Icing Algorithm Assessment:

RUC-derived and WRR-derived CIP and FIP

Prepared By:

Quality Assessment Product Development Team

NOAA/ESRL/GSD

Andrew F. Loughe[2], Joan E. Hart[2], Geary J. Layne[2],
Melissa A. Petty[3], and Jennifer L. Mahoney[1]

December 1, 2011

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:
J.L. Mahoney NOAA/ESRL/GSD, 325 Broadway, Boulder, CO 80303; Jennifer.Mahoney@noaa.gov

Table of Contents

List of Tables

List of Figures

Executive Summary

This report summarizes a formal comparison of icing analysis and forecast fields derived from the operational RUC model and the newly-developed WRF Rapid Refresh model (WRR). On behalf of the Federal Aviation Administration's Aviation Weather Research Program, the Quality Assessment Product Development Team analyzed icing algorithm output made available over a two-week period in September 2011, and the team has determined that:

Algorithm output derived from the WRR yields performance results (for MOG severity) that are similar to those of the RUC while providing for more efficient use of the available airspace.

Specifically, when comparing output from operational algorithms that use RUC data with output from enhanced versions that use WRR data, the following results have been determined–

For CIP:

- WRR production of icing probability, severity, and SLD is less than that of the RUC, especially at night and in mid-to-upper vertical levels. RUC production is much greater in eastern portions of the domain and in the northwest corner of the domain.
- For probability and severity, grid-to-grid agreement and correlation are highest in low-to-mid vertical levels. CIP SLD agreement is quite high throughout all vertical levels.
- The WRR detection rate for MOG icing events is *slightly lower* than for the RUC. The WRR detection rate for non-icing events is *slightly higher* than for the RUC.

For FIP:

- WRR production of icing probability and severity is similar to that of the RUC, but WRR production of probability is greater at the higher probability values. WRR production is also greater along mountain ranges in the west.
- WRR SLD production is ***much less*** than it is for the RUC. This is likely due to differences in the way these models handle convection in the humid southeast.
- Grid-to-grid agreement and correlation ***are poor*** for FIP SLD at upper levels. As forecast lead time increases, so do differences in the amount of SLD produced by the two models.
- The WRR detection rate for MOG icing events is *slightly higher* than for the RUC. For non-icing events, the detection rate is nearly identical, with WRR performance slightly exceeding that of the RUC.

While performing this evaluation using all available data received from the Aviation Weather Center, approximately 5% of the WRR/CIP 2-D grids were found to possess anomalies in the algorithm output. For these grids, early analysis suggests that algorithm output was created even though the input data were unavailable or incomplete.

In highlighting similarities and differences between icing fields obtained from RUC-derived and WRR-derived algorithm output, it is important for primary users of these products to understand the fundamental differences in the output, so that they can quickly and effectively adapt to icing fields that are produced using input data from the new WRR model.

1. Introduction

This report presents a comparison of icing analysis and forecast elements derived using input data from both the operational Rapid Update Cycle weather prediction system (RUC; Benjamin et al. 2004), and the newly-developed WRF Rapid Refresh model (WRR; Benjamin 2006). These two models provide input to icing algorithms that have been developed to produce fields of:

1) Icing probability.

2) Potential for the existence of Super-cooled Large Droplets (SLD).

3) Icing severity information (none, trace, light, moderate, heavy).

The Current Icing Product algorithm (CIP) provides current (analysis) information on icing conditions aloft, and the Forecast Icing Product algorithm (FIP) provides equivalent information out 1 - 12 hours into the future.

These icing algorithms were developed and are maintained by the In-Flight Icing Product Development Team (IFI PDT) at the National Center for Atmospheric Research (NCAR). The algorithms have been installed at the Aviation Weather Center, where operational versions run using input data from the RUC model, and recently enhanced versions run in evaluation mode using input data from the WRR model.

1.1 Motivation and approach

In the very near future, the operational RUC will be replaced by the WRR as the official National Weather Service model providing short-range forecast information in support of aviation. Since CIP and FIP algorithm output are valuable sources of supplemental information for in-flight icing conditions, the Quality Assessment Product Development Team (QA PDT) has been tasked with quantifying the degree of similarity and difference between RUC-derived and WRR-derived icing algorithm output, prior to operational implementation of the new WRR model. This report summarizes results by highlighting the comparative:

1) Spatial coverage and production of icing quantities.

2) Production over a range of acceptable values– a distributions approach.

3) Grid averages, correlation, agreement, and error (grid differences) for the icing fields.

4) An assessment of the skill of each model when compared with voice-recorded pilot reports (PIREPs).

In highlighting similarities and differences in icing fields obtained from CIP and FIP, it is important to communicate significant differences to the users of these data. Once the RUC data stream is discontinued, users will need to quickly and effectively adapt to icing fields that are produced using input data from the new WRR model.

2. Data

This section describes data products that were used in the assessment, including the operational RUC model, the WRR model, icing algorithms that utilize data from these models, and observations that were assembled to conduct a limited skill assessment. The data collection period for this study is 10-25 September 2011, which is a somewhat warm time of the year, not as prone to expansive icing conditions.

2.1 The Rapid Update Cycle (RUC)

The Rapid Update Cycle (RUC) is a NOAA/NCEP operational weather prediction system developed by the NOAA/ESRL Global Systems Division. The system comprises an analysis/assimilation component, along with a forecasting component that is initiated hourly. A primary advantage of the RUC model is its frequent update cycle, which is particularly beneficial to U.S. aviation interests and to forecasters of severe weather (http://ruc.noaa.gov/). Over the years, the RUC has been improved in many ways, including enhancements to model initialization, assimilation of new observation sources including radar data, and refinements to model parameterization schemes (Benjamin et al. 2004).

2.2 The WRF Rapid Refresh (WRR)

The WRF Rapid Refresh is the latest version of the 1-h rapid update cycle, and it is slated for operational deployment in 2012 (Benjamin et al. 2006). The WRR uses as its forecasting component a recent version of the Weather Research Forecast model (WRF, v3.2+), together with RUC-like physics including Thompson/NCAR micro-physics, and a Gridpoint Statistical Interpolation (GSI) assimilation system. The model performs computations on a rotated latitude/longitude Arakawa C-grid (http://rapidrefresh.noaa.gov/). The actual version of the dynamical solver used in the WRR is the Advanced Research WRF (ARW), which was primarily developed by NCAR (Skamarock et al. 2008).

The WRR is intended to be the backbone of future aviation products developed by FAA Product Development Teams (PDTs). The use of a community-developed dynamic forecast model with recognized performance advantages, along with the GSI with its benefits over oceanic regions, are two notable reasons for development of the WRR. Additionally, updates to the way the model handles microphysical quantities is intended to improve the forecast of hydro-meteor species in this new model (Wolf and McDonough 2010).

2.3 CIP and FIP icing algorithms

CIP is a multi-sensor diagnostic designed to provide guidance to pilots on the potential for encountering icing during a flight. CIP combines numerical model output from the RUC or WRR, useful data from satellites, radars, and lightning detectors, along with available surface station data (METARs) and pilot reports. Icing fields are generated hourly, are output in final form to a 20-km grid, and provide three-dimensional diagnoses of icing severity, icing probability, and the potential for encountering super-cooled large droplets (SLD, >50 micrometers). Icing severity ranges from 0-4: none, trace, light, moderate, heavy, and the range

for probability and SLD is from 0.0 to 1.0 (although the algorithms do not produce icing probability greater than 0.85).

CIP identifies locations where droplets or precipitation might accumulate as ice onto the frame of a moving aircraft. It is important to note that aircraft are normally susceptible to icing only when traveling through a layer of super-cooled liquid water, not through a layer of actual ice, which does not typically adhere to a fast moving aircraft. In an effort to determine the likelihood of icing, RUC and WRR pressure-level temperature, relative humidity, and vertical velocity data are utilized by the icing algorithms. Resultant icing fields are output at 1,000-ft vertical increments from 1,000 to 30,000 ft in altitude. Bernstein et al. (2005) have noted that the relationship between model relative humidity (RH) and icing is strongly affected by characteristics of the underlying model, and less-so by icing physics itself. A new RH interest map was developed for WRR in order to facilitate the effective transfer of these algorithms to a new modeling environment. In the new model, a nighttime correction for CIP volume calculations was implemented for the probability and severity fields. This enhancement to the algorithm provides for nighttime volume results that are now more consistent with results analyzed during the daytime.

Both the FIP and CIP provide the aforementioned quantities each hour of the day on a domain covering the Continental United States (CONUS). Consistent with its current operational use, the forecast lead-times for the FIP are 1, 2, 3, 6, 9, and 12 hours. FIP uses only model forecast output to determine icing conditions aloft, as no sensor input is obviously available for future dates and times. As the name implies, CIP provides information on *current* conditions, but in reality uses data provided from the 3-h forecast period of the most recent model run. The 0-h data are not used because moisture variables and a model's cloud microphysical properties need time to "spin-up" before they become useful.

While performing this evaluation using all available data received from the Aviation Weather Center, approximately 5% of the WRR/CIP 2-D grids were found to contain anomalies in the algorithm output. Early analysis suggests that there may have been a problem with the completeness of the input data when producing these grids. See the appendix for an example of these issues with the WRR/CIP grids.

2.4 Pilot Reports (PIREPs)

Voice-recorded pilot reports may include information on the severity of encountered icing. In the absence of specialized, on-board instrumentation, PIREPs are effectively the only source of in-flight icing observations, and are considered valuable and somewhat rare. More than 2,800 PIREPs were assembled for the skill assessment portion of this study.

It is important to note that PIREPs are subjective and inconsistently reported. Not every encounter with icing generates a report, and there are no guidelines for when to issue a report of "no icing". Consequently, non-events are under-represented in the observation data set. Since PIREPs are issued after encountering an event, temporal and spatial errors are also inherent in the data (Wandishin et al. 2011). Icing PIREP intensities are typically identified by nine unique

severity categories. Reports that indicate an intensity equal to or greater than the Light-to-Moderate category, are utilized in this study to designate observations of Moderate or Greater (MOG) icing. Icing of this intensity is likely to be impactful on the flight of an aircraft (Madine et al. 2008).

3. Analysis Methodology

This section describes techniques employed to identify similarities and differences between CIP and FIP algorithm output derived from the WRR and RUC models. Four analysis methods are explained, along with a description of how each contributes to quantifying algorithm differences. For each of these techniques, data from the study period are event equalized so that corresponding grids are available from each model.

3.1 Icing production: spatial coverage

An evaluation of the spatial coverage and the amount of icing produced by WRR and RUC versions of CIP and FIP addresses whether each version produces similar amounts of icing in the same geographic locations. Icing fields are first thresholded, using values that are consistent with those utilized by the Aviation Digital Data Service (ADDS). The threshold for probability is ≥ 0.25; SLD is ≥ 0.01; and severity is ≥ 3 (moderate or greater, MOG). Other thresholds have been investigated for the analysis phase of this study, but are not shown here. Once each field is thresholded, a tally of threshold exceedance is recorded for each grid point over the full domain. As data for the study period are processed, tallies are organized by issuance time, lead time (for FIP), and by vertical level. Tallies are then aggregated to provide an overall view of similarities and differences between algorithm output. These results are displayed in a series of color-coded horizontal maps, depicting geographic regions where icing production differs the most (WRR production *minus* RUC production is the convention used for generating the difference maps).

