August 2018

Using Advanced Post-processing Methods with the HRRR-TLE to Improve the Prediction of Cold Season Precipitation Type

Timothy Thielke
University of Wisconsin-Milwaukee

USING ADVANCED POST-PROCESSING METHODS WITH THE HRRR-TLE TO IMPROVE THE PREDICTION OF COLD SEASON PRECIPITATION TYPE

by

Timothy Thielke

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Atmospheric Science

at

The University of Wisconsin – Milwaukee

August 2018

ABSTRACT

USING ADVANCED POST-PROCESSING METHODS WITH THE HRRR-TLE TO IMPROVE THE PREDICTION OF COLD SEASON PRECIPITATION TYPE

by

Timothy Thielke

The University of Wisconsin-Milwaukee, 2018
Under the Supervision of Professor Paul Roebber

In this study we explore advanced statistical methods with the operational High-Resolution Rapid Refresh Model (HRRR) Time-Lagged Ensemble (TLE) to improve the prediction of cold season precipitation type. TLEs are a computationally efficient method to provide a slightly improved probabilistic forecast as the differences between model runs are an approximation of initial condition uncertainty. We apply evolutionary programming, weight-decay bias correction, and Bayesian Model Combination with fifteen HRRR forecast variables that potentially relate to precipitation type for station locations in the contiguous United States that are along and to the east of 100° W longitude to obtain probabilistic precipitation type forecasts. These methods are shown to provide improved probabilistic information for both the areal distribution of cold season precipitation and the timing and location of phase transitions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I'd like to thank my advisor, Dr. Paul Roebber, for all of his help and guidance throughout my thesis work. Your expertise in machine learning and in the field of meteorology is extraordinary and I look forward to our work in the future. I would also like to thank the UW-Milwaukee atmospheric science department for allowing me to come to Milwaukee and providing funding for me through teaching assistantships. This project was in part supported by the UCAR Developmental Testbed Center Visitors program and benefited from collaboration with NOAA/ESRL scientists Trevor Alcott and Isidora Jankov. Next, a big thank you to Dr. Vince Larson and Dr. Clark Evans for agreeing to be on my committee and providing useful insight for my thesis work. I would like to thank Andrea Honor for putting up with me and standing by me through the good and bad times. I know for certain that I couldn't have done this without your support. Finally, I would like to thank my family. I would not be where I am today without their guidance, care, and support they have selflessly provided since day one.

I. INTRODUCTION

Freezing rain, although usually short-lived, brings significant societal and economic risks by causing hazardous travel conditions and also can damage to the power infrastructure. In the United States, freezing rain can occur anywhere. In February of 1994 an intense ice storm struck several states in the Southeastern region of the United States causing over 3 billion dollars' worth of damage, killing nine people, and bringing mass power outages the effected over 2 million people, Lott and Sittel (1996). Costly ice storms are not just limited to the southeastern U.S. In January of 1998 southeastern Canada and a number of states in Northeastern U.S. and were also impacted by a significant ice storm that left some areas with over an inch of ice that brought 3 billion dollars in damages to Canada and at least 1.4 billion dollars in damages for the U.S. while a total of 56 lives were lost, Lott et al. (1998). These two ice storms, amongst many others, show the importance providing an accurate forecast so actions can be taken to save lives, property, and money. For example, with a better forecast, energy companies would be able to prepare for mass outages by placing company employees in strategic locations to return power to their customers much more quickly. Unfortunately, forecasting precipitation type continues to be a challenge due in part to the need to specify the details of the temperature and moisture profile. According to Cortinas (2000), for freezing rain cases in the Great Lakes region, temperature anomalies warmer than 0°C are typically observed between 850 and 750 hPa while sub-freezing temperatures are observed at the surface (Figure 1.). For such a structure to persist, warm-air advection is needed aloft to counteract the cooling caused by the melting of snow and evaporation of rain. At the same time, cold air advection is needed at the surface to offset latent heat release from the liquid water that freezes on contact with surfaces near the ground.

In this study we focus on forecasting the areal coverage and duration of three precipitation types: rain, snow, and freezing rain. Spatially and temporally, snow and rain are the most common - freezing rain usually occurs only for a short duration and over a relatively small area constituting a transition between snow and rain (Stewart 1992). The longer freezing rain persists, however, the costlier the storms become such as with the long-duration 1998 ice storm, Lott et al. (1998). Recent improvements in precipitation type have been seen with advances in numerical weather prediction (Ikeda et al. 2013).

The High-Resolution Rapid Refresh (HRRR) provides short-range, rapid updates with hourly forecasts. The National Oceanic and Atmospheric Administration's (NOAA) Earth System Research Laboratory (ESRL) generate a Time Lagged Ensemble (TLE), hereafter referred to as the HRRR-TLE, using the three most recent hourly model runs for its ensemble members. The idea in generating ensembles is to represent the uncertainty in the initial conditions of the atmospheric state and thus provide the range of possible outcomes, ideally sufficiently calibrated to provide a reliable probability distribution (e.g. Grimit and Mass 2002). However, there are a number of methods for applying initial condition perturbations and many of these impose substantial additional computational cost. The concept of a TLE, first proposed by Hoffman and Kalnay (1983), is to provide a low-cost alternative that still is sufficient to provide useful probabilistic information.

Unfortunately, ensembles are known to be under dispersive (Hamill and Whitaker 2007; Novak et al. 2008) suggesting that post-processing methods may help to produce improved information. One such approach is Model Output Statistics (MOS; Glahn and Lowry 1972) which uses multiple linear regression to fit model variables to observed quantities. In section 2, we will review a few such approaches that can be applied to the HRRR-TLE to improve the

prediction of precipitation type, while maintaining low computational requirements so that those procedures can be applied to operational forecasts. Additionally, we will discuss the observations and forecasting data used in these approaches. Section 3 presents the results, including a example through presentation of a cold season cyclone case study, while section 4 provides a discussion of said results and proposes potential areas of future work.

II. METHODS

The HRRR-TLE dataset was provided by NOAA ESRL for the purpose of this study and covers the period November 2013 through February 2017. This data was restricted to the cold season months (November-March) for the purposes of this study. During that period, 29 cyclones were identified and constitute the set of cases to be consider. We focus on cyclones that track from the Gulf of Mexico or Colorado and only consider those events in which more than one precipitation type was observed during the cyclone lifetime. Observations were gathered from the regular synoptic surface observing network, and include National Weather Service Offices, manual stations, and automated surface observing system (ASOS) stations. Only those stations located along and to the east of the 100°W longitude were used in this study to verify precipitation type. The HRRR-TLE uses the three most recent hourly model runs for its ensemble members to forecast the next six hours. For example, in Figure 2. there are three HRRR temperature forecasts for 1800 UTC on the 10th of January 2016. These forecasts represent the 1200 UTC HRRR-TLE model run where the 0800, 0900, and 1000 UTC HRRR forecasts are used as the ensemble members where the most lagged member, 0800 UTC, is member one, and the least lagged member, 1000 UTC, is member 3. As there are 4 initiating time periods for the hourly HRRR-TLE (e.g. 0000 UTC, 0600 UTC, 1200 UTC, 1800 UTC),

there is a forecast for each hour in a day. Over the span of our dataset, the HRRR-TLE model has been updated and modified which could present some errors due a mismatch in the training and testing data. However, the research done in this study is meant to be adaptive so that it can be applied to a model that is being updated in real-time. The HRRR-TLE provides forecasts for twenty surface variables of which fourteen are selected for use in this study, based on their potential relevance to determining precipitation type. Only surface variables were considered for simplicity, although it would prove useful to have a 3-D dataset as we know that temperatures and moisture profiles in the vertical do have an influence on precipitation type probabilities. That being said, there are several variables that are integrated from vertical profiles and thus we are capturing some of that information. An additional, derived variable is included which categorizes the temperature relative to the freezing point (Table 1.). For example, if the temperature is between 0°C and 1°C it would be categorized with a value of 1. On the other hand, if the temperature at the ground was below -1°C then the categorical temperature would be -2. After determining the categorical temperature, we standardize each HRRR-TLE variable based on the given variable's mean value and standard deviation determined by the subset of data used for training. Standardization was not applied to categorical, probabilistic, percentage variables. The resulting standard anomalies were then applied as inputs in our upcoming methods.

To directly compare the 3-km gridded HRRR-TLE forecasts to a station observation, a bi-linear interpolation was applied, based on the nearest four grid point locations. All cases where no precipitation was reported are removed. Next, a random filtering process is applied to each observation and forecast in order to thin the number of snow and rain cases such that an approximate balance in numbers between freezing rain, rain, and snow cases remain. After the filtering was complete approximately 1,000 observations and associated forecast of each

precipitation type are left for the study. The first half of these are used for training and the following 20% are used for cross validation while the remaining 30% are applied as testing cases. Only dates were used to determine what is training, cross validation, and testing.

Evolutionary programing (EP), first introduced by Roebber (2010, 2013), is the method which is used to "map" the HRRR forecasts to the observations. In the present study, the EP algorithms are two logistic regression equations, formed from two sets of five IF-THEN equations, each composed of standardized variables, three operators, and three variable coefficients that are structured like algorithms seen in Table 5. One set of If-THEN equations are used to determine the probability for freezing rain, while the other determines snow probabilities. Any residual probability is then classified as rain. If there is an instance where an IF statement is never true, it is excluded from analysis. The two sets of five IF-THEN algorithms are then used in a logit equation for each of the precipitation types to determine probability:

$$P = \frac{e^x}{(1 + e^x + e^y)} \tag{1}$$

Where *x* and *y* represent the sum of results from two sets IF-THEN equations used to determine the probability of a given precipitation type, *P*. After this the sum of the probabilities for freezing rain, snow, and rain sum up to 1.

Initially we generate 10,000 random algorithms and then allow them to train following the evolutionary protocol: evaluate, thin, reproduce (with mutations), repeat. The measure of success in this instance is the Brier Score, Brier (1950). In this protocol, the worst performing 20% of the algorithms are removed from consideration while the top 20% are reproduced (through cloning and mutation) and the progeny are used to replace the worst performers.

