

On the Value of Time-Lag-Ensemble Averaging to Improve Numerical Model Predictions of Aircraft Icing Conditions

MEI XU, GREGORY THOMPSON, DANIEL R. ADRIAANSEN, AND SCOTT D. LANDOLT

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 22 May 2018, in final form 6 March 2019)

ABSTRACT


The High-Resolution Rapid Refresh (HRRR) model with its hourly updating cycles provides multiple weather forecasts valid at any given time. A logical combination of these individual deterministic forecasts is postulated to show more skill than any single forecast for predicting clouds containing supercooled liquid water (SLW), an aircraft icing threat. To examine the potential value of using multiple HRRR forecasts for icing prediction, a time-lag-ensemble (TLE) averaging method of combining a number of HRRR forecasts was implemented for a multiple month real-time test during the winter of 2016/17. The skills of individual HRRR and HRRR-TLE aircraft icing predictions were evaluated using icing pilot reports (PIREPs) and surface weather observations and compared with the operational Forecast Icing Product (FIP) using the Rapid Refresh (RAP) model. The HRRR-TLE was found to produce a higher capture rate of icing PIREPs and surface icing conditions of freezing drizzle or freezing rain than single deterministic HRRR forecasts. As a trade-off, the volume of airspace warned in HRRR-TLE increased, resulting in a higher false detection rate than in the deterministic HRRR forecasts. Overall, the HRRR-TLE had similar probability of detection and volume of airspace warned for icing as the operational FIP prediction for the icing probability of 25% or greater. Alternative techniques for composing TLE from multiple HRRR forecasts were tested in postseason rerun experiments. The rerun tests also included a comparison of the skills of HRRR and HRRR-TLE to the skills of RAP and RAP-TLE.

1. Introduction

Accurate forecasts of aircraft icing and supercooled large drops (SLD; freezing drizzle and freezing rain) conditions are of great importance to aviation safety. In fact, new aircraft icing regulations were enacted in January 2015 to introduce a new icing certification rule, section 25.1420 (FAA 2014), and a FAA engineering standard (“Appendix O”) that defines SLD environments for certification of affected aircraft. Past icing forecast algorithms were developed by applying temperature and humidity thresholds to the output of numerical weather prediction (NWP) models (Schultz and Politovich 1992) and by refinement using vertical thermodynamic profiles (Forbes et al. 1993) and merging observations from various platforms (Bernstein et al. 2005). The currently operational icing forecast product that the aviation community relies upon, Forecast Icing

Product (FIP; <https://aviationweather.gov/icing/fip>), issues short-term forecasts of icing threat based on the thermodynamic and water-phase variables from the 13-km Rapid Refresh (RAP) model (McDonough et al. 2004). FIP provides valuable information including icing probability, icing severity, and potential for SLD.

Moving toward smaller grid spacing and more sophisticated physical parameterizations, the current High-Resolution Rapid Refresh (HRRR; Benjamin et al. 2016) has shown promising results for directly predicting clouds with an icing threat (Thompson et al. 2017). The rapid hourly cycling of HRRR also makes available multiple forecasts valid at any given time. Past experience and previous studies (e.g., Wolff and McDonough 2010; Cintineo et al. 2014; Thompson et al. 2017) have shown that cloudy regions predicted by single deterministic NWP model forecasts tend to be spatially smaller than is observed. This low bias in cloud amount is seen for relatively low-altitude clouds such as boundary layer clouds, as well as midtroposphere clouds with cloud-top temperature in the range of 0°–20°C. Therefore, it is logical to presume that aircraft icing is underpredicted by the current deterministic HRRR forecasts. This led us to

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Mei Xu, meixu@ucar.edu

postulate that a logical combination of individual deterministic forecasts may improve the prediction of aircraft icing beyond any single forecast. Furthermore, given the uncertainties in the model's initial conditions and imperfections in the model's physical parameterizations, the ensemble approach may be a viable way to reduce forecast uncertainties. Thompson et al. (2017) described an attempt to construct the time-lag-ensemble (TLE) average for the prediction of supercooled liquid water (SLW) using a set of individual HRRR-TLE forecast members. For one case study of a high-impact weather event, it was shown that the TLE technique increased, sometimes substantially, the number of correctly captured icing reports. Also as expected, there was a trade-off with the prediction of negative icing reports, due to the increase in the total predicted icing airspace volume by the TLE averaging procedures.

To test systematically the effectiveness of the TLE procedures, an algorithm similar to that developed in Thompson et al. (2017) was implemented and automated to run in a 3.5-month real-time test beginning 1 December 2016 and ending 15 March 2017. During the real-time test, individual HRRR forecasts from hourly cycles were received from the National Centers for Environmental Prediction (NCEP). These deterministic HRRR forecasts were combined to form 1–12-h HRRR-TLE forecasts (with the HRRR-TLE forecast length refers to the shortest HRRR forecast used in the TLE). Observations of icing conditions aloft from pilots (PIREPs) and surface weather conditions from Meteorological Aerodrome Reports (METARs) were used to assess the icing conditions aloft and at the surface. Finally, in order to compare the icing predictions of the HRRR individual and HRRR-TLE forecasts to the existing operational forecast, real-time FIP data were collected and evaluated as well. For brevity, only the evaluation results of 3- and 6-h forecasts of HRRR, HRRR-TLE, and FIP are discussed in this paper. Following the real-time test, rerun experiments were conducted to test alternative TLE averaging methods for HRRR and to evaluate the TLE forecasts composed of RAP members. In the sections to follow, this paper describes the TLE procedures (section 2), observation datasets and verification methods (section 3), results from the systematic real-time evaluation (section 4), postseason rerun experiments (section 5), and intermodel comparisons (section 6). Section 7 gives the summary and conclusions.

2. The TLE methodology

The TLE method implemented for the real-time test in this study is similar to that described in Thompson

et al. (2017), but applied to the operational HRRR forecasts. At each grid point, a set of fractional weights were assigned to the explicitly predicted liquid water content (LWC; sum of cloud water and rain) from the individual HRRR forecasts to create a weighted average as the resultant HRRR-TLE forecast of LWC. Icing conditions in both HRRR and HRRR-TLE were diagnosed from LWC and temperature, with the ice accretion rate calculated following Thompson et al. (2017). A threshold of $\text{LWC} > 10^{-6} \text{ g m}^{-3}$ (and temperature $< 0^\circ\text{C}$) was used in the real-time test as the criteria for icing forecast on any grid point. No attempt was made to discriminate the various icing intensities with model-predicted ice accretion rates in this work.

The operational HRRR model is run every hour with hourly forecasts to 18 h over a domain that covers the contiguous United States (CONUS) with 3-km grid spacing. The numerical core of HRRR employs the Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008), version 3.6.1, utilizing the Thompson and Eidhammer (2014) aerosol-aware microphysics. Additional details regarding physical parameterizations used in HRRR can be found in Benjamin et al. (2016). Because of the nature of real-time experiments and prototype software to process HRRR data, the availability of any specific forecast for any specific time of day was not always guaranteed. Infrequent data outage and occasional data transfer delay occurred. The real-time processing system took in what was available each hour after approximately 2-h wait time to construct HRRR-TLE forecasts of various forecast lengths.

Each HRRR-TLE forecast used as few as 3 or as many as 9 time-lagged ensemble members, all valid at the same time. Only the 3- and 6-h HRRR-TLE forecasts were evaluated in this paper. For the 3-h HRRR-TLE, the most recent forecast used is the 3-h HRRR forecast, and similarly for the 6-h HRRR-TLE, the most recent member forecast is the 6-h HRRR. The eligible member forecasts for 6-h HRRR-TLE is schematically shown in Fig. 1. However, during the real-time test, no processing information was recorded to know exactly how many and which HRRR members actually entered into a specific HRRR-TLE forecast.

