Final Report: An Adaptive Bayesian Model Combination Post Processor for the HRRR-TLE Forecast System

Paul Roebber and Tim Thielke
University of Wisconsin at Milwaukee

1. Introduction

Time-lagged ensembles (TLE), originally proposed by Hoffman and Kalnay (1983) as a computationally inexpensive alternative to full initial condition perturbation ensembles, combine a (usually small) sample of deterministic model runs initialized at different times. Experience with TLEs has shown that they can provide some useful probabilistic information since run-to-run differences in initial conditions can be used as a measure of the uncertainty of the initial atmospheric state. For short-cycle (i.e., hourly) models such as the High Resolution Rapid Refresh (HRRR), however, the resulting time-lagged estimate of the analysis uncertainty is too small. Since more sophisticated ensembles are also underdispersive (e.g., Hamill and Whitaker 2007 and references therein; Novak et al. 2008), underdispersion is not a new problem, but merely a problem of degree, and as with other types of ensembles, post-processing is a useful method for at least partially accounting for this limitation and producing more reliable probabilistic forecasts.

The HRRR-TLE (Alexander et al. 2014) run by NOAA's Earth System Research Laboratory (ESRL) currently uses the three most recent runs of the experimental "HRRRx" and provides forecasts out to 24 hours. Given the ~2h latency, the result is that the 1200 UTC HRRR-TLE, for example, is based on the HRRRx forecasts from the 0800, 0900, and 1000 UTC initializations and provides forecasts out for the next six hours. ESRL made the choice to limit ensemble size in order to maximize ensemble lead time, understanding that for a TLE there are rapidly diminishing returns as the number of members is increased. Although ESRL has been moving towards a high-resolution ensemble that is not based on time-lagging, the process of extracting maximum information from ensemble members used in this study is not dependent on how those members are produced. For the purposes of this study, the effort was focused on improving the time-lagging approach in the context of winter precipitation phase, in collaboration with NOAA/ESRL scientists Isidora Jankov and Trevor Alcott. In the next section, we will describe the results from these investigations, and in section 3, we will discuss what these findings suggest about potentially profitable future research and application directions.

2. Approach and Findings

a. Methods

The HRRR-TLE dataset was provided by NOAA/ESRL, covering the period November 2013 through February 2017. These data were restricted to the cold season months (November-March) for the purposes of this study. We focused on cyclones that tracked from the Gulf of Mexico or Colorado and only considered those events in which more than one precipitation type was observed during the cyclone lifetime. We identified such cases in order to make the analysis tractable (i.e., considering cases that are directly involved with cyclone precipitation processes and not fundamentally dependent on orographic features). The focus on multiple precipitation types was made in order to highlight the ability of the approach to account for phase transitions. Additional work/models may need to be developed to account for orographically-dependent precipitation events but the method should be transferable to cases where no precipitation phase transitions occur. With these restrictions, 29 cyclones were identified and constitute the set of cases that were considered.

In order to verify precipitation type, observations were obtained from the regular synoptic surface network (including National Weather Service Offices, manual stations, and automated surface observing system stations along and to the east of 100°W longitude). The HRRR-TLE dataset included forecasts for twenty surface variables of which fourteen were used in this study, based on their potential relevance to determining precipitation type. An additional, derived variable is included which categorizes the temperature relative to the freezing point (Table 1). After determining the categorical temperature, we normalize each HRRR-TLE variable based on the given variable's mean value and standard deviation determined by the subset of data used for training. Normalization was not applied to categorical, probabilistic, and percentage variables. The resulting standard anomalies were then applied as inputs in our future methods.

In order to compare the 3 km gridded HRRR-TLE forecasts to a station observation, a bilinear interpolation was applied, based on the nearest four grid point locations. All cases where no precipitation was reported are removed, such that the algorithm predictions are conditional on the expectation of precipitation. (the model also can be applied operationally under the expectation that no precipitation occurs at the site of interest, since the precipitation amount is an input to the equations and the probabilities will account for the zero value). In order to balance cases of freezing rain, rain, and snow for training, a randomized filtering process is applied to each observation and forecast, leaving approximately 1,000 observations and associated forecasts of each precipitation type. Slightly over half of the filtered data are used for training with another 20% used for validation. The remaining 25% are applied for independent testing. Only the dates of the cyclones were used to determine which data will be used for training, validation, and testing.

Evolutionary Programing (EP; Roebber 2010, 2013, 2015abc, 2018) is the method which is used to "map" the HRRR forecasts to the observations. The EP algorithms take the form of logistic regression equations, formed from two sets of five IF-THEN equations, each composed of five variables, three operators, and three variable coefficients. The two sets of IF-THEN algorithms are then used in the logistic function for each of the three precipitation types to determine probability.

Forecast Variables	Units	Identifiers
2 m Temperature	k	T
2 m Dew Point Temperature	k	TTD
Low Level Cloud Coverage	0/1	CL
Middle Level Cloud Coverage	0/1	CM
Upper Level Cloud Coverage	0/1	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: The forecast variables provided as input to the Evolutionary Program (EP) algorithms.

The probability *P* of freezing rain is then given by:

$$P_{freezing\ rain} = \frac{e^x}{(1+e^x+e^y)} \qquad (1)$$

where *x* and *y* represent the results from the two sets of five IF-THEN equations (*x* for freezing rain and *y* for snow). The probability for snow is then:

$$P_{snow} = \frac{e^{y}}{(1+e^{x}+e^{y})} \tag{2}$$

and the probability for rain is 1-$P_{freezing\ rain}$-P_{snow}.

The EP protocol is similar to prior studies (Roebber 2015abc; Roebber 2018), in which the process is to start by random initialization of a large set of algorithms, evaluate their performance [in this case, using the Brier Score (Brier 1950) as the measure of success], and reproduce the so-defined best fit algorithms via cloning with some mutation to introduce innovation (where these new algorithms replace the bottom 20% based on the performance measure). The evaluation and reproduction steps continue until a convergence criterion is met. Here, we maintain a master list of the top 100 performers at any stage in the training, based on performance on the validation dataset – if an algorithm at generation *N* is found that outperforms any one of the members of the master list, it replaces that algorithm, and training continues until a fixed number of generations have been studied (set to a number larger than is expected to be needed for optimization; in this study, we use 300 generations).

Although Roebber (2015a) introduced cloning and mutation in this process to allow the propagation of the best solutions while still providing for innovation, here we apply mutations more aggressively such that mutation occurs in the production of every new algorithm. Although mutation can often be destructive (Roebber 2015a), the cloning process along with the master list ensures that the best solutions are not lost. Once the training is completed for 300 generations, a full random initialization and training is run again for another 300 generations except that the master list is retained, and updated only when a new algorithm performs sufficiently well 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. 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. A decaying average bias correction following Cui et al. (2012) is applied to the probabilities produced by the EP algorithms, with a 5 percent weighting of the current error. Finally, after the bias correction is applied, the probabilities are normalized to sum to unity.

In a final step, a model selection process is used to allow the application of Bayesian model combination (BMC; Monteith et al. 2011; Roebber 2015a). The model selection starts by ranking each of the 100 members from the master list for each of the three HRRR-TLEs, based on Brier Score. Then, proceeding down the ranked list, we find the next best ranking algorithm that has a Brier Score Difference (BSD) from the best performing member that is above the 25th percentile; the third selection is the next highest ranked algorithm that has a BSD greater than the 50th percentile. This selection procedure is repeated for each of the three TLEs to produce a set of 9 EP algorithms (three from each TLE). When applying BMC to this group, we use 4 raw weights, thus requiring that the posterior probability of 4^9 combinations be evaluated on the cross-validation dataset. The combination with the least amount of error is the final selected weighting scheme.