Cloning and mutation were first implemented by Roebber (2015) and is an important process to allow for the propagation of the best solutions while still allowing for the introduction of innovations. Here, we apply mutations more aggressively than in Roebber (2015), with mutations occurring with the production of every new algorithm.

After the full set of 10,000 algorithms are restored through this process, the cross-validation Brier Score is used to define and store the top 100 performing algorithms. This list is kept and updated throughout the training process. A total of 300 generations are processed, and the full initialization and training is run again for a subsequent 300 generations with the exception that the top performer list is maintained and updated only when a new algorithm has sufficient performance to make this list. This procedure is followed for a total of 5 sets of 300 generations. The rationale for this procedure is to allow for a robust search of the phase space in order to define the best algorithms. In a parallel study, M.S. student Jesse Schaeffer, under the direction of Professors Roebber and Evans, is using a similar procedure in order to train algorithms to forecast tropical cyclone intensity as part of the Joint Hurricane Testbed (Roebber 2018, personal communication). The overall procedure is applied to each of the three HRRR-TLE members, yielding a total of 300 EP algorithms to be used in the next step.

Next, a decaying average bias correction following Cui et al. (2012) is applied to the probabilities produced by the EP algorithms. The bias correction equation simply applies weights to both the previous bias correction and the current error with the majority of the weight on the previous bias correction, but still considering current error by placing a small amount of weight on the current bias. In this study, we apply 95 percent of the weight on the previous bias with only a 5 percent weight placed on the current error. Cui was able to find improvement by applying this decaying bias correction to the Global Ensemble Forecast System up to the 7^{th} day

forecast. As we are only interested in a six-hour forecast, applying such a method to our EP ensemble members yields promising results. After the bias correction is applied to the forecast, we normalize the probabilities once again so the sum of three precipitation type forecasts do not exceed a value of 1.

Finally, Bayesian model combination (BMC) is applied to a select few bias corrected ensembles. BMC is similar in many ways to the more commonly known Bayesian model averaging (BMA; Raftery et al. 2005), procedure in that both BMA and BMC search through the available algorithms and place weights on said algorithms to provide the best forecast. The difference between the two is that BMA tries to locate the data generating ensemble member, or truth, and optimizes the weights so that the forecast reflects that ensemble member. BMC, on the other hand, does not assume that one of the available members is the data generating model, but instead tries to find the best combination of the available ensemble members to find the optimal forecast. Monteith et al. (2011) found that BMC outperforms BMA across a variety of datasets considered. Roebber (2015) found that BMC in combination with bias correction, when applied to members of the GFS MOS ensemble, substantially improved forecasts for surface temperature.

One limitation of BMC is that the computation costs increase exponentially with the number of ensemble members, such that, for example, if ten ensemble members are evaluated using four possible raw weights, over 1,000,000 (4^{10}) combinations need to be evaluated. In order to reduce the ensemble members to a tractable number, we use the Brier Score to rank each of the 100 members from each of the three HRRR-TLEs. Since, as noted previously, ensembles tend to be under dispersive, and the same is true of EP ensembles (Roebber 2015), we choose the best ranking EP algorithm from each TLE member and then we choose the next best ranking

algorithm that has a Brier score difference from the best performing member that is greater than the 25th percentile. The third selected algorithm is the next best ranked that has a difference greater than the 50th percentile. By repeating this for each of the three TLE members we obtain a total of nine EP algorithms that will be used as our forecasts with the weights applied to each as determined by the BMC method. While running through all the possible weights that can be applied to these nine algorithms, we systematically calculate the posterior probability for a given combination using training data, similar to that of Monteith et al. (2011). The combination with the least amount of error is the final selected weighting scheme.

We compare the performance of the BMC to any individual EP algorithm, or to the HRRR-TLE forecast member, using the Brier Score (for probabilities), the Heidke Skill Score (HSS; Panofsky and Brier 1958) for the full 3x3 deterministic forecasts (obtained from the maximum individual category probability) and standard 2x2 contingency measures for individual precipitation type forecasts. These latter measures are the critical success index (CSI), probability of detection (POD), bias, and false alarm rate (FAR). The equations used for CSI, POD, bias, and FAR are:

$$POD_{type} = \frac{HITS_{type}}{(HITS_{type} + MISS_{type})} \tag{2}$$

$$CSI_{type} = \frac{HITS_{type}}{(HITS_{type} + MISS_{type} + FA_{type})} \tag{3}$$

$$BIAS_{type} = \frac{HITS_{type} + FA_{type}}{(HITS_{type} + MISS_{type})} \tag{4}$$

$$FAR_{type} = \frac{FA_{type}}{(HITS_{type} + FA_{type})} \quad (5)$$

The definition for *Hits*, *Miss*, and *FA* used in the aforementioned measures can be seen in Table 2. The HSS associated with chance was also applied to the forecast based on its performance across all precipitation types.

$$HSS = \frac{Hits_{all} - Chance}{(Total\ Fcst. - Chance)} \quad (6)$$

$$Chance = \frac{(O_{ZR} * F_{ZR}) + (O_{SN} * F_{SN}) + (O_{RN} * F_{RN})}{(Total\ Fcst.)} \quad (7)$$

Here chance is determined by the multiplying total observation of a given precipitation type by the total forecasts of that given precipitation type and summing for all precipitation types. That value is then divided by the total number of forecasts generated by that member. Chance is then applied in the HSS formula by subtracting from all forecasts that were correctly observed, $Hits_{all}$, and then divided by the total number of forecasts that are also subtracted by chance. The results of this analysis are found in Table 3 and are based solely on the testing dataset. To find these scores, we convert the probabilistic results into a deterministic forecast by selecting the max probability at a given time and location. For the HSS, POD, and CSI the higher values

represent the better forecasts. In the case of Brier score and FAR the opposite is true so that better forecast is the one with the smaller values. As for bias, if a value equals 1 then it shows and unbiased forecast while greater than 1 represents an over-forecast and thus values less than are an under-forecast.

III. RESULTS

a. Performance and analysis of ensemble members

The results shown in this section are based on the independent test data only, not the training and cross validation data which were used in various stages of EP algorithm development and selection. The procedures described in section 2 produced a total of 300 algorithms, most of which independently outperformed the individual HRRR-TLE member forecast from which they were derived, but usually at the cost of losing skill in forecasting one of the three precipitation types. Before being bias corrected, EP member 40 (M40B), derived from HRRR-TLE member 2 (the member with a lag ranging from 3 to 9 hours), performs well with forecasts in rain and snow, but it's ability to predict freezing rain decreases. After bias correction, EP member 40 (M40A) gains a large increase in its ability to forecast freezing rain while also keeping the POD and CSI of rain and snow relatively high and lowering the FAR across all three precipitation types. The general increase in skill from M40A is also seen in its HSS and Brier score (Table 3).

After the bias correction, M40A becomes the best performing EP member out of all 300 possible algorithms, but it is not selected by the BMC process (Table 4; note that the BMC selection process cannot reference the test data, which is kept strictly segregated from all aspects of training and calibration). Interestingly, the BMC weighting process only chose the three EP

members that were derived from HRRR-TLE member 3, the least time-lagged member with forecasts ranging from 2 to 8 hours. In practice, that means that the EP post-processed HRRR is not strictly a time-lagged ensemble but reflects the reality that the best forecasts are often the most updated. Further, we note that since NOAA is moving towards a full (non time-lagged) HRRR ensemble, the technique successfully employed here can likely be profitably applied to that modernized version of the HRRR ensemble.

For the present study, an equal weighting of algorithms 15 (M15), 18 (M18), and 72 (M72) of HRRR-TLE member 3 was selected as the most optimal combination of algorithms. Figure 3 shows a performance diagram (Roebber 2009) that directly compares the success ratio, POD, bias, and CSI for all three precipitation types of M40B, M40A, HRRR-TLE member 3, and the weighted combination of M15, M18, and M72. Using these diagrams, the improvements, or lack thereof, can be seen from the applications of the bias correction and the weighted combination determined by BMC. For example, the aforementioned increased ability in M40 after bias correction can be seen as the member's POD value drastically increases without creating more false alarms seen in Figure 3a. In Figure 3c., however, a direct comparison of the weighted combination versus M40A, shows that the weighted combination performs only slightly worse than M40A. Altogether this is a positive result, as it indicates that without *a priori* knowledge the weighting process is able to largely match the best performance, which is in itself considerably superior to that of the HRRR.

In Table 5 we break apart M15, M18, and M72 into the associated IF-THEN equations allowing for a more in-depth analysis of how the members produce their forecasts. The ability to interpret the forecast logic is one major advantage of this form of EP relative to many other types of machine learning. M15 appears to specialize in probabilistic snow forecasts (POD=0.9116,

CSI=0.6696) depending largely on surface temperatures being at or below freezing (based on the temperature category) modulated by several variables, most importantly moisture availability, with drier conditions promoting higher probabilities. This can be seen by noting that with all variables at zero anomaly but TCAT = -1, the snow probability increases from 0.217 to 0.781, and further to 0.809 with a negative anomaly in precipitable water (-1).

M18 and M72, on the other hand are oriented more towards freezing rain probabilities, with POD ~ 0.63 and CSI ~ 0.45, both superior to M15 in that category, but less effective than M15 in the other two. In M18, an anomalous (northerly) wind strongly affects the probability of freezing rain. For example, with all anomalies set to zero the probability of freezing rain increases from 0.253 to 0.418 when V=-1. Notably, if the HRRR is forecasting a probability of ice pellets, the freezing rain probability increases further. Thus, M18 appears to be emphasizing conditions north of a warm frontal boundary in the presence of warm air advection but where a cold layer is present.