Ideally, the weights assigned to the individual members that compose an ensemble forecast should be dependent on the relative skills of the individual forecasts. However, without reliable a priori knowledge of the skills of the individual HRRR forecasts for icing, the weights used in this test were arbitrarily set. Table 1 gives the weights used in the real-time test. Depending on the number of available forecasts, one of the lines in Table 1 was used to combine the member forecasts into the resultant TLE. Another crucial aspect of the

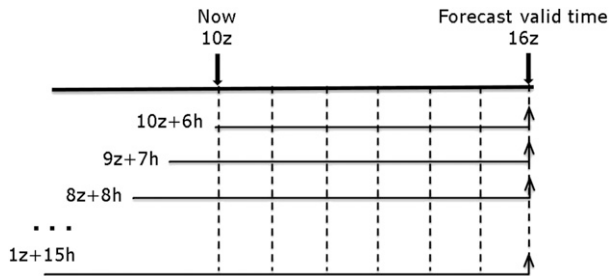


FIG. 1. Schematic showing the 10 eligible HRRR member forecasts that comprise a 6-h TLE ensemble forecast. Which members were actually used in a TLE depends on the availability of files and the maximum number of members allowed. A maximum of 9 were allowed in the real-time test.

real-time system was the selection of member forecasts. To retain maximum flexibility in a real-time operating environment, the table of weights was used in conjunction with a priority ordering of the forecast hours shown in Table 2. Take 6-h TLE for example, 10 HRRR forecasts are eligible for composing 6-h TLE (having a lead time of 6 h or longer) and they were ordered as 6, 9, 12, 7, 11, 8, 10, 13, 14, and 15 h (Table 2). If all 10 forecasts were received when the 6-h TLE was composed, the first 9 files would be used. The 9 HRRR forecasts would be given the weights in line 1 of Table 1 to form the 6-h TLE. If less than 9 HRRR files were received at the time the TLE was composed, weights in line 2 or below of Table 1 would be used. If less than 3 files were received, the 6-h TLE would be missing. Again, the priority rankings used in the real-time test (Table 2) were arbitrarily set.

Following the real-time test, a set of postseason rerun experiments were conducted using a subset of the real-time cases. In the rerun mode, the weights and members were varied in a controlled manner, and their impact on the resultant TLE forecast was explored (section 5). The TLE method described above was also used to construct RAP-TLE forecasts from selected deterministic RAP forecasts, which will be discussed in section 6.

3. Validation of model forecasts

a. Validation of HRRR model forecasts aloft using PIREPs

Pilot reports (PIREPs) are verbally relayed from pilots to various aviation support personnel on the ground and generally contain reports of upper-air conditions of cloud cover, temperature, wind, weather, turbulence, and icing. The icing information from PIREPs includes direct reports of either no icing (NEG) or subjective levels of icing experienced such as trace (TRC), light (LGT), moderate (MOD), or heavy/severe (SEV).

TABLE 1. The weights given to individual forecasts to create a HRRR-TLE forecast depending on the number of available files to include in the TLE average.

No. of files	Weights
9	0.1, 0.1, 0.1, 0.1, 0.2, 0.1, 0.1, 0.1, 0.1
8	0.1, 0.1, 0.2, 0.1, 0.1, 0.2, 0.1, 0.1
7	0.1, 0.1, 0.2, 0.15, 0.15, 0.2, 0.1
6	0.5, 0.1, 0.1, 0.1, 0.1, 0.1
5	0.50, 0.15, 0.15, 0.10, 0.10
4	0.55, 0.20, 0.15, 0.10
3	0.60, 0.25, 0.15
2	0.65, 0.35

Occasionally, pilots will use intermediate categories such as LGT-MOD, in which case we placed them into the bin with the more severe category. Since TRC reports may not be operationally relevant to end users and the number of SEV reports is relatively small, we focused on the verification statistics calculated for the two intermediate icing severity thresholds: *LGT and above* and *MOD and above*. Various inherent uncertainties and limitations are associated with PIREPs (Schwartz 1996). As a result, the verification statistics are not to be interpreted in an absolute sense, but as relative assessment for comparing the capabilities of different forecasting systems.

The validation was done for mostly daytime hours of 1200–0300 UTC, when PIREPs are most prevalent, for all 105 days of the real-time test. To match the observations with forecasts in time, PIREPs from 30 min prior to 30 min after the top of each hour were used to verify the forecasts valid at the hourly output interval of the HRRR. Spatially, the HRRR model values at a set of 8×8 grid points surrounding the report and within 1000 ft (305 m) below and above the reported altitude

TABLE 2. The ranking of each HRRR forecast hour used in the time-lag-ensemble average.

Priority	Forecast hour
1	6
2	9
3	12
4	1
5	3
6	4
7	5
8	7
9	11
10	8
11	10
12	2
13	13
14	14
15	15

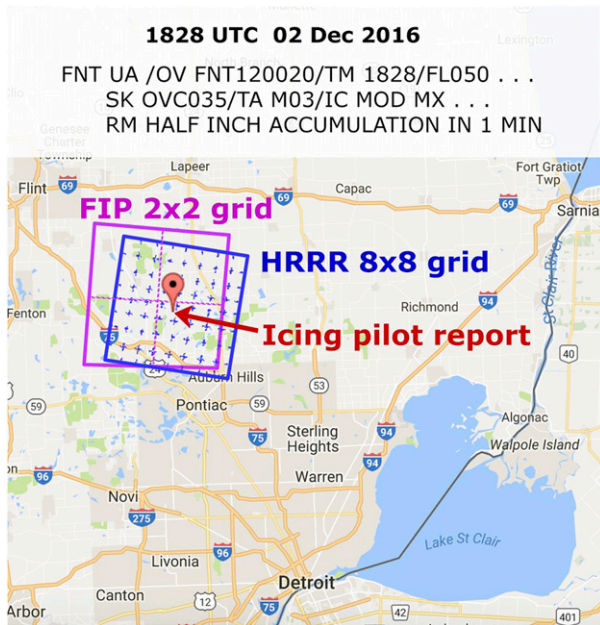


FIG. 2. Schematic showing the HRRR 8×8 ($24 \times 24 \text{ km}^2$) gridbox region and the FIP 2×2 ($26 \times 26 \text{ km}^2$) gridbox region where the model forecasts are validated by the indicated PIREP observation.

were retrieved. The 8×8 model grid points are taken as 4 consecutive rings surrounding the PIREP covering a region of approximately $24 \times 24 \text{ km}^2$ (see Fig. 2). In the validation of forecast for a positive icing PIREP, a model prediction was scored as a hit if any one of the 64 model points surrounding the PIREP on any model level within 1000 ft from the report altitude contained icing (SLW above the threshold of 10^{-6} g m^{-3}). However, in the case of negative icing, all of the 64 grid points surrounding the no-icing PIREP at the model level nearest to the reported altitude were required to have no model-predicted icing to be considered a hit (correct negative).