3.2 Icing production: distribution of values

In studying the distribution of values from each algorithm field, it is possible to determine just how similar these algorithms are at diagnosing icing within discrete bins throughout the full range of possible values. This approach does not provide information on the geographic location of icing, but rather summarizes important information on the overall characteristics and intensity of the algorithm output. Data for each field is binned to determine the number of times particular values occur. For both probability and SLD, bins range from 0 to 1 by 0.01. For severity, bins range from 0 to 4 by 1.

Data for the study are processed, with tallies incremented for a particular bin when a data value falling within that bin is encountered in the data stream. Final results are then stratified by issuance and lead time. The degree to which WRR and RUC versions of CIP and FIP indicate icing within the same bins is compared using this distributions approach. Bar graphs of these results effectively represent the probability density function (PDF) of values derived from the algorithm output of each model. By comparing ratios of the tallies in each bin (i.e., WRR count

divided by RUC count), a measure of relative over- and under- forecasting of icing production from each field and model is determined.

3.3 Grid-to-grid comparisons: averages, agreement, correlation, differences

Grid-to-grid comparisons help determine the degree of similarity and difference between any two corresponding grids, or sets of grids. Quantities that are computed for these grids include: grid averages, agreement, correlation, and error (grid differences). These measures, which are summarized in **Table 4.1**, give an overall indication of the degree of correspondence between WRR and RUC versions of CIP and FIP.

Grid average is simply the average of values over a 2-D grid. For convenience in referencing this quantity, the grid average is often referred to as the "mass", as in, there is more mass from one model compared to the mass of the other model. For a given model, if the average is taken from all of the grids over all of the valid times and all of the vertical levels, the overall mass is obtained. To compute the mass ratio between two models, the overall mass from one model is divided by the overall mass of the reference model (i.e., Mass Ratio = WRRmass / RUCmass, see **Table 4.1**).

Grid agreement measures the extent to which any two thresholded grids indicate that icing will occur in the same location, at or above the intensity level specified by the threshold. Agreement is the intersection (where *both models* indicate icing will occur) divided by the union (where *either model* indicates icing will occur). The thresholds used for computing agreement are consistent with those utilized for the icing production/spatial coverage analysis described in section 3.1 (probability ≥ 0.25; SLD ≥ 0.01; severity ≥ 3).

Correlation provides a measure of the applicability of a linear mapping between two grids.

Error (or grid difference) between output grids is determined by computing the root mean squared error (RMSE). This quantity is calculated by taking the difference between fields at each grid point, squaring the differences, and then averaging all of the squared differences. This yields the mean squared difference. Finally, the square root of mean squared differences is taken to preserve the original units of the input fields. Since the differences are squared before they are averaged, the RMSE is especially useful for identifying large errors between the output grids.

3.4 Skill assessment using PIREPs

A limited skill assessment of CIP and FIP MOG icing severity, masked by probability and verified against PIREPs, is performed in this report. Such an analysis facilitates the determination of the impact that algorithm differences have on overall skill. If differences in algorithm output are encountered, it is important to know if these differences matter; that is, do the WRR and RUC versions of the algorithms yield comparable performance results. Consistent with how icing algorithms are viewed on ADDS, the severity field is masked using probability values of 0.0, 0.25, and 0.5.

3.4.1 Matching the algorithm output to PIREPs

To perform the skill assessment, algorithm output and available PIREPs are associated with one another in both time and space. PIREPs are matched temporally to the analysis times of CIP output, and to the valid times of FIP output. This step provides a means for determining observations that contribute to the overall verification of the algorithms. PIREPs that were issued 30 minutes *after* the analysis time of the CIP are utilized in this study, as well as PIREPs that were issued from 30 minutes *prior to* and 30 minutes *after* FIP valid times. Spatially, CIP and FIP severity grids are matched to PIREPs using a 12-point matching scheme (**Figure 3.1**). If one of the 12 surrounding values from the CIP or FIP grid matches a PIREP value, given the MOG criteria, then a correct diagnosis of icing conditions is noted.

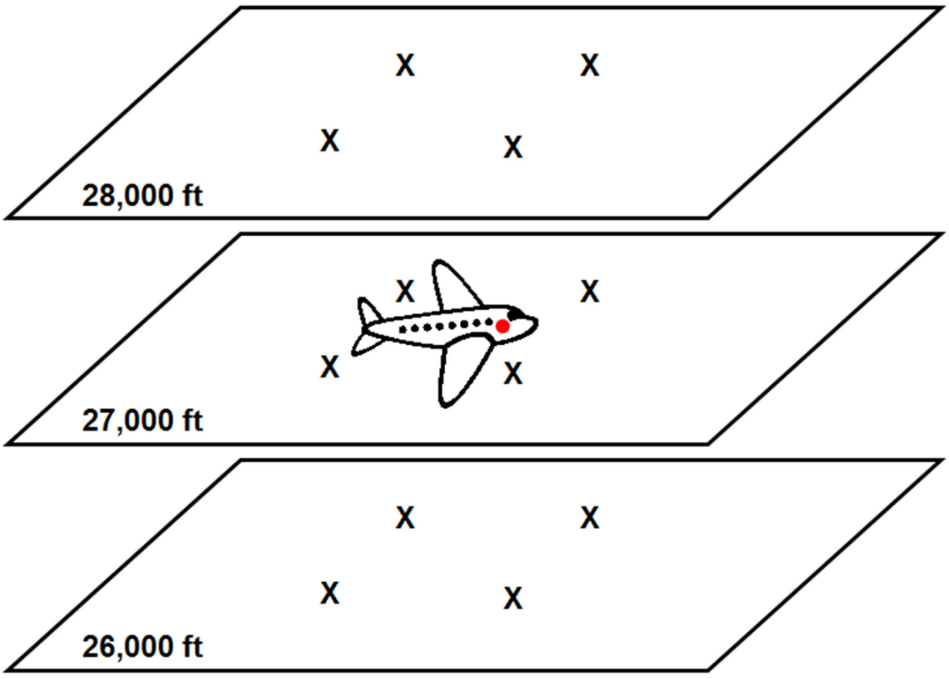

Figure 3.1: 12-point scheme for matching a pilot report (red dot) with algorithm output stored on two-dimensional grids (X = model grid points). If any of the surrounding 12 points matches the pilot report, given the MOG criteria, a "hit" is recorded.

3.4.2 Skill measures

Algorithm skill is measured by evaluating the level of agreement between algorithm output and PIREPs, in the context of MOG icing severity. Situations where both the algorithm and the PIREP denote MOG icing are classified as a *hit*. When the algorithm *does not* indicate MOG icing but the PIREP does, it is classified as a *miss*. A *false alarm* is when an algorithm indicates MOG icing, but the PIREP does not. *Correct negatives* are when both the algorithm and the PIREP report no MOG icing. These values of hits (H), misses (M), false alarms (FA), and correct negatives (CN) are accumulated to compute the probability of detecting MOG icing events (PODy), and the probability of detecting non-icing events (PODn). For both PODy and PODn, a value of 1.0 indicates a perfect detection rate (**Table 3.1**).

Finally, volume efficiency is derived by relating the detection rate to the amount of overall icing volume that is indicated by the algorithm. Volume efficiencies are relative quantities, with large values being better (greater efficiency). Larger volume efficiency means that there is less icing coverage over the airspace to achieve a given accuracy in detecting icing conditions. CIP and FIP skill assessment results are presented in the next section of this report.

Table 3.1: Dichotomous summary statistics used in this report (PODy and PODn).

Statistic	Formula	Description
PODy	H / (H + M)	Proportion of MOG icing events that were correctly detected
PODn	CN / (FA + CN)	Proportion of non-icing events that were correctly detected

4. Results

4.1 Brief summary of results

In carrying out this comparison of WRR and RUC versions of CIP and FIP, it has been determined that algorithm output derived from the WRR generally yields performance results (for MOG severity) that are similar to those of the RUC while providing for more efficient use of the available airspace.

For CIP:

- WRR production of icing is generally less than it is for the RUC, especially at night and in mid-to-upper vertical levels.
- For probability and severity, grid-to-grid agreement and correlation are highest in low-to-mid vertical levels. CIP SLD agreement is quite high throughout all vertical levels.
- For the WRR, the detection rate for MOG icing severity (PODy) is *slightly lower* than it is for the RUC, but the WRR detection rate for non-icing events (PODn) is *slightly higher* than it is for the RUC.

For FIP:

- WRR production of probability and severity is similar to that of the RUC, but WRR production for the probability field is greater than for RUC at higher probability values.
- WRR SLD production is ***much less*** than it is for the RUC.
- Grid-to-grid agreement and correlation for FIP SLD is very poor at upper levels.
- The WRR detection rate for MOG icing events is *slightly higher* than for the RUC. For non-icing events, the detection rate is nearly identical, with WRR performance slightly exceeding that of the RUC.

In pursuing the analysis approaches that are outlined in the methodology section of the report, subsequent sections highlight specific findings. The presentation of results is consistent with the goal of providing a series of side-by-side comparisons of WRR-derived and RUC-derived icing algorithm output.

4.2 Icing production: spatial coverage

This section focuses attention on the production of icing over the full domain, addressing both the location of icing and its rate of production above a stated threshold. Tallies of icing conditions at individual grid points are aggregated over all valid times and all vertical levels. Difference maps (WRR tallies *minus* RUC tallies) depict locations where RUC production is greatest over the entire time period (in shades of blue) and where WRR production is greatest (in shades of yellow-to-red).

4.2.1 CIP results

Probability and severity – Figure 4.1a,b

- There is less production of icing probability and severity from the WRR model, most notably in the east and in the far northwest (regions with blue shading).
- There is greater production of icing probability and severity from the WRR model along western mountain ranges (Coastal Mountains of Canada, Rocky Mountains, and the Sierra Madres in Mexico).
- Differences in probability and severity are more spatially correlated than with SLD.

SLD – Figure 4.1c

- Differences in SLD are not as great. It is also worth noting that SLD is somewhat rare.
- Differences are more randomly distributed throughout the entire domain, and edge effects are not as great as they are for probability and severity.

