The 2013 HMT-HPC Winter Weather Experiment

Final Report

May 1, 2013

1. INTRODUCTION

The Hydrometeorological Testbed at the Hydrometeorological Prediction Center (HMT-HPC[1]) hosted 23 forecasters, researchers, and model developers (Appendix A) at its third annual Winter Weather Experiment from January 15 – February 15, 2013. This year's experiment continued to focus primarily on exploring methods to better quantify and communicate uncertainty in winter weather forecasts. Specifically, the experiment aimed to address five goals:

- Explore the use of ensembles to better quantify uncertainty in winter weather forecasts.
- Explore the use of explicit snowfall accumulations from model microphysics schemes.
- Explore the utility of ensemble data-mining and post-processed products in the forecast process.
- Explore the utility of longer range winter weather outlook forecasts.
- Explore how to more effectively communicate uncertainty in winter weather forecasts to improve decision support services.

This report summarizes the activities, findings, and operational impacts of the experiment.

2. EXPERIMENT DESCRIPTION

Data

The 2013 experiment featured three experimental ensemble systems (Table 1). The Air Force Weather Agency (AFWA) provided two different 10-member ensembles. Both are multi-physics, multi-initial condition, Advanced Research WRF (ARW) ensembles that use the data assimilation schemes from each of their member models to generate initial and boundary condition diversity (Table 2). The AFWA-WRF is a 20-km ensemble over the Northern Hemisphere, while the AFWA-HR is a 4-km convection-allowing ensemble over the CONUS. HPC provided the HPC Autoensemble (HPCENS), an internally generated 28-member 32-km ensemble consisting of all 21 members of the Short Range Ensemble Forecast System (SREF), two versions of the GFS Ensemble System (GEFS) ensemble mean, the European Center for Medium Range Weather Forecasting (ECMWF) ensemble mean (ECENS), and the deterministic runs of the North American Model (NAM), GFS, Canadian Global Environmental Multiscale model (CMC), and ECWMF. The two versions of the GEFS ensemble mean use different snowfall accumulation methods, while the NAM uses the 4-km CONUS nest through 60 hours

[1] On March 5, 2013 the Hydrometeorological Prediction Center (HPC) was renamed the Weather Prediction Center (WPC). As a result, the HMT-HPC was also renamed the HMT at the Weather Prediction Center (HMT-WPC).

Table 1. Models and ensembles used during the 2013 HMT-HPC Winter Weather Experiment. All models were initialized at 00 UTC except the SREF (21 UTC) and the 20-km AFWA (06 UTC). The SREF snow-to-liquid ratios were capped at 28:1. Experimental guidance is shaded.

Provider	Model	Resolution	Forecast Hours	Snow-to-Liquid Ratio
EMC	SREF (21 members)	16 km	87	For temperatures < 5°C: $SLR = (273.15 - T_{2m}) + 8$
AFWA	WRF-ARW (10 members)	20 km	144	For temperatures < 4°C: $SLR = 5\sqrt{5 - (T_{2m} - 273.15)}$
AFWA	WRF-ARW (10 members)	4 km	72	For temperatures < 4°C: $SLR = 5\sqrt{5 - (T_{2m} - 273.15)}$
HPC	Autoensemble (28 members)	32 km	72	Average of: 11:1 Climatology Roebber Technique applied to the NAM Roebber Technique applied to the GFS
EMC	NAM	12 km	84	Roebber Technique
EMC	NAM	12 km	84	Rime factor-modification to Roebber Technique

Table 2. AFWA-WRF (20 km) and AFWA-HR (4 km) ensemble membership. The convective scheme is only used in the AFWA-WRF. Initial and lateral boundary conditions are provided by the UKMET (UM), the Global Forecast System (GFS), and the Canadian Global Model (GEM).

Member	IC/LBC	LSM	Microphysics	PBL	Convection
1	UM	NOAH	WSM5	YSU	KF
2	GFS	NOAH	Goddard	YSU	BMJ
3	GEM	NOAH	Ferrier	MYJ	Grell
4	GEM	NOAH	Thompson	YSU	KF
5	UM	NOAH	Thompson	YSU	BMJ
6	GFS	NOAH	Thompson	MYJ	Grell
7	GEM	NOAH	Goddard	YSU	BMJ
8	GEM	NOAH	WSM5	YSU	BMJ
9	UM	RUC	Ferrier	MYJ	KF
10	GFS	NOAH	WSM5	YSU	Grell

before transitioning to the 12-km parent model. Data from the operational SREF were used for comparison to the experimental ensembles.

The 2013 experiment also featured a new snowfall accumulation technique that incorporates information directly from the NAM's microphysics scheme. Currently, snowfall accumulations are derived from the NAM using a snow-to-liquid ratio (SLR) provided by the Roebber Technique (Roebber et al. 2003) and the instantaneous precipitation type from the NCEP dominant precipitation type method (Manikin 2005) every six hours. The new technique, called the rime factor-modified snowfall accumulation, modifies the initial SLR provided by the Roebber Technique by incorporating information about the amount of riming on an individual ice particle due to riming and liquid water accretion and the percentage of precipitation that reaches the ground frozen. In addition to accounting for these microphysical parameters, the rime factor-modified snowfall accumulation also provides data at an increased temporal frequency, with hourly data available through 36 hours and three hourly data available from 36-84 hours. Additional information about the computation of the rime factor-modified snowfall accumulations is available in Appendix B. The rime factor-modified snowfall accumulations were compared to snowfall accumulations derived using the Roebber Technique in the operational NAM.