Following Brown et al. (1997), we computed the probability of detection (POD) or hit rate by the model forecasts for icing reports at the reported icing severity levels, and the probability of detection for no-icing (NEG) reports, POD_{no} . By definition, POD_{no} is the fraction of the no-icing PIREPs that were correctly forecast as no. As discussed in Brown et al. (1997), the pilots usually do not make systematic reports, and typically in PIREPs, there are many more icing reports than no-icing reports. The event counts in PIREPs overestimate the relative frequency of icing conditions, and do not represent the “true” distribution of icing conditions. As a result, it is inappropriate to combine the yes- and no-icing reports together to estimate joint probabilities. For example, using PIREPs to estimate the standard false alarm ratio, defined as the probability of a

no-observation given a yes-forecast, would produce misleading results of values that are much too small. The limitations of PIREPs also make it difficult to adopt some other commonly used statistical evaluation methods for dichotomous forecasts, such as the equitable threat score (ETS) or the true skill statistic (TSS). The POD and POD_{no} used in this study are observation-based probabilities. The complement of POD_{no} , the probability of false detection ($\text{POFD} = 1 - \text{POD}_{\text{no}}$), can be viewed as a limited measure of false alarms, since it is the fraction of the observed “no” events that were incorrectly forecast as “yes.” In the discussions to follow, the phrase, false alarm rate, will be used to refer to the observation based probability of false detection (POFD).

From the real-time test, daily, monthly and 105-day aggregated statistics of POD and POD_{no} (POFD) were compiled. In addition, a total volume of icing impacted airspace (VOL) was computed for each forecast. The method for computing VOL follows that from Thompson et al. (1997a) and Brown et al. (1997) and VOL is defined as the sum of all model gridbox volumes containing forecast icing conditions for each HRRR or HRRR-TLE forecast. Therefore, the VOL is a measure of the total spatial extent of the predicted SLW field over the model domain in increments of HRRR gridbox volume. An increase of VOL from HRRR to HRRR-TLE is clearly expected as a result of the ensemble spread.

To facilitate comparing HRRR validation results with validation results using the FIP and RAP, one may define VOL for the HRRR grid in a different manner. By partitioning the HRRR grid into 4×4 gridpoint neighborhoods, each neighborhood contains an area of $12 \times 12 \text{ km}^2$, which is very close to the FIP/RAP individual grid cell area of $13 \times 13 \text{ km}^2$. This partitioning method results in 16 individual HRRR grid cells representing a $12 \times 12 \text{ km}^2$ area. If any one of the individual HRRR grid cells within the $12 \times 12 \text{ km}^2$ neighborhood had SLW above the threshold (10^{-6} g m^{-3}), then the entire $12 \times 12 \text{ km}^2$ neighborhood was considered to have icing for the VOL calculation. This methodology effectively degrades the HRRR grid resolution from $3 \times 3 \text{ km}^2$ to $12 \times 12 \text{ km}^2$ to allow close comparison with the FIP/RAP (discussed further in section 6). It should also be noted that the HRRR and FIP/RAP have similar vertical grid increment.

b. Validation of RAP model and FIP forecasts aloft using PIREPs

The current FAA-sponsored operational forecast icing product provides 2–18-h forecasts of icing probability, icing severity, and potential for supercooled large drops (including freezing drizzle and freezing rain) on a grid with 13-km horizontal spacing and 500-ft vertical

spacing. FIP icing probability was calibrated against PIREPs and always remains below 100% since the forecast of icing cannot be done with absolute certainty at any given location in space and time. The maximum icing probability that is permitted is based on forecast lead time. At 3- and 6-h lead times, the maximum permitted icing probability values are 74% and 64%, respectively. FIP's forecasts of icing severity are given in five categories similar to those used in PIREPs: none, trace, light, moderate, and heavy. The real-time display of FIP can be found at <http://www.aviationweather.gov/icing/fip>.

During the real-time test and postseason rerun, FIP forecasts at various probability levels were evaluated versus PIREPs using the four grid points ($26 \times 26 \text{ km}^2$, see Fig. 2) immediately surrounding each PIREP. Given the 13-km spacing FIP grid, the four gridpoint evaluation area is nearly equivalent to what was used in the HRRR evaluation. If any of the four grid boxes had icing probability equal or greater than a given probability level (any of 5%, 15%, 25%, 35%, 45%, and 55%), then the FIP forecast corresponding to the PIREP was designated as having predicted icing at that probability level. No evaluation was made for FIP predicted icing severity categories. For FIP at each of the six probability thresholds, POD (or POD_{no}) was calculated using PIREPs of LGT and above, MOD and above, and NEG, respectively. Similar to the HRRR evaluation, when evaluating the no-icing PIREPs to compute POD_{no} for FIP, all four grid boxes surrounding the report had to contain an icing probability below the thresholds mentioned. The different thresholds of icing probability correlated with different volumes of airspace warned since the 5% and above represents a far greater volume than the 55% and above threshold.

Similarly, RAP and RAP-TLE forecasts were also validated versus PIREPs using the four grid points ($26 \times 26 \text{ km}^2$) immediately surrounding each PIREP. Aggregated statistics of POD, POFD and VOL were calculated for RAP and RAP-TLE during the rerun experiments. Since the FIP and RAP domains are larger than the HRRR domain, for comparison purposes only the PIREPs that were used in HRRR validation were used to evaluate FIP and RAP.

c. Validation of HRRR surface precipitation type using METAR

During the real-time test, surface conditions of icing (freezing rain or freezing drizzle) as well as other surface weather types were validated using METAR observations. Weather types in METAR were classified into seven categories: rain or drizzle (RA or DZ), snow (SN), fog (FG), freezing rain or freezing drizzle (FZRA or

FZDZ), freezing fog (FZFG), ice pellet (PL), and graupel/hail (GR/SG). Among the categories, the reports of rain and snow are generally the most reliable while the reports of fog and freezing fog are the least reliable. In decoding the METAR data, some classification logics were applied (e.g., when a report of rain was found with an air temperature $T < 0^\circ\text{C}$, it was placed as freezing rain). In a typical hour during cool season, there are ~ 2000 reporting stations with as many as 500–700 precipitation observations over CONUS. Most METAR locations make reports of occurrence of the various surface weather types, but do not make nonprecipitation reports. For these stations, a lack of report of a weather type does not necessarily mean that the weather type did not occur. However, a subset of 33 locations are Service Level A or B sites augmented with human observations that can provide reliable information of nonoccurrence of events. Therefore, similar to PIREPs, the METAR data were also biased toward positive precipitation reports. No skill scores that combine hits and false alarms could be evaluated.

To verify forecasts valid each hour, METARs from 15 min prior to 15 min after the hour were used with the same horizontal matching process as was done in the PIREP verification. The time window used for METARs were narrower than that used for PIREPs because there were generally many more METAR reports than PIREPs. Also different from the PIREP verification, the POD and POD_{no} were computed for each surface weather condition using two different threshold criteria: at least 1 and at least 8 of the 64 HRRR model points surrounding the report contained the observed condition. Obviously, the 8-point POD represents a more stringent test compared to the 1-point test and will always have a lower POD value.

The model forecast of precipitation type was diagnosed from the hydrometeor and temperature fields at the lowest model level. For example, in the case a METAR has RA (rain) or DZ (drizzle), then if the model has rainwater $\text{QR} > 10^{-6} \text{ g m}^{-3}$ and $T > 0^\circ\text{C}$, it is a hit. In the case a METAR has FZRA (freezing rain) or FZDZ (freezing drizzle), then if the model has rainwater $\text{QR} > 10^{-6} \text{ g m}^{-3}$ and $T < 0^\circ\text{C}$, it is a hit. It should be pointed out that the METAR reports of both FG and FZFG are based on visibility, not on the actual presence of visible droplets, so the verification of FG and FZFG using cloud water were highly unreliable. The graupel/hail reports were also not used in the verification. Since the main interest of this work is on icing prediction, the verification focused on freezing rain and freezing drizzle conditions, even though additional surface precipitation types were verified using METAR. POD for various precipitation types were computed from METAR

reports from all the ~2000 observing stations whereas POD_{no} (or $POFD$) were computed using only the limited dataset from the 33 Service Level A or B stations.

d. Statistical significance of the validation

During the 3.5-month test period, a total of 32 000 icing PIREPs of trace, light, moderate, or severe were collected (see Fig. 3a). In addition, there were 6500 explicit no-icing (NEG) reports during the period. Icing PIREPs were present for 1670 out of the total 1785 h, typically between a few and a few dozen reports per hour. Typical numbers for daily light icing reports are 100–300 across the CONUS. Figure 3b shows the number of METAR observations according to categorized surface precipitation types during the real-time test period. During the 3.5 months, there were approximately 244 000 reports of RA or DZ, 292 000 reports of SN, 10 000 reports of FZRA or FZDZ, 1000 reports of PL, 57 000 reports of FG, and 23 000 reports of FZFG. Only the observations within the evaluation time window were counted in Figs. 3a and 3b.