For evaluating the performance of the HRRR-TLE forecast member, an individual EP, or the BMC, we use the Heidke Skill Score (HSS; Panofsky and Brier 1958) for the full 3x3 deterministic forecasts (where that forecast is defined by the maximum individual category probability), and we use the standard 2x2 contingency measures (POD, FAR, CSI, bias) for individual precipitation type category forecasts. (note that in this formulation, mixed precipitation type is not allowed). As noted above, the algorithm process is not invoked if no precipitation is forecast at a given point. Additionally, we use the Brier Score for the probabilities.

The HSS is corrected for chance such that:

$$HSS = \frac{Hits_{all} - Chance}{(N - Chance)} \tag{3}$$

$$Chance = \frac{(O_{ZR} * F_{ZR}) + (O_{SN} * F_{SN}) + (O_{RN} * F_{RN})}{N} \tag{4}$$

where *N* is the total number of forecasts and *O* and *F* are the total observed and forecast events for each of the 3 categories (i.e. O_{ZR} and F_{ZR} are the observed and forecast number of freezing rain events, respectively, and ZR, SN, and RN refer to freezing rain, snow, and rain, respectively).

b. Results

The average Brier Score and Heidke Skill Score of HRRR-TLR 3 (the most recent ensemble member) across all cyclones that occurred within the testing period (split off from the original training and validation data, as noted previously) were 0.289 and 0.625, scores very slightly superior to those of the older TLE members of the ensemble. Thus, for further comparison of the BMC to the HRRR, we will consider HRRR-TLE 3.

In conducting this analysis, we noted a propensity for the EP ensemble to over forecast freezing rain, likely the result of considering relatively balanced samples of freezing rain, snow and rain events for training, when the actual occurrence of freezing rain is less frequent than the other two categories.

To remedy this over forecasting, we randomly select 2,600 rain and snow cases from our validation data while using all freezing rain cases that occur within that same dataset (N=348). We then develop a further correction to the probabilistic forecasts from the EP BMC by training an artificial neural network (ANN) with a single hidden layer and three hidden nodes. This corrected BMC, when applied to the same independent data as above, yielded improvements in BIAS, FAR, and CSI for freezing rain with an overall improvement in Brier Score (to 0.083, with an HSS of 0.830). Scores for the BMC before and after ANN bias correction and the HRRR-TLE 3 (Table 2) show these improvements and the superior performance of the EP approach across all 3 precipitation type categories (see also Fig. 1).

	Score	**BMC**	**BMC-C**	**HRRR TLE 3**
Freezing Rain	POD	0.554	0.354	0.248
	FAR	0.848	0.684	0.632
	BIAS	4.490	1.456	0.781
	CSI	0.136	0.189	0.176
Snow	POD	0.803	0.857	0.679
	FAR	0.061	0.054	0.059
	BIAS	0.838	0.907	0.731
	CSI	0.784	0.830	0.652
Rain	POD	0.931	0.957	0.773
	FAR	0.055	0.072	0.092
	BIAS	1.002	1.065	0.849
	CSI	0.883	0.895	0.737

Table 2. Average deterministic scores representing measures of success for each precipitation type for the BMC before bias correction (BMC), BMC after bias correction (BMC - C), and the third member of the HRRR-TLE. The EP BMC is created from an unequal weighting of two algorithms derived from HRRR-TLE 1, two from HRRR-TLE 2, and one from HRRR-TLE 3.

One of the advantages of the EP design as implemented by Roebber (2015a) is the interpretability of the forecast logic. Analysis of the final algorithms revealed that the temperature anomaly was the most frequently used input for generating a probability, a not surprising result given physical considerations. Secondarily important variables in that respect were precipitable water anomalies and low cloud coverage. For the conditional part of the algorithm, the HRRR TLE forecast probability for snow and high cloud coverage were most commonly invoked. For further insight into the algorithm performance, two case studies were analyzed (below).

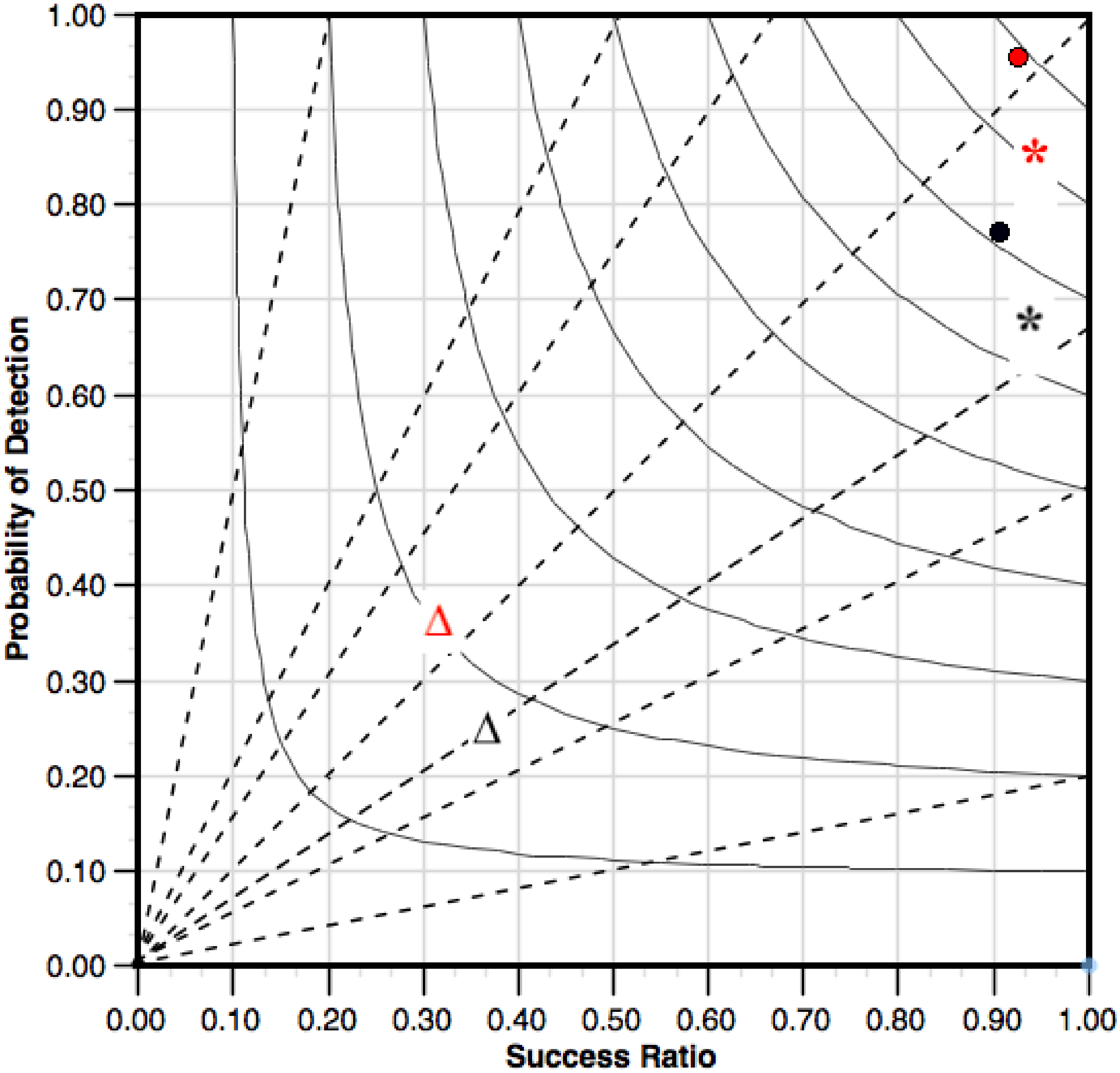

Figure 1: Performance diagram (Roebber 2009) for precipitation phase forecasts for the HRRR-TLE3 (black) and the EP BMC (red). Shown are freezing rain (triangles), snow (asterisks), and rain (circles).