Although M72 also specializes in freezing rain, it arrives at its probabilities using a different variable emphasis (e.g., the HRRR ice pellet probability and the meridional wind anomaly do not matter). Here, the focus is on overall precipitation production, particularly in the instance where the HRRR predicts some chance of freezing or frozen precipitation other than ice pellets. Consider, for example, a case where mixed precipitation is forecast by the HRRR: freezing rain (30%), ice pellets (10%), snow (40%), and rain (20%). In the absence of an anomalous meridional wind, but with the precipitation amount anomaly greater than 0.2, M18 produces an approximate 31% chance of freezing rain, 25% chance of snow, and a 44% chance of rain. M72, on the other hand, produces 51%, 21%, and 28% for these categories. If a strong negative meridional wind anomaly is present, however, indicating strong northerly flow, M18

increases freezing rain to the most likely category at 47%. This diversity in individual members of the weighted ensemble may well be a critical advantage as far as producing properly calibrated ensemble forecasts, an active area of research using evolutionary programming (Roebber, 2018, personal communication). Furthermore, as a practical matter, forecaster confidence can be increased if individual forecasts arrived at using different approaches reach similar conclusions.

b. Case Study: 16-18 December 2016

To place the overall results in specific context, we have chosen for analysis a major winter storm (16-18 December 2016) which greatly affected travel, with a fatal 55-car pileup occurring near Baltimore along with many motor vehicle accidents reported in the Midwest (TWC 2016). This case was also chosen in order to illustrate the limitations of this approach and to illustrate the ongoing challenge of making such forecasts.

At 1200 UTC 16 December 2016, an upper-level trough was positioned over the Pacific NW region of the contiguous United States (Fig. 4a). By the next day, the digging trough was bringing strong divergence aloft over the Intermountain West and southern Colorado (Fig. 4b), with surface cyclogenesis occurring in response. (Fig. 5a). By 0000 UTC 18 December, the trough had continued eastward, with a jet streak beginning to form over Illinois and stretching southwestward over Texas and New Mexico (Fig. 4c). Meanwhile at the surface, an axis of low pressure was evident along the downstream edge of the jet streak and approaching trough, producing a qausi-stationary cold front positioned from Lake Ontario southward to coastal Texas (Fig. 5b). Developing behind the cold front, a polar high-pressure system was settling over the Plains and western Great Lakes regions, with snow observed from Northern Michigan southward

into Oklahoma. A snow-to-rain transitional region also occurred along the frontal boundary with scattered reports of freezing rain over southern Illinois and northeastern Arkansas (Fig. 6a). By 1200 UTC 18 December, the 300 hPa jet streak had intensified with peak winds of 180-190 knots (Fig. 4d), producing strong divergence over Tennessee, Lake Ontario, and the Appalachians. By this time, the frontal boundary had shifted farther east and was then stretching from Upstate New York to Mississippi and Alabama (Fig. 5c). Over the southern region, the National Weather Service had issued severe thunderstorm and flood warnings for the storms forming along the cold front; lake effect snow was occurring over Western Michigan with synoptically-forced snowfall still occurring over parts of Indiana and Ohio. Over the northern Appalachians, precipitation was primarily rain, with the phase transition occurring between Ohio and Pennsylvania. At this time, there were scattered reports of freezing rain and snow in Kentucky and Tennessee, while to the north in Maine there were more widespread reports of freezing rain (Fig. 7c). By 1800 UTC the cold front was approaching coastal New England (Fig. 5d), with the 300 hPa jet streak and associated upstream trough continuing to propagate eastward (Fig. 4e). Over New England, mostly rain was occurring while the Great Lake states continued to report snow.

This winter storm produced from 3 to 14 inches of snow in the Midwest, a trace to a tenth of an inch of ice in parts of Indiana, Kansas, and Missouri and a half inch of ice in Wakeman, Ohio (TWC 2016). Meanwhile, the northeastern region of the U.S. received from 3 to 9 inches of snow along with reports of 0.3 to 0.4 inches of ice. Even places as far south as North Carolina saw trace amounts of ice accumulations.

The overall performance of the EP BMC and of the HRRR is summarized in Tables 6 and 7, respectively. This data shows that the EP BMC was relatively unsuccessful in forecasting this

event, with POD and CSI of 0.241 and 0.071 (freezing rain), 0.917 and 0.704 (snow), and 0.526 and 0.503 (rain). In comparison, the HRRR fared better, with POD and CSI of 0.381 and 0.267 (freezing rain), 0.965 and 0.950 (snow) and 0.981 and 0.929 (rain). Next, we will examine the individual hours and forecasts to gain better understanding of what happened.

At 0000 UTC on December 18th there were 6 reports of freezing rain, 2 of which were correctly forecast while the other were 4 were forecast as snow (ZR:2, SN:4, RN:0). In addition to the 6 freezing rain reports, there were 160 observations of snow (ZR:14, SN:145, RN:1) and 47 rain observations (ZR:4, SN:15, RN:28). Figure 6a depicts the observations from 0000 UTC with Figure 6b representing the BMC and Figure 6c the HRRR-TLE member 3, both of which were forecast two hours out and valid for the same time as the observations. The BMC forecast for 0000 UTC shifts the transition line farther to the east then what was observed while also placing a high probability of freezing rain to occur over Oklahoma. By 0600 UTC 8 (Figure 7a) reports of freezing rain are observed (ZR:0, SN:6, RN:2), alongside 120 snow observations (ZR:4, SN: 112, RN:4), and 74 rain observations (ZR:15, SN:25, RN:34).

Figure 8a and b compare the observations to the forecasts once again with a similar trend as seen at 0000 UTC. The transition line is forecast farther east with large portion of the rain observations being forecast as either snow or freezing rain. Moving on to 1200 UTC observations, 14 stations identified freezing rain (ZR:5, SN:3, RN:6), 50 identified snow (ZR:3, SN:47, RN:0), and 88 reported rain (ZR:17, SN:29, RN:42). The trend continues in Figures 9a and b as the transition is falsely placed farther to east. Many rain observations are falsely identified as freezing rain with a few instances where snow is forecast. We continue see the trend where the transition line is shifted to the east for the 1200 UTC BMC forecast. For our final time, 1800 UTC, 1 station reported freezing rain, which was incorrectly forecast as snow, 55

observations of snow (ZR:5, SN:49, RN:1), and 97 rain observations (ZR:7, SN:33, RN:57). Although the 1800 UTC EP BMC forecast overall had a better grasp of the transition line, it forecast snow over Georgia, Alabama, Mississippi, and Louisiana where rain was observed.

Thus, the poor performance of the EP BMC is largely tied to misplacement of the transition line. In order to understand this better, we examine the HRRR-TLE member 3 forecast, upon which the BMC EP forecast depends. Figure 6c, 7c, 8c, 9c, represent the HRRR forecast corresponding to the observations and BMC forecasts for 0000, 0600, 1200, and 1800 UTC on December 18th, respectively. Comparing the HRRR to the observations, it is evident that the HRRR better forecast the position of the transition line.

Since the BMC selected algorithms are based in the HRRR variables and the HRRR performed well, we need to dissect the individual forecast probabilities and their variable drivers to understand this failure. At 1200 UTC, there were six instances when rain was forecast instead of freezing rain and three instances when snow was forecast instead of the observed freezing rain. A closer look at the probabilities shows that in these instances freezing rain was forecast as the second largest rather than the largest of the three possible precipitation types. In fact, in three of these forecasts the freezing rain probability was less than 1% lower than the larger probability with two more being within 10% of the largest probability. In most cases, the close miss was associated with a rain forecast as compared to a snow forecast.

Next, we look at the algorithms associated with a close miss (where the freezing rain probability was less than 1% lower than the larger rain probability) and also a large miss (where the freezing probability was greater than 10%). At the weather observing station at Rochester, NH (KDAW) at 1200 UTC 18 December 2016 the BMC forecast a 42.6% chance for freezing rain and a 43.3% chance for rain. Our forecast selection simply chooses the largest probability

and thus the forecast was wrong. M15 placed 81.5% chance for freezing rain and a 18.5% chance for rain at KDAW while M18, on the other hand, had a 14.8% chance for freezing rain and 79.5% chance for rain while M72 had much lower probabilities with a 31.6% chance for freezing rain and a 32.0% chance for rain. The BMC selection process places an equal weight on all three thus producing the final probabilities. Sensitivity tests show that a combination of anonymously low visibility alongside temperatures more than 1°C above freezing and low accumulating liquid precipitation forecasts from the HRRR yielded a 65.1% increase in the probability for freezing rain in M15. In contrast, anonymously southerly winds and a HRRR forecast for rain caused M18 to forecast low freezing rain probabilities. M72 had the most equal probabilities across all categories and thus forecasts from the input variables from there weren't substantial. M72 represented the uncertainty in the forecast providing probabilities near equal to each other between rain (32.01%) and freezing rain (31.59%). Sensitivity test reveal that the anomalies the HRRR forecast at KDAW didn't provide a substantial evidence that one precipitation was more likely than the other thus balancing the probabilities out. The weather station at Frenchville, ME (KFVE) reported freezing rain at 1200 UTC, but snow was the highest probability forecast at 55.3% with the freezing rain forecast of 23.7% by the BMC members. M15 produced a large probability of snow (91.5%) and a small probability of freezing rain (5.4%). The source of such drastic separation in probabilities in M15 stems from the HRRR forecast snow probability, which in M15 has the effect of decreasing the probability of freezing rain to zero and increasing the snow probability by 56% when holding all other variables at 0. M18 also produces the highest probability for snow at 41.0%, with a 26.3% chance for freezing rain. Similar to KDAW, the HRRR forecast anonymously southerly winds acting to hinder the chance for freezing rain in M18. Again, M72 provided roughly equal probabilities for snow, rain, and freezing rain.

In this case the EP BMC forecasts for freezing rain were quite poor, but further analysis shows that although freezing rain wasn't forecast as the max probability it still a comparatively high probability. Applying a better means for selecting probabilities may improve the forecast in instances when similar to this case. A look into the algorithms shows that two members, M15 and M18, were forecasting a more definitive probability while M72 represented the uncertainty that was present in the forecast. At KDAW, where rain was forecast, M15 had a large probability for freezing rain, but M18 had a large probability for rain and since the BMC places equal a third of the weight on each of the members, a close miss probability was forecast. At KFVE we once again see that M15 and M18 both place a moderate to large probability on snow while M72 forecast a 33% chance for rain, freezing rain, and snow further showing the uncertainties in this forecast. In these instances, it shows the importance of the members agreeing with one another to provide a confident forecast. On the other hand, if a forecaster analyzes the members individually then they can get a grasp of the uncertainty of the forecast and could still provide useful information.

