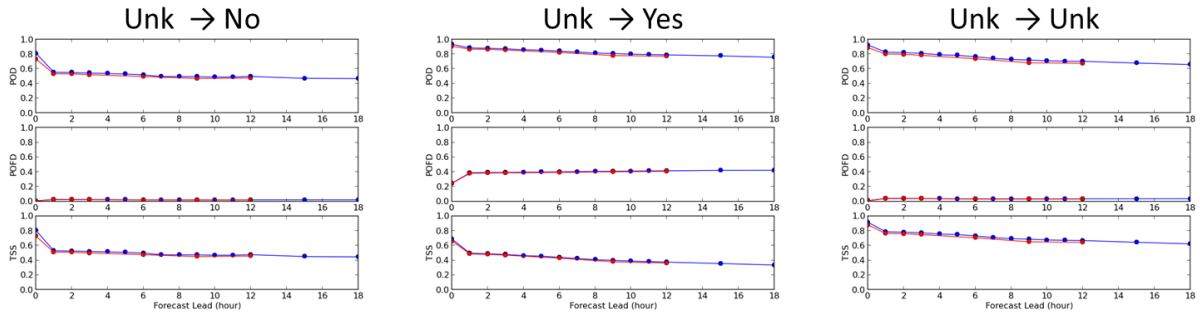

Figure 5.13: As in Fig. 5.12, but for forecast verified by METAR.

While it cannot be claimed that the METAR data represents a completely accurate representation of SLD occurrence in the atmosphere, it does have the advantage of being a temporally continuous, fixed-in-location observation set. Therefore, it does not suffer from the adverse selection bias where pilots pointedly avoid flying through areas where SLD is believed to be present, reducing the number of positive observations. Nevertheless, an amount of uncertainty remains in determining the skill of CIP and FIP in forecasting SLD occurrence.

5.2.3 Satellite

In situ measurements of in-flight icing are of great value, but they also suffer from a large, well-known defect: pilots avoid flying through areas strongly believed to have significant icing potential. Together with the idea that the non-existence of an icing report is not to be equated with a report of a lack of icing, the result is a lack of information from regions of the most interest. Furthermore, certain basic characteristics of icing in the atmosphere, such as the frequency of occurrence, remain unknown.

Remote sensing of the atmosphere circumvents the problem of biased observations, but at the cost of a lack of precision in those observations. Because of the uncertainties in the location and magnitude of satellite-based icing observations, a method is employed combining data from two different satellites (CloudSat and CALIPSO) in two different ways so as to form upper and lower bounds for the presence of icing in the atmosphere. The details are contained in Section 4.2.3, but, briefly, the result is a 'strict' field in which there is a high confidence of the presence of icing, i.e., CIP/FIP should indicate icing in this region; and a 'relaxed' field in which there is at least a low confidence of the presence of icing, or put conversely, there is a high confidence of a lack of icing outside this region, i.e., CIP/FIP should not indicate icing outside of the relaxed region.

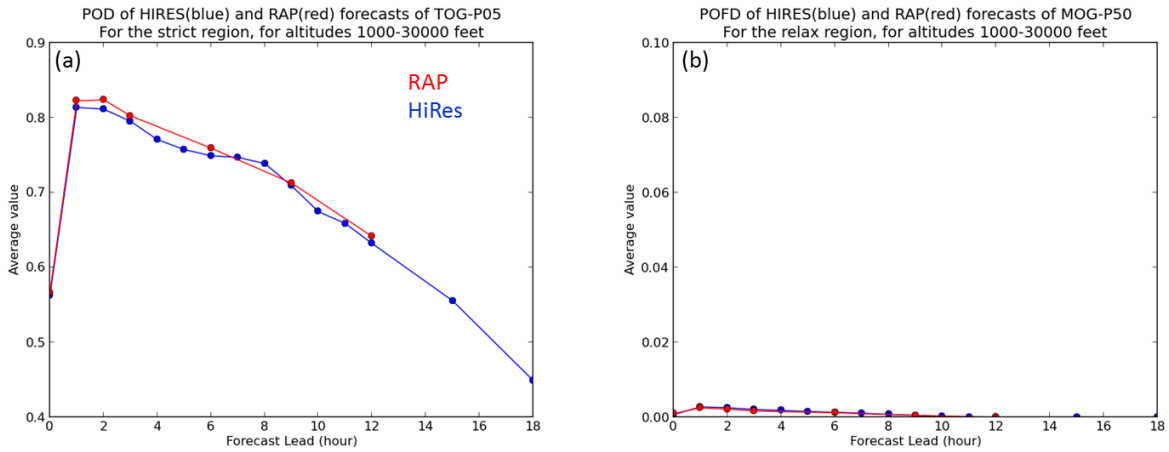

Figure 5.14: POD (a) and POFD (b) for RAP (red) and HiRes (blue) measured against CloudSat and CALIPSO-derived icing observations. POD calculations are for forecasts of Trace and above with the 5% probability mask; POFD for the moderate and above with the 50% mask.