i. Case Study: 17 February 2015

To place the overall results in specific context, we have chosen for analysis a major winter storm (16-17 February 2015). This storm produced up to 0.50-0.75 inches of freezing rain in portions of Tennessee, Georgia, North Carolina, and South Carolina, and widespread snowfall with impacts in New England and also some municipalities in Kentucky and Tennessee that struggled to keep up with the accumulating snow (NBC 2015; NOAA 2015). There was disruption to both ground and air transportation (e.g., 1200 flights were canceled), and widespread power outages causing five states to declare a state of emergency (NBC 2015). This case is an example where the EP BMC performed well.

At 1200 UTC 15 February 2015, an upper-level trough propagated across the Pacific NW region into the intermountain west of the United States (Figs. 2-3) with developing surface low-pressure on the leeward side of the Rocky Mountains near the Panhandle of Texas. By 1200 UTC the following day, the digging trough and strengthening divergence aloft over the Ohio Valley region resulted in a surface cyclone positioned over east

Texas with a stationary front extending into the area of strong divergence aloft. Over the next 24 hours, the cyclone continues towards the east with a surface low redevelopment off the coast of Virginia and North Carolina.

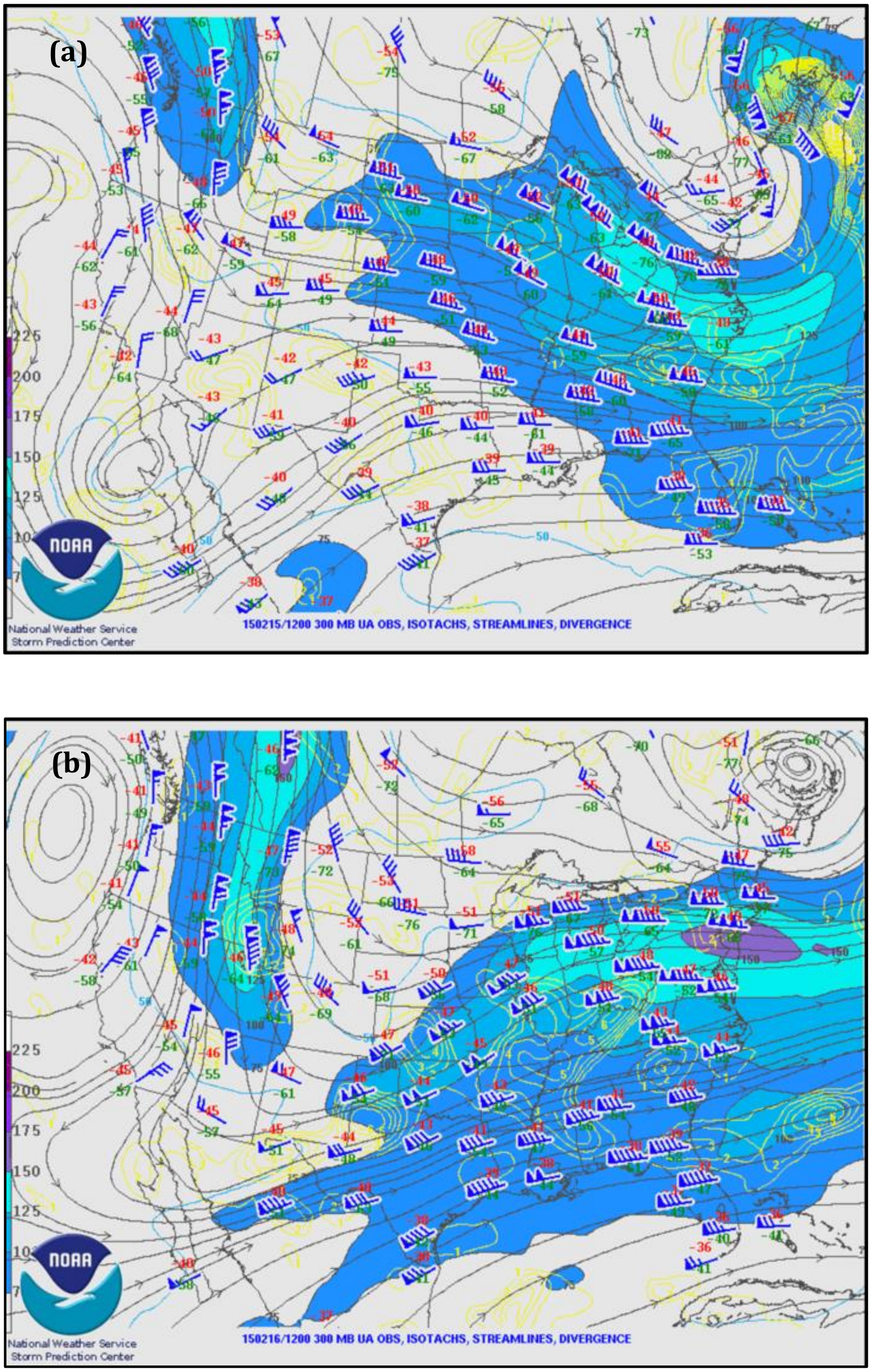
(a)
NOAA
National Weather Service
Storm Prediction Center
150215/1200 300 MB UA OBS, ISOTACHS, STREAMLINES, DIVERGENCE
(b)
NOAA
National Weather Service
Storm Prediction Center
150216/1200 300 MB UA OBS, ISOTACHS, STREAMLINES, DIVERGENCE

Figure 2. 300 hPa analysis for (a) 1200 UTC 15 February 2015 and (b) 1200 UTC 16 February. These analyses were obtained from https://www.spc.noaa.gov/obswx/maps/