In addition to the traditional model output, several other experimental forecast tools were available during the experiment. The operational SREF featured a weighted mean that is calculated by determining the difference between the individual ensemble member solutions and the ensemble mean solution. The goal of this technique is to provide improved ensemble mean predictions by taking into account current ensemble member performance (Du and Zhou 2011). In collaboration with the Environmental Modeling Center (EMC), Stony Brook University (SBU) provided a real time ensemble sensitivity analysis tool. This tool aims to identify how changes in the initial conditions may ultimately impact the forecast outcome by identifying the upstream features that are driving the forecast differences (Colle and Chang, 2011; Chang et al. 2013). The experiment also featured the Extreme Forecast Index (EFI) developed by the ECMWF. This tool relates the current forecast probabilities from the ECMWF ensemble to probabilities derived from a model climatology for several sensible weather variables with the goal of alerting forecasters to anomalous events (Zsoter 2006; Lalaurette 2003). Additional details about each of these tools can be found in Appendix B.

Snowfall verification was based on the 20-km gridded HPC snowfall analysis. This analysis is generated through a two-step Barnes objective analysis that incorporates data from COOP, CoCoRaHS, and METAR observations. Where METAR observations indicate that the precipitation type is snow, quantitative precipitation estimates (QPE) from the Climatology-

Calibrated Precipitation Analysis (CCPA; Hou et al. 2013) and climatological snow-to-liquid ratios (Baxter et al. 2005) are used to determine snowfall accumulations. Freezing rain forecasts were verified based on METAR observations at points selected in advance by the forecast team.

Daily Activities

The 2013 experiment featured three complimentary activities. A detailed version of the daily schedule can be found in Appendix C.

a. Experimental Forecasts and Subjective Model Evaluations

Each morning, participants used a combination of operational and experimental 00 UTC guidance (21 UTC for the SREF and 06 UTC for the AFWA-WRF) to create 24 hour experimental probability of exceedance forecasts for a storm of interest during either the Day 1 (24 – 48 hour) or Day 2 (48 – 72 hour) time period. These forecasts covered the 00 – 00 UTC period and highlighted areas with a slight (10%), moderate (40%), and high (70%) probability of exceeding 2 in, 4 in, and 8 in of snowfall (4 in, 8 in, and 12 in for high-end events) during the 24 hour period (Fig. 1). During events that also featured freezing rain, one or more of the snowfall thresholds were replaced with an appropriate freezing rain threshold (0.01 in, 0.10 in, or 0.25 in). In addition to the graphical forecasts, participants also wrote a brief forecast discussion and rated their overall forecast confidence as above average, average, or below average. When choosing the daily forecast area, priority was given to storms presenting the greatest forecast challenge.

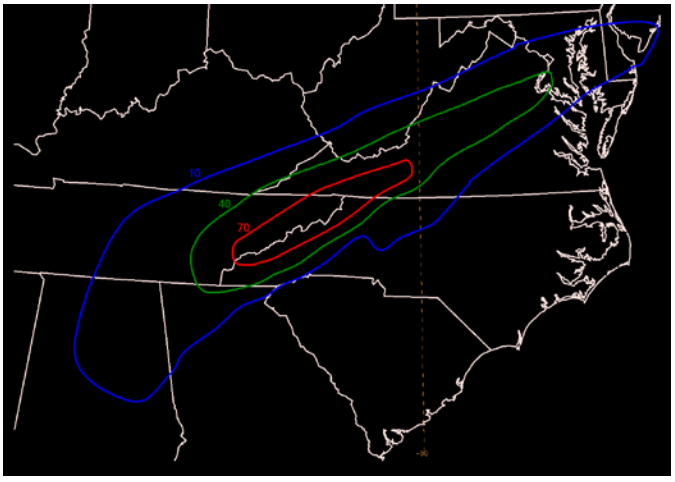

Figure 1. Experimental Day 1 forecast from 16 January 2013 indicating the probability of exceeding 2 inches of snowfall during the 24 hour period ending 00 UTC 18 January 2013.

During the afternoon, participants were asked to subjectively evaluate the performance of both the experimental forecasts and the experimental model guidance for events from the previous week of the experiment. The subjective evaluations consisted of a series of survey questions designed to determine whether the experimental models provided additional value to forecasters compared to either the operational SREF (ensemble guidance) or the operational NAM (rime factor-modified snowfall accumulations).

b. Decision Support

In addition to the experimental forecasts, participants prepared a public forecast graphic that highlighted the anticipated winter weather hazards (Fig. 2); they then used this graphic to conduct a mock decision support briefing for a regional emergency management group. The goal of the graphic was to highlight *where* the winter weather event would occur within the daily domain, *when* the event would occur, and *what* specific hazards would be associated with it. While preparing the graphic, participants were asked to consider additional hazards beyond the snowfall and freezing rain accumulations that were the focus of the morning experimental forecasting activities such as the effects of temperature, wind, and event timing.

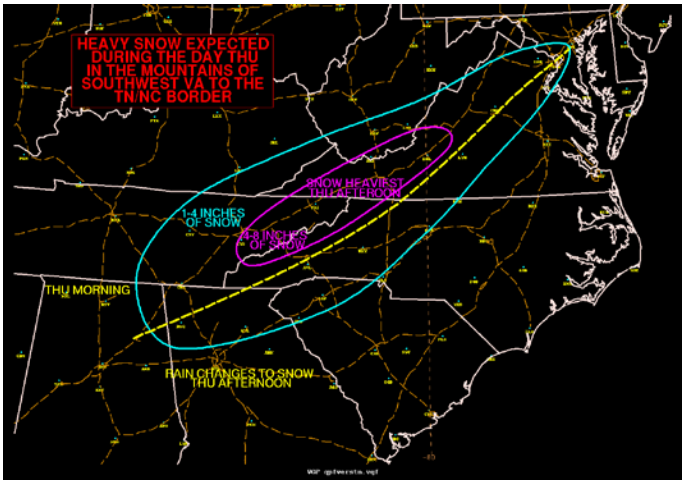

Figure 2. Experimental Day 1 public forecast graphic from 16 January 2013 highlighting the anticipated hazards over the 24 hour period ending 00 UTC 18 January 2013.

Each week, participants were initially encouraged to explore different ways to clearly convey the key forecast information to the public through the use of different colors, line types, etc. Later in the week they were introduced to a proposed event-type classification scheme based on Rooney (1967) to guide the creation of the graphic (Table 3). After being introduced to both approaches, by the end of the week participants were able to choose which approach (or combination of approaches) to use.