As a result of the satellite method providing only bounds on icing occurrence, POD and POFD can be computed separately but cannot be combined to produce a skill score. FIP achieves a strong POD for all but the longest lead hours (Fig. 5.14a) while CIP captures over a quarter fewer events than the 1-h FIP. Indeed, the FIP POD is higher than that for CIP out through 12 hours. Note also that the POD for RAP and HiRes are nearly indistinguishable. (The apparent differences in the curves occur primarily for forecast leads not present in RAP, that is, where the RAP curve is interpolated between data points.) Both RAP and HiRes have an extremely low rate of false detection (Fig. 5.14b). In part, this is because of an inflated count of 'correct no' forecasts that serves to inflate the denominator of the POFD. For example, an icing algorithm is rarely necessary to determine that there will be no icing near the surface in the southern portions of the domain where temperatures are well above freezing or higher in the atmosphere and further north where temperatures are too cold for any liquid water to be present. No attempt was made to eliminate these "easy" correct-no forecasts in this evaluation, but an approach for doing so may be included in future assessments.

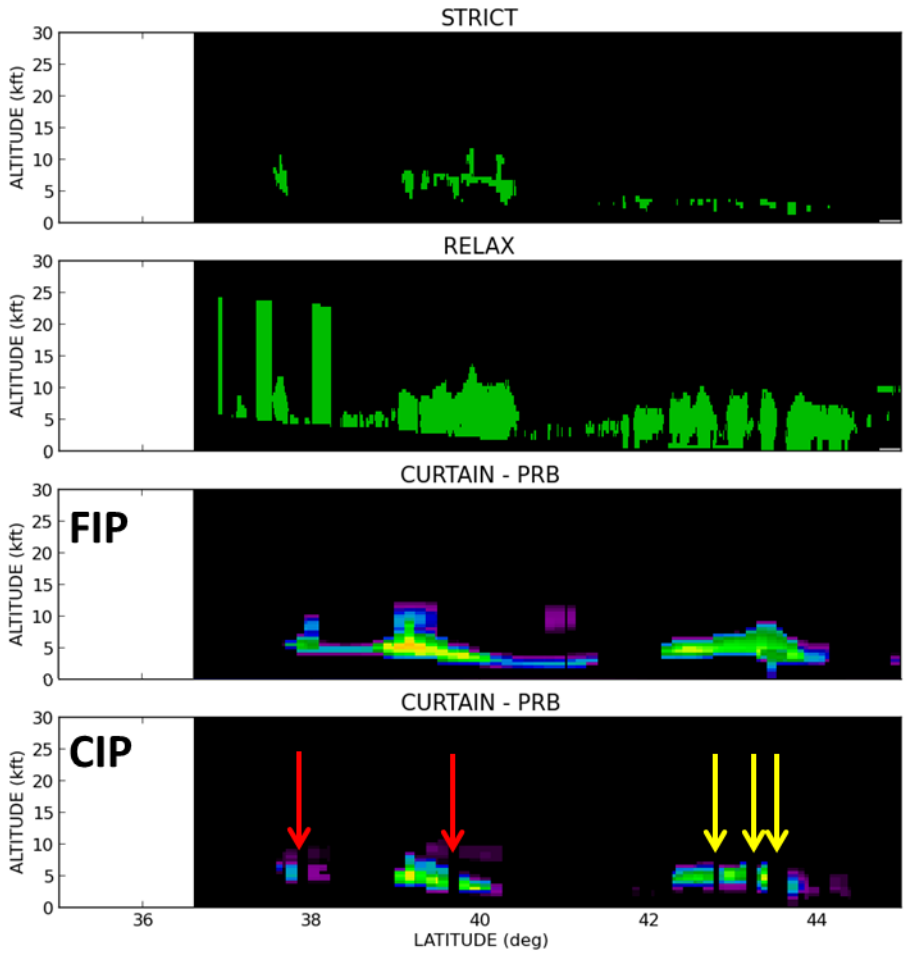

Figure 5.15: A sample satellite swath showing, from top to bottom, the strict satellite icing field, the relaxed satellite icing field, FIP, and CIP. Red and yellow arrows highlight empty vertical columns in the CIP field.

The discrepancy between CIP and FIP is larger when verified against satellite than it is when verified by PIREPs. One explanation is the greater coverage in the FIP icing fields than in CIP, as was noted in Section 5.1. This difference in coverage is expected to have a greater impact in the satellite verification because of its continuous swath as opposed to the PIREPs whose discrete, isolated nature reduces the likelihood of finding a location where FIP predicts icing and CIP does not. An example of larger FIP icing coverage from the satellite verification perspective is shown in Fig. 5.15. In addition to the overall greater coverage, note the holes, or empty columns (indicated by the red and yellow arrows), present in the CIP field but not the FIP. For the locations indicated by the yellow arrows, there are cloud/icing-free columns in the satellite data as well, but the character of these holes is different in the CIP field, particularly the purely vertical edge to the holes. The CIP holes indicated by the red arrows are absent altogether from the satellite field.