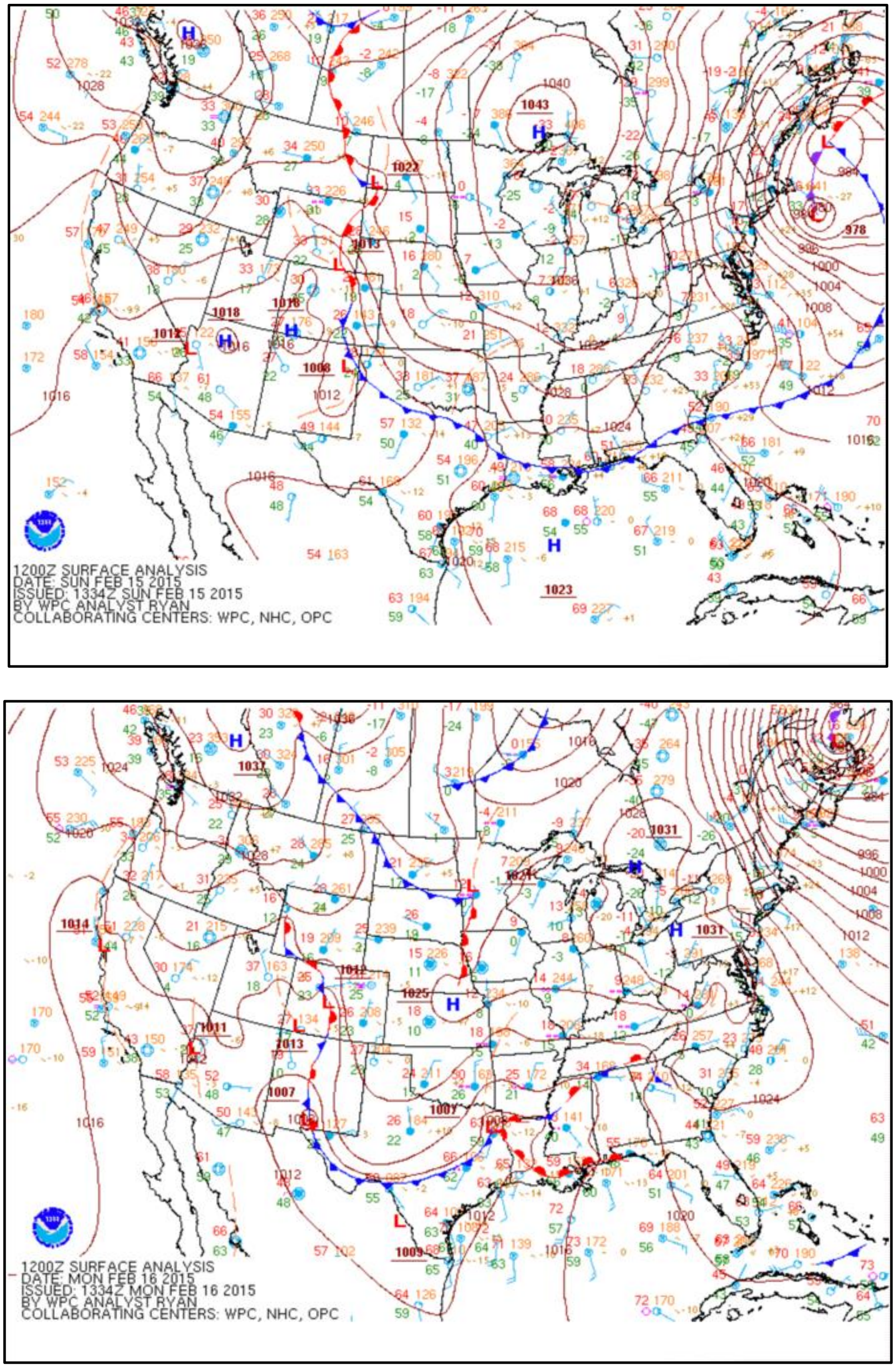
1200Z SURFACE ANALYSIS
DATE: SUN FEB 15 2015
ISSUED: 1334Z SUN FEB 15 2015
BY WPC ANALYST RYAN
COLLABORATING CENTERS: WPC, NHC, OPC
1200Z SURFACE ANALYSIS
DATE: MON FEB 16 2015
ISSUED: 1334Z MON FEB 16 2015
BY WPC ANALYST RYAN
COLLABORATING CENTERS: WPC, NHC, OPC

Figure 3. Observed surface conditions for (a) 1200 UTC 15 February 2015 and (b) 1200 UTC 16 February. These analyses were obtained from https://www.wpc.ncep.noaa.gov/archives/web_pages/sfc/sfc_archive_maps.php?

Figures 4-5 shows the observations (marker types) and forecasts (colors; blue, snow; red, freezing rain; green, rain) from the calibrated BMC and the latest member of the HRRR-TLE at 0000 and 0300 UTC on 17 February 2015, which is at the beginning and in the middle of the period of focus owing to the occurrence of freezing rain. At 0000 UTC, there is snow to the north of the warm front rain to the south, and a transitional zone in between. This pattern persists through 0500 UTC, but the number of freezing rain observations are somewhat variable at each time.

For the 0000-0500 UTC period, the EP BMC outperformed the most recent member of the HRRR-TLE (Table 3). Of note, the BMC's POD for freezing rain is 0.616 compared to 0.306 for the HRRR. On the other hand, there is still a prominent over-forecast bias in the BMC such that the CSI is somewhat less than the more conservative HRRR (however, the Brier Skill Score for the BMC, computed relative to the HRRR, is 0.224 indicating better probabilistic performance). Overall, the BMC is clearly superior for this period.

	Metric **17 February 2015** **0000 – 0500 UTC**	**EP BMC**	**HRRR** **TLE 3**
ALL Types	HSS	0.819	0.691
	CSI	0.826	0.772
	FAR	0.096	0.080
	POD	0.904	0.828
Freezing Rain	CSI	0.327	0.418
	FAR	0.590	0.433
	POD	0.616	0.306
Snow	CSI	0.879	0.821
	FAR	0.021	0.045
	POD	0.896	0.855
Rain	CSI	0.899	0.811
	FAR	0.077	0.084
	POD	0.972	0.876

Table 3. Average deterministic and probabilistic forecasting scores for both the bias corrected BMC and HRRR TLE-3 from 0000-0500 UTC 17 February 2015.

At 0000 UTC, the BMC forecast (Fig. 4a) correctly indicates a precipitation phase transition zone over North Carolina, Tennessee, South Carolina, and Georgia. However, in the area north of the warm front, the BMC indicates freezing rain rather than the snow that was observed. Though incorrect, this is near the location of precipitation transition along the warm front, whereas at the same valid time, the HRRR-TLE3 does not predict widespread freezing rain in the area. In contrast, along the cold front (extending southwest over Alabama and Mississippi), there is no transitional precipitation zone despite indications of such in the BMC forecast.

Over time, the warm frontal boundary and the precipitation phase transition zone shifts northward. At 0300 UTC, the BMC continues its over-forecast for freezing rain (rather than snow) along this frontal boundary, but also captures those locations where freezing rain was occurring (Fig. 4b). The HRRR, however, does not start predicting more widespread freezing rain until 0400 UTC (not shown).

Since the BMC provides probabilistic forecasts, we next examine these for this case. At 0000 UTC, the freezing rain probabilities are the highest right along the warm front (0.50 to 0.75) and cold front boundaries with rapidly decreasing probabilities with distance away from the boundaries (not shown). The probabilities indicate diminishing chances for freezing rain to the north (e.g., near the northern state lines of Tennessee and North Carolina) with rapid drop-offs thereafter. This drop-off is even sharper to the south of the warm front.

As with the warm front, the freezing rain probabilities are highest along the cold front (04.0-0.60), with much higher rain probabilities with distance from the boundary. These characteristics continue throughout the period and could represent useful information to forecasters looking for the location of a precipitation phase transition line.