Table 3. Proposed event-type classification scheme used to create some of the public forecast graphics during the decision support component. This scheme was updated several times over the course of the experiment based on participant feedback.

Type	Description	Color	Notes
1*	Low-end plowable snowfall	Light Blue	▪ 1-4 inch snowfall ▪ Traffic delays expected ▪ Possible school delays and cancelations
2*	High-end plowable snowfall	Medium Blue	▪ 4-12 inch snowfall ▪ Road closures possible ▪ Numerous school cancelations
3*	Crippling snow event	Dark Blue	▪ 12-24+ inch snowfall ▪ Most roads closed
4	White Rain	Green	▪ 1-3 inch snowfall on unpaved surfaces only ▪ Temperatures at or above freezing
5	≤ 1" snow with extremely cold or rapidly falling temperatures	Purple	▪ Snow accumulates quickly on paved surfaces
6	Wintery Mix	Pink	▪ Mixture of snow, sleet, and freezing rain ▪ Snow and sleet accumulations > 2 inches ▪ Possible school delays and cancelations
7	Ice Storm	Red	▪ Ice accumulations > 0.10 inches ▪ Hazardous travel
8	Ice Glaze	Light Orange	▪ Ice accumulations < 0.05 inches ▪ Hazardous travel
9	Snow to Rain	Yellow	▪ Several inches of snow followed by rain
10	Rain to Snow	Brown	▪ Rain followed by several inches of snow
11	Highly uncertain and/or two or more scenarios possible	White	

*impacts exacerbated when accompanied by strong winds

As in the 2012 experiment, the Weather for Emergency Management Decision Support (WxEM) team provided daily feedback on both the mock briefings and the public forecast graphics. In addition, the WxEM team also provided a weekly orientation that introduced participants to the emergency management community, discussed the specific roles of both the forecaster and the emergency management community in the decision support process, and provided tips for successful decision support briefings.

During the last week of the experiment (February 11 – 15), the decision support briefing was conducted in collaboration with the Aviation Weather Testbed's (AWT) Winter

Weather Experiment and included discussions of both the ground-based (HMT-HPC) and in-flight (AWT) winter weather hazards. This coordination provided an opportunity to explore both cross-testbed collaboration and remote experiment participation.

c. Day 4-5 Winter Weather Outlook Forecasts

A new aspect of this year's experiment was the addition of experimental Day 4-5 winter weather outlook forecasts. The goal of this activity was to explore the utility of longer range winter weather forecasts and initiate discussion about the potential for future expansion of HPC's winter weather product suite. Using the most recently available guidance, participants were asked to highlight areas across the country with a threat for winter weather during the Day 4-5 (00 – 00 UTC) period (Fig. 3). These forecasts evolved during the experiment from a single forecast for the entire 48 hour period during Week 1 to separate Day 4 and Day 5 forecasts with short text descriptions for each highlighted area by the end of the experiment.

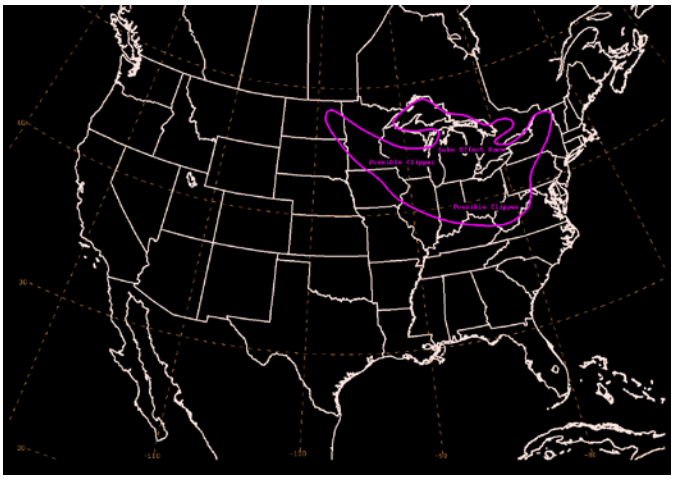

Figure 3. Experimental Day 4 outlook forecast from 29 January 2013 highlighting the area expected to receive winter weather during the 24 hour period ending 00 UTC 3 February 2013.

3. CASES

The experiment period was characterized by a broad mean trough over much of the central and eastern United States with a pronounced ridge just off the west coast (Fig. 4a). While the 2012-2013 winter was characterized overall by slightly warmer than normal temperatures across the central and eastern U.S. and cooler than normal temperatures in the western U.S. (not shown), the experiment period was characterized by near normal temperatures across the Great Lakes and Northeast, with the anomalously warm temperatures largely confined to the central and southern plains (Fig. 4b).