The POD shown in Fig. 5.14 is calculated using the least restrictive CIP/FIP icing forecast, i.e, trace-and-above with the 5% probability mask, giving the icing products the best possible chance to capture observed icing events. It is instructive to examine the sensitivity of the accuracy measure

to this choice. Fig. 5.16 shows the POD of CIP and FIP verified against the strict satellite field for all combinations of severity (up to MOG) and the three probability masks. The top left panel is identical to that shown in Fig. 5.14a. The results show a strong sensitivity to both the severity category and the probability mask, with the peak (1-h lead) POD falling from greater than 0.8 for the trace-and-above with the 5% mask to less than 0.2 % for the MOG with the 50% mask. Results are qualitatively similar for POFD (not shown) but the range of values is greatly constrained for the reasons noted above.

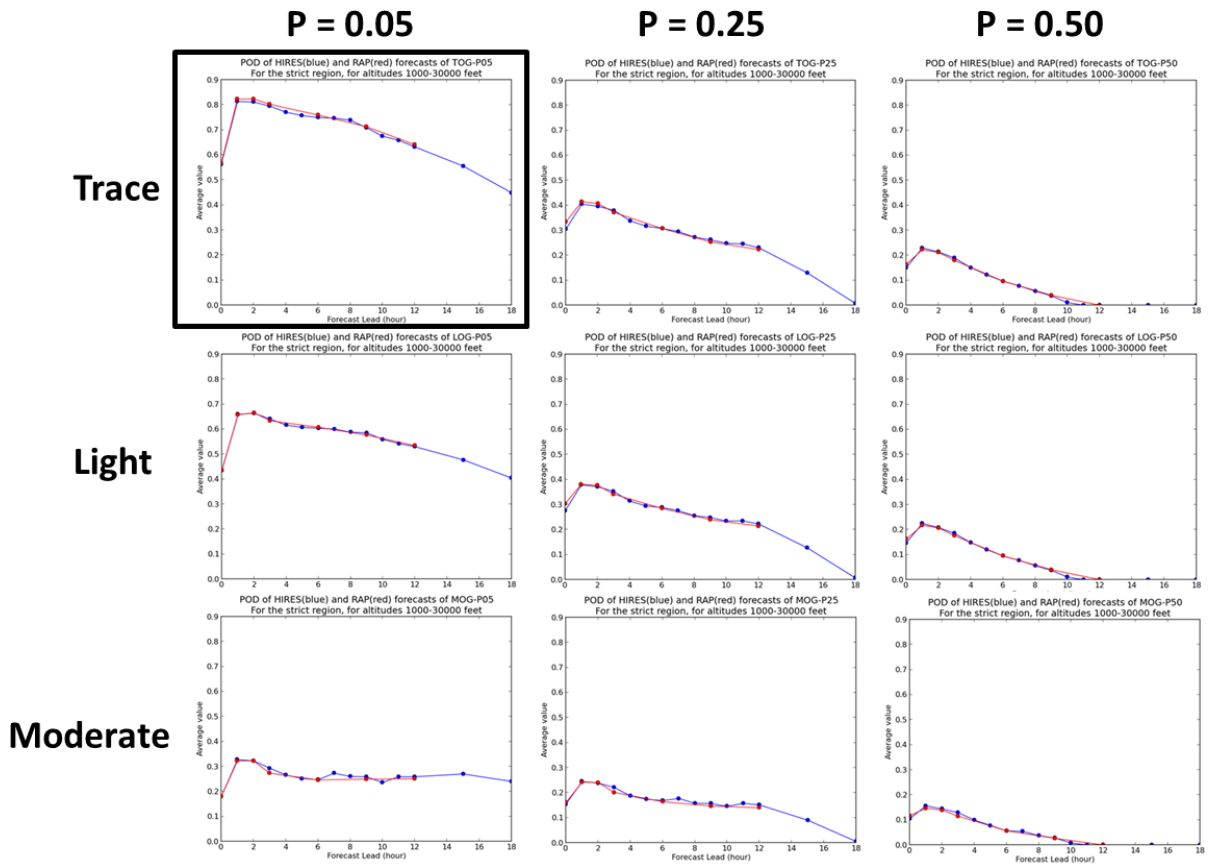

Figure 5.16: As in Fig. 5.14a, but for all combinations of severity category (rows: trace-, light-, moderate-and-above) and probability mask (columns: 5, 25, 50%).

5.3 CIP/FIP compared to G-AIRMET

In addition to CIP and FIP, MOG-icing forecasts are provided by G-AIRMETs, though the latter consist of snapshots issued four times a day (0300, 0900, 1500, 2100 UTC) at three-hourly forecast leads out to 12 h. In order to facilitate a fair comparison between the two products, only CIP/FIP issuance and lead times matching those of the G-AIRMETs are included in this section of the evaluation. G-AIRMETs also operate under a minimum size requirement: they must cover an area of at least 3000 mi^2. No attempt is made to place similar size requirements on the CIP/FIP forecasts.

G-AIRMETs capture nearly 60% of all MOG-icing events for the 0-h lead (because of the minimum-size restriction, G-AIRMETs are not expected to capture all icing events), declining only to around 50% at twelve hours, but this comes at the cost of a high false alarm rate (~0.40) resulting in fairly low skill (Fig. 5.17). CIP captures fewer events than the G-AIRMETs, but with a large reduction in false alarms (by a factor of four) so that the CIP skill is substantially higher than G-AIRMET skill. The RAP FIP has a nearly identical POD to the G-AIRMETs, while the HiRes FIP captures slightly more events; both nearly match CIP in reducing the number of false alarms leading to much higher skill for FIP.