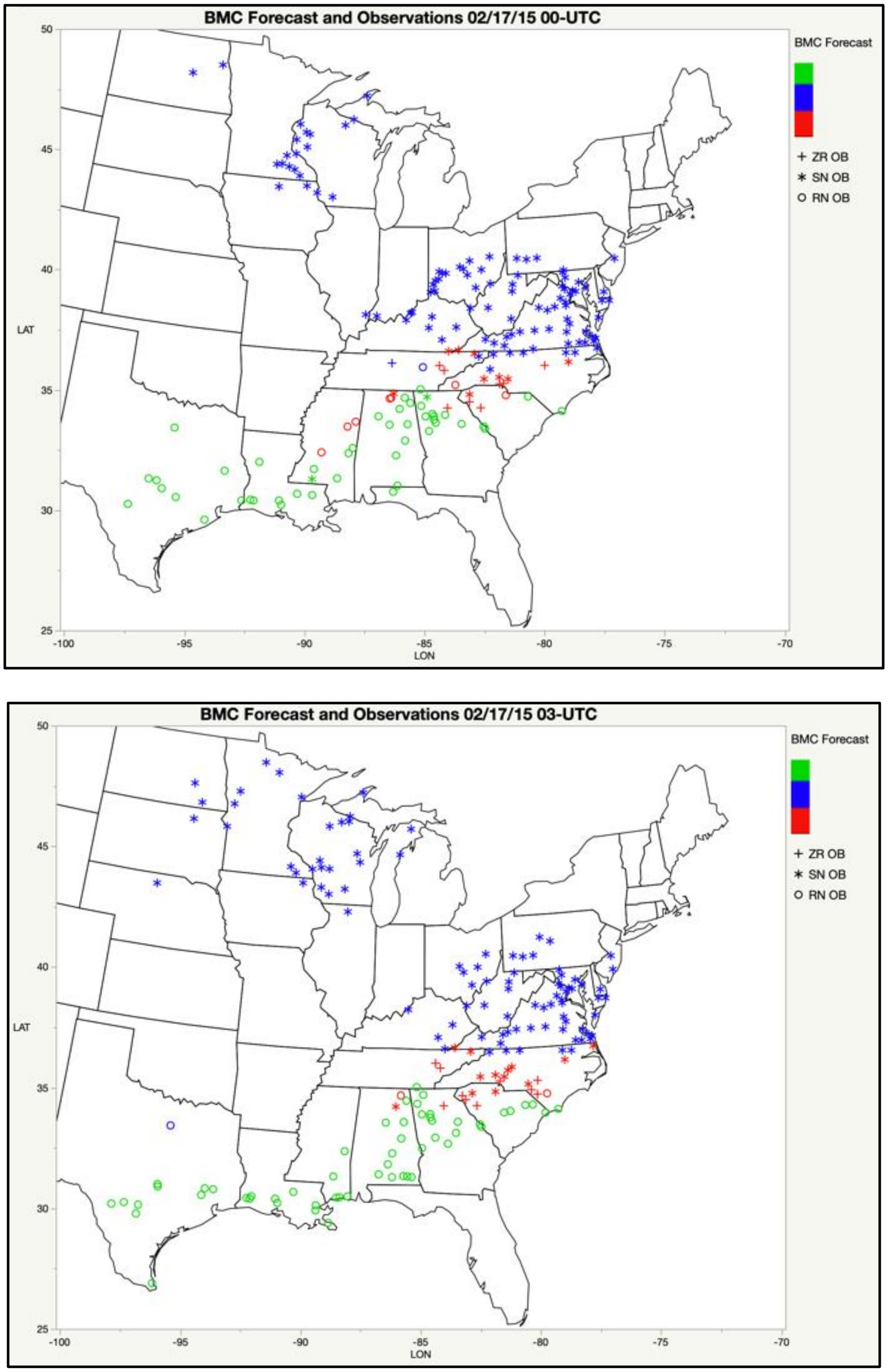

Figure 4. Observations from stations reporting freezing rain (plus), snow (star), rain (circle), or freezing rain and snow (square) and BMC forecast of freezing rain (red), snow (blue), and rain (green) on 17 February 2015 at (a) 0000 UTC and (b) 0300 UTC.

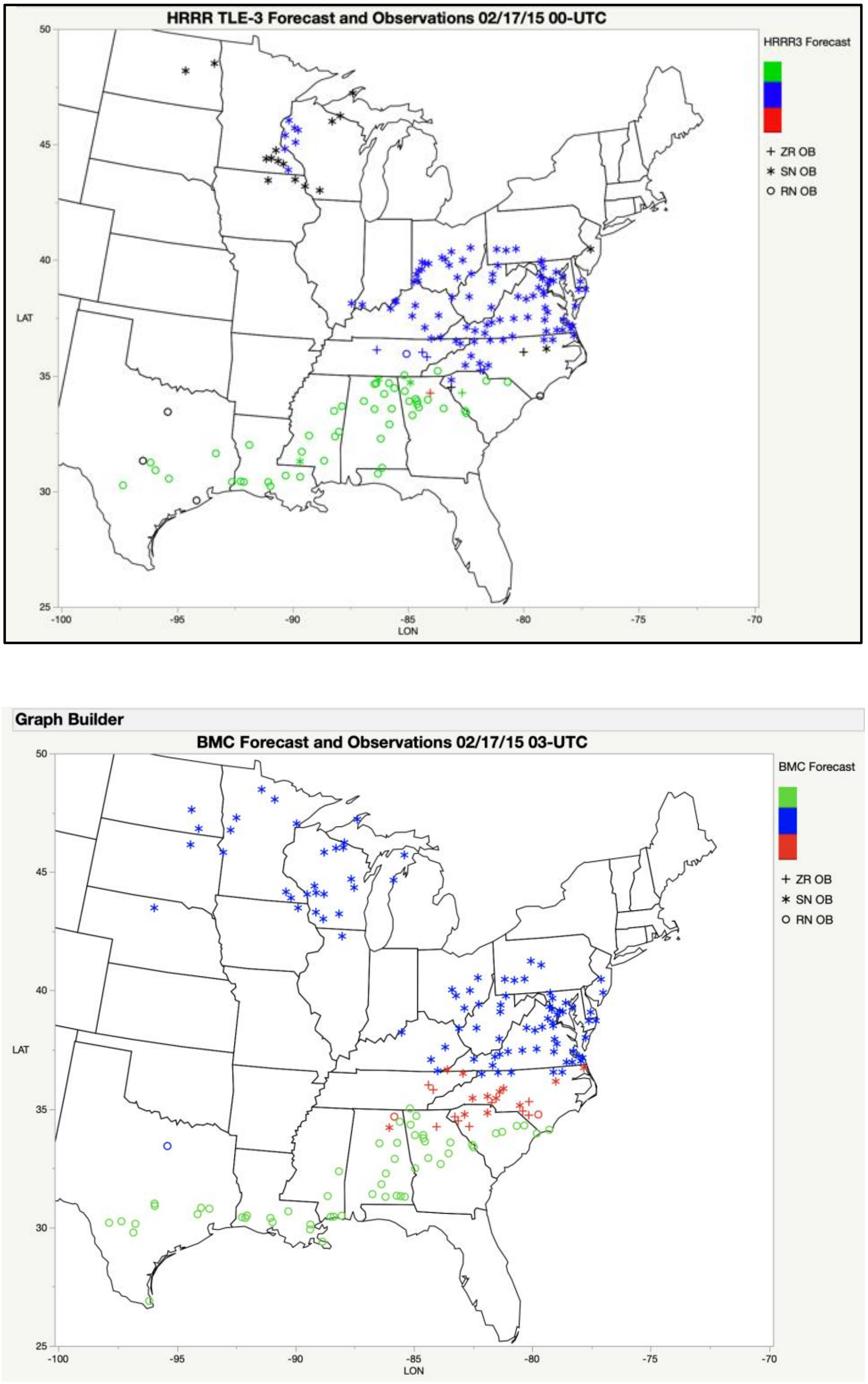

Figure 5. Observations from stations reporting freezing rain (plus), snow (star), rain (circle), or freezing rain and snow (square) and HRRR TLE-3 forecast of freezing rain (red), snow (blue), and rain (green) on 17 February 2015 at (a) 0000 UTC and (b) 0300 UTC. If the symbol is black then that precipitation phase was observed but no precipitation was forecast at that time and location.