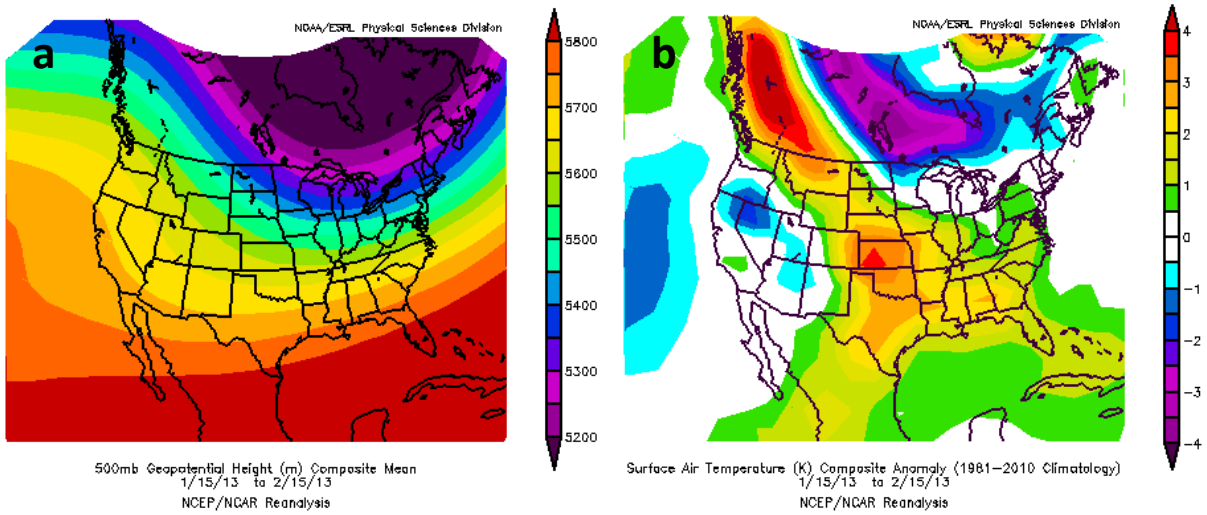

Figure 4. Composite mean (a) 500 mb heights and (b) surface air temperature anomalies for the 15 January – 15 February, 2013 period. Images generated from the NCEP/NCAR Reanalysis provided by NOAA/ESRL/Physical Sciences Division (http://www.esrl.noaa.gov/psd/data/composites/day).

With only limited influence from the subtropical jet, the first half of the experiment period was dominated by a series of clipper systems that brought light snowfall to the Great Lakes and Ohio Valley. The weather became more active later in the experiment, beginning with a strong cold front that moved across the country at the end of January bringing widespread 4-6 inch snowfall totals to Iowa and Wisconsin and numerous reports of severe weather farther south. Early February brought blizzard conditions to the northeast, with snowfall totals of 24 inches or more reported across parts of New York and New England. This storm was quickly followed by a blizzard across the northern plains and upper Midwest that brought 10 or more inches of snow to Minnesota and the Dakotas. Finally, although the event occurred after the experiment ended, the subjective verification was extended to include a system in late February that produced widespread 6 inch snowfall totals across the central plains into the upper Midwest. A complete list of the snowfall events investigated during this year's experiment can be found in Table 4.

Table 4. Experimental forecasts and subjective verification for the 2013 HMT-HPC Winter Weather Experiment. D1 and D2 refer to Day 1 (24 – 48 hr) and Day 2 (48 – 72 hr) forecasts, respectively. Supplemental verification was completed by HPC forecasters in the weeks following the experiment to provide a more robust evaluation. Events included in the objective verification are marked with an asterisk (*).

Forecast Valid Time	Forecast		Verification		Forecast Area	Notes
00Z 5 Jan 2013			D1	D2	Southern Plains	
00Z 12 Jan 2013			D1	D2	Northern Rockies	
00Z 17 Jan 2013	D1*		D1	D2	Mid Atlantic to Northeast	
00Z 18 Jan 2013	D1*		D1	D2	Southeast to Mid Atlantic	
00Z 19 Jan 2013	D1*		D1	D2	Mid Atlantic to Northeast	
00Z 21 Jan 2013		D2*	D1	D2	Great Lakes	
00Z 24 Jan 2013	D1*		D1	D2	Great Lakes	
00Z 26 Jan 2013	D1	D2*	D1	D2	Upper Midwest to Mid Atlantic	
00Z 28 Jan 2013		D2	D1	D2	Mid-Mississippi Valley to Upper Midwest	
00Z 30 Jan 2013	D1		D1	D2	Pacific Northwest to Northern Rockies	
00Z 31 Jan 2013	D1*		D1	D2	Central Plains to Upper Great Lakes	Significant snowfall in upper Midwest; severe weather across central and southern U.S.
00Z 2 Feb 2013		D2*	D1	D2	Mid Atlantic to Northeast	
00Z 3 Feb 2013		D2*	D1	D2	Ohio Valley	
00Z 4 Feb 2013		D2*	D1	D2	Mid Atlantic to Northeast	
00Z 6 Feb 2013	D1*		D1	D2	Upper Midwest to Mid Atlantic	
00Z 8 Feb 2013		D2*	D1	D2	Great Lakes	
00Z 9 Feb 2013		D2*	D1	D2	Great Lakes to Northeast	Northeast blizzard
00Z 10 Feb 2013		D2*	D1	D2	Northeast	Northeast blizzard
00Z 11 Feb 2013		D2*	D1	D2	Central Rockies to Upper Midwest	Northern plains/upper Midwest blizzard
00Z 13 Feb 2013	D1*		D1	D2	Central and Southern Plains	
00Z 15 Feb 2013	D1*	D2*	D1	D2	Mid Atlantic to Northeast	
00Z 17 Feb 2013	D1*	D2*	D1	D2	Great Lakes and Mid Atlantic to Northeast	
00Z 22 Feb 2013			D1	D2	Central Plains	Significant snowfall across central U.S.
00Z 23 Feb 2013			D1	D2	Central Plains to Upper Midwest	Significant snowfall across central U.S.

4. EXPERIMENTAL MODEL PERFORMANCE

Through a combination of evaluations completed during the experiment and supplemental evaluations completed after the experiment ended, a total of 48 cases were evaluated, 24 on Day 1 and 24 on Day 2.

In addition to these subjective evaluations, objective verification statistics were compiled for events in which an experimental forecast was issued (23 cases). The verification used observed accumulations from the gridded HPC snowfall analysis and was performed for 4 in and 8 in snowfall forecasts over the area of interest identified each day by participants. All model forecasts were re-gridded to a common 20-km grid. Two cases that did not include either a 4 in or 8 in snowfall forecast were removed from the dataset, and one case was removed because of the limitations inherent in compiling a gridded snowfall analysis in mountainous terrain, reducing the total number of cases to 20. Day 1 and Day 2 forecasts were combined for the purposes of the objective verification because of the small sample size. Prior to calculating the Gilbert Skill Score (GSS; equitable threat score), the frequency bias of the forecasts was removed by matching the frequency distribution of the forecast snowfall accumulations to that of the analysis in order to eliminate the sensitivity of the equitable threat score to the frequency bias.