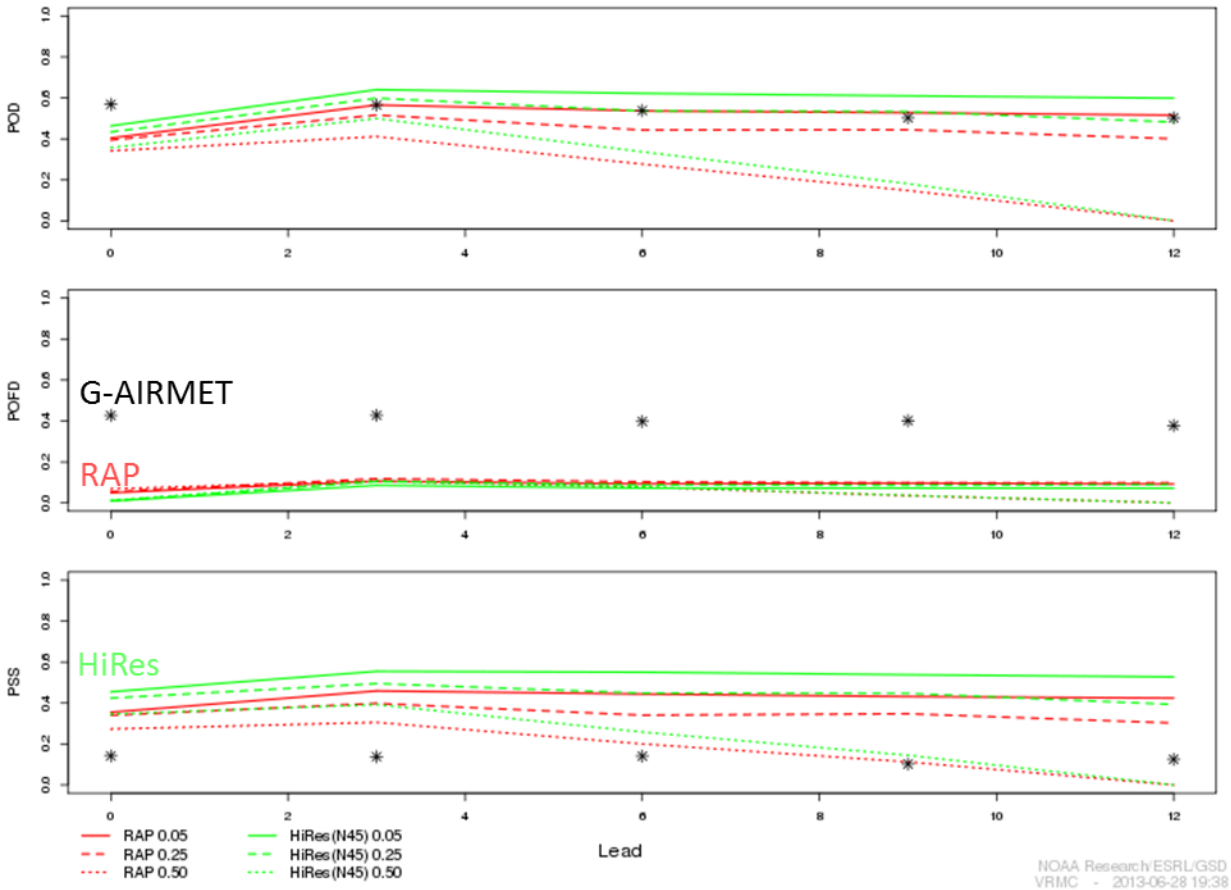

Figure 5.17: As in Fig. 5.9, but for RAP (red), HiRes (green), and G-AIRMET (asterisk).

Another perspective of the forecast performance is the volume required to capture MOG icing events. Figure 5.18 presents the G-AIRMET and CIP/FIP POD as a function of the percent volume covered by the forecasts, where the volume is for all forecasts valid at a given time, not the volume of a single polygon, and the 5% probability mask is used for CIP and FIP. Once again, it is seen that the CIP POD is less than that for the G-AIRMETs, but CIP reduces the forecast volume by nearly half. It would be possible to increase the number of MOG-icing events captured by CIP by using a lower

forecast severity threshold, but the resulting forecast would consume nearly 50% more volume than the G-AIRMETs. FIP forecasts of MOG-icing are very similar to the G-AIRMETs in both the POD and volume. This with the reduction in POFD—that is, FIP reduces the number of false alarms without reducing the POD or total forecast coverage—indicates that FIP is considerably more accurate with the placement of its icing forecasts. One caveat for this interpretation is that pilots may be more motivated to report less-than-moderate icing conditions inside a G-AIRMET than outside it, such that false detections outside of G-AIRMETs may be underreported. Reducing the forecast severity threshold for FIP produces an equally large increase in the volume but with a smaller gain in POD.