Given an icing report of any type in time and space, the model forecast is either a “hit” or a “miss.” The collection of the forecasts for all reports forms a binomial distribution of size N . Assuming the modeling system has a hit rate of p , then the success proportions would follow an approximate normal distribution with the mean equal to the true proportion p , and with standard deviation:

$$\sigma = \sqrt{\frac{p(1-p)}{N}}.$$

The upper limit is $\sigma = 0.5/\sqrt{N}$ for $p = 0.5$ (50%). For $p = 0.8$, $\sigma = 0.4/\sqrt{N}$, and for $p = 0.9$, $\sigma = 0.3/\sqrt{N}$.

The above analysis gives us a way to roughly estimate the confidence interval of the calculated POD or $POFD$ for given sample sizes. To ensure the standard deviation of the calculated POD or $POFD$ is within 1% (i.e., with 95% confidence interval at -1.96% and $+1.96\%$ for approximate Gaussian distribution), N would need to be 900–2500 (assuming $p = 0.1$ – 0.9). Given that for PIREPs reports of LGT and NEG, $N = 32\,000$ and 6500 , respectively, the uncertainties of the aggregated mean POD and $POFD$, in terms of standard deviation, will be approximately 0.2% – 0.3% and 0.4% – 0.6% . For the estimation of daily POD and $POFD$, the values will increase to about 2% – 3% and 4% – 6% . The mean POD for severe icing (SEV) also has a larger uncertainty, with a standard deviation of about 2% – 3% . For the surface verification using METAR icing reports, $N = 10\,000$, and therefore, $\sigma \approx 0.3\%$ – 0.5% .

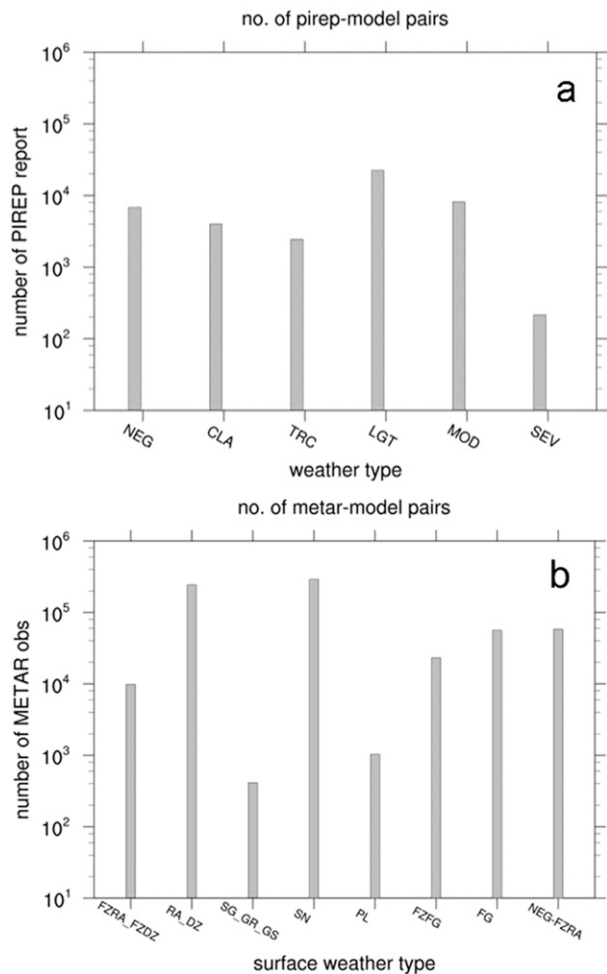


FIG. 3. (a) Total number of PIREPs vs weather types during the 3.5-month testing period. NEG indicates explicit no-icing reports at locations, and CLA is for reports of “clear sky above,” which imply no-icing conditions. CLA reports are not used in this study. (b) Number of METAR reports vs surface weather types during the 3.5-month test period. The value for NEG-FZRA is the number of negative freezing rain/freezing drizzle reports.

4. Results from the winter 2016–17 real-time test

a. Forecast skills for icing aloft

The 3.5-month aggregated probability of detection by the 3- and 6-h forecasts of HRRR and HRRR-TLE with respect to the PIREP icing category is shown in Figs. 4a and 4b. Clearly, an icing forecast that uses the time-lag-ensemble (TLE) procedure is shown to capture more icing pilot reports than its corresponding deterministic forecast. For icing PIREPs of light and greater severity (indicated as LGT in Fig. 4a), the probability of detection by the 6-h forecasts increased from 59% to 75% using the TLE method. For PIREPs of moderate and greater severity (indicated as MOD in Fig. 4a), the increase in probability of detection was from 63% to 79%

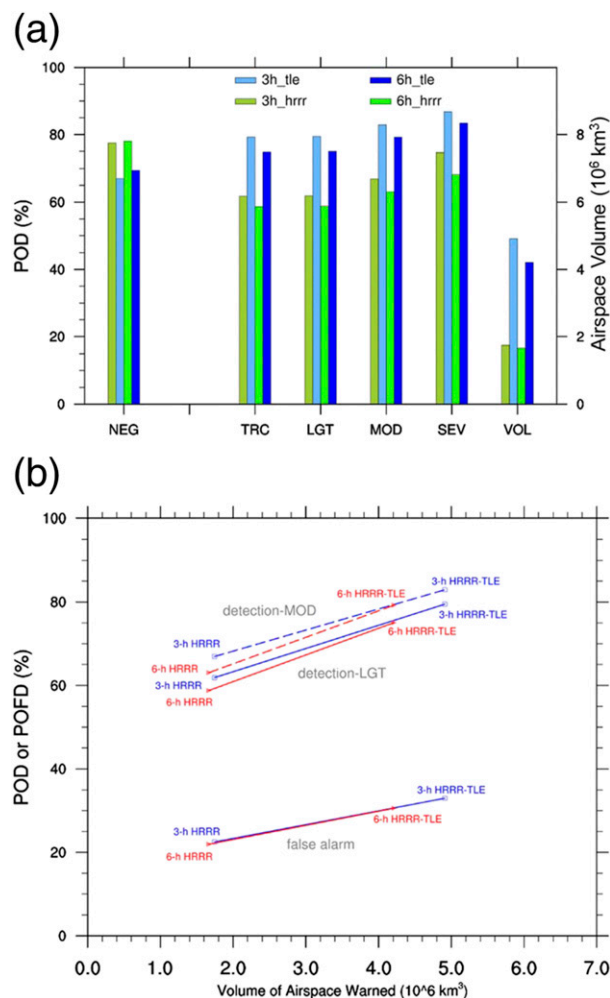


FIG. 4. (a) POD of the aloft icing reports by 3-h HRRR and HRRR-TLE and 6-h HRRR and HRRR-TLE forecasts. The bars labeled TRC (LGT, MOD) are for the indicated and greater severity. The last cluster of bars are for the VOL in the forecasts. (b) POD and POFD vs VOL by 3-h HRRR and HRRR-TLE (blue) and 6-h HRRR and HRRR-TLE (red) forecasts, for light and above (solid lines), and moderate and above (dashed lines) icing severity. The lines connecting HRRR and HRRR-TLE are plotted for clarity only.

for 6-h forecasts. The improvement of POD by TLE in the 3-h forecasts are similar to that in the 6-h forecasts, with the general POD skills of the 3-h forecasts slightly higher than those of the 6-h forecasts. Also evident in Fig. 4 is the general trend of increasing POD as PIREP icing severity increases, with the highest POD for severe icing conditions.