IV. CONCLUSION AND FUTURE WORK

Freezing rain continues to be a major forecast challenge. Often times freezing rain only lasts for an hour or less in a region where transitions from rain to snow is occurring (Cortinas 2000). Despite this transitional nature, even short-lived freezing rain events can cause treacherous travel conditions and put a strain on the power distribution system.

One way to improve the ability to forecast for freezing rain, given the uncertainty, is to generate probabilistic forecasts using ensembles. With the TLE version of the HRRR as input, we have used Evolutionary Programming (Roebber 2010, 2013) to generate 300 algorithms

potentially to be used as ensemble forecast members. We next corrected for forecast bias using the decaying average bias correction of Cui et al (2012). The best performing members (based on Brier score) that also exhibit sufficient differences from each other (based on algorithm-to-algorithm Brier Score difference) are then weighted using the process of Bayesian Model Combination (BMC) in order to estimate the probability of snow, rain, and freezing rain.

The BMC process placed an equal amount of weight on 3 EP members that originated from the HRRR-TLE member 3, which is the least time lagged member. This suggests that as NOAA moves from the TLE version of the HRRR to a full-fledged HRRR ensemble, these techniques can be readily applied to that forecast system. Contingency tables were created to compare how the application of EP, bias correction, and BMC affected the skill of the probabilistic forecast. Figure 3 shows how the ability to forecast freezing rain was increased relative to the HRRR without compromising the rain or snow forecasts.

Given that the structure of the EP, algorithms were deliberately designed with forecast interpretation in mind (Roebber 2010, 2013), we were able to consider the forecast logic of the 3 selected EP members (Table 5). Member 15 (M15) specializes in snow forecasts with a primary focus on temperatures being at or below freezing with a secondary focus on drier conditions bringing higher probabilities. Members 18 (M18) and 72 (M72) each specialize more for freezing rain than M15, but each also place their weighting on different variables, which suggests that analyzing the members individually may provide additional forecast insight. For example, M18 places high probabilities for freezing rain when winds are anonymously northerly and the HRRR is forecasting ice pellets while M72 places a higher chance for freezing rain when the HRRR is forecasting a mixture of precipitation types.

Applying the BMC weighted forecast to a winter storm that effected the Mid-West, Great Lakes, southern U.S., and New England (Figure 7a-l), we were able to see that while the occurrence of freezing rain, snow, and rain overall were handled well in the test cases, as suggested by the skill scores, this particular event was a challenge to the system owing to a mis-location of the transition line, with the EP system placing the transition from rain to snow farther to the east than was observed. A closer look into a few stations that observed freezing but forecast either rain or snow showed some interesting results. In these select cases, freezing rain was always the second highest probability and in some instances was less than a percent lower than the max probabilistic value. Looking into two stations, one of which had a freezing rain probability less than one percent lower than the max probabilistic value selected (KDAW) and the other had a larger separation in probabilities (KFVE), we were able to see that M15 and M18 were forecasting opposite of one another. For KDAW, M15 placed high probabilities on freezing rain and moderate probabilities on rain while M18 had the opposite with higher probabilities for rain. Instead of helping to decide which forecast may be more likely, M82 forecast nearly equal probabilities for rain and freezing rain. In this case, a combination of anonymously low liquid precipitation accumulation, low visibility, and temperatures greater than freezing which originate from the HRRR forecast placed higher chances on freezing rain in M15 while M18 picked up on anonymously southerly winds and thus reducing freezing rain chances for that member. KFVE turned out to have larger probabilities in snow for both M15 and M18 while M72 didn't provide much support by forecasting equal probabilities for all precipitation types. In this instance M15 placed very large probability (91%) on snow while freezing rain kept on the low side. At KFVE the HRRR was forecasting snow and temperatures below freezing which greatly influenced M15's probability for snow. M18 gave moderate probabilities for snow and freezing rain

showing more uncertainty in the forecast. The HRRR variable that had the most influence on M18 turned out to be anonymously low precipitable water lowering freezing probabilities.

As might be expected, there are plenty of opportunities for future work. For example, more case studies should be done to understand more fully the performance of the system in a variety of synoptic contexts. What are the major sensitivities that limit predictions? In what circumstances does the system excel? Further insights into how to select EP algorithms to be used in the BMC process are needed – we have employed one reasonable approach but there is no guarantee or expectation that this is necessarily optimal. Additionally, the BMC weighting for those EP algorithms that were selected ended up discounting the information from earlier time-lagged members, but there may still be useful information contained in those forecasts. Does this relate to the initial EP selection process? Does this relate to the metric of "correctness" used in the BMC process? Would including more members in the BMC produce more robust probabilistic performance? Enlargement of the training dataset is needed – machine learning techniques are critically dependent on the training data and in general are better at interpolation than extrapolation. Having more examples for the EP to train on would allow it to take the broader variety of circumstances in which transitional precipitation in association with winter storm events occur. Would the addition of a terrain height variable improve the forecast as we know that certain topographies can improve or inhibit the chance for freezing rain? In analyzing probabilities where the forecast was wrong, were able to see that the freezing rain forecast probabilities were slightly lower than the largest probability that was selected as the forecast. It may be a good idea to find a better means to select the precipitation type forecast based on the forecast probabilities. Finally, is there a way to better use all the data that are available? In current machine learning training practice, exemplars are approximately balanced across

categories (in this case, rain, snow, and freezing rain), regardless of the underlying climatological frequency of those categories, an approach necessitated by the particular way that the "rules" are learned. Unfortunately, this sacrifices data that might be useful for training if it could be better exploited.

V. FIGURES

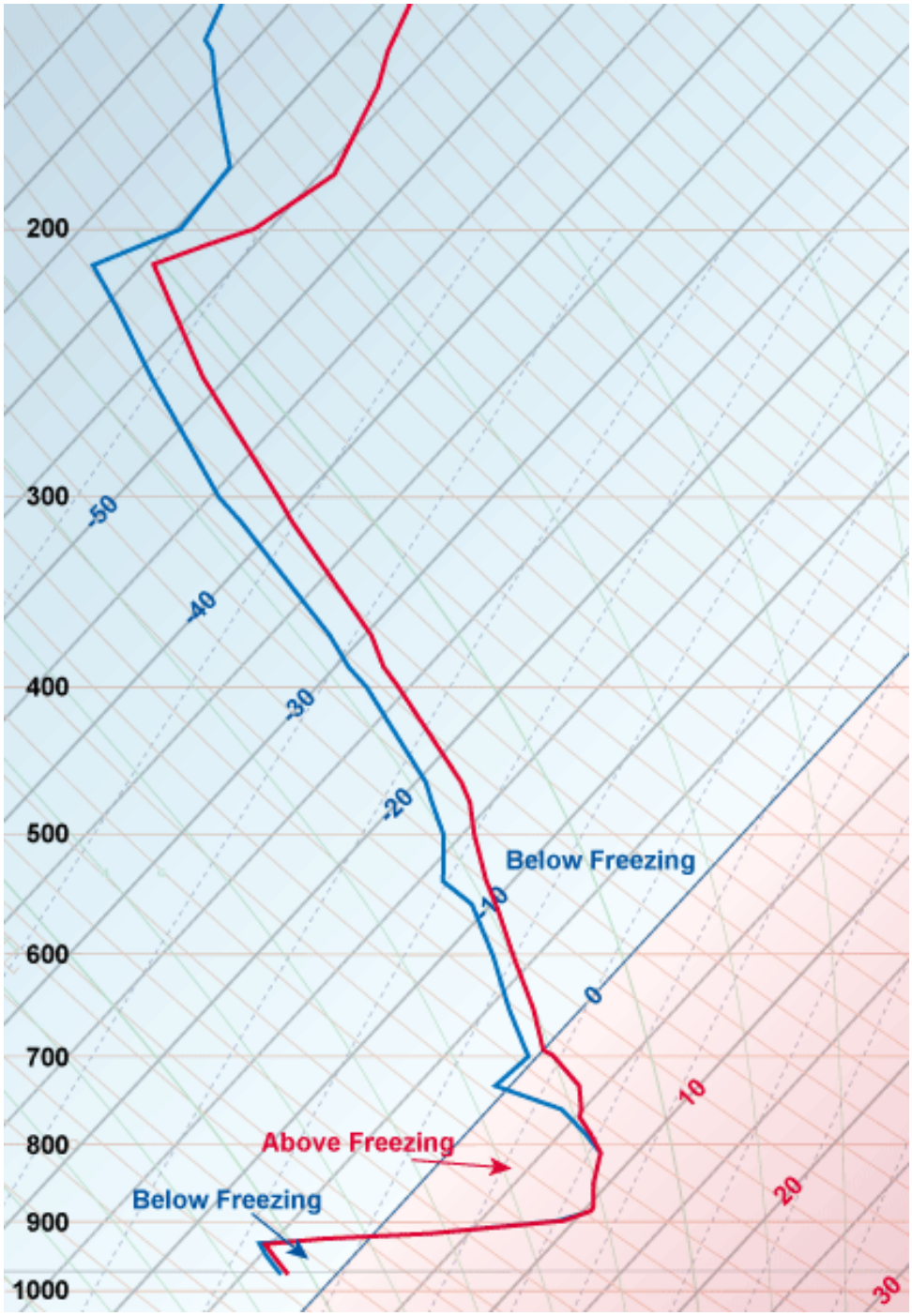


Figure 1. Conceptual skew-T diagram depicting ideal vertical profile for freezing rain. Figure gathered from https://www.weather.gov/jetstream/skewt_samples.