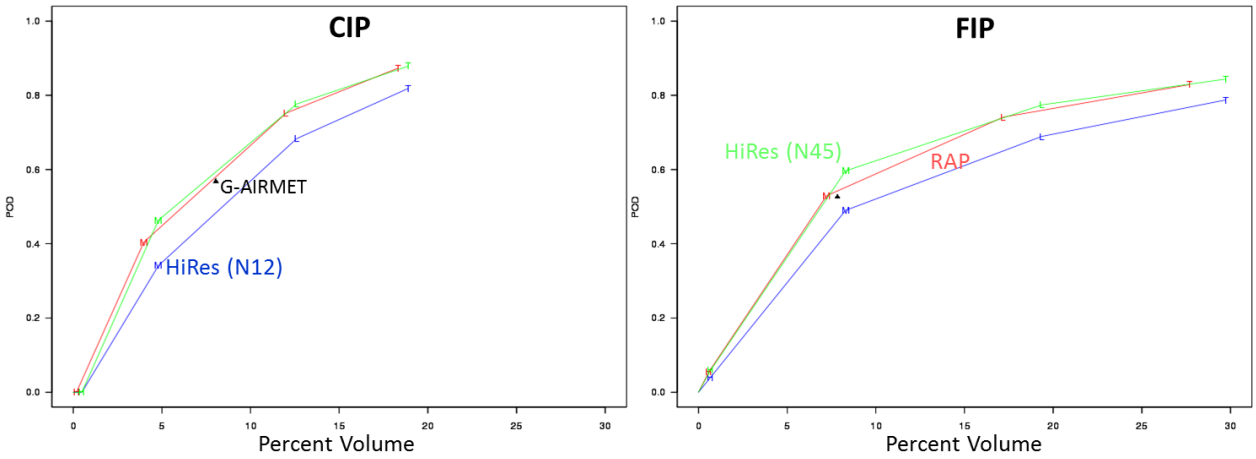

Figure 5.18: POD as a function of the forecast percent volume from CIP (left panel) and FIP (right panel) for G-AIRMET (black triangle), RAP (red), HiRes (N45) (green), and HiRes (N12) (blue).

Increasing the probability mask both moves the curves to the left (i.e., smaller volume) and moves the points on the curve downward (i.e., lower POD). Instead of examining the POD as a function of volume, one can also look at skill as a function of volume. Using the 25% probability mask, the gap in skill between CIP and FIP nearly disappears and CIP uses a smaller volume to achieve that skill (Fig. 5. 19). Both products are clearly superior to the G-AIRMET in that they yield higher skill with smaller volumes. In fact, using the Light severity threshold as a forecast of MOG icing for CIP gives a large increase in skill with a volume still smaller than G-AIRMETs. Using the Light threshold with the 25% mask compared to the Moderate threshold with the 5% mask (not shown) increases the PSS from around 0.4 to around 0.6 while increasing the volume from around 5% to around 7%. Different users will make different choices as to an acceptable tradeoff between skill and the volume required to achieve that skill, and the various combinations of severity level and probability mask allow for a wide range of choices in that space.

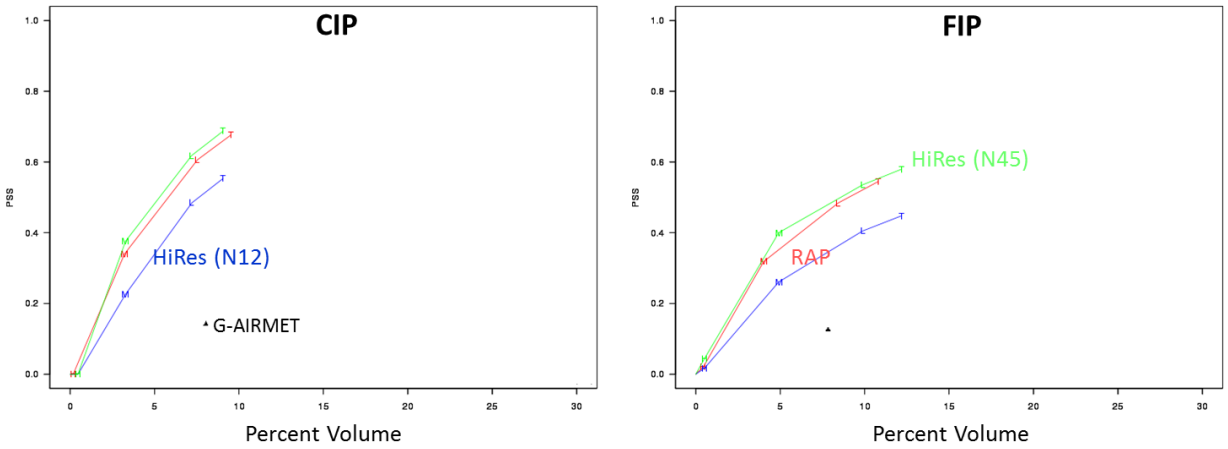

Figure 5.19: As in Fig. 5.18, but for PSS as a function of forecast percent volume.

5.4 CIP/FIP as a supplement to G-AIRMET

Along with comparing the performance of CIP and FIP to G-AIRMETs, one can consider their use as a supplement to the G-AIRMETs. That is, inside the G-AIRMETs, does CIP/FIP reduce the number of false alarms while still capturing most of the icing events? Conversely, does CIP/FIP capture events missed by G-AIRMETs (i.e., events that occur outside of the G-AIRMET polygons) without unduly increasing the number of false alarms?