On the other hand, as the POD increased with TLE, the POD_{no} decreased, resulting in increased POFD due to false alarms. The percentage of correctly captured negative icing (POD_{no}) was 78% for the 6-h deterministic HRRR forecast and decreased to 69% with HRRR-TLE.

This is to say that the POFD increases from 22% to 31% when using the TLE method. Since the predicted cloudy regions are increased using TLE, the volume of total airspace impacted increased by 147%, from approximately 1.7×10^6 to $4.2 \times 10^6 \text{ km}^3$, by TLE on average during the 3.5 months. This trade-off of volume of airspace warned for increasing POD was mentioned in prior studies and remains a significant challenge for any forecast system to gain the greatest increase of POD while retaining the lowest POFD. Again it should be pointed out that, the VOL values in Figs. 4a and 4b were defined at 3-km resolution as the volume sum of all HRRR grid boxes having explicitly predicted SLW. The 6-h deterministic HRRR forecasts consistently have slightly smaller VOL than the 3-h HRRR forecast, pointing to a possible small effect of cloud reduction in longer forecasts. The notably smaller VOL in 6-h TLE ($4.2 \times 10^6 \text{ km}^3$) as compared to 3-h TLE VOL ($4.9 \times 10^6 \text{ km}^3$) are presumably mainly because generally fewer members are used in 6-h TLE, given that HRRR forecasts of 3–5-h lead time could be included in 3-h TLE, but not in 6-h TLE.

The daily values of POD and POFD are shown in Fig. 5. As discussed in section 3d, the daily POD/POFD calculation is associated with a greater degree of uncertainty due to the smaller sample sizes. As an example, for the 6-h forecasts, the POD values for the deterministic HRRR spread from 40% to 80% with the 10th and 90th percentiles at 47% and 70%, respectively, while the POD for HRRR-TLE ranges approximately from 55% to 90% with the 10th and 90th percentiles at 65% and 83%. The POFD for the deterministic HRRR ranges from 5% to 40% with the 90th percentile around 31%, while the values for HRRR-TLE are 15%–55%, and the 90th percentile is at 42%.

b. Forecast skills for surface precipitation type

Figure 6 shows the 3.5-month aggregated statistics of the HRRR and HRRR-TLE forecasts validated against categorical surface weather conditions reported in METARs. For each of the six groups of weather conditions, two sets of four bars are plotted, for POD using the 1-point threshold and 8-point threshold, respectively, for the 3- and 6-h individual and HRRR-TLE forecasts. As expected, for all surface weather types, the POD for the more stringent 8-point threshold is lower than the POD for the 1-point threshold; however, the decreases are generally small among each weather category. An exception is the POD for FZRA/FZDZ, for which the fractional decrease going from the 1-point to 8-point threshold is relatively large. More noteworthy is the additional benefit of the TLE method to match the observed weather more consistently. For instance,

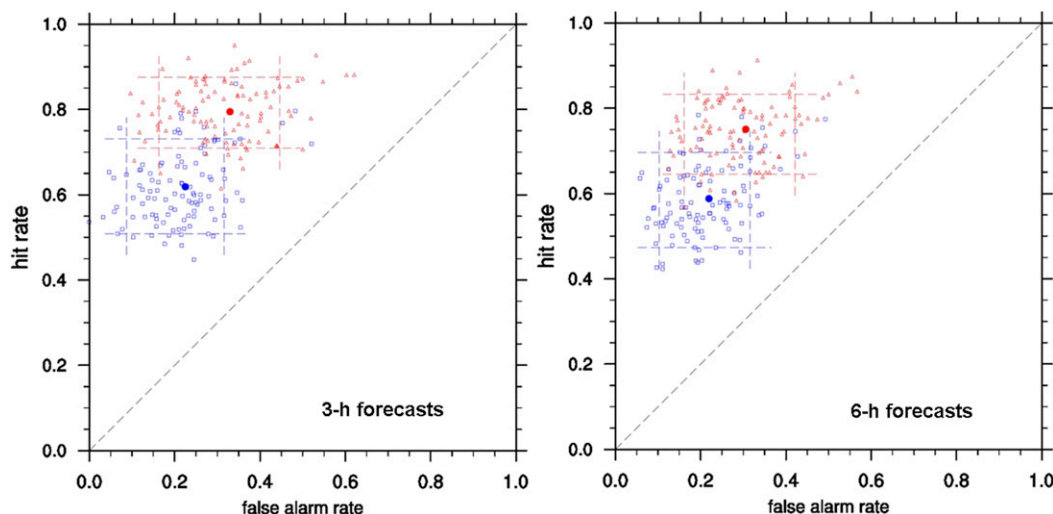


FIG. 5. Scatterplots of daily hit rate (POD) vs false alarm rate (POFD) of icing aloft by HRRR (blue) and HRRR-TLE (red) forecasts at the light and above severity level (LGT) during the real-time test. The two filled dots are the aggregated values during the 3.5 months for HRRR and HRRR-TLE, respectively. The horizontal and vertical dashed lines mark the upper and lower 10th percentiles of the POD and POFD, respectively.

the POD for 6-h forecasts at 1-point threshold increases from 75% to 86% for snow and from 44% to 55% for the combination of either FZRA or FZDZ using the TLE method. The POFD values for FZRA/FZDZ for all forecasts remain low, from 1%–2% for HRRR to 3%–7% for HRRR-TLE. When a false detection is defined as having only 1 of 64 model grid points showing a nonobserved precipitation type instead of 8, the POFD values increase considerably. The highest POFD values are seen in the forecasts of rain and drizzle (RA/DZ).

Generally, the deterministic HRRR forecasts underpredict the occurrence of FZRA/FZDZ conditions. With TLE, the cloudy and precipitating regions are increased to improve the POD values for icing conditions aloft as well as at the surface, however, the POFD values also increase. One significant caveat with the analysis of surface weather conditions is the known deficiency of freezing drizzle and ice pellet reports in the existing METARs from Automated Surface Observing System (ASOS; Ramsay 1999; Landolt et al. 2017),