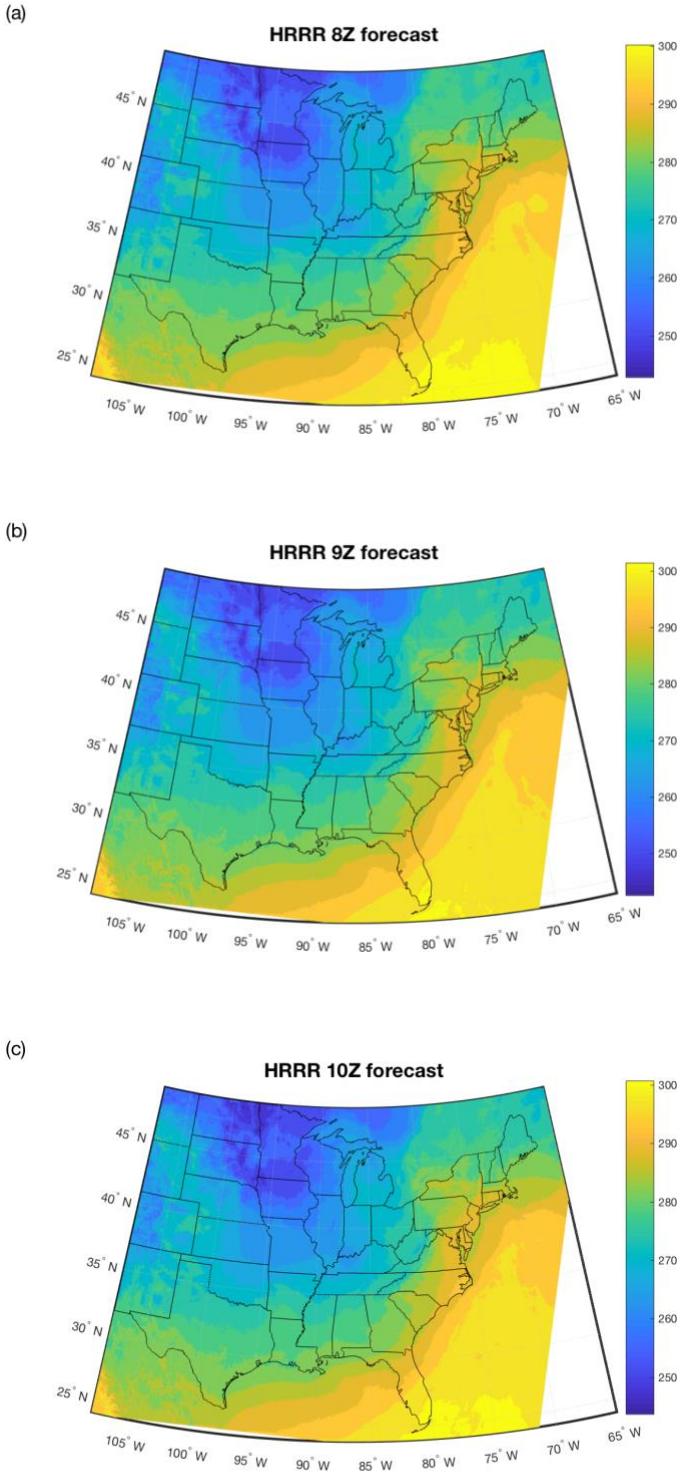


Figure 2. The 0800 (a), 0900 (b), and 1000 (c) UTC HRRR 2-m forecasts, valid for 1800 UTC on January 10th, 2016, that make up the 3 members of the 1200 UTC HRRR-TLE forecast.

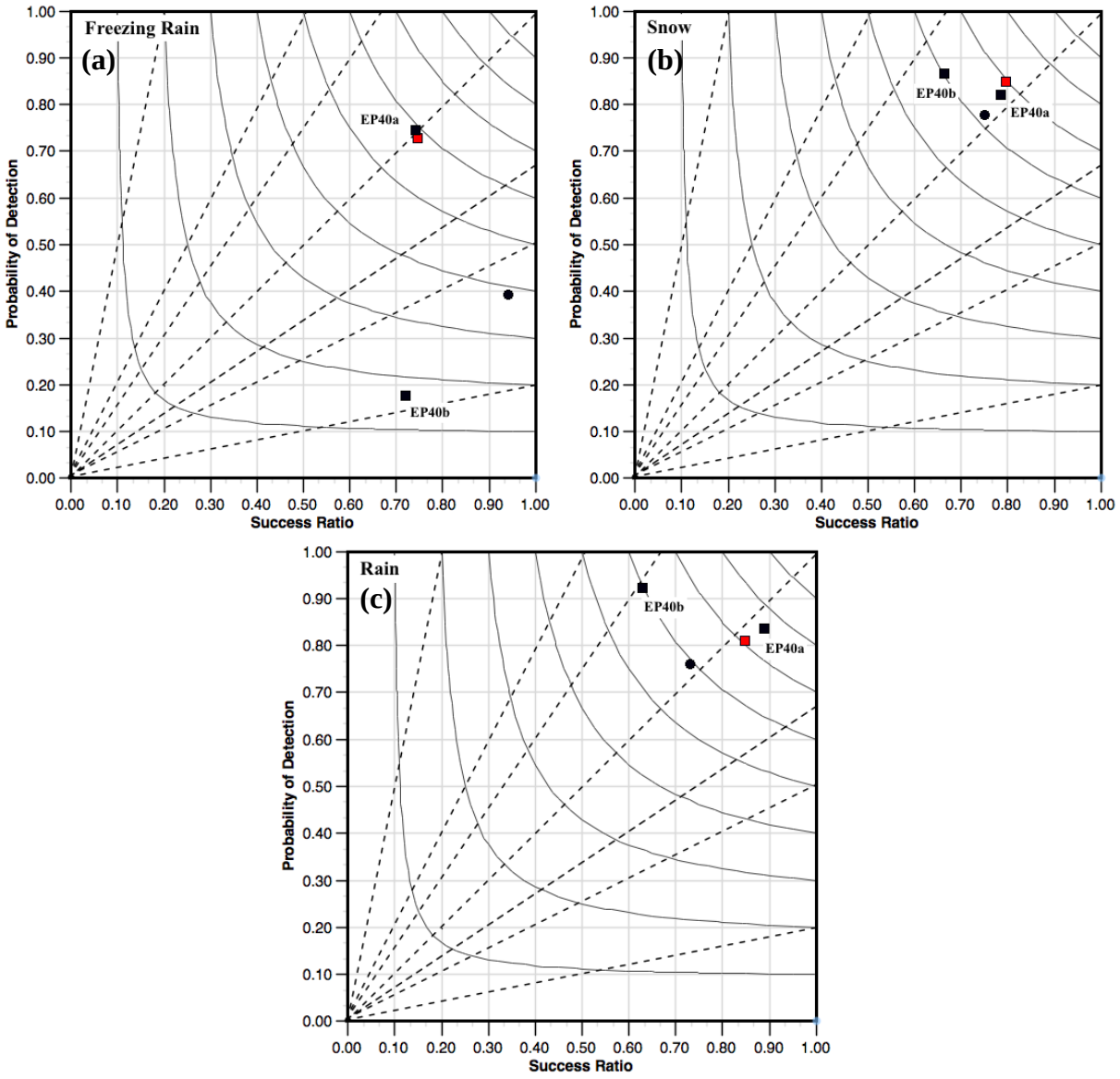


Figure 3. Performance diagrams for Freezing rain (a), Snow (b), and Rain (c) based on the success ratio, POD, CSI, and bias of the given members. The raw HRRR-TLE member 3 is represented by the black circle, with EP member 40, both before and after bias correction, marked with the black box. The red circle represents the final weighted combination of members 15, 18, and 72 all of which have been bias corrected.