Focusing first on the airspace inside the G-AIRMETs, once again FIP captures more events than CIP at the cost of somewhat more false alarms (Fig. 5.20). In the context of being used as a supplement to the G-AIRMETs, both HiRes products eliminate over 80% of all false alarms located inside G-AIRMET polygons, but FIP does so with a smaller penalty, that is with a smaller number of missed events. For most of the results presented in previous sections, the improvement of the HiRes over the RAP is steady but slight. CIP/FIP performance inside the G-AIRMET presents a case where

HiRes is clearly better than RAP. RAP is slightly worse at avoiding false alarms (~20% vs. 15% for FIP) while missing more MOG icing events.

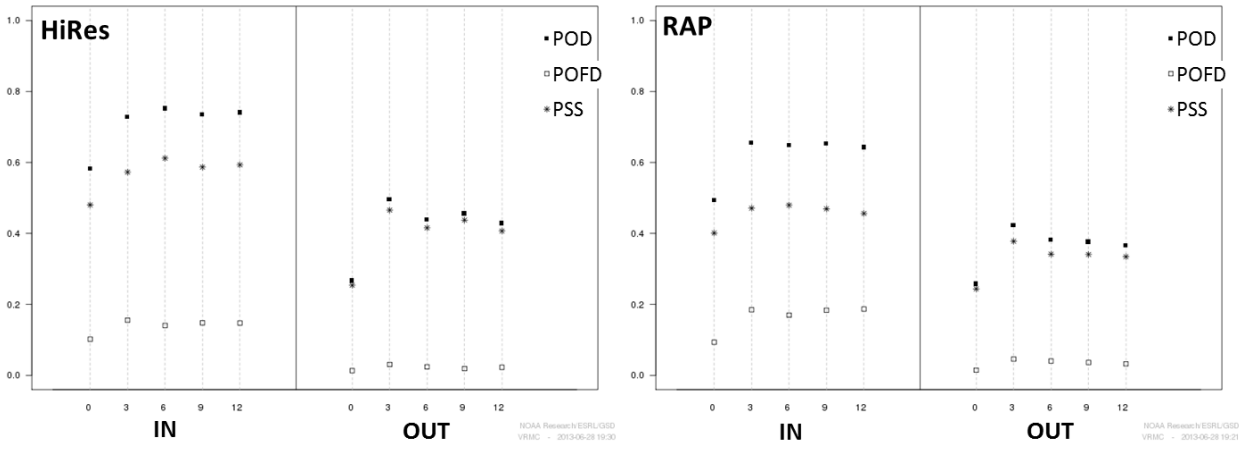

Figure 5.20: POD (filled square), POFD (hollow square), and PSS (asterisk) for the HiRes and RAP with the 5% probability mask for the regions inside and outside G-AIRMETs.

The HiRes FIP captures nearly half of all icing observations outside of the G-AIRMETs at almost no cost, i.e., the POFD values are under 0.05. (Note that the lower POD outside of the G-AIRMET is to be expected; if we assume that G-AIRMET forecasts are skillful, then the areas outside the G-AIRMETs are by definition more isolated and more difficult to forecast.) The improvement of the HiRes over the RAP is again apparent and the improvement of the FIP over the CIP is more pronounced than it is inside the G-AIRMETs.

5.5 Consistency

Two approaches to evaluating the consistency of a forecast product are used: how much does the forecast field change relative to the previous issuance valid at the same time and how much does it change relative to a separate, independent field. A truly independent field is not available for icing, so the FIP is measured against itself and against the CIP field, as described in Section 4.3.4 for SLD and for MOG severity. Both fields are converted to binary yes/no fields before measuring agreement.

For MOG severity, FIP is much more like itself than it is like CIP (Fig. 5.21), as is evident from the big jump in agreement between hours one and two in the FIP-to-FIP plots (hour one consists of the agreement between the 1-h FIP and CIP; hour two consists of the agreement between the 2-h FIP and the 1-h FIP). The spike at 3 hours in the FIP-to-CIP comparison is because CIP includes data from the most recent RAP forecast, which due to the latency of the RAP model is typically the 3-h forecast. For both FIP-to-CIP and FIP-to-FIP there is a slight reduction in agreement as the probability mask is increased. This is likely a function of the sensitivity of the agreement measure to the event base-rate and forecast bias: As the probability mask is increased the number of 'yes'

forecasts is reduced. Typically, this in turn reduces the total number of hits, the numerator in the agreement measure. The steeper drop in the higher probability masks is a result of the probability cap and its dependence on the lead hour.