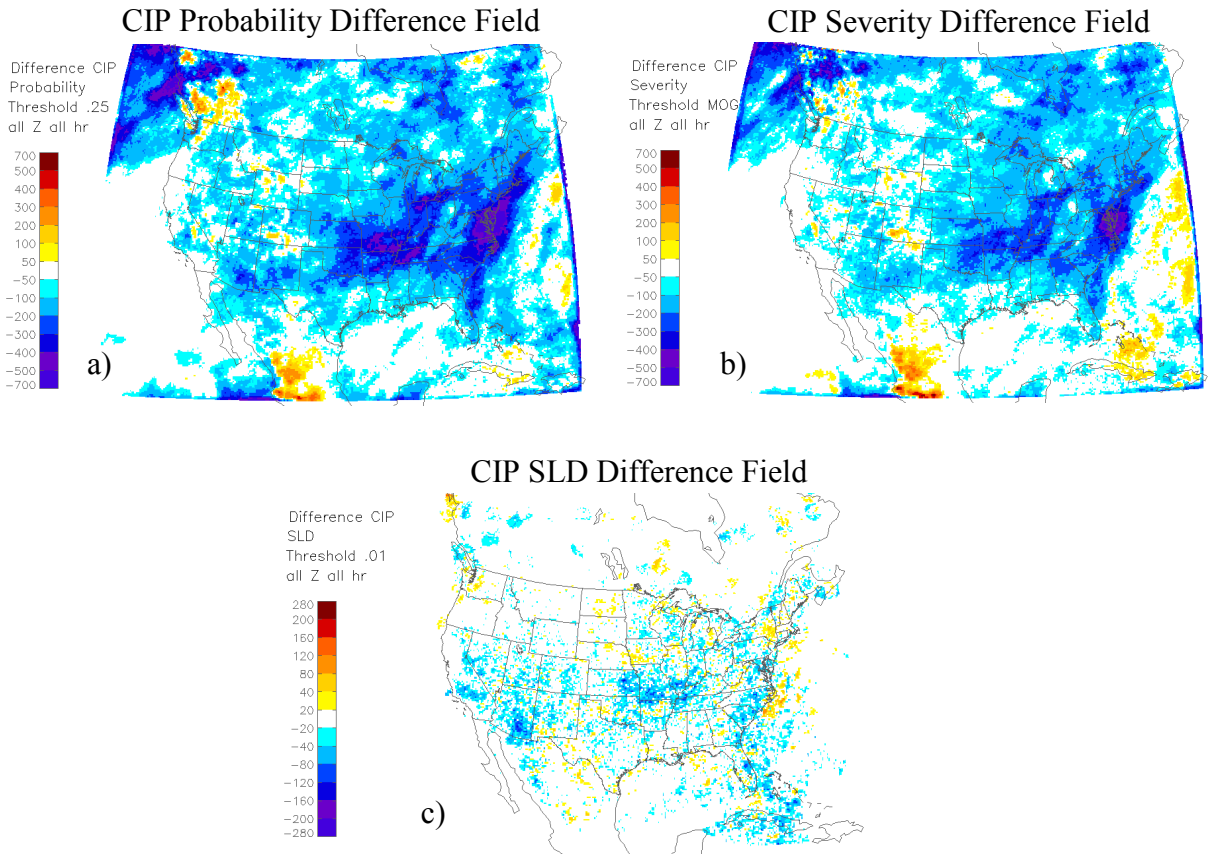

Figure 4.1: Difference in tallies for CIP icing conditions at each grid point over the full time period and over all vertical levels for (a) probability, (b) severity, and (c) SLD. The convention for computing these difference fields is WRR (tallies) – RUC (tallies), so blue shaded areas depict where RUC production is greatest, and yellow-to-red shaded areas depict where WRR production is greatest.

4.2.2 FIP results

Probability and severity – Figures 4.2a,b

- There is less production of icing probability and severity from the WRR throughout areas of the south, especially in the humid southeast (regions with blue shading).
- There is greater production of icing probability and severity from the WRR model (regions with yellow-to-red shading) throughout Canada and along western mountain ranges (Coastal Mountains of Canada, Rocky Mountains, Sierra Nevada, and the Sierra Madres in Mexico).

SLD – Figure 4.2c

- Differences in SLD are very large in the southeast: WRR production is much less than it is for the RUC model, possibly due to dramatic differences in how the models handle convection in the humid southeast.
- WRR production of SLD is greater in some areas of Canada.

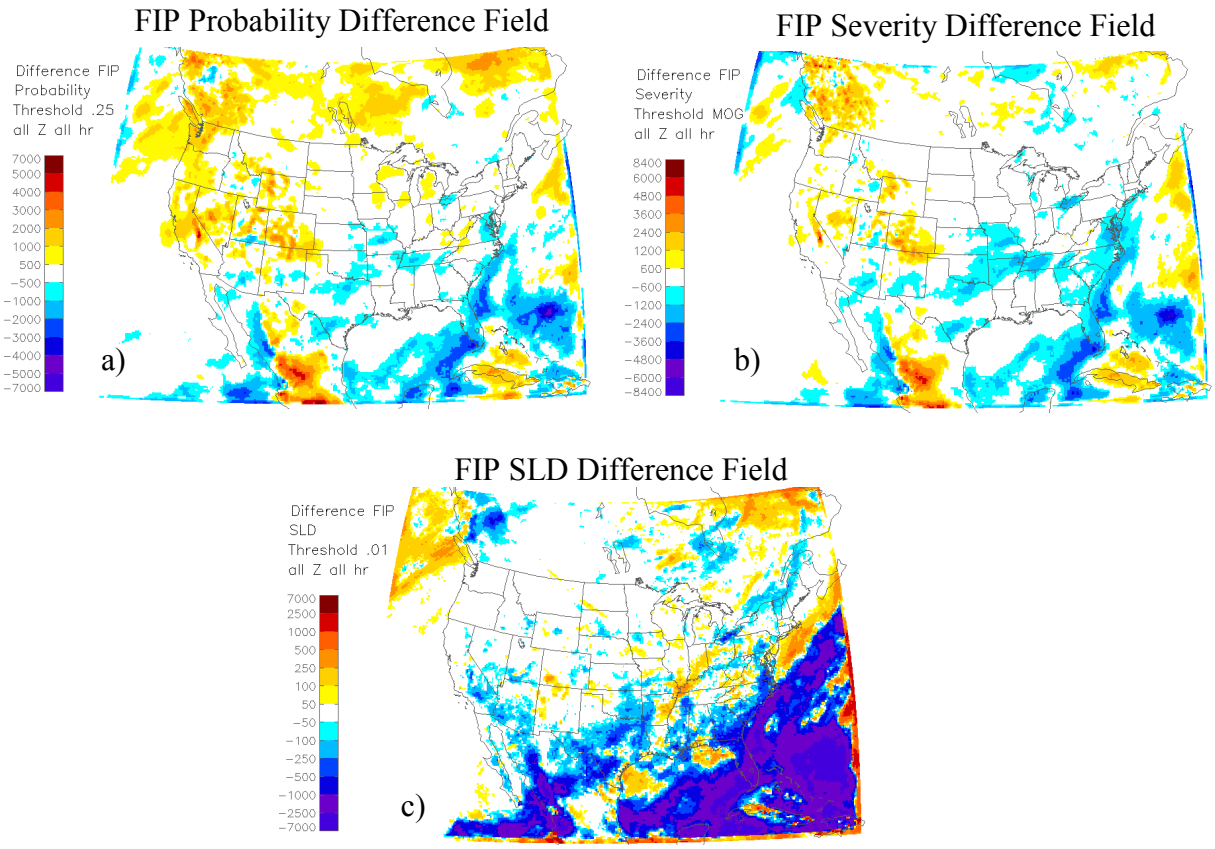

Figure 4.2: Differences in tallies for FIP icing conditions at each grid point over the full time period and over all vertical levels for (a) probability, (b) severity, and (c) SLD. The convention for computing these difference fields is WRR (tallies) – RUC (tallies), so blue shaded areas depict where RUC production is greatest, and yellow-to-red shaded areas depict where WRR production is greatest.

4.3 Icing production: distribution of values

This section focuses attention on the production of icing over the range of possible values for each field. Tallies of icing conditions within discrete bins were aggregated over all valid times and all vertical levels. Ratio plots depict the relative over- and under- production by each model (WRR tallies *divided by* RUC tallies). This approach summarizes important information on the characteristics and intensity of the algorithm output.

4.3.1 Probability field

- For CIP, WRR production of icing probability is noticeably less than it is for the RUC (ratio of tallies < 1.0 for nearly all probability bins, see **Figure 4.3a**). Differences are even larger for higher probability values (ratio decreases).

- For FIP, WRR production is very similar to that of the RUC (ratio of tallies near 1.0, see **Figure 4.3b**). For probability values above 0.56, WRR production is greater than it is for the RUC (ratio of tallies > 1.0). This field's maximum value is actually 0.85.

- For FIP probability ≥ 0.50, WRR production is greater for lead times 1,2,3,6 hours, reversing at forecast hour = 9 (**Figure 4.3c**). The absolute percent difference in production between these two models increases at forecast hour 9 (**Figure 4.3d**).

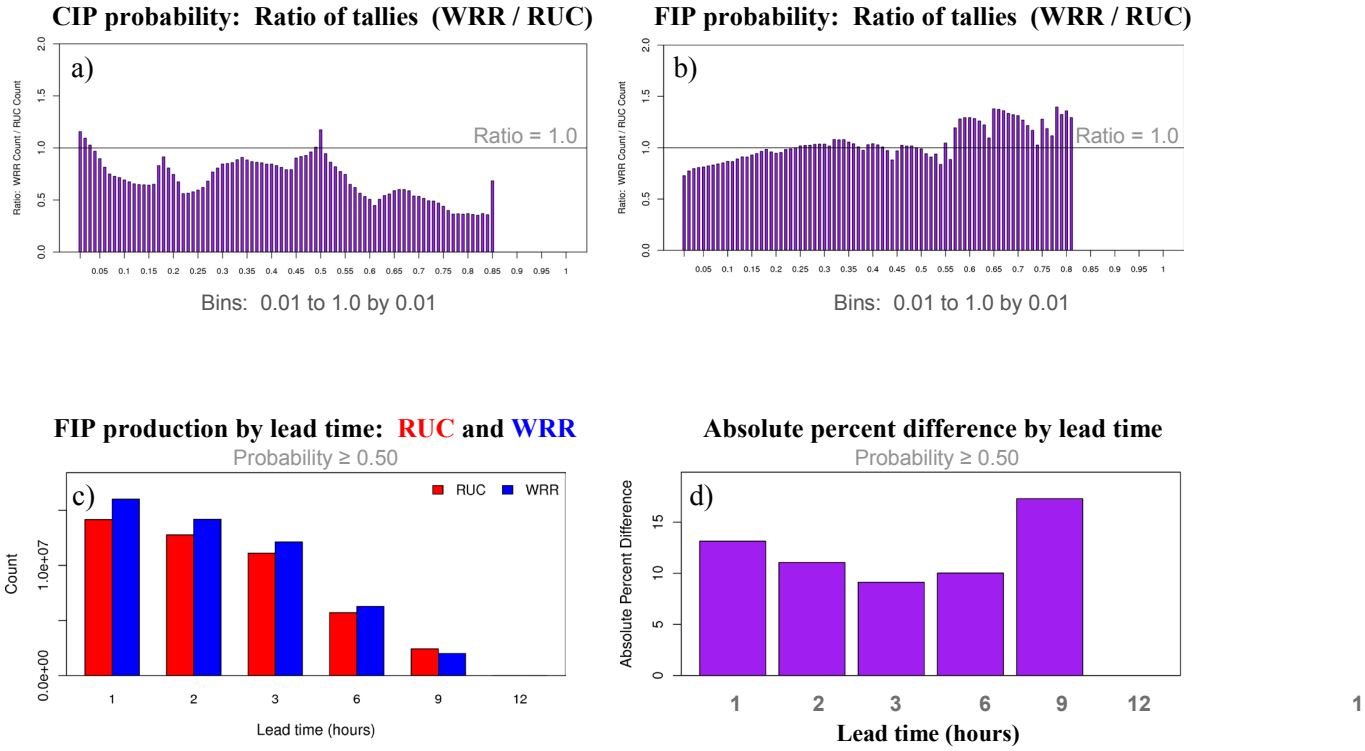

Figure 4.3: The top figures present the ratio of tallies for icing production (WRR tallies *divided by* RUC tallies, thin purple bars) for the CIP algorithm and the FIP algorithm (**Figures 4.3a,b**). Bins are 0.01 in width and span the range of possible values. In the lower figures, FIP production tallies from the RUC and WRR models are presented by lead time (red & blue bars), and the absolute percent differences in production by lead time are shown (thick purple bars) in **Figures 4.3c,d**.