AFWA-WRF and AFWA-HR Ensembles

Overall, subjective evaluations indicate that the operational SREF generally provided better winter weather forecast guidance than both the AFWA-WRF and the AFWA-HR, but performance varied considerably between the Day 1 and Day 2 forecasts (Fig. 5). Both AFWA ensembles struggled during the Day 1 period, with the majority of forecasts subjectively rated as either “worse” or “much worse” than the SREF. During the Day 2 period, however, the ensemble performance was mixed, with both ensembles receiving a similar number of “worse” or “much worse” ratings as they received “better” or “much better” ratings.

While participants appreciated the additional forecast details that the 4-km AFWA-HR was able to provide, both ensembles often appeared to overforecast snowfall accumulations (Fig. 6), which contributed to some of the poor ratings in the subjective evaluation surveys. Although the ensemble configuration was slightly different, a similar tendency was also observed in the AFWA-HR during the 2012 Winter Weather Experiment. While not specifically investigated this year, the tendency to overforecast snowfall accumulations during the 2012 experiment was attributed to a combination of the snow-to-liquid ratio algorithm used in the ensemble and potential biases in the ensemble mean quantitative precipitation forecast (QPF). Since the

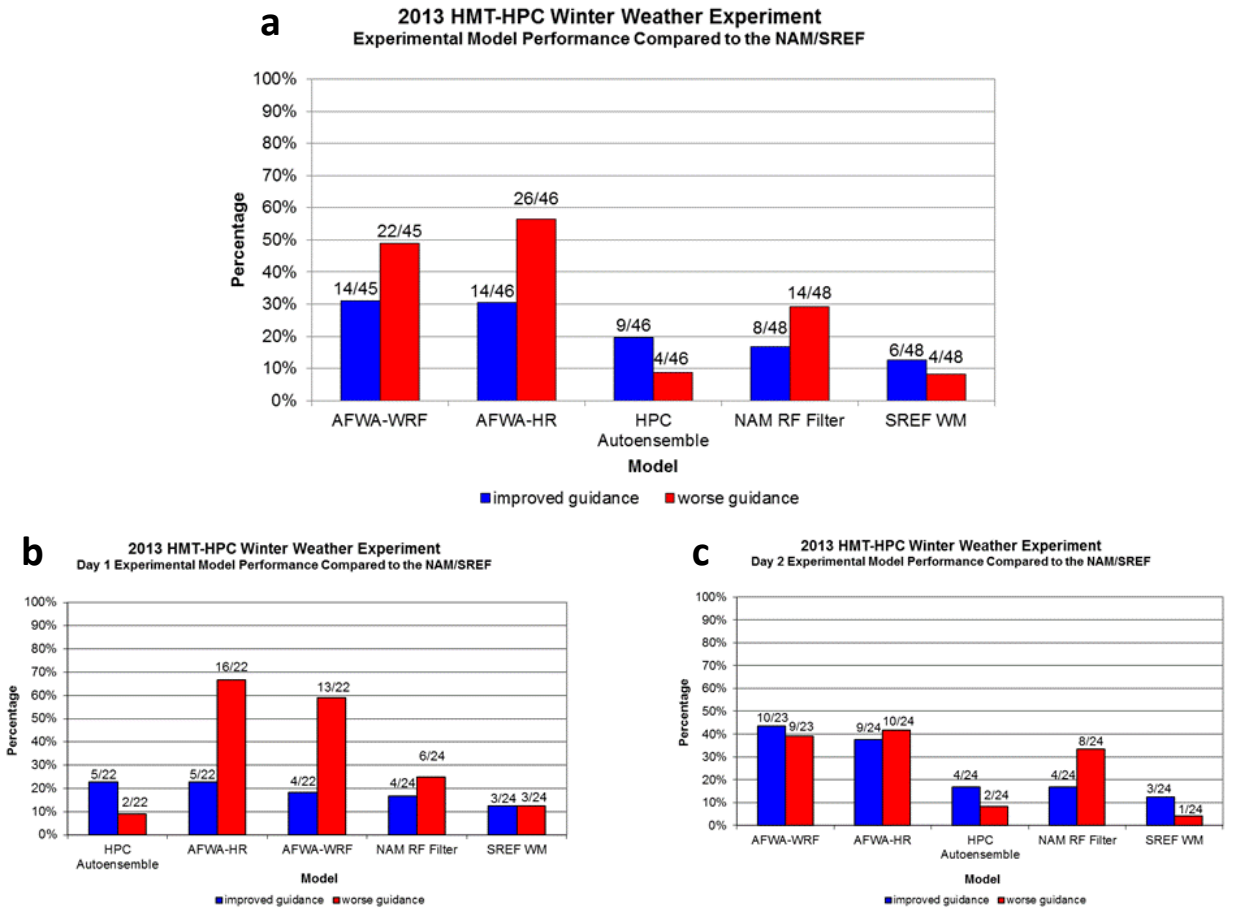

Figure 5. Experimental model performance for (a) the combined Day 1-2 period, (b) Day 1, and (c) Day 2 based on participant feedback from subjective model evaluations conducted during the 2013 HMT-HPC Winter Weather Experiment. Participants were asked to determine whether the forecasts from the 00 UTC experimental guidance (06 UTC AFWA-WRF, 21 UTC SREFWM) were much better, better, about the same, worse, or much worse than the corresponding operational guidance based on observations from the gridded HPC snowfall analysis. The experimental ensembles were compared to the operational 21 UTC SREF, while the rime factor-modification (RF Filter) was compared to the operational 00 UTC NAM using the Roebber Technique.

snow-to-liquid ratio used in the AFWA ensembles this year is identical to the ratio used during last year's experiment, it is likely that this is again a contributing factor.