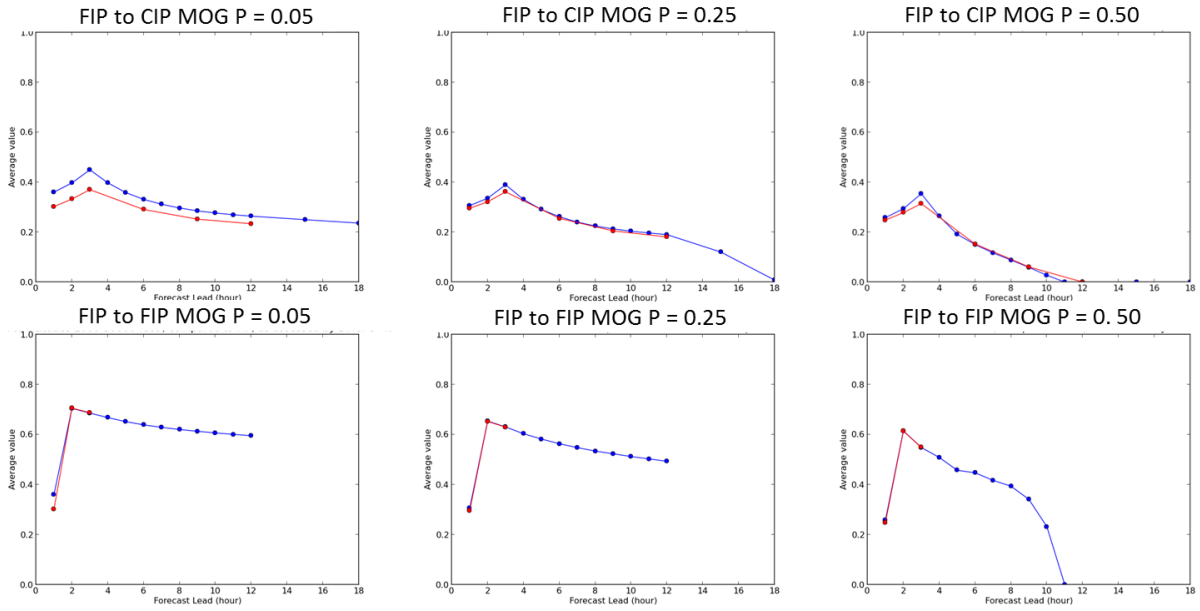

Figure 5.21: Agreement between FIP and CIP (top row) and FIP and previous FIP issuances (bottom row) for the HiRes (blue) and RAP (red) MOG severity field using the 5% (left column), 25% (middle column), and 50% (right column) probability masks.

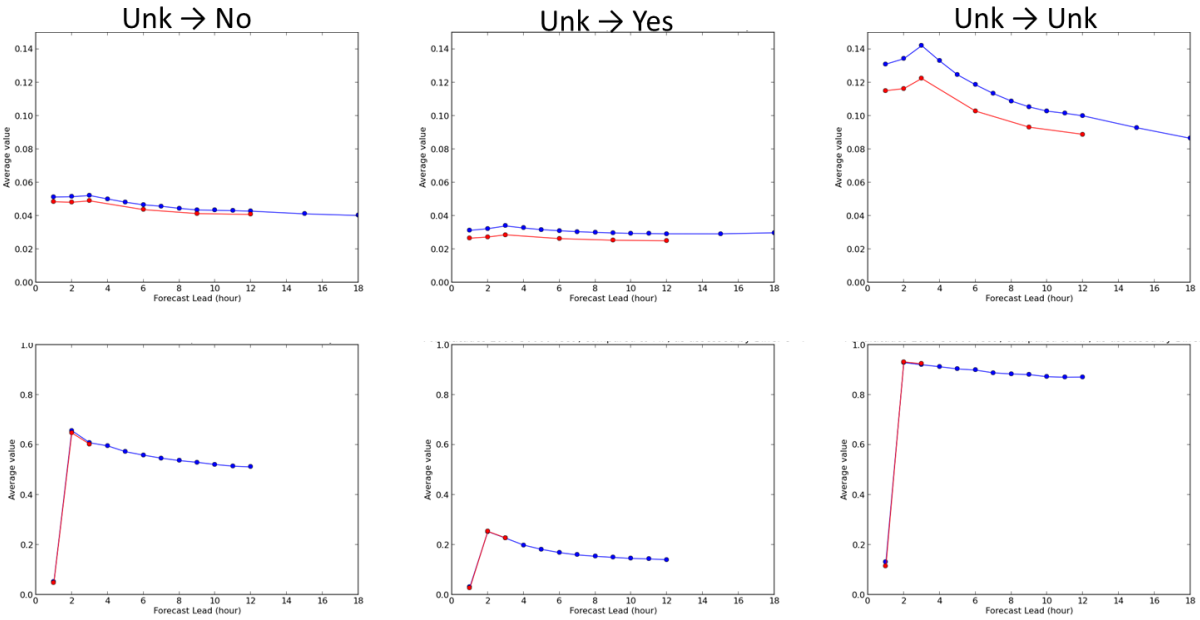

Figure 5.22: As in Fig. 5.21, but for SLD with the unknown values treated as 'no' (left column), 'yes' (middle column), and 'unknown' (right column).

For SLD, the agreement between FIP and CIP is much lower than it is for MOG severity, even when the unknown SLD forecasts are held out as unknowns (Fig. 5.22). The agreement between successive FIP forecasts is highly dependent on how the unknown forecasts are handled. Treating the unknowns as 'yes' forecasts results in very low agreement (though still higher than the FIP to CIP agreement). The agreement increases substantially when the unknowns are treated as 'no' forecasts (as is done in ADDS), while leaving the unknowns as unknowns results in very strong agreement. The exception to this is CIP: While there is sensitivity in the CIP fields to the treatment of the unknowns, the agreement remains very low even when the unknowns are left as unknowns. This is likely a result of the large difference in the extent of the number of 'unknown' forecasts in CIP compared to FIP.