ii. Case Study: 24 February 2016

Here, we present aspects of a poor BMC forecast, which occurred for the winter storm of 23-25 February 2016 (for this case, we focus on the period from 1200-1700 UTC 24 February). The synoptic pattern was set with an upper level trough propagating over the midwestern states on 23 February and consequent surface low pressure development over southeastern Texas (Fig. 6). By 1200 UTC 24 February, the maturing low-pressure system is undergoing occlusion, but with a secondary boundary forming along a line from the center over Kentucky to Vermont (this boundary is subsequently classified as a warm front). This set-up provided a favorable snow band environment over Indiana and Illinois, and over the ensuing 24 hours the system continues eastward to the Atlantic Ocean.

Although not widespread, freezing rain did occur over portions of the northeastern United States (Fig. 7). Rain was observed across most of the midwestern and southeastern states, and over much of the northeast. This storm also produced steep gradients in accumulated snowfall, with ranges of a few inches to 15 inches in some tightly located areas (NOAA 2016). (notably, severe thunderstorms including a tornadic outbreak occurred with this storm)

Based on Brier Skill Score relative to the HRRR-TLE, the BMC was less skillful during 1200-1700 UTC 24 February for freezing rain (-0.119), although overall, its probabilistic forecasts are superior (0.133). This result is reflected in the deterministic forecasts as well, with overall high FAR, largely resulting from over forecasts of freezing rain (Table 4).

At 1200 UTC, most areas are experiencing rain with some transition to snow over Illinois, northern Indiana and southern Michigan and a second area over New York and New England (Fig. 7). The transitional area in the Great Lakes region is largely direct from rain to snow, whereas the northeastern region experiences more freezing rain in the transition. Over time, there is an increase in snow observations in the midwest and more rain/snow and less freezing rain in the northeast. At 1200 UTC, the BMC highlights both transition areas but with a substantial over-forecast of freezing rain. By 1500 UTC, more snow is forecast (as observed) but there is also an offset in the position of the rain-to-snow transition line. Over the northeast, there continue to be problems both in the number of freezing rain locations and the location of the transition.

	Metric **24 February 2016** **1200 – 1700 UTC**	**EP BMC**	**HRRR** **TLE 3**
ALL Types	HSS	0.653	0.680
	CSI	0.770	0.806
	FAR	0.131	0.077
	POD	0.869	0.863
Freezing Rain	CSI	0.145	0.120
	FAR	0.472	0.104
	POD	0.243	0.138
Snow	CSI	0.566	0.709
	FAR	0.124	0.116
	POD	0.613	0.779
Rain	CSI	0.864	0.862
	FAR	0.115	0.067
	POD	0.974	0.919

Table 4. Average deterministic and probabilistic forecasting scores for both the bias corrected BMC and HRRR TLE-3 from 1200-1700 UTC 24 February 2016.

The superior performance of the HRRR-TLE can be attributed in part to its under-forecasting bias, which for this particular case proved to be largely correct (few freezing rain locations; Fig. 8). As with the BMC, the HRRR transition line is offset and there are misforecasts of rain in southern and central Michigan.

In this case, both the BMC and the HRRR-TLE provided poor forecasts failing to capture two regions of interest where a transition to rain and snow is expected and freezing rain is possible. One possible reason for this failure could be due to the complexity of the mid-latitude cyclone. During this time the cyclone is nearing the end of rapid intensification and is responsible for multiple frontal boundaries in addition to intense snow bands on the northeast side of the cyclone. Over the Great Lakes region, the HRRR-TLE largely forecast anomalously warm temperatures (also reflected in the categorical temperatures). The BMC post-processor is not designed to correct for problems with the HRRR-TLE's variables – rather the "matching" is of inputs (HRRR-TLE variables) to outputs (probability of three precipitation phase types). Given this matching, it is possible to partially correct for bias in the HRRR, as with Model Output Statistics. Surface temperatures have a large influence on the BMC probabilities, however, and thus high level performance requires better forecasts for variables like temperature. In other words, improvements in HRRR forecasts can be further leveraged through post-processing of this kind.

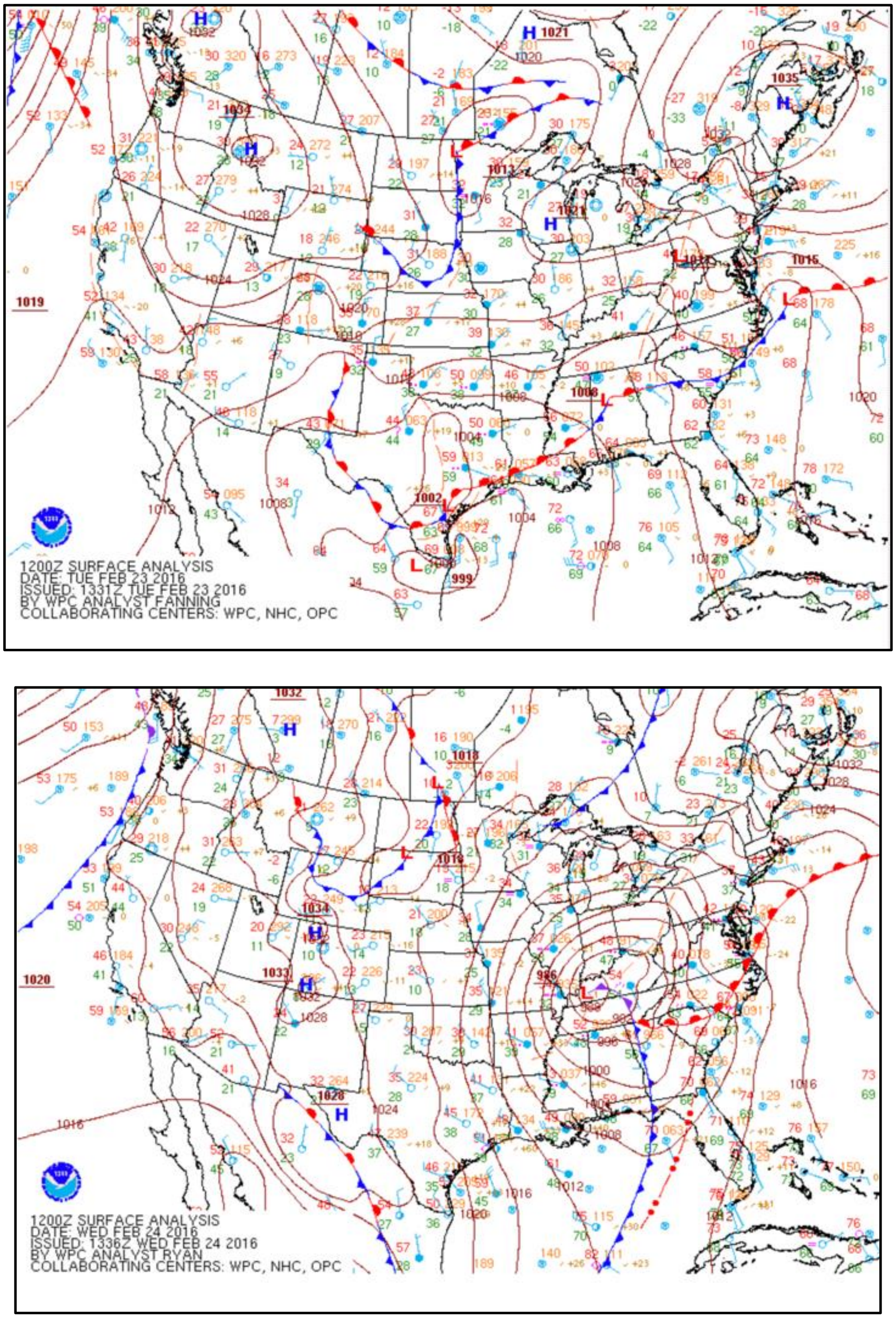
1200Z SURFACE ANALYSIS
DATE: TUE FEB 23 2016
ISSUED: 1331Z TUE FEB 23 2016
BY WPC ANALYST FANNING
COLLABORATING CENTERS: WPC, NHC, OPC
1200Z SURFACE ANALYSIS
DATE: WED FEB 24 2016
ISSUED: 1336Z WED FEB 24 2016
BY WPC ANALYST RYAN
COLLABORATING CENTERS: WPC, NHC, OPC