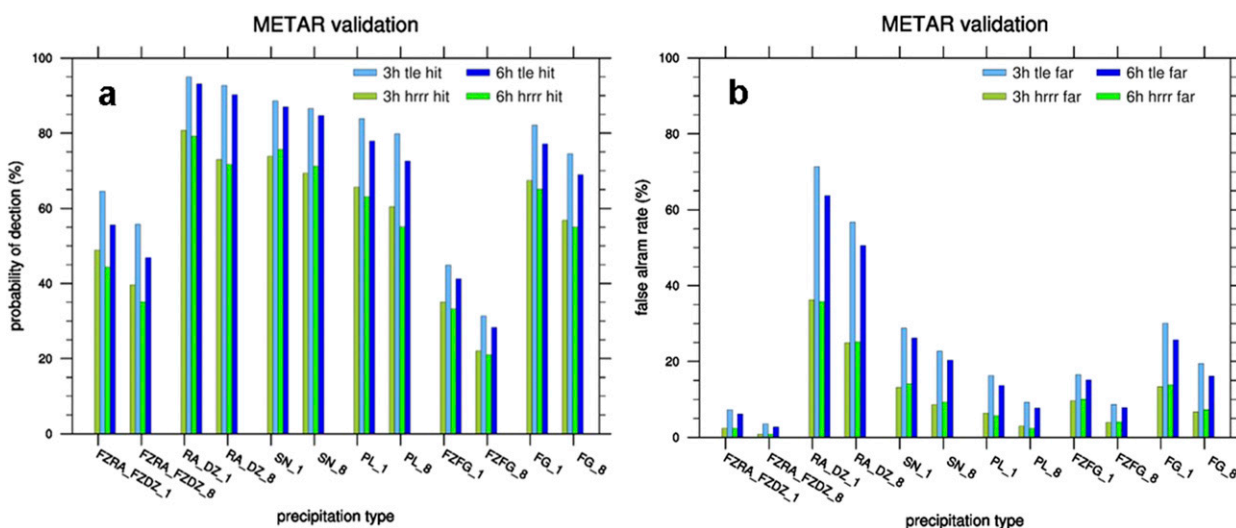


FIG. 6. (a) POD and (b) POFD of 6 different surface weather type combinations by 3-h HRRR (olive), 3-h HRRR-TLE (light blue), 6-h HRRR (green) and 6-h HRRR-TLE (blue) forecasts validated against METAR. Two levels of detection are defined as having at least one and at least eight matching grid points within a $24 \times 24 \text{ km}^2$ box surrounding the observation location.

which are sometimes incorrectly diagnosed as snow, reported as unknown precipitation (UP) or missed altogether. The improvement of automatically diagnosed surface weather conditions using a new precipitation-type algorithm is a subject of current research.

5. Alternative TLE methods

To explore further the ability of HRRR-TLE for icing forecast, alternative TLE methodologies are tested in a postseason rerun and analysis, using a subset of the real-time cases. The rerun was done for 4 validation times daily, at 1500, 1800, 2100, and 0000 UTC, during December 2016. For each validation time, seven deterministic HRRR forecasts, with a lead time of 2, 3, 4, 6, 9, 12, and 15 h, were combined in alternative ways to form a list of TLE forecasts (Table 3). The same fractional weights as given in Table 1 were used in the rerun test when applicable (i.e., in TLE-7Member, TLE-6Member and TLE-3Member). These alternative TLE forecasts, as well as the individual forecasts, were evaluated against PIREPs in the same way as in the real-time test (section 4).

The POD and POFD of the TLE and deterministic forecasts of the rerun test are shown in Fig. 7. The first thing to note in Fig. 7 is that the 3-h deterministic HRRR and 3-h HRRR-TLE by weighted averaging 6-member forecasts have very similar skills to those in the real-time test (Fig. 4b). Given that the rerun has a smaller sample size, this argues for the reliability of the section 4 results. Among the deterministic forecasts, while the POD generally increases with shorter forecast lead time, the probability of false detection and the volume of warned airspace increases with shorter lead time. This may again indicate that there exists a slight trend of decreasing cloudiness as the model integrates forward in time.

The next thing to note is that TLE-7Member, the TLE by weighted averaging 7 members (Table 3), has the same skill as TLE-MAX, the TLE by taking the maximum SLW value of the 7 members at each grid point. This is presumably because a relatively low SLW amount (10^{-6} g m^{-3}) is used as the threshold above which a detection by the model is signified, such that if any member predicts icing at a grid point, the TLE by weighted average would predict icing. This renders the weights used in the TLE relatively unimportant, as long as the resultant average SLW is above the threshold value. The POD and POFD of a forecast have more to do with the extent or regions, but not the specific values of SLW in the clouds. Essentially both TLE-7Member and TLE-MAX have cloudy regions encompassing all cloudy regions in the seven individual HRRR forecasts. Tests have shown that when the icing threshold of SLW is increased to 10^{-3} g m^{-3} , the magnitude of SLW in the forecasts will become important.

TABLE 3. List of the TLE experiments conducted for December 2016. The methodology explains how the TLE hydrometeor fields are composed from those of the deterministic HRRR forecasts.

Experiment	TLE methodology
TLE-7Member	Weighted average of 2-, 3-, 4-, 6-, 9-, 12-, and 15-h HRRR
TLE-6Member	Weighted average of 3-, 4-, 6-, 9-, 12-, and 15-h HRRR
TLE-MAX	Maximum value of 2-, 3-, 4-, 6-, 9-, 12-, and 15-h HRRR
TLE-3Member	Weighted average of 2-, 3-, and 6-h HRRR
TLE-Median	Median value (fourth largest) of 2-, 3-, 4-, 6-, 9-, 12-, and 15-h HRRR
TLE-Median2	Second largest value of 2-, 3-, 4-, 6-, 9-, 12-, and 15-h HRRR
TLE-Median3	Third largest value of 2-, 3-, 4-, 6-, 9-, 12-, and 15-h HRRR

When only 3 members (2, 3, and 6 h) are used in the TLE forecast (TLE-3Member), its skill and volume-warned decreased proportionally from the 7-member TLE. The TLE-Median, formed by taking the median value of the 7 forecasts, essentially represents the icing prediction by “vote of the majority.” Both the POD and VOL in TLE-Median are considerably lower than those of the deterministic forecasts, indicating that there is a relatively small region where any four (or more) of the seven HRRR cloud fields overlap. This has lead us to experiment with TLE-Median2 and TLE-Median3, in which the TLE icing fields are composed by taking, respectively, the second or third largest LWC values forecast by the seven members at each grid point. That is, the icing volume in TLE-Median2 (TLE-Median3) represents the regions where at least two (three) members predict icing. The skills and volumes of TLE-Median3 are similar to those of the deterministic HRRR forecasts, while TLE-Median2 gives higher POD and POFD/VOL. That is, the overlapping cloudy regions by at least three of the seven members are comparable in size to the cloudy region in any arbitrary deterministic member, while the overlapping cloudy regions by at least two members are generally larger than that in any member forecast.

Also indicated in Fig. 7 are the POD and POFD values by the *a posteriori best and worst* member. These are the evaluation scores achieved by selecting the most and least skillful deterministic forecasts for each hour. The knowledge as to which member will perform best (or worst) for a certain forecast hour, of course, is not *a priori* available. The *a posteriori best and worst* skills are calculated here only as proxies to the upper and lower limits of the predictability of the current deterministic