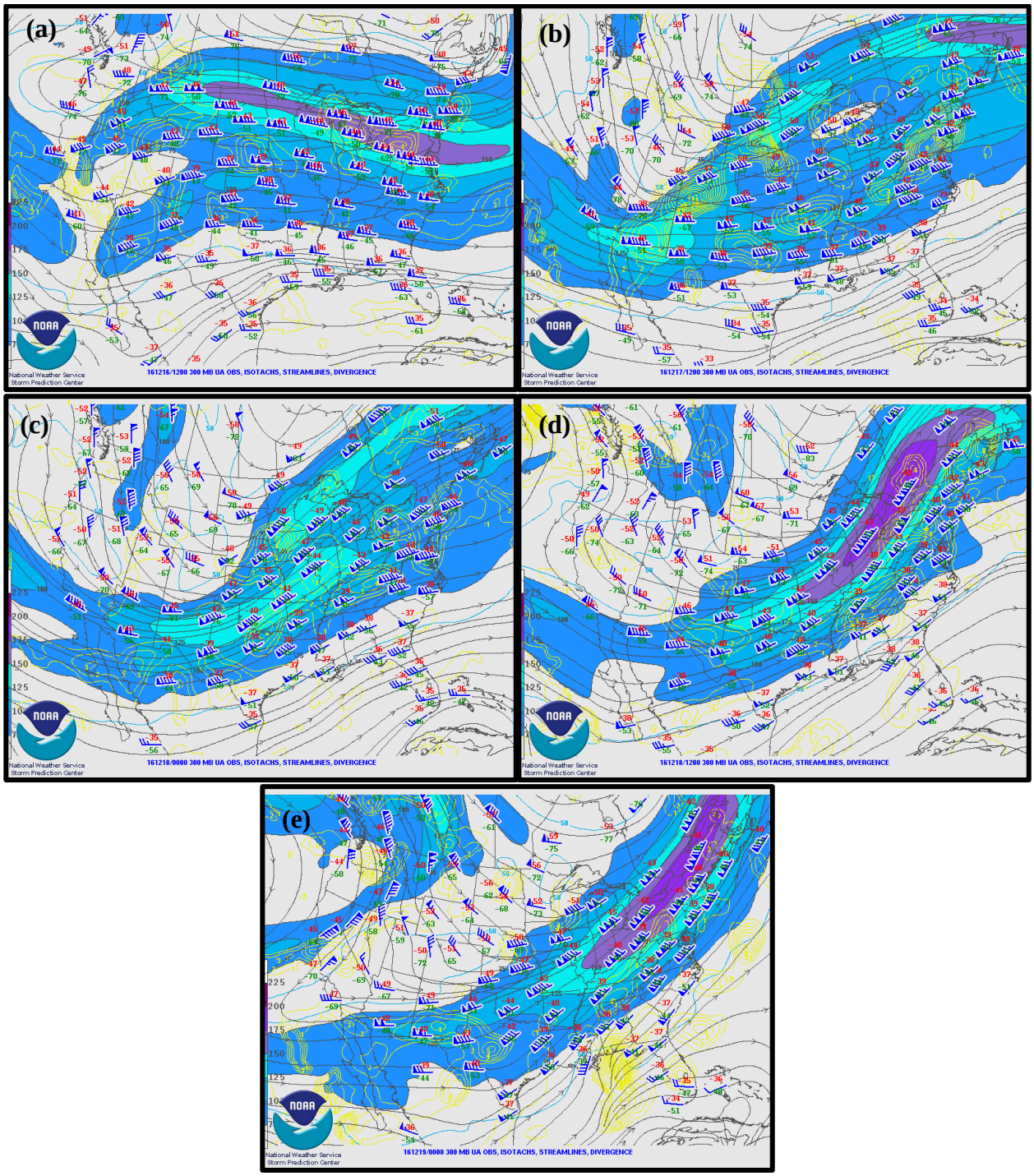


Figure 4. (a)-(e) 300 hPa observations, isotachs, and divergence for (a) 16 December 2016 at 1200 UTC, (b) 17 December at 1200 UTC, 18 December (c) at 0000 UTC and (d) 1200 UTC, and (e) 19 December 0000 UTC. These figures were gathered from https://www.spc.noaa.gov/obswx/maps/

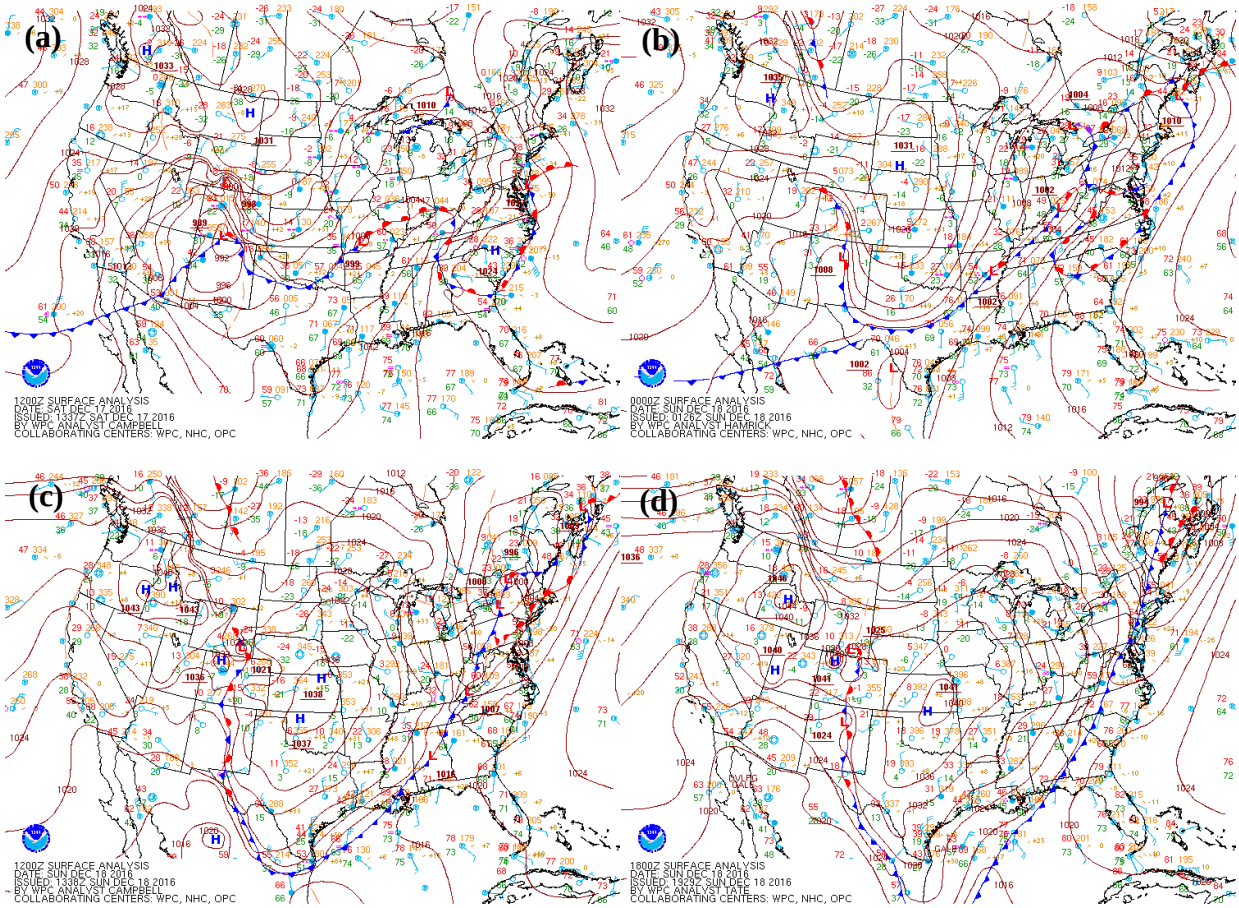
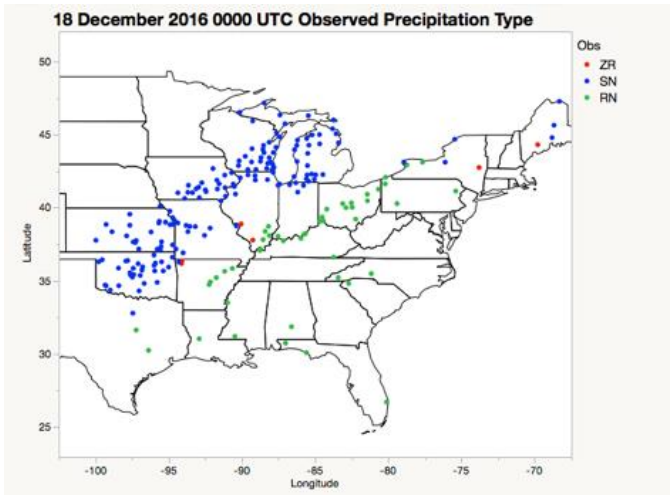
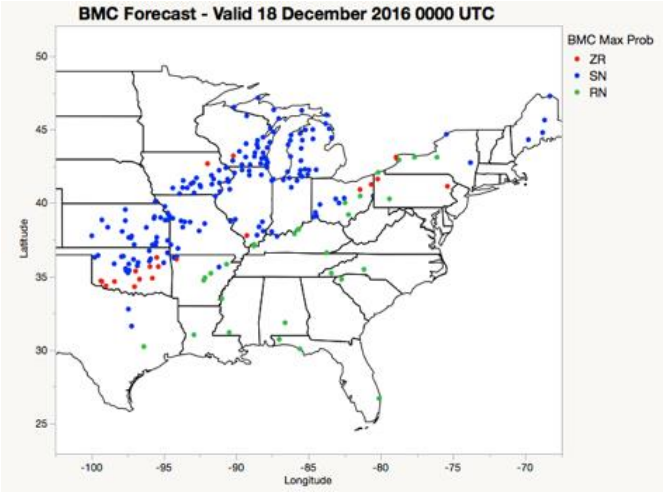


Figure 5. Observed surface conditions for (a) 17 December 2016 at 1200 UTC and18 December (b) at 0000 UTC, (c) 1200 UTC, and (d) 1800 UTC. These figures were gathered from https://www.wpc.ncep.noaa.gov/archives/web_pages/sfc/sfc_archive_maps.php?

(a)



(b)



(c)

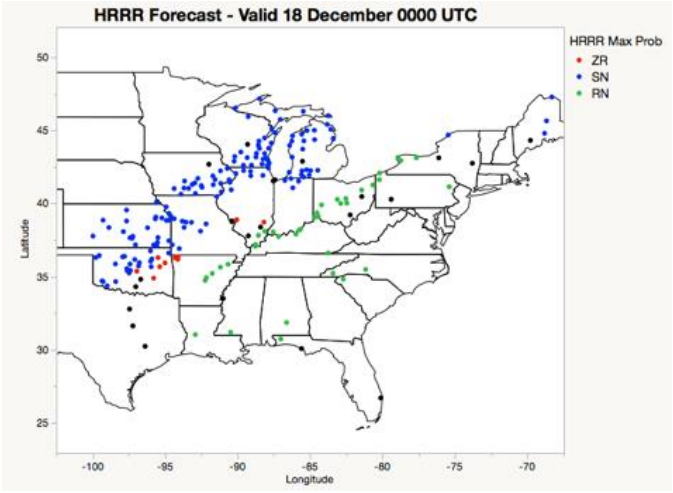


Figure 6. Observations (a) from 18 December 2016 at 0000 UTC from stations reporting freezing rain (red), snow (blue), or rain (green) alongide the EP BMC 2-hour forecast (b) and HRRR-TLE member 3 2-hour forecast (c), valid for observing time, of freezing rain, snow, and rain with black dots representing locations where the HRRR forecast no precipitation.