Figure 6. Observed surface conditions for 1200 UTC (a) 23 February 2016 and (b) 24 February 2016. These analyses were obtained from https://www.wpc.ncep.noaa.gov/archives/web_pages/sfc/sfc_archive_maps.php?

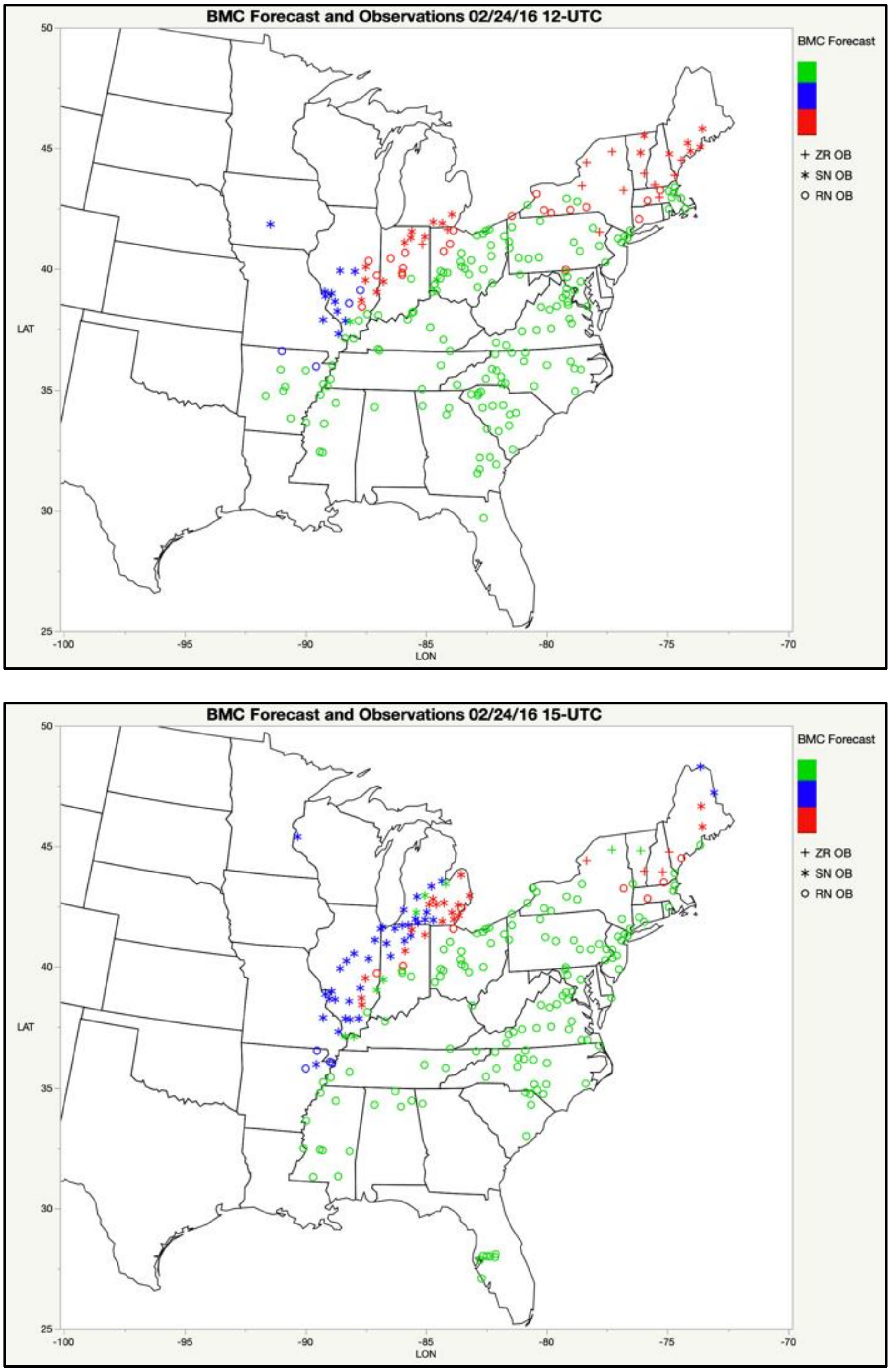

Figure 7. Observations from stations reporting freezing rain (plus), snow (star), rain (circle), or freezing rain and snow (square) and BMC forecast of freezing rain (red), snow (blue), and rain (green) on 24 February 2016 at (a) 1200 UTC and (b) 1500 UTC.