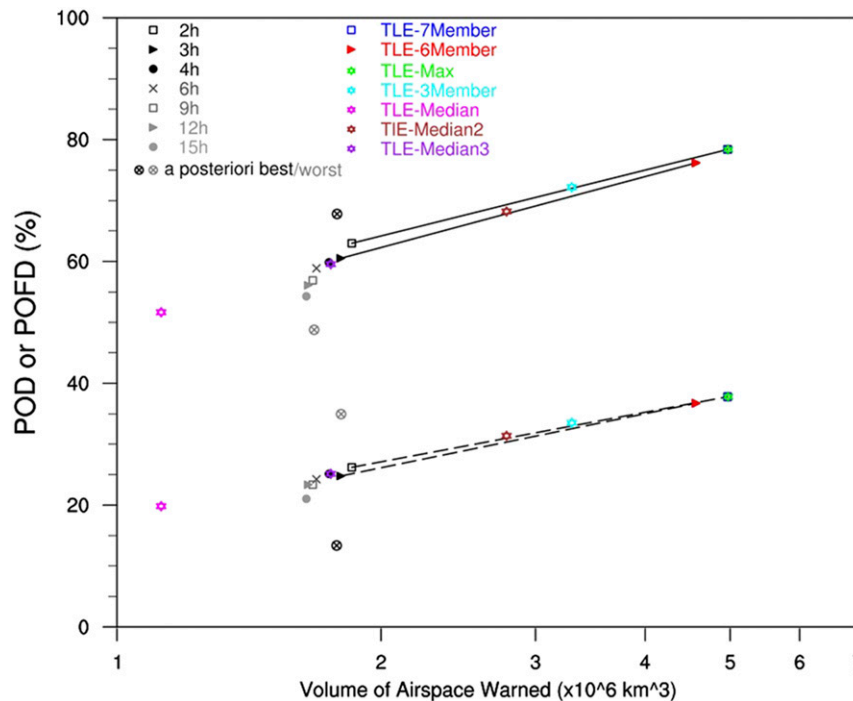


FIG. 7. POD and POFD vs the forecast icing volumes (VOL) for 2-, 3-, 4-, 6-, 9-, 12-, and 15-h HRRR forecasts (gray), and various HRRR-TLE forecasts (colors) using alternative TLE methods, for the month of December 2016. The a posteriori best (worst) skills are obtained by selecting the best (worst) performing member forecast at each hour. The lines connecting the deterministic and TLE points illustrate the trends and do not represent data points in the interim.

HRRR forecasts. As at any given validation hour the forecasts with shorter lead time were more likely the ones with higher POD, the 2-h forecast was most frequently found to be the a posteriori *best* member for POD. In general, all the TLE methods tested showed trade-offs between POD and POFD, and future icing prediction methods should be sought to achieve better POD while maintaining POFD as low as possible.

6. HRRR-TLE versus FIP and RAP

To illustrate the value of HRRR-TLE for icing forecasting, its skills were examined in the context of the current operational icing forecasts: FIP. The skills of HRRR forecasts were also compared with those of RAP forecasts, both deterministically and with TLE applied, to give insight to how the model grid resolution affects the icing prediction capabilities. This exercise was done in the rerun mode since additional evaluations beyond those in the real-time test were required. Similar to the rerun of HRRR, RAP forecasts were obtained and evaluated at four validation times daily, at 1500, 1800, 2100, and 0000 UTC, during the month of December 2016. The same TLE methodology as for HRRR-TLE,

the weighted averaging method, was used to combine deterministic RAP forecasts with a lead time of 2, 3, 4, 6, 9, 12, and 15 h to form RAP-TLE.

Figure 8 shows the PIREPs validation results of 3-h FIP forecasts at various thresholds of icing probability, as well as results for 3-h HRRR and HRRR-TLE, and 3-h RAP and RAP-TLE. Given that FIP/RAP and HRRR have different domain and grid spacing, care was taken to make their skill representations as consistent as possible. When the FIP and RAP forecasts were validated against PIREPs, only the domain area that matches the HRRR domain was considered. The POD and POFD of FIP/RAP were evaluated by searching 4 grid points surrounding each PIREP, and the VOL for FIP/RAP was defined as the sum of all 13-km grid boxes having $SLW > 10^{-6} \text{ g m}^{-3}$. Once again it should be noted that the VOL for HRRR and HRRR-TLE in Fig. 8 are defined differently from those presented in Figs. 4–7 (as previously described in section 3a). However, the POD and POFD calculations for HRRR were not modified for comparison to FIP/RAP, and still required any of the 64 HRRR grid cells surrounding the PIREP to have SLW above the threshold to be scored as a hit. Likewise, for negative icing all 64 points had to

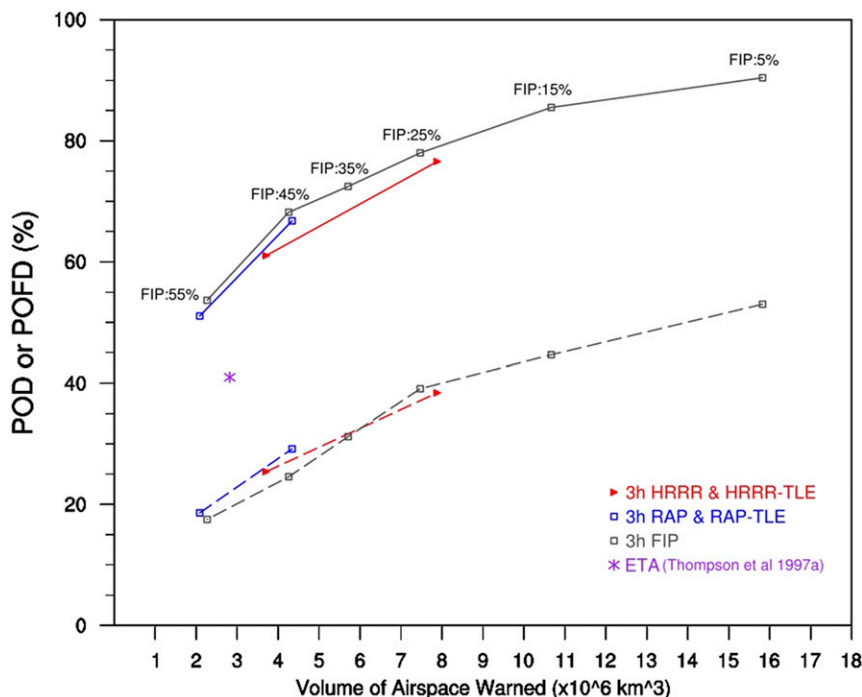


FIG. 8. POD (solid lines) and POFD (dashed lines) vs volume of airspace warned (VOL) by 3-h HRRR and HRRR-TLE (red), 3-h RAP and RAP-TLE (blue), and 3-h FIP (gray) at a number of thresholds of icing probability, for the month of December 2016. The purple asterisk marks the POD for the Eta model from Thompson et al. (1997a).

be free of SLW to be considered a hit (POD_{no}). No modifications were made to this methodology because the scoring occurs over roughly the same volume between the two model grids ($24 \times 24 \text{ km}^2$ from the HRRR and $26 \times 26 \text{ km}^2$ from the FIP/RAP).

Apparent from Fig. 8 is that for all of the models or forecast methods, as POD increases, so does VOL. The result is a trade-off between POD and POFD. The TLE averaging increases the POD and POFD for both RAP and HRRR, with HRRR-TLE giving considerably higher POD and POFD than RAP-TLE. The volume of airspace warned in 3-h HRRR-TLE nearly doubles the value in 3-h RAP-TLE. For the probability based FIP, as the probability threshold gets higher, the VOL/POFD decreases and so does the POD. The FIP at 55% probability yields similar though slightly higher POD and VOL as RAP, while FIP at 45% probability slightly outperforms the RAP-TLE with a similar VOL but lower POFD.

Between 3-h HRRR-TLE and 3-h FIP, the skill of HRRR-TLE is roughly equivalent to FIP at 25% icing probability threshold in terms of both POD and POFD. Even though the TLE procedure increased the overall regions of predicted icing conditions compared to the deterministic HRRR, the total volume of airspace warned in the HRRR-TLE forecast remain reasonable

without being overly exaggerated in the context of FIP icing probability forecasting. On the other hand, in order for HRRR-TLE to add more significant value to the FIP operational forecasts at higher icing probability thresholds, methods should be developed to filter out the regions of low icing probability and to lower the POFD in HRRR-TLE. Thompson et al. (1997a) have evaluated the formerly operational Eta model versus PIREPs in a very similar manner and found a POD of 41% with a corresponding volume of $2.86 \times 10^6 \text{ km}^3$ (shown with a star in Fig. 8). Explicit numerical weather prediction of supercooled liquid water has since come a long way.