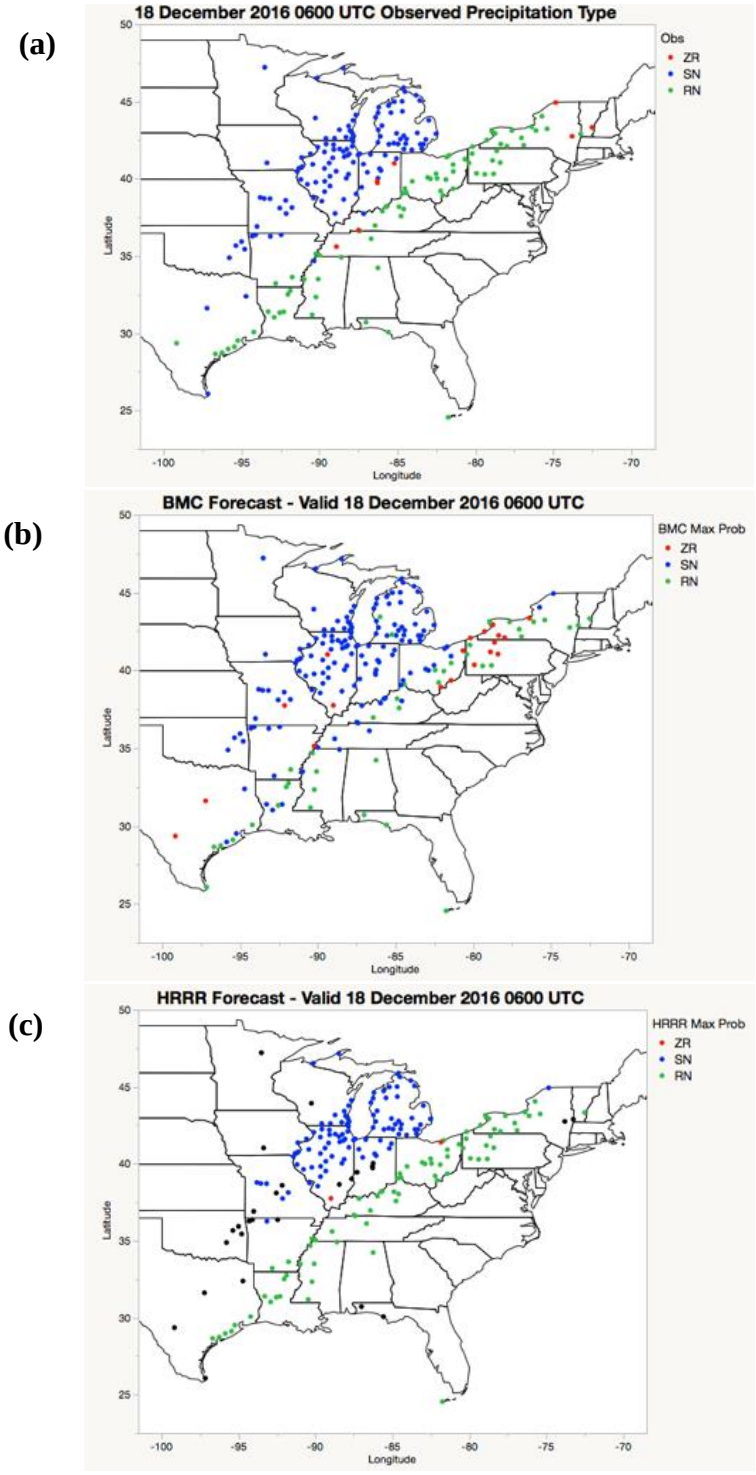


Figure 7. Observations (a) from 18 December 2016 at 0600 UTC from stations reporting freezing rain (red), snow (blue), or rain (green) alongide the EP BMC 6-hour forecast (b) and HRRR-TLE member 3 6-hour forecast (c), valid for observing time, of freezing rain, snow, and rain with black dots representing locations where the HRRR forecast no precipitation.

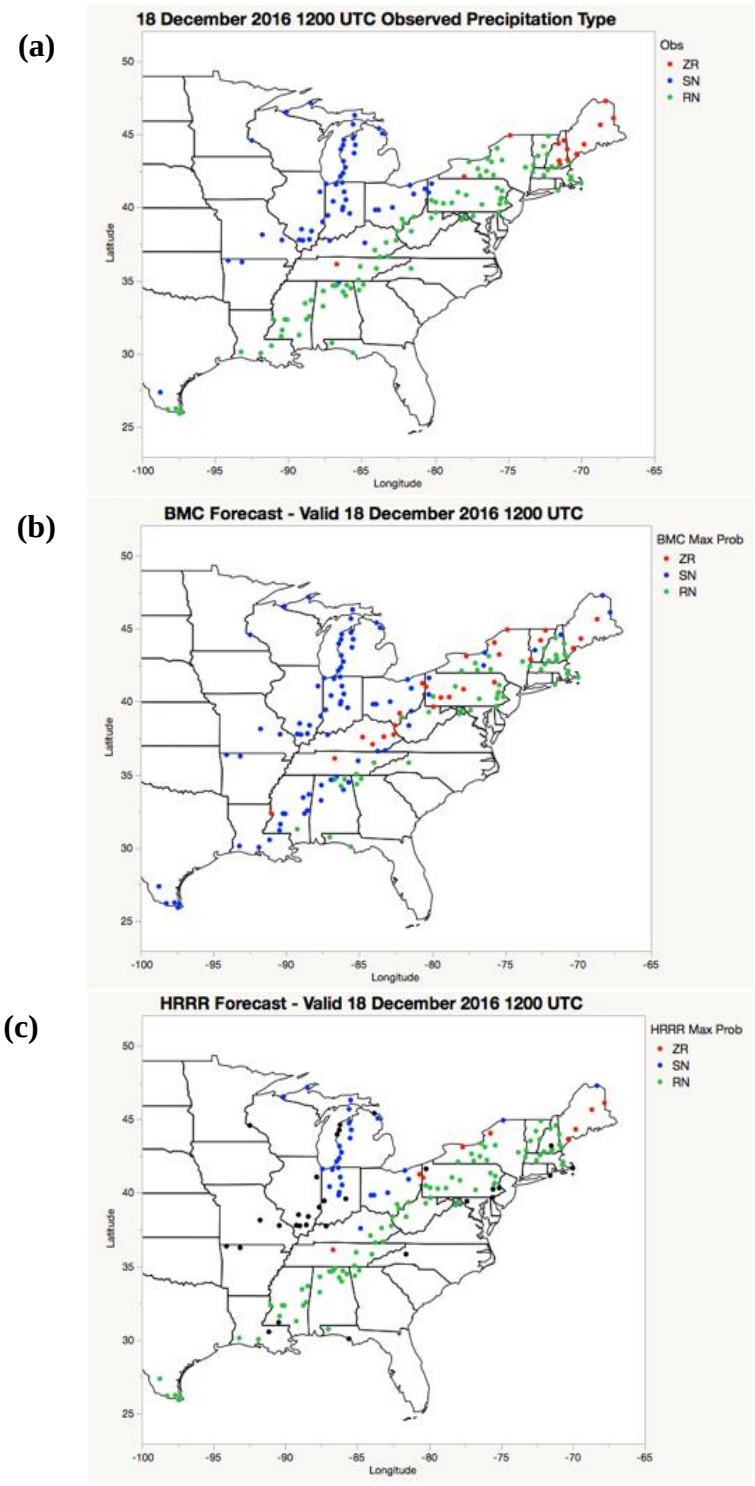


Figure 8. Observations (a) from 18 December 2016 at 1200 UTC from stations reporting freezing rain (red), snow (blue), or rain (green) alongide the EP BMC 2-hour forecast (b) and HRRR-TLE member 3 2-hour forecast (c), valid for observing time, of freezing rain, snow, and rain with black dots representing locations where the HRRR forecast no precipitation.

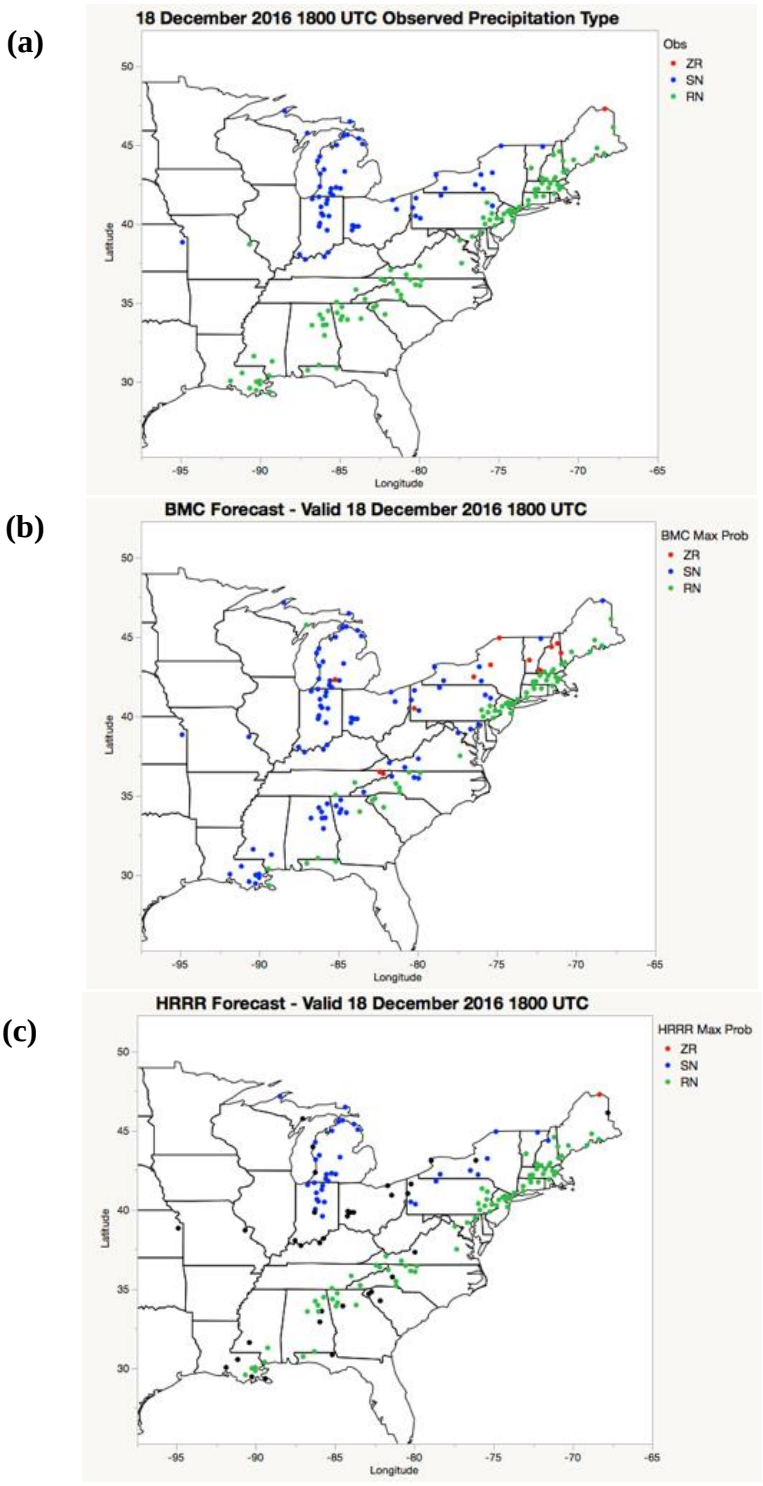


Figure 9. Figure 6. Observations (a) from 18 December 2016 at 0000 UTC from stations reporting freezing rain (red), snow (blue), or rain (green) alongide the EP BMC 6-hour forecast (b) and HRRR-TLE member 3 6-hour forecast (c), valid for observing time, of freezing rain, snow, and rain with black dots representing locations where the HRRR forecast no precipitation.