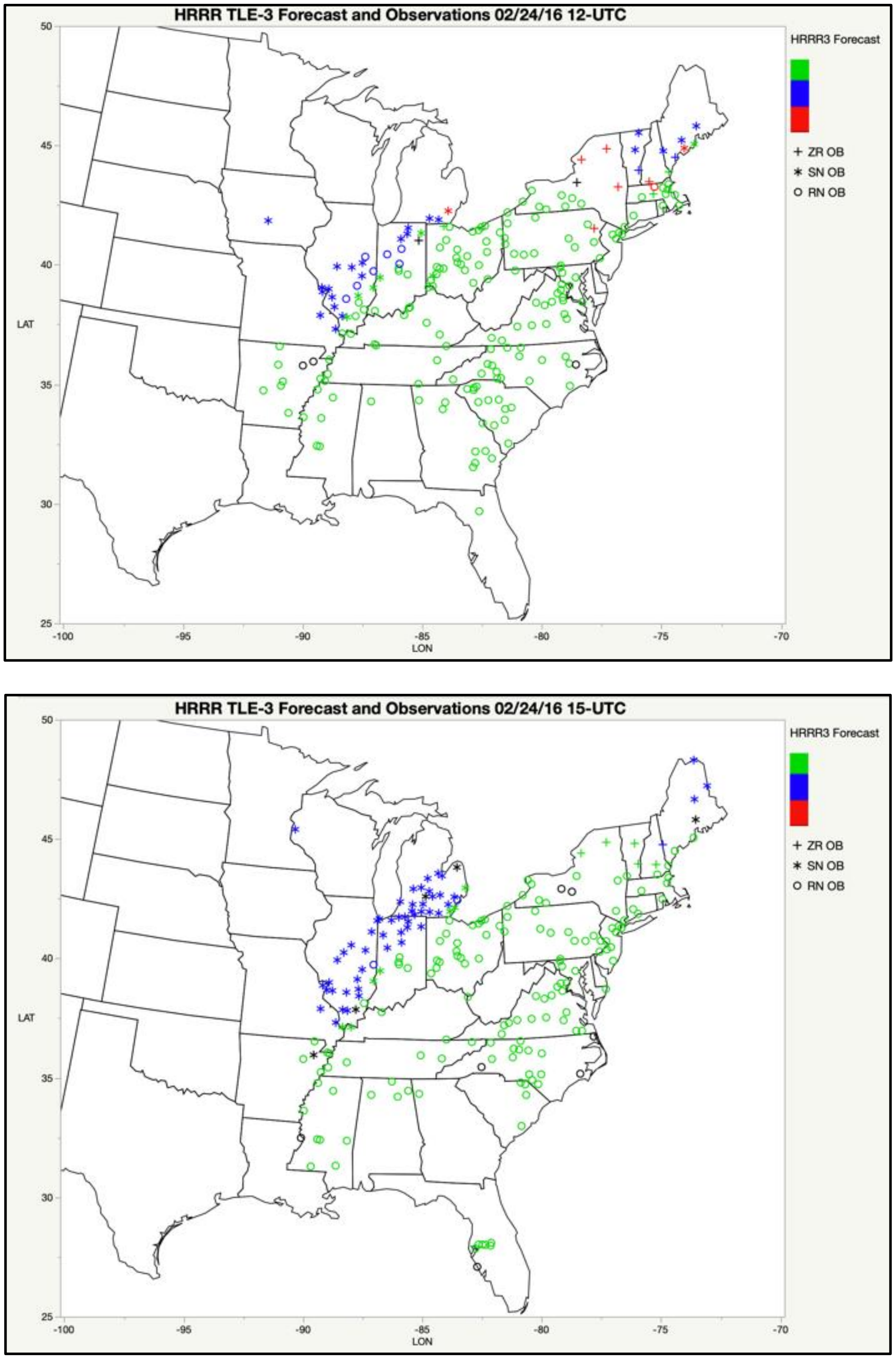

Figure 8. Observations from stations reporting freezing rain (plus), snow (star), rain (circle), or freezing rain and snow (square) and HRRR TLE-3 forecast of freezing rain (red), snow (blue), and rain (green) on 24 February 2016 at (a) 1200 UTC and (b) 1500 UTC. If the symbol is black then that precipitation phase was observed but no precipitation was forecast at that time and location.

3. Future Directions

Freezing rain continues to be a major forecast challenge. In this study, with the HRRR as input, Evolutionary Programming is used to generate algorithms that produce probabilistic forecasts for wintertime precipitation phase (rain, freezing rain, and snow). Further adjustments based on bias correction and Bayesian Model Combination provide additional improvements.

The algorithms outperform the HRRR on independent test data, but nevertheless, there are forecast failures, as illustrated by the second test case. The approach used here is readily applicable to any HRRR forecast, not simply the time-lagged ensemble. By considering the details of the individual forecast algorithms, we have been able to illustrate the sensitivity of the forecast to particular inputs – this information may be used by forecasters to adjust their own view of the likelihood of freezing rain events, based upon their assessment of the reliability of a particular HRRR input in a particular situation.

Future work includes answering the following questions and concerns, some specific to this work and others of a more general interest. Some of the former include: What are the major sensitivities that limit precipitation phase predictions when employing this method? Effective machine learning requires ample data for training, validation, and testing. For this work, we have used a relatively limited sample and intentionally restricted it to cases associated with winter cyclones. Thus, more data and data applicable to broader circumstances would likely improve performance and provide better generalization to a variety of scenarios.

Issues arising in this work which are broadly applicable to machine learning efforts include the following. What is the best procedure for selecting a set of algorithms to be used in the BMC process, to insure both accuracy and precision? We have employed one reasonable and defensible approach here but there may be better ways. In training, one must employ a measure of success, but experimentation is needed to determine which measure (or combinations) will produce the best performance. A general problem in machine learning is training with imbalanced data – the situation where one or more categories of interest are infrequent owing to their inherent climatologies. This situation is generally managed using over- or under-sampling, but in either case, some data are lost. Adaptive EP modes have been developed but not fully explored. Clearly in operational situations, a machine learning approach that takes advantage of adaptation would be preferred and requires further development. Finally, the EP mode used in this study is only one of several developed by PI Roebber, and others are under active development (e.g., Roebber and Crockett 2019). Further development of these approaches is needed.

References

Alexander, C., S.G. Benjamin, S.S. Weygandt, D.C. Dowell, and E.P. James, 2014: Time-Lagged 3-km ensemble high-resolution rapid refresh (HRRR) forecasts for key convective storm, fire weather and wind energy events in 2013. *26th Conference on Weather Analysis and Forecasting*, 2-6 February 2014, Atlanta, GA.

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

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

Hamill, T. M., and J. S. Whitaker, 2007: Ensemble calibration of 500-hPa geopotential height and 850-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.

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*, San Jose, CA, IEEE, 2657–2663.

NBC, 2015: Winter Weather: Ice Storm Leaves 300,000 Without Power in South. NBC News, 27 September 2019, https://www.nbcnews.com/news/weather/winter-weather-ice-storm-leaves-300-000-without-power-south-n307306.

NOAA, 2015: 2015 February 16-17 Winter Storm. NWS, 27 September 2019, https://www.weather.gov/ffc/20150216_winterstorm

NOAA, 2016: Recap of February 24, 2016 Heavy Snow. NWS, 8 October 2019, https://www.weather.gov/lot/2016feb24_snow

Novak, D. R., D. R. Bright, and M. J. Brennan, 2008: Operational forecaster uncertainty needs and future roles. *Wea. Forecasting*, **23**, 1069–1084.

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.

Roebber, P.J., 2010: Seeking consensus: A new approach. *Mon. Wea. Rev*., **138**, 4402-4415.

Roebber, P.J., 2013: Using evolutionary programming to generate skillful extreme value probabilistic forecasts. *Mon. Wea Rev*., **141**, 3170-3185.

Roebber, P.J., 2015a: Evolving ensembles. *Mon. Wea Rev*., **143**, 471-490.

Roebber, P.J., 2015b: Using evolutionary programs to maximize minimum temperature forecast skill. *Mon. Wea. Rev*., **143**, 1506-1516.

Roebber, P.J., 2015c: Adaptive evolutionary programming. *Mon. Wea. Rev*., **143**, 1497-1505.

Roebber, P.J., 2018: Using Evolutionary Programming to Add Deterministic and Probabilistic Skill to Spatial Model Forecasts. *Mon. Wea. Rev*., **146**, 2525-2540.

Roebber, P.J., and J. Crockett, 2019: Using a coevolutionary post-processor to improve skill for both forecasts of surface temperature and nowcasts of convection occurrence. *Mon. Wea. Rev*., **147**, 4241-4259.