6 Conclusions and Discussion

Distributions of field values were evaluated to understand general product characteristics. Findings reveal the CIP and FIP HiRes and RAP versions are very similar in behavior for all three fields. The correlation between severity and probability is apparent in both. There is a small but consistent shift in the FIP HiRes version, however, toward higher severity and probability. The characteristics of the CIP are different from those of the FIP, for both HiRes and RAP: specifically, the CIP has a strong diurnal signal in severity/probability coverage in the high layer that is not found in the FIP.

A skill comparison was performed between the CIP and FIP HiRes and RAP versions, using PIREP- and satellite-based techniques for severity, and PIREP- and METAR-based techniques for SLD. When evaluating the severity field using PIREPs, the HiRes version performs slightly better than the RAP version, but only when using a neighborhood of similar areal extent. Using a smaller neighborhood consistent with the finer resolution of the HiRes results in lower skill, indicating that the increase in information resolution of the HiRes does not match the increase in grid resolution. For SLD, performance improves when treating the 'unknown' portion of the field as unknown (i.e., excluding unknown values from skill computation), rather than as 'no' as implicitly done in the ADDs display. This is true when verifying with either PIREPs or METARs. For severity, FIP performs better than CIP, while the opposite is true for SLD. The discrepancy between CIP and FIP severity is larger when verified against satellite than it is when verified by PIREPs. This could be due to the greater coverage in the FIP icing fields than that of the CIP, and exacerbated by the holes, or empty vertical columns present in the CIP but not the FIP, which were discovered during application of the satellite-based technique.

The skill of the CIP and FIP severity field was also compared to that of the G-AIRMET, using PIREPs. Findings show that FIP achieves higher POD values, with a much smaller POFD. As the volume for FIP is comparable to that of the G-AIRMET, this is an indication that FIP is considerably more accurate in the placement of its icing forecasts. While CIP has a lower POD than G-AIRMET, it reduces the forecast volume by almost half, and remains more skillful overall. These performance results are similar for both the HiRes and RAP versions.

In considering CIP and FIP as a supplement to the G-AIRMET, the FIP HiRes version captures nearly 80% of the MOG icing inside G-AIRMETs, while excluding nearly 80% of the non-MOG icing reports.

Outside the G-AIRMETs, the FIP HiRes captures nearly half of the MOG icing reports. CIP/FIP performance inside the G-AIRMET presents a case where HiRes is clearly better than RAP. RAP is slightly worse at avoiding false alarms (~20% vs. 15% for FIP) while missing more MOG icing events.

FIP consistency was evaluated both across successive issuances of FIP valid at the same time, and relative to the CIP valid at the same time. The consistency is found to be much greater among FIPs from successive issuances than that between FIP and CIP. It can also be seen that the behavior of the additional lead hours in the HiRes is in line with the behavior of hours that are also in the RAP.

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 or position of the FAA. The authors would like to thank the In-Flight Icing Product Development Team for providing the forecast data that was needed for the evaluation.

REFERENCES

Benjamin, S.G., T.G. Smirnova, K.J. Brundage, S.S. Weygandt, T.L. Smith, B. Schwartz, D. Devenyi, J.M. Brown, and G.A. Grell, 2004: A 13-km RUC and beyond Recent developments and future plans. 11th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis, MA, October 2004, 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. Appl. Meteor.*, **44**, 969-986.

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, Sep. 11-15, American Meteorological Society (Boston), 393-398.

Carriere, J.M., S. Alquier, C. LeBot, and E. Moulin, 1997: Statistical Verification of Forecast Icing Risk Indicies, Meteorological Applications, Vol 4, Issue 2, p.115-130.

Chapman, M., M. Pocernich, A. Holmes, P. Boylan, P. Kucera, B.G. Brown, J.L. Mahoney, and J.T. Braid, 2007: Quality Assessment Report: Forecast Icing Product (FIP) – Severity. Report to the FAA Aviation Weather Technology Transfer Board. Available from the Quality Assessment Research Team (Jennifer.Mahoney@noaa.gov).

Kay, M.P., C. Lu, S. Madine, J. L. Mahoney, and P. Li, 2009: Detecting cloud icing conditions using CloudSat datasets. Preprints, 23rd Conference on Weather Analysis and Forecasting, 1-5 June, Omaha, NE, Amer. Met. Soc.

Madine, S., S. A. Lack, S. A. Early, M. Chapman, J. K. Henderson, J. E. Hart, and J. L. Mahoney, 2008: Quality Assessment Report: Forecast Icing Product (FIP).

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), 30 pp.

Murphy, M. P., 2010: Product Description Document, Graphical Airman's Meteorological Advisory (G-AIRMET). Available at: http://aviationweather.gov/static/docs/gairmet/G-AIRMETPDD_2010.pdf

NWS, 2007: Aviation Weather Services, Advisory Circular AC 00-45F. U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, and U.S. Department of Transportation, 393 pp.