VI. TABLES

Forecast Variables	Units	Identifier
2 m Temperature	k	T
2 m Dew Point Temperature	k	TTD
Low Level Cloud Coverage	%	CL
Middle Level Cloud Coverage	%	CM
Upper Level Cloud Coverage	%	CH
U-Component Wind	m/s	U
V-Component Wind	m/s	V
Precipitable Water (PWAT)	mm	PWAT
Total Accumulated Precipitation	mm	PP
Visibility	m	VIS
Precipitation Type - Rain	0/1	RN
Precipitation Type - Snow	0/1	SN
Precipitation Type –Ice Pellets	0/1	IP
Precipitation Type –Freezing Rain	0/1	ZR
Categorical Temperature	-2, -1, 0, 1, 2	TCAT

Table 1. A list of the 14 HRRR-TLE forecast variables that are used in this study alongside one derived variable.

2x2 Contingency Table		Event Observed	
		YES	NO
Event Forecast	YES	*HIT*	*FA*
	NO	*MISS*	*Correct Negative*

Table 2. The 2x2 standard contingency table defining a *HIT*, *FA*, and *Miss*.

Model/Member	Brier Score	HSS
HRRR-TLE 1	0.541	0.514
HRRR-TLE 2	0.523	0.521
HRRR-TLE 3	0.553	0.508
M40B	0.425	0.482
M40A	0.366	0.702
M15	0.379	0.597
M18	0.449	0.532
M72	0.450	0.489

Table 3. Brier scores and Heidke Skill Scores, eq. 6, calculated for the raw HRRR-TLE members alongside member 40 before bias correction (M40B), member 40 after bias correction (M40A), and members 15, 18, and 72 after bias correction.

Bayesian Model Combination Selection	
EP Member	Weights
Member 67 (TLE-1)	0.000
Member 63 (TLE-1)	0.000
Member 14 (TLE-1)	0.000
Member 40 (TLE-2)	0.000
Member 2 (TLE-2)	0.000
Member 73 (TLE-2)	0.000
Member 15 (TLE-3)	0.333
Member 18 (TLE-3)	0.333
Member 72 (TLE-3)	0.333

Table 4. The 9 EP members selected, 3 from each HRRR-TLE member, with their associated weights determined by Bayesian Model Combination.

EP Member – Precip Type	IF	THEN
Member 15 - ZR	TTD ≤ VIS	-0.3578*PWAT^2 - 0.3921*PWAT
Member 15 - ZR	PP ≤ V	-0.6598*VIS - 0.1283*V*VIS
Member 15 - ZR	PWAT ≤ V	0.0938*PP*PWAT - 0.1027*TTD
Member 15 - ZR	PP > SN	2.7787*TCAT
Member 15 - SN	TCAT ≤ CL	-0.9571*PWAT + 0.3592*TCAT*TTD
Member 15 - SN	SN > TCAT	-0.1097*PP*V + 3.1114
Member 15 - SN	V ≤ TCAT	-0.2771*CL*V*TCAT
Member 18 - ZR	ALWAYS	1.50620*IP - 0.51434*V
Member 18 - ZR	PWAT ≤ RN	-0.01893*PP*VIS + 0.07842*RN
Member 18 - ZR	VIS ≤ V	-0.97460*IP - 0.02775*RN*PP
Member 18 - ZR	T > IP	-0.73121*PWAT + 0.13029*RN*PP
Member 18 - SN	IP > U	-0.14146*IP*TCAT - 0.50582*PWAT
Member 18 - SN	RN > PP	0.43864*TCAT*IP - 0.76860*PWAT
Member 18 - SN	PWAT ≤ PP	-0.71475*PP - 0.71503*PWAT + 0.43085*V
Member 18 - SN	IP ≤ PWAT	0.41672*VIS - 0.10510*PP - 0.40116*RN
Member 18 - SN	U > TCAT	-0.03734*IP*RN*PWAT
Member 72 - ZR	PP > RN	0.45272*ZR + 4.12538*SN
Member 72 - ZR	VIS > ZR	-0.57231*T - 0.04166*CM*PP
Member 72 - ZR	CM > PP	0.04899*CM*ZR - 0.01334*RN
Member 72 - ZR	ZR > SN	0.81232*VIS + 0.63438*U + 0.10746*RN
Member 72 - SN	VIS ≤ PP	0.21945*RN*VIS*U
Member 72 - SN	U > ZR	-3.27548*T
Member 72 - SN	CM ≤ SN	-0.35009*SN*PP - 0.86478*T
Member 72 - SN	RN ≤ VIS	-0.02163*RN - 0.34822*U*ZR
Member 72 - SN	ZR > RN	0.41740*SN*T - 0.31507*SN

Table 5. The series of IF-THEN algorithms generated by the EP that were selected by BMC as the most optimal from the 9 possible options.

BMC 3x3 Contingency Table		Event Observed		
		ZR	SN	RN
Event Forecast	ZR	7	26	43
	SN	14	352	102
	RN	8	6	161

Table 6. A 3x3 contingency table showing the relationship between the precipitation type that was observed (column) vs. the precipitation type forecast (rows) by the EP BMC members.

HRRR 4x3 Contingency Table		Event Observed		
		ZR	SN	RN
Event Forecast	ZR	8	5	4
	SN	4	304	1
	RN	9	6	262
	None	8	69	39

Table 7. A 4x3 contingency table showing the relationship between the precipitation type that was observed (column) vs. the precipitation type forecast (rows) by the HRRR-TLE member 3 with the last row representing where the HRRR forecast no precipitation.

VII. REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, 78, 1-3.

Cortinas, J., 2000: A Climatology of Freezing Rain in the Great Lakes Region of North America. *Mon. Wea. Rev*., 128, 3574-3588.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, 27, 396-410.

Glahn H. R. and D. A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *J. Appl. Meteor.* 11, 1203-1211.

Grimit E. P., and C. F. Mass, 2002: Initial Results of a Mesoscale Short-Range Ensemble Forecasting System over the Pacific Northwest. *Wea. Forecasting*, 17, 192-205.

Hamill, T. M., and J. S. Whitaker, 2007: Ensemble Calibration of 500-hPa Geopotential Height and 820-hPA and 2-m Temperatures Using Reforecasts. *Mon. Wea. Rev.*, 135, 3273-3280.

Hoffman, R.N., and E. Kalnay, 1983: Lagged Average Forecasting, an Alternative to Monte Carlo forecasting. *Tellus A*, 35A, 100-118.

Ikeda, K., M. Steiner, and J. Pinto, 2013: Evaluation of Cold-Season Precipitation Forecasts Generated by the Hourly Updating High-Resolution Rapid Refresh Model. *Wea. Forecasting*, 28, 921-939.

Lott J. N, M. C. Sittel, 1996: The February 1994 Ice Storm in the Southeastern U.S. 7pp, https://www1.ncdc.noaa.gov/pub/data/special/iwais96.pdf

——, D. Ross, A. Graumann, 1998: Eastern U.S. Flooding and Ice Storm January 1998. 6pp, ftp://ftp.ncdc.noaa.gov/pub/data/extremeevents/specialreports/Eastern-US-Flooding-and-Ice-Storm-January1998.pdf

Monteith, K., J. Carroll, K. Seppi, and T. Martinez, 2011: Turning Bayesian Model Averaging into Bayesian Model Combination. *Proc. Int. Joint Conf. on Neural Networks (IJCNN'11)*, San Jose, CA, IEEE, 2657–2663.

Novak, D. R., D. R. Bright, and M. J.Brennan, 2008: Operational Forecaster Uncertainty Needs and Future Roles. *Wea. Forecasting*, 23, 1069-1084.

NOAA, 2018: The Air Up There: Skew-T Examples. NWS, 6 August 2018, https://www.weather.gov/jetstream/skewt_samples

Panofsky, H.A., and G. W. Brier, 1958: Some Applications of Statistics to Meteorology. *The Pennsylvania State University Press*, 200 pp.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, 24, 601-608

——, 2010: Seeking consensus: A New Approach. *Mon. Wea. Rev.*, 138, 4402–4415.

——, 2013: Using evolutionary Programming to Generate Skillful Extreme Value Probabilistic Forecasts. *Mon. Wea. Rev.*, 141, 3170–3185.

——, 2015: Evolving Ensembles. *Mon. Wea. Rev.*, 143, 471-490.

Stewart R. E., 1992: Precipitation Types in the Transition Region of Winter Storms. *Bul. Amer. Met. Soc.*, 73, 287-296.

The Weather Channel, 2016: Winter Storm Decima a Cross-Country Snow and Ice Storm (RECAP). Accessed 7 August 2018, https://weather.com/storms/winter/news/winter-storm-decima-forecast-northwest-rockies-midwest-east.