Another consideration that impacts the evaluations is the quality of the HPC gridded snowfall analysis. This analysis is calculated on a 20-km grid using a two-step Barnes objective analysis. In cases with limited snowfall observations or snowfall amounts that vary considerably over short distances (ex: sparse population, terrain, lake-effect, etc.), the analysis often struggles to accurately depict the true snowfall accumulations. It is possible that some of the higher

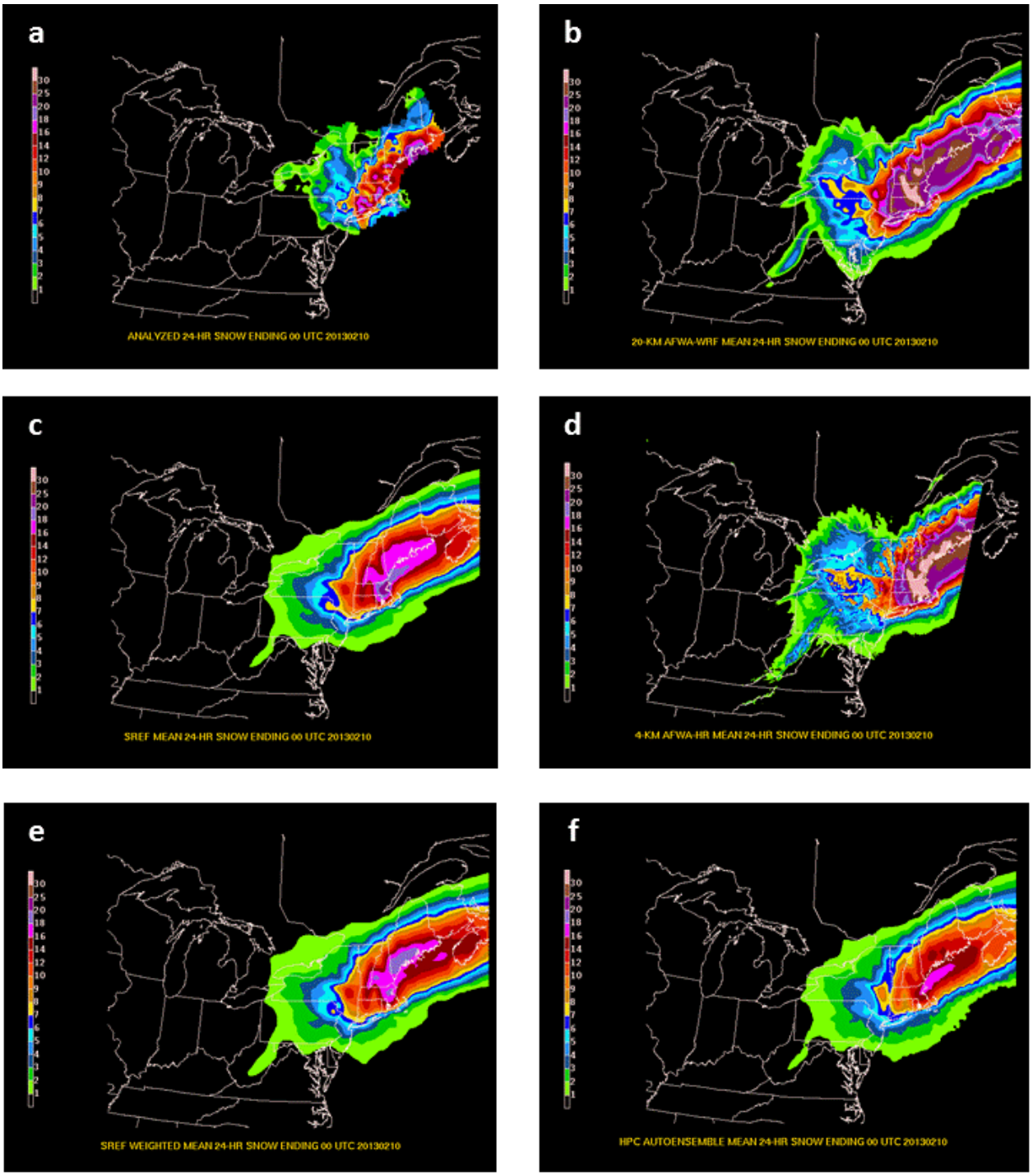

Figure 6. (a) Observed 24 hour snowfall ending 00 UTC 10 February 2013 from the HPC snowfall analysis and the corresponding Day 2 ensemble mean forecasts from the (b) AFWA-WRF, (c) SREF, (d) AFWA-HR, (e) SREF Weighted Mean, and (f) HPC Autoensemble.

snowfall amounts depicted in the AFWA ensembles were realistic on small scales, but were not captured by the HPC snowfall analysis, resulting in the forecast being penalized during the

evaluation process. While this undoubtedly occurred in some cases, the high snowfall bias was not restricted to mesoscale events and was also seen in large synoptic scale systems that were well sampled with snowfall observations.

Probabilistically, both AFWA ensembles were generally able to capture the range of possible solutions (Fig. 7). While the subjective verification results are fairly similar between the two ensembles, on several occasions participants noted that the AFWA-WRF produced a probability field that was less dispersed and contained noticeably higher probabilities than the AFWA-HR (Fig. 8). Since both ensembles contain the same number of members, the reason for these differences isn't clear, although like in last year's experiment the high resolution nature of the AFWA-HR had a tendency to result in unrealistically discontinuous probability areas.