4.3.2 SLD field

- For CIP, WRR production of SLD is nearly the same as it is for the RUC throughout the entire range of values (ratio of tallies is very nearly 1.0 in all bins, **Figure 4.4a**).
- For FIP, WRR production of SLD is ***much less*** than it is for the RUC throughout the entire range of possible values (**Figure 4.4b**). There is one minor anomaly for bin=0.67, where the total production from each model is abnormally large, and where the WRR production of SLD exceeds that of the RUC by a large margin.
- For FIP SLD ≥ 0.01, WRR production is ***much less*** for all lead times (**Figure 4.4c**), the absolute percent differences are very large, and these differences increase with increasing lead time (**Figure 4.4d**). Note by comparison with other fields the large magnitude of the y-axis in the last figure (maximum of 70%).

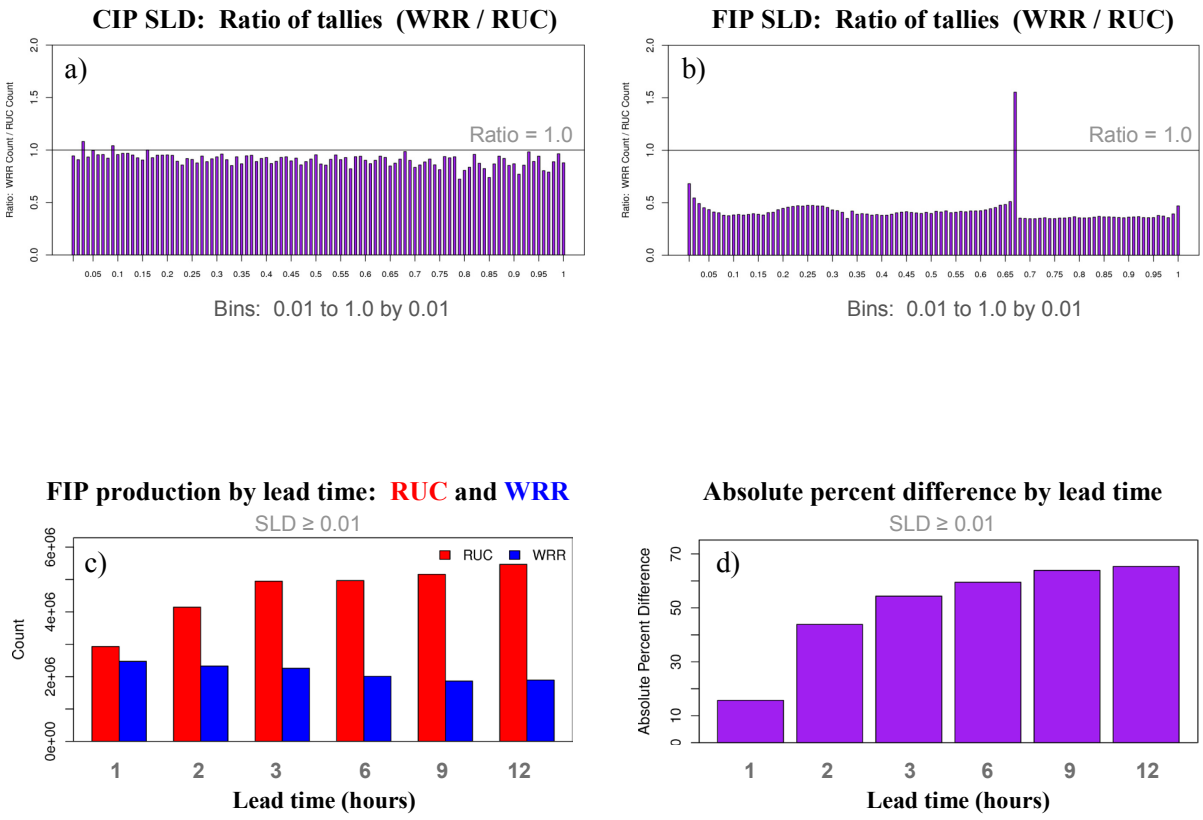

Figure 4.4: The top figures present the ratio of tallies for icing production (WRR tallies *divided by* RUC tallies, thin purple bars) for the CIP algorithm and the FIP algorithm (**Figures 4.4a,b**). Bins are 0.01 in width and span the range 0.01 to 1.0. In the lower figures, FIP production tallies from the RUC and WRR models are presented by lead time (red & blue bars), and the absolute percent differences in production by lead time are shown (thick purple bars) in **Figures 4.4c,d**.

4.3.3 Severity field

- For CIP, WRR production of icing severity is similar to that of the RUC across most severity values (**Figure 4.5a**). Relative production from WRR does decrease with increasing severity, as RUC production is much greater at SEV=4 (ratio of 0.42).
- For FIP, WRR production of icing severity is nearly identical to that of the RUC across all severity values (SEV=1 to 4, see **Figure 4.5b**). Relative production from WRR does decrease with increasing severity, as RUC production is greater at SEV=4 (ratio of 0.77). The relative decrease for WRR/FIP is not as large as for WRR/CIP.
- For FIP, WRR production of MOG severity is very similar to that of the RUC at all lead times (**Figure 4.5c**). Differences do increase with increasing lead time, but the overall differences are still relatively small (**Figure 4.5d**). Note by comparison with other fields the small magnitude of the y-axis in the last figure (maximum of 8%).

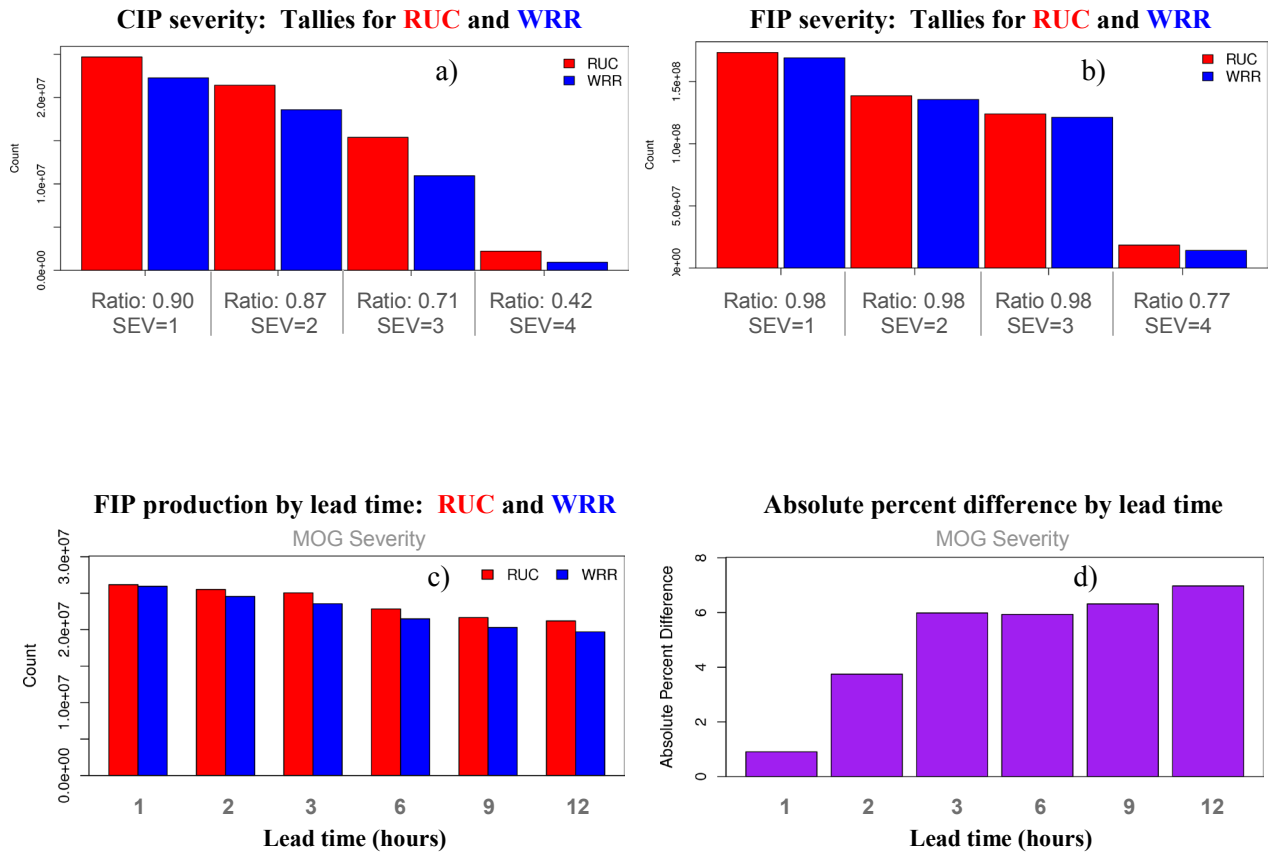

Figure 4.5: Tallies for icing production (RUC tallies and WRR tallies) by severity value (1 to 4). Note for all of the results, that WRR production is less than that of the RUC (**Figures 4.5a,b**), and that this signature increases with increasing severity value, especially for the CIP algorithm. In the lower figures, FIP production tallies from the RUC and WRR models are presented by lead time (red & blue bars), and the absolute percent differences in production by lead time are shown (thick purple bars) in **Figures 4.5c,d**.

4.3.4 CIP results by issuance hour

This sub-section presents results stratified by CIP issuance hour, to measure the degree of similarity and difference in icing production between RUC and WRR for daytime vs nighttime.

- Consistent with previous findings for CIP, WRR production of icing is less than it is for the RUC model for each field, and throughout all issuance times of the day (**Figure 4.6a-f**).
- The largest absolute percent differences in production are noted for the CIP probability and severity fields (**Figures 4.6b,d)**. For SLD, differences are large only during a few nighttime hours (**Figure 4.6f)**. Take note of the y-axis range of values.
- It is of interest to note that CIP differences are generally larger at night, consistent with the understanding that changes have been applied to the WRR/CIP probability and severity volume calculations, as highlighted in section 2.3 of this report.