It should be pointed out that part of the reason for HRRR forecasts to have predicted considerably larger VOL than RAP is because that, in calculating the VOL values shown in Fig. 8, the horizontal resolution of the HRRR LWC field was effectively downgraded from 3 to 12 km. The VOL of HRRR forecasts defined as the total volume of all 3-km grid boxes with predicted icing (Fig. 7) is much lower than that of the total volume of all 12-km boxes with icing, and is more comparable to the VOL of RAP. The HRRR and HRRR-TLE forecasts could be evaluated at higher horizontal resolution in the future. It would also be interesting to test whether a HRRR-based FIP will be able to improve upon the skill of the RAP-based FIP.

7. Summary and conclusions

With relatively high resolution (3-km grid spacing) and explicitly predicted supercooled liquid water, the HRRR model is being tested on its capability for providing direct icing forecasts for aircraft and ground icing applications. Multiple HRRR forecasts available at any valid time from consecutive cycles make it possible to construct time-lag-ensemble forecasts. A TLE method based on weighted averaging was implemented in a 3.5-month-long real-time test and icing forecasts were evaluated using observational datasets from PIREPs and METARs. The HRRR-TLE forecasts were shown to produce a higher capture rate of icing PIREPs and METAR surface precipitation type than the deterministic HRRR forecasts, even using very ad hoc duration ranks and weighting factors. However, a trade-off of the TLE technique was the increase of the false alarm rates and broadened regions of predicted overall icing areas. The volume of airspace warned by HRRR-TLE is similar to that of the 25% probability threshold in the existing operational FIP. The TLE method was also tested on RAP forecasts and shown to increase both the capture rates and false alarm rates. Alternative TLE methods were also shown to produce higher POD with higher POFD and vice versa.

For future work, satellite screening techniques similar to that in Thompson et al. (1997b) may be applied to very short-term forecasts of HRRR-TLE as possible ways to reduce POFD while maintaining high POD. Such techniques diagnose cloud-free regions and remove predicted icing regions in the forecasts where there are no subfreezing cloud tops in the satellite analysis. Another logical application of the TLE method is the inherent probabilistic forecast information that could be derived from the ensemble member forecasts for icing prediction. Additionally, a true HRRR ensemble system (HRRR-E) with multiple forecast members generated from the same cycle could easily be substituted for the TLE ensemble members (or combinations thereof). In fact, current plans call for a HRRR-E to be developed in the near future that utilize different data assimilation methods, different physical parameterizations, or different lateral boundary conditions. Theoretically, such ensembles will be able to better represent the probabilities and uncertainties in high-resolution NWP. The lessons learned and techniques developed with TLE will be valuable to the future effort of using the HRRR-E to provide more skillful icing forecasts.

Acknowledgments. This research is in response to requirements and funding by the Federal Aviation Administration (FAA) Grant DTFWA-15-D-00036. The

views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA. The authors are grateful to our colleague Sarah Tessendorf for helpful discussions, constructive suggestions, and careful review of the initial draft of the paper. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

REFERENCES

- Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, <https://doi.org/10.1175/MWR-D-15-0242.1>.
- 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 observation. *J. Appl. Meteor.*, **44**, 969–986, <https://doi.org/10.1175/JAM2246.1>.
- Brown, B. G., G. Thompson, R. T. Bruintjes, R. Bullock, and T. Kane, 1997: Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Wea. Forecasting*, **12**, 890–914, [https://doi.org/10.1175/1520-0434\(1997\)012<0890:IOIFIA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1997)012<0890:IOIFIA>2.0.CO;2).
- Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 2014: An empirical model for assessing the severe weather potential of developing convection. *Wea. Forecasting*, **29**, 639–653, <https://doi.org/10.1175/WAF-D-13-00113.1>.
- FAA, 2014: U.S. Code of Federal Regulations, Title 14 (Aeronautics and Space), Part 25 (Airworthiness Standard: Transport Category Airplanes), Section 25.1420—Supercooled large drop icing conditions. Office of the Federal Register, <http://www.ecfr.gov>.
- Forbes, G. S., Y. Hu, B. G. Brown, B. C. Bernstein, and M. K. Politovich, 1993: Examination of conditions in the proximity of pilot reports of icing during STORM-FEST. Preprints, *Fifth Int. Conf. on Aviation Weather Systems*, Vienna, VA, Amer. Meteor. Soc., 282–286.
- Landolt, S., A. J. Schwartz, A. Gaydos, and S. DiVito, 2017: Impacts of the implementation of the Automated Surface Observing System (ASOS) on the reports of precipitation type in airport terminal areas around the United States. *18th Conf. on Aviation, Range, and Aerospace Meteorology*, Seattle, WA, Amer. Meteor. Soc., 8.4, <https://ams.confex.com/ams/97Annual/webprogram/Paper313277.html>.
- McDonough, F., B. C. Bernstein, M. K. Politovich, and C. A. Wolff, 2004: The forecast icing potential (FIP) algorithm. Preprints, *20th Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Seattle, WA, Amer. Meteor. Soc., 231–238.
- Ramsay, A. C., 1999: A multi-sensor freezing drizzle algorithm for the automated surface observing system. Preprints, *15th Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Dallas, TX, Amer. Meteor. Soc., 193–196.
- Schultz, P., and M. K. Politovich, 1992: Toward the improvement of aircraft icing forecasts for the continental United States. *Wea. Forecasting*, **7**, 491–500, [https://doi.org/10.1175/1520-0434\(1992\)007<0491:TTIOAI>2.0.CO;2](https://doi.org/10.1175/1520-0434(1992)007<0491:TTIOAI>2.0.CO;2).

- Schwartz, B., 1996: The quantitative use of PIREPs in developing aviation weather guidance products. *Wea. Forecasting*, **11**, 372–384, [https://doi.org/10.1175/1520-0434\(1996\)011<0372:TQUOPI>2.0.CO;2](https://doi.org/10.1175/1520-0434(1996)011<0372:TQUOPI>2.0.CO;2).
- Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., <https://doi.org/10.5065/D68S4MVH>.
- Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.*, **71**, 3636–3658, <https://doi.org/10.1175/JAS-D-13-0305.1>.
- , R. Brientjes, B. Brown, and F. Hage, 1997a: Intercomparison of in-flight icing algorithms. Part I: WISP94 real-time Icing Prediction and Evaluation Program. *Wea. Forecasting*, **12**, 878–889, [https://doi.org/10.1175/1520-0434\(1997\)012<0878:IOIFIA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1997)012<0878:IOIFIA>2.0.CO;2).
- , R. Bullock, and T. F. Lee, 1997b: Using satellite data to reduce spatial extent of diagnosed icing. *Wea. Forecasting*, **12**, 185–190, [https://doi.org/10.1175/1520-0434\(1997\)012<0185:USDTRS>2.0.CO;2](https://doi.org/10.1175/1520-0434(1997)012<0185:USDTRS>2.0.CO;2).
- , M. K. Politovich, and R. M. Rasmussen, 2017: A numerical weather model's ability to predict characteristics of aircraft icing environments. *Wea. Forecasting*, **32**, 207–221, <https://doi.org/10.1175/WAF-D-16-0125.1>.
- Wolff, C., and F. McDonough, 2010: A comparison of WRF-RR and RUC forecasts of aircraft icing conditions. *14th Conf. on Aviation, Range, and Aerospace Meteorology*, Atlanta, GA, Amer. Meteor. Soc., 6.2, <https://ams.confex.com/ams/90annual/webprogram/Paper160083.html>.