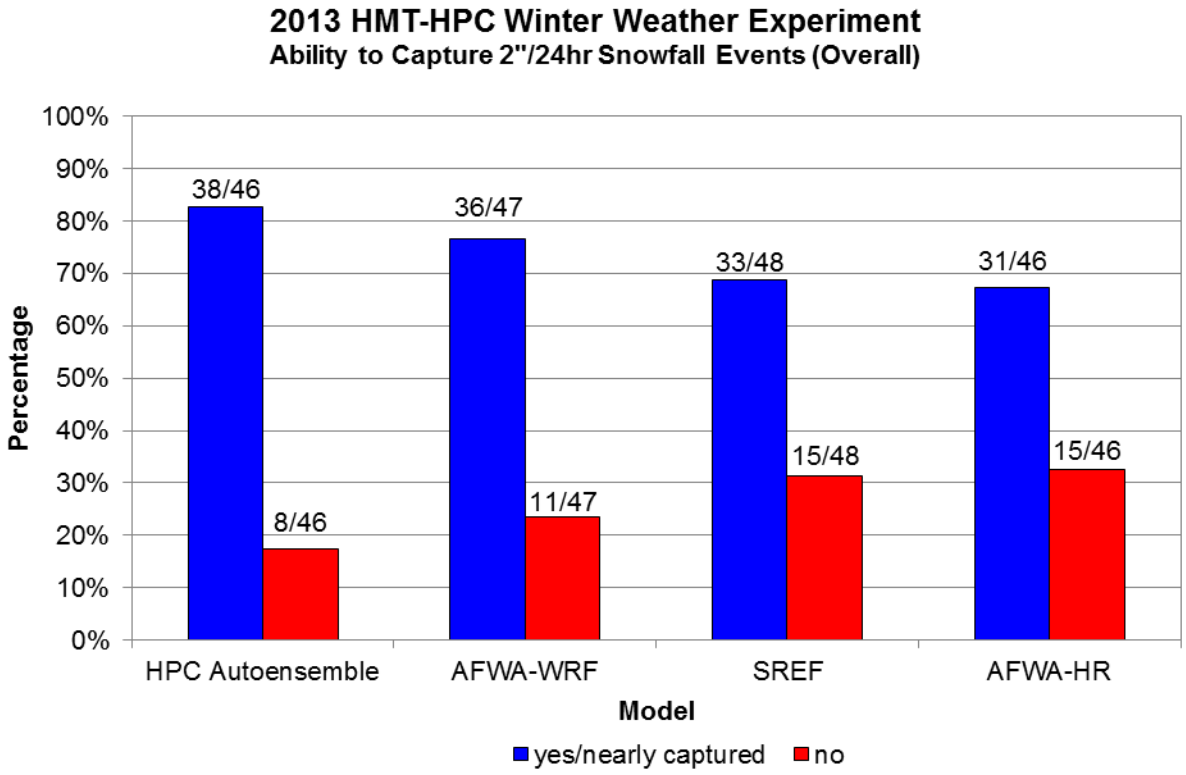

Figure 7. Subjective evaluation of the ability of the experimental ensembles to capture the 2 in/24 hr snowfall events with the model 1% probability contour. Participants were asked to determine whether the observed 2 inch snowfall area fell entirely within the 1% probability contour from the model. "Nearly captured" represents cases in which there were only very small areas of observed 2 inch snowfall outside of the 1% probability contour.

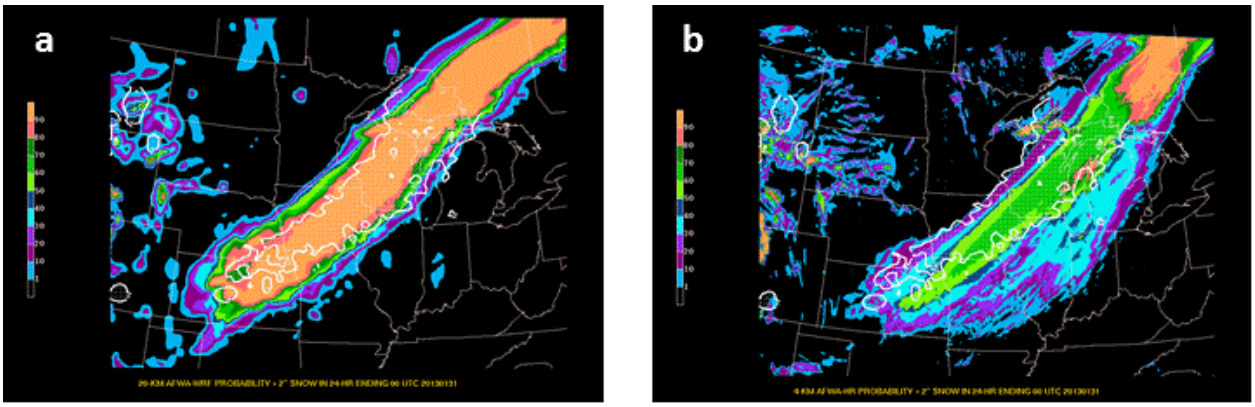

Figure 8. Day 1 forecast of the probability of exceeding 2 inches of snowfall in the 24 hour period ending 00 UTC 31 January 2013 from the (a) AFWA-WRF and (b) AFWA-HR. The white hatched area indicates the observed 2 inch snowfall.

The results from these subjective evaluations are largely supported by the objective verification. Figure 9 summarizes the equitable threat score and frequency bias for the ensembles used during the experiment. Although not statistically significant, the AFWA ensembles provided slightly better guidance than the SREF at the 4 in threshold but slightly worse guidance at the 8 in threshold. However, as was noted in the subjective evaluations, the snowfall forecasts from both AFWA ensembles have a high bias, with the frequency bias values exceeding 2.0 for both the 4 in and 8 in snowfall thresholds. In contrast, the operational SREF has a frequency bias near 1.0 for both snowfall thresholds. This high bias likely contributes to some of the improvement seen at the 4 in threshold, but overall limits the utility of the AFWA guidance.