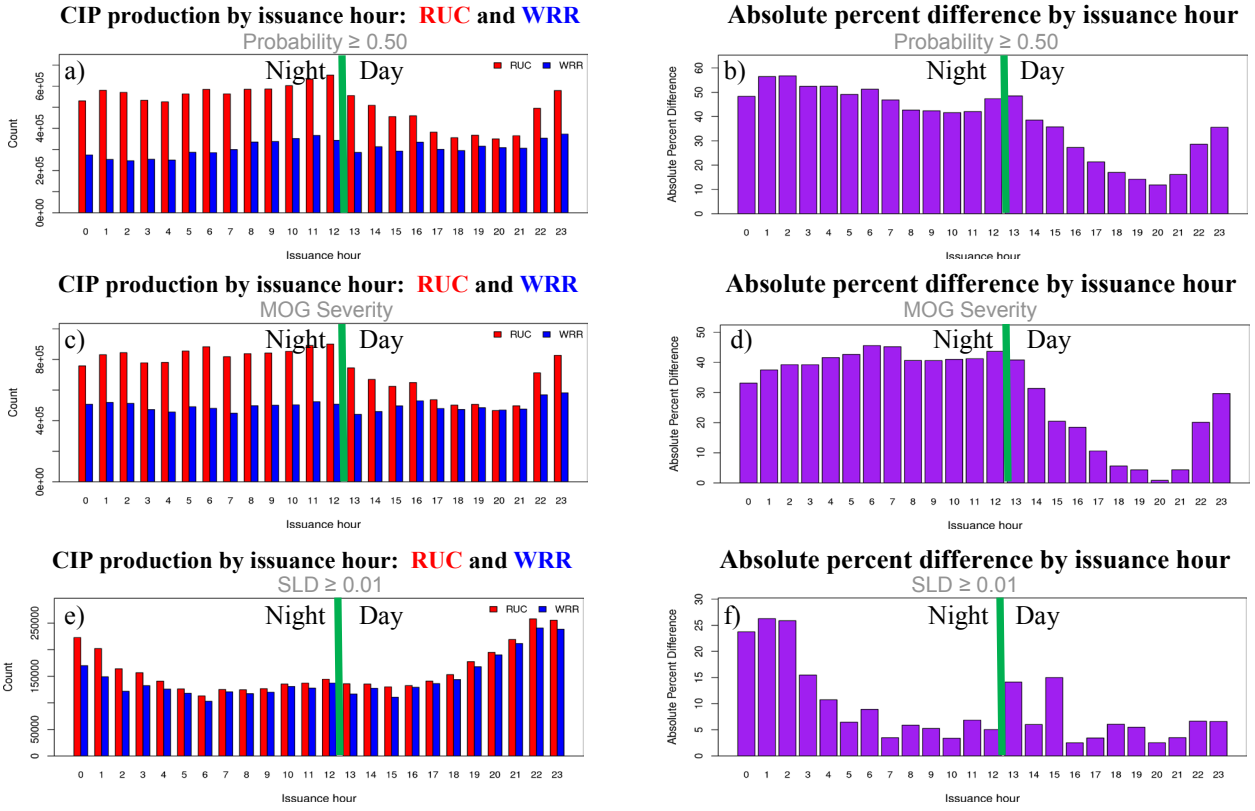

Figure 4.6: Production of icing from the CIP algorithm for the RUC and WRR models, along with the absolute percent difference between output from the two models (in purple). This analysis is presented by issuance hour, thus facilitating the understanding of just how much the products differ with time-of-day. Nighttime hours are roughly indicated by those issuance hours left of the thick, green line, and daytime is roughly indicated to the right of the green line. For CIP, satellite reflectance data is an active component of the algorithm during the daytime.

4.3.5 FIP results by issuance hour

This sub-section presents results stratified by FIP issuance hour, to measure the degree of similarity and difference in icing production between RUC and WRR for daytime vs nighttime.

- For FIP, WRR production of icing is similar to that of the RUC for the probability and severity fields (**Figures 4.7a-d**). Where there are differences, they reflect greater production from the WRR model for probability, and reversed for severity.
- For the SLD field, WRR production is ***much less*** than from the RUC model. These differences are substantial, and do not vary much with time-of-day (**Figures 4.7e,f**).
- It is important to note that there is no strong, consistent diurnal signal in the difference fields, likely because FIP forecasts do not use any satellite data. The overall counts at hours 10 and 16 UTC are lower due to some problems with data transmission. Take note of the y-axis range of values, especially for SLD.

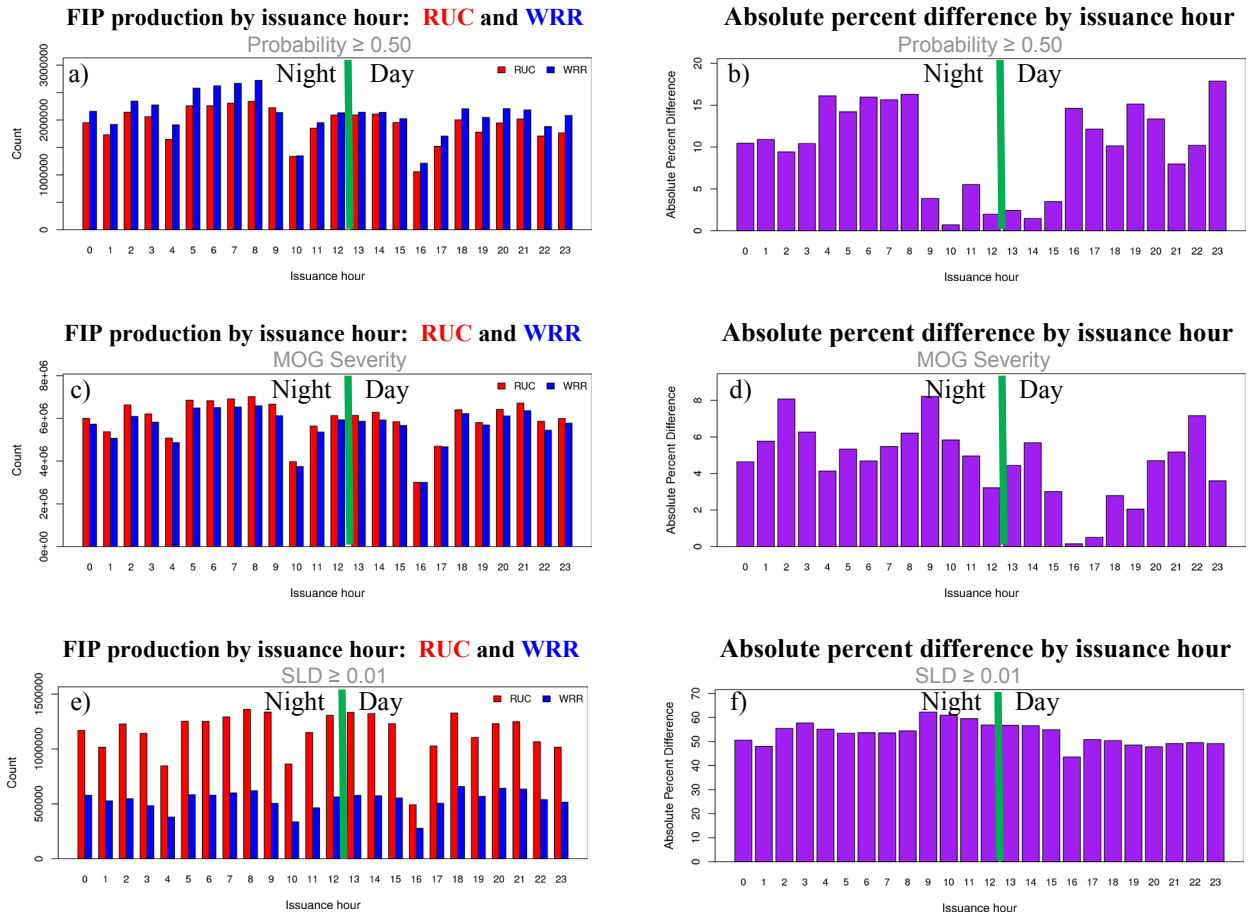

Figure 4.7: Production of icing from the FIP algorithm for the RUC and WRR models, along with the absolute percent difference between output from the two models (in purple). This analysis is presented by issuance hour, thus facilitating the understanding of just how much the products differ with time-of-day. Nighttime hours are roughly indicated by those issuance hours left of the thick, green line, and daytime is roughly indicated to the right of the green line. For FIP, which is a forecast product, there is no daytime/nighttime issue with satellite data, since it is not used. Results do not vary substantially with longer or shorter lead times included (or removed) from the processing.

4.4 Grid-to-grid comparisons: averages, agreement, correlation, differences

This section presents results that are designed to determine the degree of similarity and difference between two corresponding grids, or sets of grids. The measures to be presented: averages, agreement, correlation, and error (**Table 4.1**), give an indication of the amount of correspondence between WRR and RUC versions of CIP and FIP gridded output.

Table 4.1: Table describing quantities that are computed when performing a grid-to-grid assessment of the similarities and differences between WRR and RUC versions of CIP and FIP icing algorithm output.

Quantity	Description
Mass	Field average over a 2-D grid.
Mass ratio	WRR_{mass} / RUC_{mass} A comparative measure of mass. The mass from the operational RUC-derived field is used as the reference field (denominator). A value of 1.0 indicates an identical amount of mass from the two models. The further the ratio is from unity (above or below), the more dissimilar are the grids.
Agreement	Ratio of the intersection divided by the union of two thresholded fields. A value of 1.0 indicates perfect overlap of the thresholded fields.  **Intersection: Purple** **Union: Red and Blue**
Correlation	A measure of the applicability of a linear mapping between two fields. A value of 1.0 indicates that a perfect linear mapping exists between two fields.
RMSE **(root mean squared error)**	A measure of the "average" error (difference) between two fields. A value of 0.0 indicates no difference between the fields.

Results from each computed measure for the grid-to-grid assessment are presented in a series of graphs that provide multiple pieces of information in a single figure. Results are stratified by vertical level in the atmosphere, and where appropriate, by FIP forecast lead time (along the x-axis). An example of these graphs is presented and explained in **Figure 4.8**.

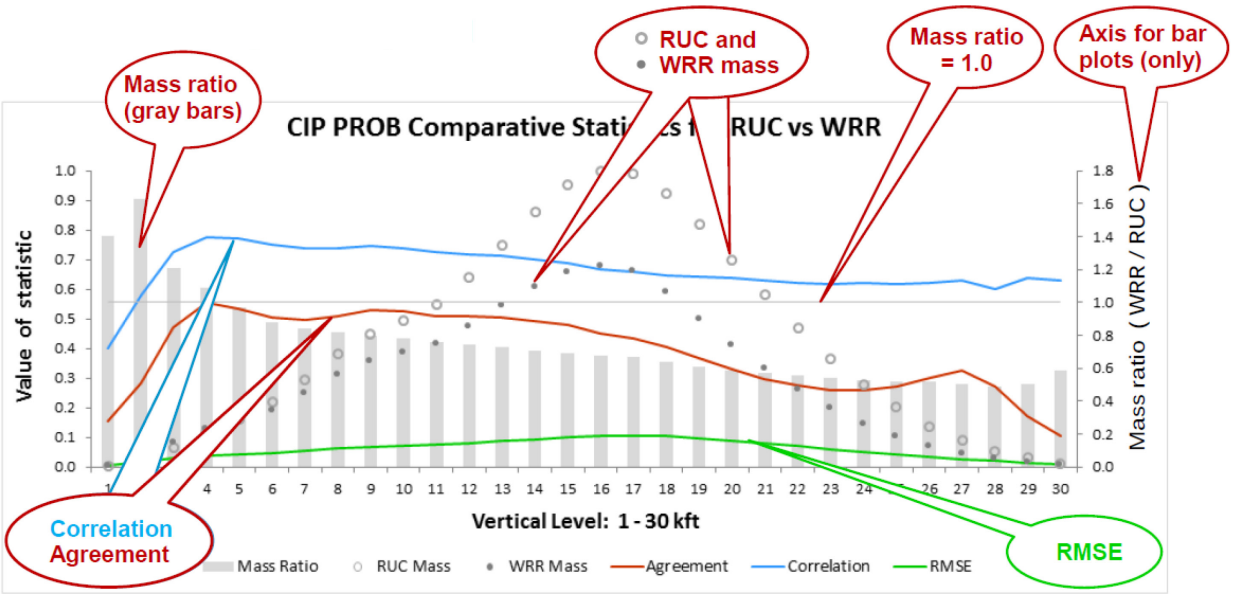

Figure 4.8: Sample figure depicting grid-to-grid results stratified by vertical level. Gray features in this figure are associated with mass relationships: individually for each model by the circles and dots, and in a ratio context from the gray bars (y-axis on the right is for the mass ratios). The colorful lines depict measures of grid correspondence: agreement, correlation, and RMSE (y-axis to the left).

Important elements of this graph can be divided into two parts, (1) mass measurements are associated with the gray-scale features, and (2) measures of grid correspondence are associated with the brightly colored features (red, blue, green):

1) The open circles represent mass measures from the RUC model, and the gray dots represent mass measures from the WRR model. If open circles are higher on the graph than corresponding gray dots, then $RUC_{mass} > WRR_{mass}$.

2) The vertical bars are associated with the y-axis on the right, and represent the ratio of mass from the two models (WRR_{mass} *divided by* RUC_{mass}). If there is less mass from the WRR model, the bars will lie below the horizontal gray line that represents a ratio value of 1.0 (equal mass from each model at that level).

3) The colorful lines are associated with measures of agreement, correlation, and RMSE. Higher values of agreement and correlation indicate a higher degree of correspondence between the compared grids, and for RMSE, a lower value indicates smaller differences between the grids.

4.4.1 Probability field

- CIP agreement and correlation between the WRR and RUC models are highest in the low-to-mid levels. These measures decrease gradually aloft (**Figure 4.9a**).
- CIP mass for each model is greatest in the mid-levels where icing is most prevalent. At most vertical levels, $RUC_{mass} > WRR_{mass}$. Mass exceedance peaks in mid-levels where the RMSE is also largest (**Figure 4.9a**).
- FIP agreement and correlation are highest in the low (but not lowest) levels, and then these measures decreases steadily aloft (**Figure 4.9b**). Disagreement is high in the lowest levels where there is low mass, but a very high mass ratio.
- FIP mass from each model is very similar ($RUC_{mass} \approx WRR_{mass}$), with mass ratios dipping below 1.0 in the mid-levels (a switch in relative mass between the models). The RMSE is largest where the mass measurements are greatest (**Figure 4.9b**).

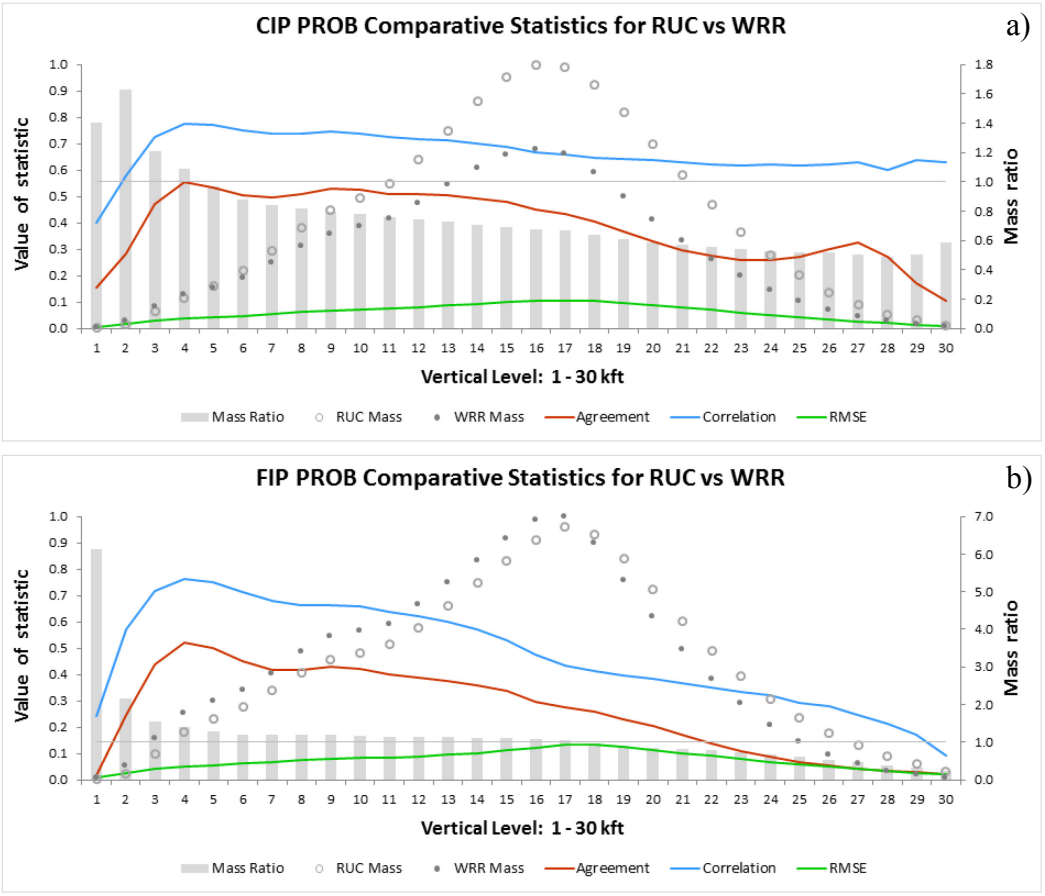

Figure 4.9: Grid-to-grid results for CIP probability (top) and FIP probability (bottom) over a range of vertical levels (1,000 to 30,000 ft). Gray features are associated with mass relationships: individually for each model by the circles and dots, and in a ratio context from the gray bars (y-axis on the right is for the mass ratios). The colorful lines depict measures of grid correspondence: agreement, correlation, and RMSE (y-axis to the left).

4.4.2 SLD Field

- For CIP, agreement and correlation are very high, and $RUC_{mass} \approx WRR_{mass}$ throughout all vertical levels. The RMSE is also very small at all vertical levels (**Figure 4.10a**).

- For FIP, agreement and correlation are high in lower levels where there is little SLD, (this is not true at the very lowest level where the mass ratio is also quite high). Agreement and correlation are ***extremely low*** in the upper-levels, where it can be seen that $RUC_{mass} >> WRR_{mass}$ (**Figure 4.10b**). Differences aloft are dramatic.

- Differences between RUC_{mass} and WRR_{mass} grow steadily with increasing forecast lead time for the FIP (**Figure 4.10c**).

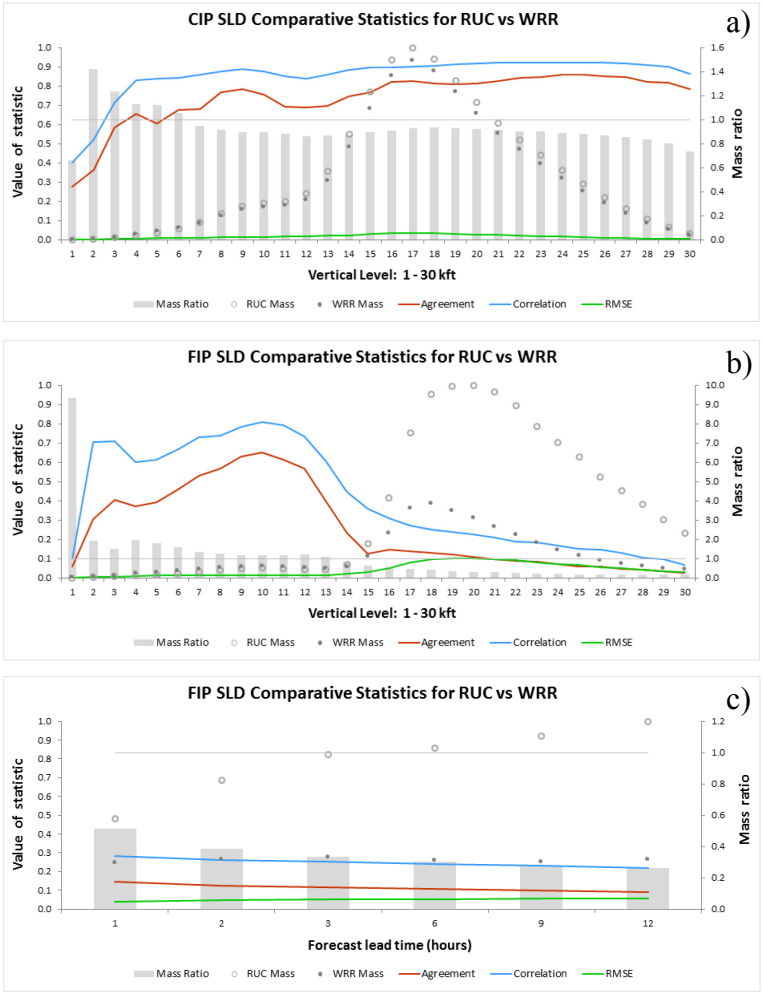

Figure 4.10: Grid-to-grid results for CIP SLD (top) and FIP SLD (middle) over a range of vertical levels (1,000 to 30,000 ft), and by FIP forecast lead time (bottom). Gray features are associated with mass relationships: individually for each model by the circles and dots, and in a ratio context from the gray bars (y-axis on the right is for mass ratios). Colorful lines depict measures of correspondence: agreement, correlation, and RMSE (y-axis to the left).

4.4.3 Severity field

- CIP agreement and correlation are highest in the low-to-mid levels. These measures increase sharply aloft, where overall mass is relatively low (**Figure 4.11a**).

- CIP mass for each model is greatest in the mid-levels where icing is most prevalent. At most vertical levels, $RUC_{mass} > WRR_{mass}$. Mass exceedance peaks in mid-levels where the RMSE is also largest (**Figure 4.11a**).

- FIP mass from each model is very similar ($RUC_{mass} \approx WRR_{mass}$), with mass ratios dipping below 1.0 in the mid-levels (a switch in relative mass between the models). The RMSE is largest where the mass measurements are greatest (**Figure 4.11b**). RMSE appears high because the raw severity values are scaled higher, from 1 to 4.

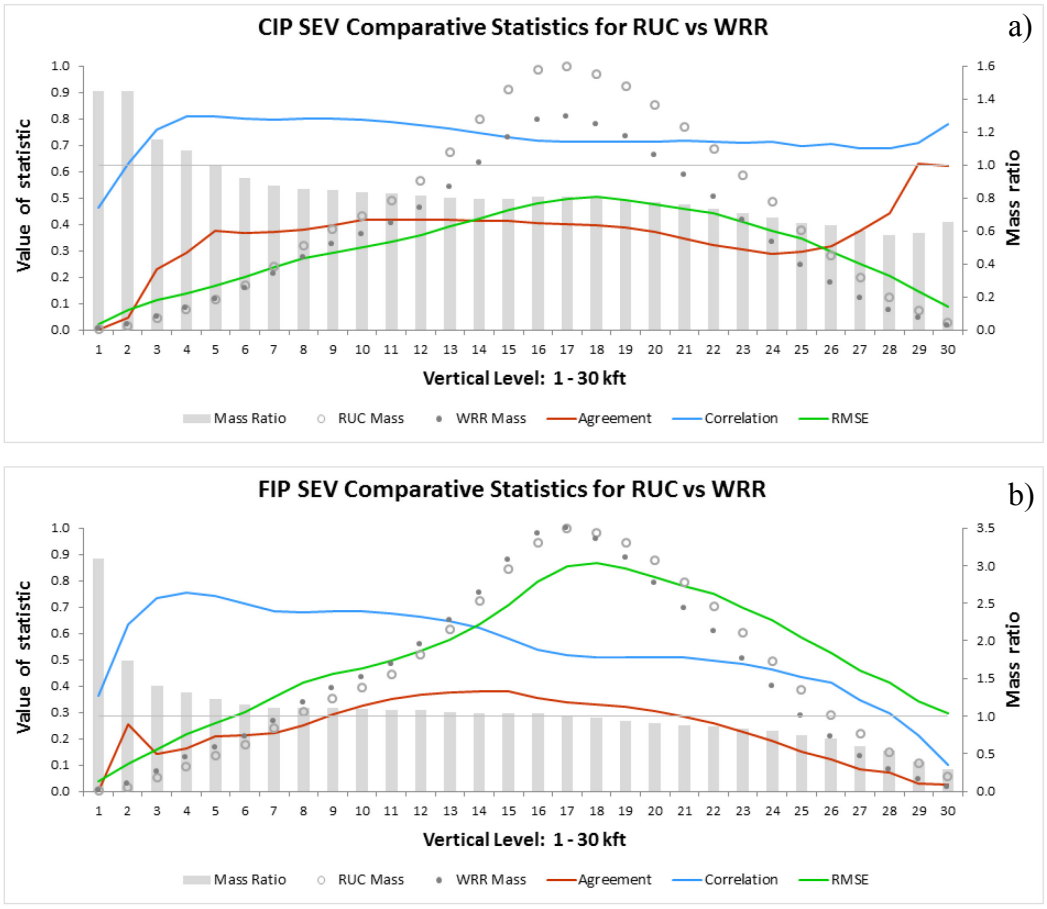

Figure 4.11: Grid-to-grid results for CIP severity (top) and FIP severity (bottom) over a range of vertical levels (1,000 to 30,000 ft). Gray features are associated with mass relationships, individually for each model by the circles and dots, and in a ratio context from the gray bars (y-axis on the right is for the mass ratios). The colorful lines depict measures of grid correspondence: agreement, correlation, and RMSE (y-axis to the left). For severity, RMSE values appear high because raw values for this field are scaled higher, from 1 to 4, not from 0 to 1.

4.5 Skill assessment

Now that differences in algorithm output from the WRR and RUC have been identified, this section highlights how much these differences matter, in the context of the relative skill in detecting MOG icing severity events and non-icing events. Severity output is compared with reports of MOG icing from PIREPs, and detection rates are then computed and compared (see **Table 3.1**). In these results it is always important to consider the associated confidence intervals.

4.5.1 FIP and CIP detection rates

- For detection of MOG icing severity events (PODy), the WRR/FIP performance is *slightly better* than it is for RUC/FIP (**Figure 4.12a**, left). For WRR/CIP, performance is *slightly lower* than it is for RUC/CIP (**Figure 4.12a**, right).
- For the detection of non-icing severity events (PODn), results are comparable for FIP and CIP, with WRR performance measured to be *slightly better* than that of the RUC (**Figure 4.12b**). All results are presented by probability mask (0.0, 0.25, 0.50).

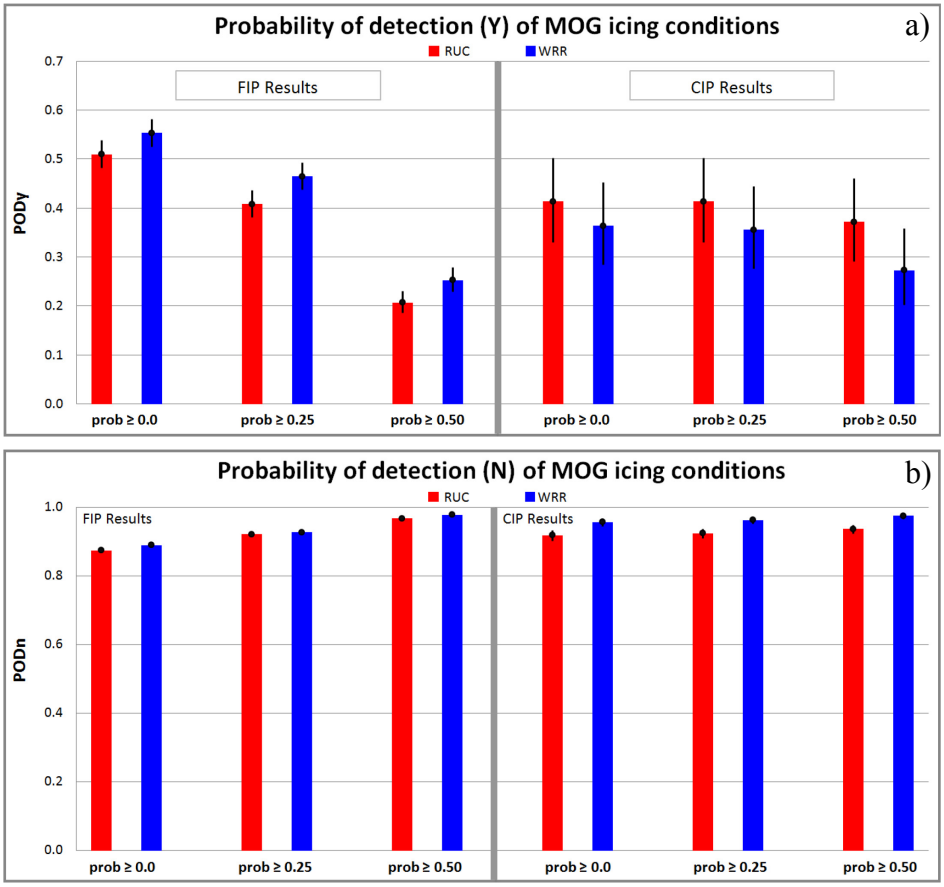

Figure 4.12: Probability of detection for MOG icing severity events (PODy, top), and the probability of detection for non-icing events (PODn, bottom), for FIP (left) and CIP (right). Results are by probability mask (0.0, 0.25, 0.50). Note the confidence intervals located in the middle of each bar by a dot and a black vertical line. Confidence intervals are tighter for the FIP because there are more data values included in the analysis (six forecast lead times). Confidence intervals are also tighter for non-icing events because they exceed the number of MOG icing events.

4.5.2 FIP detection rates by forecast lead time

- For FIP detection of MOG icing severity events (PODy stratified by probability mask and forecast lead time), WRR performance is *slightly better* than that of the RUC (**Figure 4.13a**).
- For the detection of non-icing severity events (PODn stratified by probability mask and forecast lead time), results are comparable for FIP and CIP, with WRR performance measured to be *slightly better* than that of the RUC (**Figure 4.13b**).

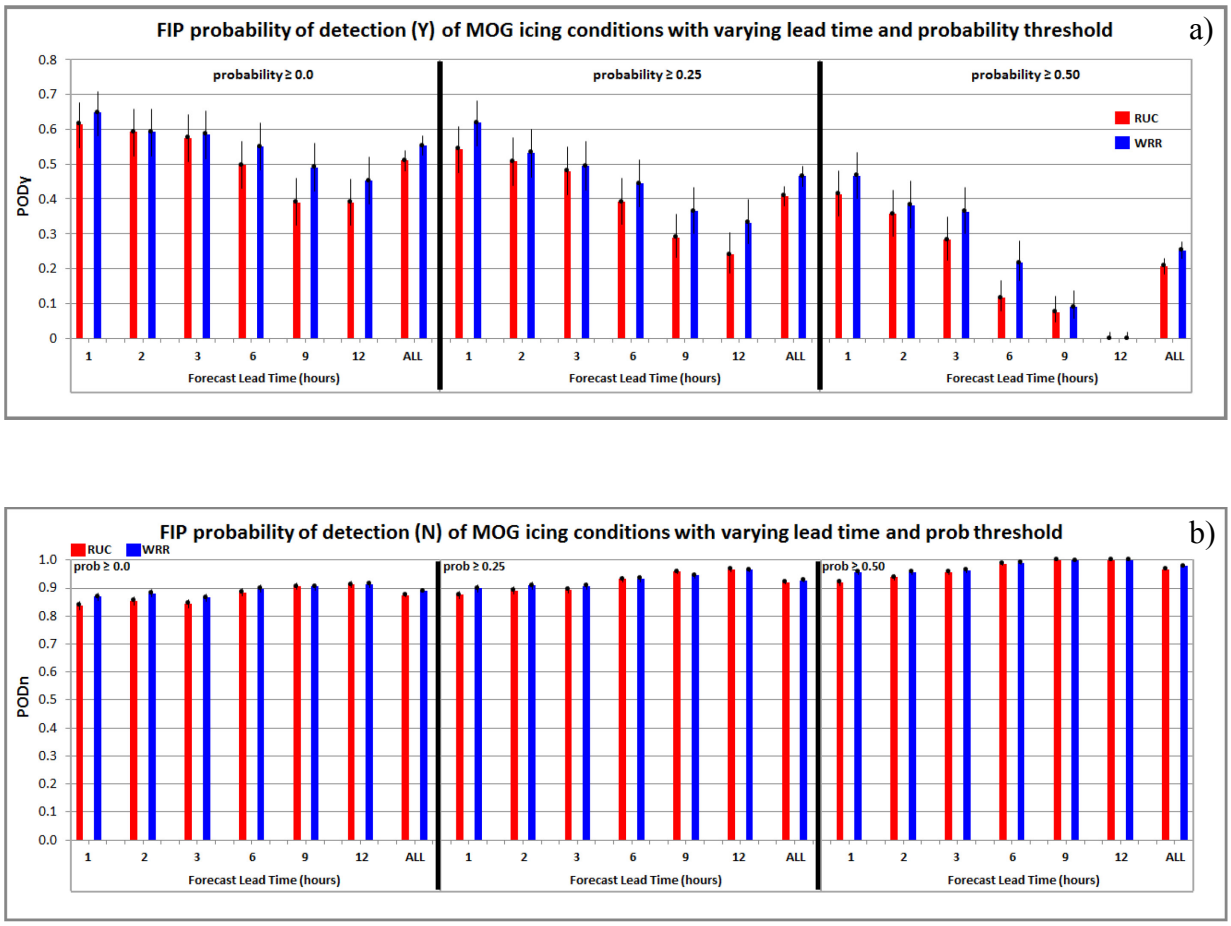

Figure 4.13: Probability of detection for MOG icing severity events (PODy, top), and the probability of detection for non-icing events (PODn, bottom). These FIP results are divided into three sections by probability mask (0.0, 0.25, 0.50). Each section then contains results by lead time (1,2,3,6,9,12, and all hours combined). Note the confidence intervals located in the middle of each bar by a dot and a black vertical line. Confidence intervals are tighter for non-icing events because there are more non-icing events than MOG icing events.

4.5.3 Volume efficiency

Volume efficiency relates the detection rate of a given model to the amount (volume) of icing that was diagnosed. Greater volume efficiency is desirable, indicating that there is less icing coverage over the airspace to achieve a given accuracy in detecting icing conditions.

- Since WRR production of icing severity is less than that of the RUC, and since detection rates for MOG icing events are quite comparable, as described in the previous sections, the volume efficiency for the WRR is always greater (better) than that of the RUC.

Table 4.2: CIP volume efficiency = (Detection Rate * 100) / (Percent Volume Coverage).

Probability Mask	RUC	WRR
0.00	16.82	**21.09**
0.25	19.20	**28.27**
0.50	26.56	**35.46**

Table 4.3: FIP volume efficiency = (Detection Rate * 100) / (Percent Volume Coverage).

Probability Mask	RUC	WRR
0.00	13.55	**15.32**
0.25	16.61	**19.53**
0.50	21.70	**26.85**

5. Summary and conclusions

This report presents a comparison of icing analysis and forecast fields derived from the operational RUC model and the newly-developed WRF Rapid Refresh model. Communicating the most significant differences in icing algorithm output from these two models is critically important to primary users of the algorithm data. Once the RUC data stream is discontinued, users will need to quickly and effectively adapt to characteristics of icing fields produced using input data from the new WRR model.

For the time period of the data acquired for this report, it has been determined that algorithm output derived from the WRR model yields performance results (for MOG severity) that are similar to those of the RUC model while providing for more efficient use of the available airspace.

For CIP: WRR production of icing is less than that of the RUC, especially at night and in mid-to-upper vertical levels. WRR production is much less in eastern portions of the domain and in the northwest corner of the domain. For CIP probability and severity, grid-to-grid agreement and correlation are highest in low-to-mid vertical levels. CIP SLD agreement is high throughout all vertical levels. While the WRR detection rate for MOG icing events is *slightly lower* than for the RUC, the WRR detection rate for non-icing events is *slightly higher* than for the RUC.

For FIP: WRR production of icing probability and severity is similar to that of the RUC, but WRR production of probability is greater at the higher probability values. WRR production is also greater along mountain ranges in the west. WRR SLD production is ***much less*** than it is for the RUC, likely due to differences in the way the forecast models handle convection in the humid southeast. For FIP SLD, grid-to-grid agreement and correlation ***are poor*** at upper levels. As forecast lead time increases, so do differences in the amount of SLD produced by the two models. While the WRR detection rate for MOG icing events is *slightly higher* than for the RUC, the detection rate for non-icing events is nearly identical.

Acknowledgements

This research is in response to requirements and funding provided by the Federal Aviation Administration. The views expressed are those of the authors and do not necessarily represent the official policy and position of the U.S. Government.

6. References

Benjamin, S.G., D. Devenyi, S.S. Weygandt, K.J. Brundage, J.M. Brown, G.A. Grell, D. Kim, B.E. Schwartz, T.G. Smirnova, T.L. Smith, and G.S. Manikin, 2004: An hourly assimilation/forecast cycle: The RUC**.** *Mon. Wea. Rev.,* **132,** 495-518.

Benjamin, S.G., D. Devenyi, R. Smirnova, S. Weygandt, J.M. Brown, S. Peckham, K. Brundage, R.L. Smith, G.Grell, and T. Schlatter, 2006: From the 13-km RUC to the Rapid Refresh. *12th Conf. on Aviation, Range, and Aerospace Meteorology,* Atlanta, GA, Amer. Meteor. Soc., CD-ROM. 9.1.

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.

Madine, S., S.A. Lack, S.A. Early, M. Chapman, J.K. Henderson, J.E. Hart, J.L. Mahoney, 2008: Quality Assessment Report: Forecast Icing Product (FIP). Submitted to Aviation Weather Technology Transfer (AWTT) Technical Review Panel.

Skamarock, W.C., J.B. Klemp, J. Dudhia, D.O. Gill, D.M. Barker, M.G. Duda, X.Y. Huang, W. Wang, J.G. Powers, 2008: A Description of the Advanced Research WRF Version 3. *NCAR Technical Note, NCAR/TN–475+STR*.

Wandishin, M.S., B.P. Pettegrew, M.A. Petty, J.L. Mahoney, 2011: Quality Assessment Report: Graphical Turbulence Guidance, Version 2.5. Submitted to Aviation Weather Technology Transfer (AWTT) Technical Review Panel.

Wolff, C.A., and F. McDonough, 2010: A comparison of WRR-RR and RUC forecasts of aircraft icing conditions. *14th Conf. on Aviation, Range, and Aerospace Meteorology,* Atlanta, GA, Amer. Meteor. Soc.

Appendix

For approximately 5% of all WRR/CIP 2-D grids acquired for this study, small anomalies were noted in the algorithm output. For the case presented in this section, results for CIP probability at 06 UTC show a grid with values that are either distinctly 0 or 1 (no other values present) for the WRR model (**Figure A.1a**), while the corresponding RUC grid appears to possess reasonable values (**Figure A.1c**). At the same time, the WRR SLD field contains all zeroes throughout (**Figure A.1b**), while the corresponding RUC SLD grid appears to possess reasonable values (**Figure A.1d**).

At the very next analysis time (07 UTC, **Figure A.2**), each grid for the two models appear to possess reasonable values throughout.

CIP data issue for the WRR

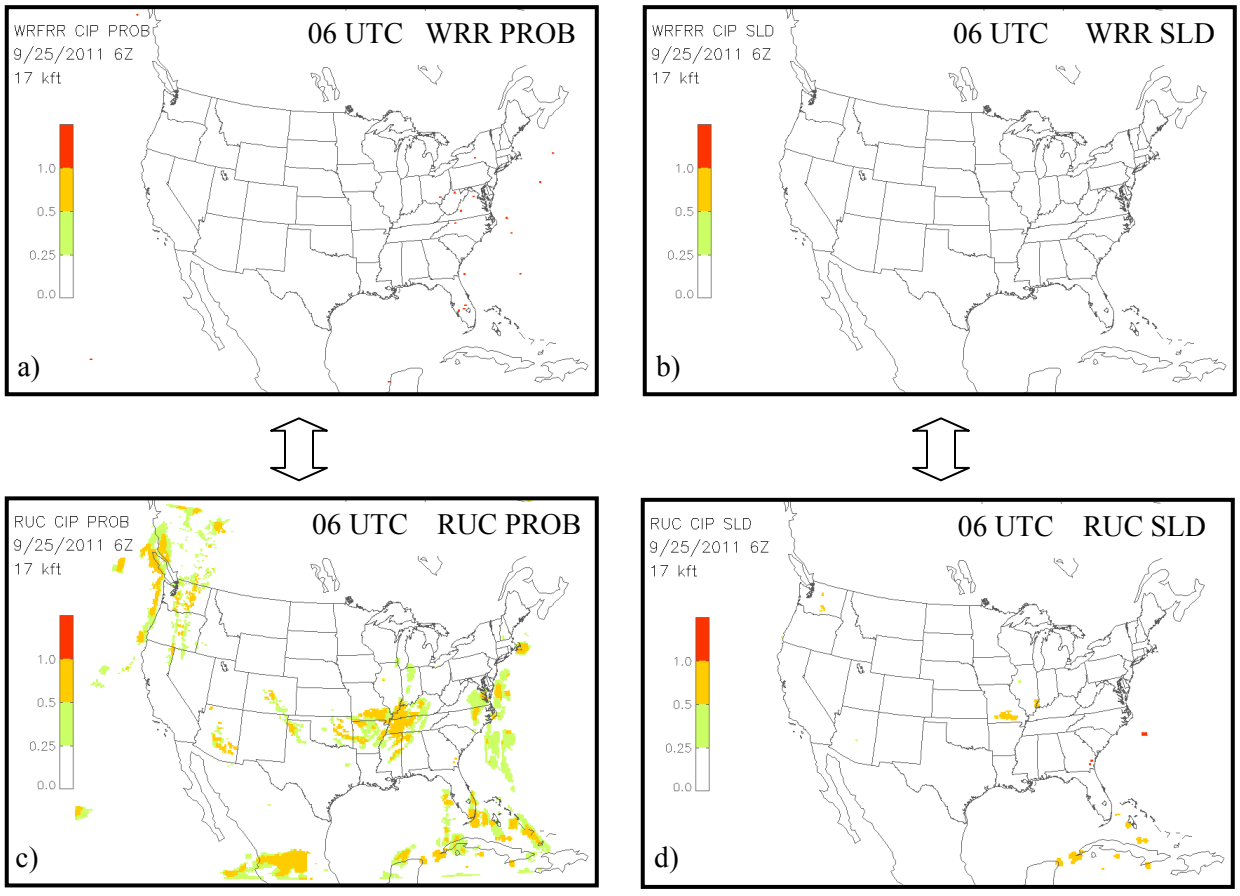

Figure A.1: Example of a data anomaly issue that was noted in approximately 5% of all WRR/CIP 2-D grids acquired for this study. Images are from 06 UTC on 25 SEP 2011. See the next set of figures (**Figure A.2**) from output that was generated just one hour later.

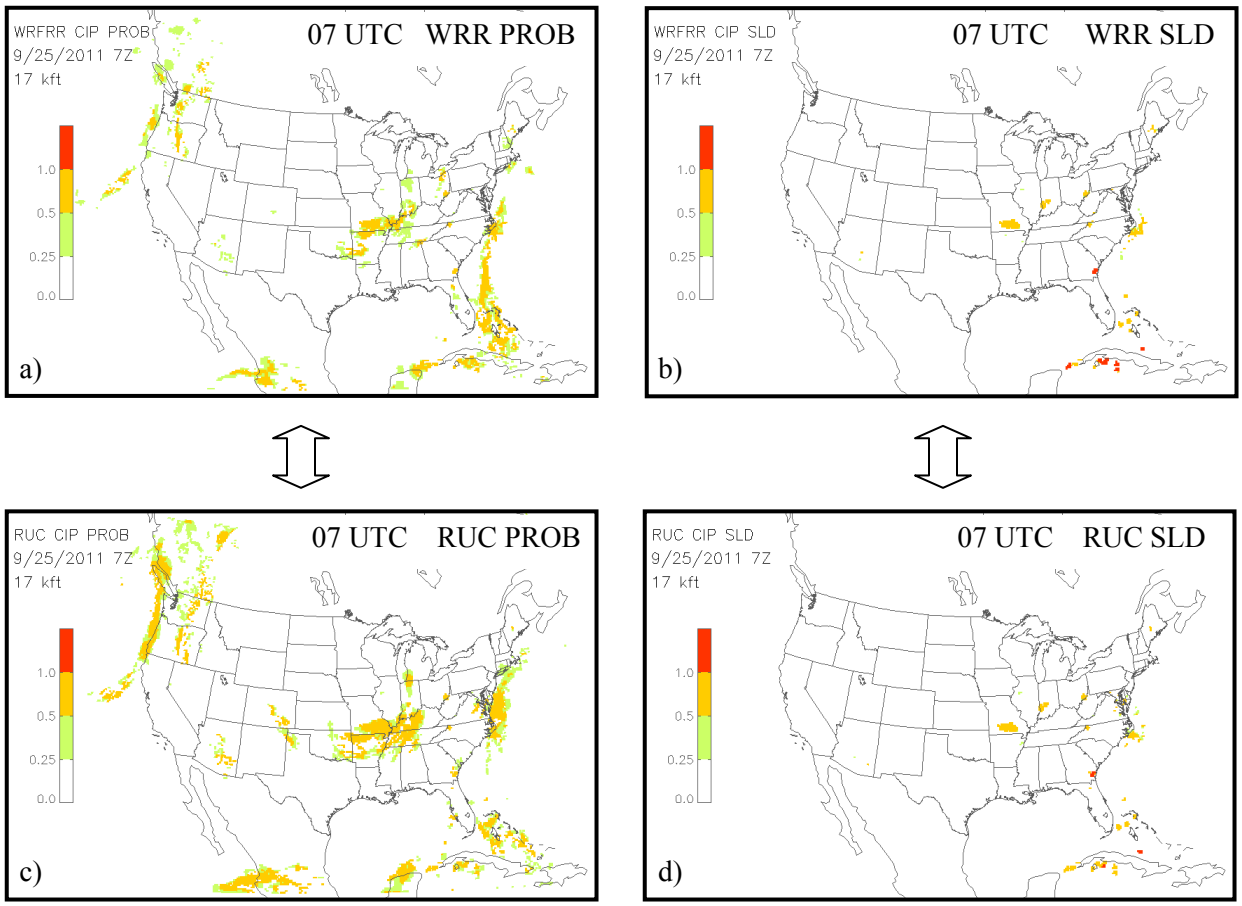

Figure A.2: Similar array of images as found in the previous figure, except there is no data anomaly issue in these CIP results that were generated just one hour later. Images are from 07 UTC on 25 SEP 2011.