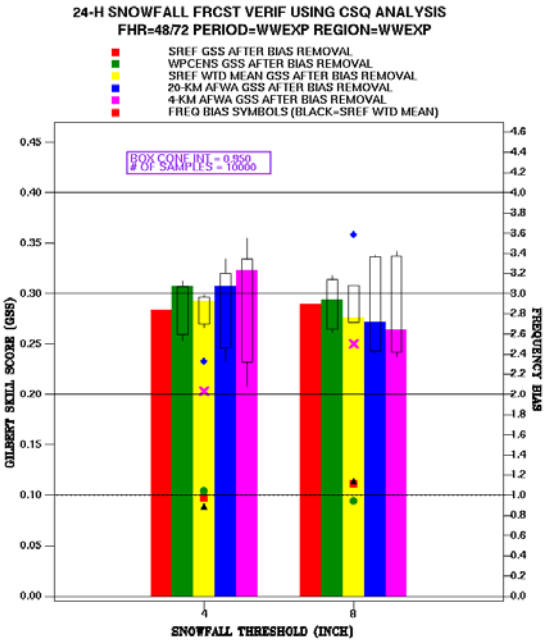

Figure 9. Bias corrected Gilbert Skill Score (GSS; shaded bars with error bars) and frequency bias (symbols) for 4-inch and 8-inch ensemble mean snowfall forecasts during the 2013 HMT-HPC Winter Weather Experiment.

The objective verification of the probabilistic forecasts from the AFWA ensembles provides additional information about the performance of these forecasts that is not available from the subjective evaluations alone. The reliability diagrams for both the 4 in and 8 in snowfall thresholds (Fig. 10) show that the AFWA ensembles are slightly less reliable than the SREF. Like the SREF, the AFWA ensembles tended to be somewhat overconfident, predicting that events would occur more frequently than they were observed in the snowfall analysis. This result is consistent with the results of the subjective evaluations of the probabilistic forecasts, which showed that the AFWA ensembles were generally able to capture the range of possible forecast solutions. The AFWA-WRF, which participants noted sometimes produced a less dispersive and more highly confident probability field, struggled the most with skill, falling below the no-skill line for some probability thresholds at both the 4 in and 8 in snowfall thresholds. When comparing the results of the subjective and objective probabilistic verification, it is important to note that the subjective evaluation focused on relatively light accumulations (2 in), while the objective verification focused on the 4 in and 8 in snowfall thresholds.

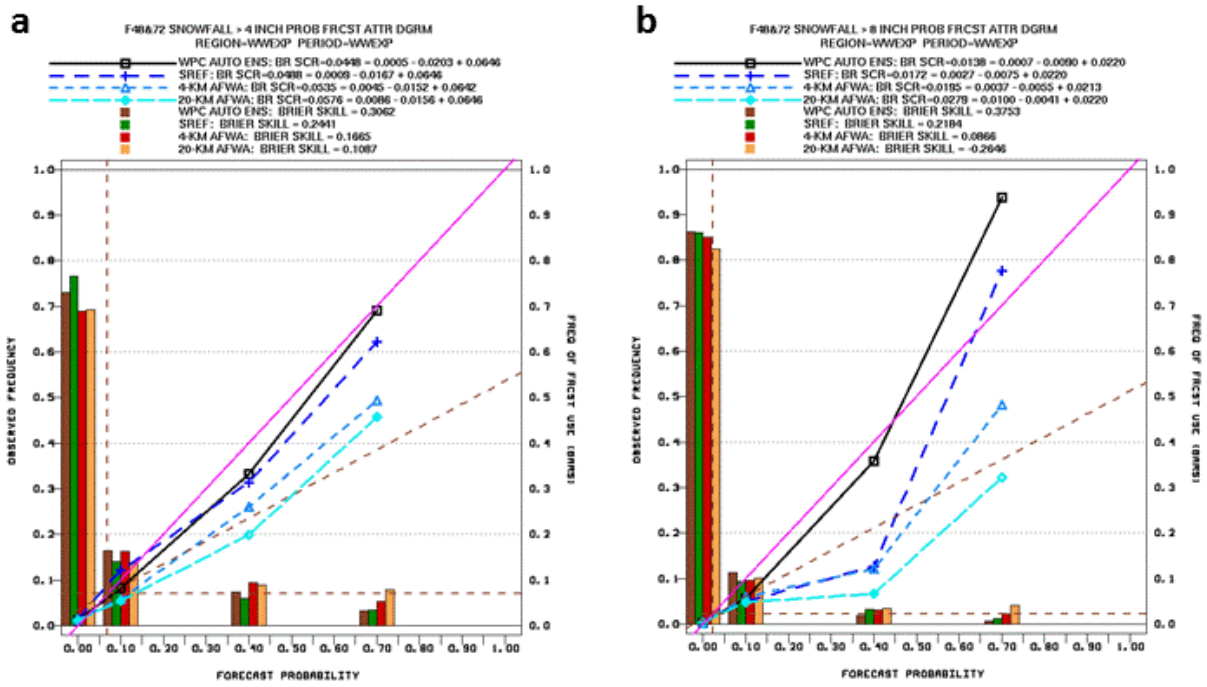

Figure 10. Reliability diagrams for probabilistic (a) 4-inch and (b) 8-inch snowfall forecasts from the four ensemble systems used during the 2013 HMT-HPC Winter Weather Experiment. Shaded bars indicate the frequency distributions for the 0%, 10%, 40%, and 70% forecasts. Perfect reliability is indicated by the solid magenta line. Points below (above) this line indicate overconfident (underconfident) model forecasts. No skill is indicated by the dashed brown line.

While the overall performance of both of the AFWA ensembles was similar, it is difficult to compare their performance on a case by case basis since the ensembles were not run on the same model cycle (00 UTC AFWA-HR and 06 UTC AFWA-WRF). Thus, it was not possible to distinguish between differences caused by resolution and differences caused by initialization. In addition, throughout the experiment it appeared that surface-based fields in the AFWA-WRF (precipitation, snowfall, etc.) were displaced approximately 100 km to the southwest. This discrepancy was particularly noticeable in terrain-driven events (not shown), although it is unclear how much this shift impacted the evaluation results.

HPC Autoensemble

The mean snowfall forecasts from the HPC Autoensemble were generally very similar to those from the SREF, which is not surprising since the SREF accounts for 21 of the ensemble's 28 members. When differences did exist between the two forecasts, the HPC Autoensemble tended to provide subjectively better forecast guidance (Fig. 5a). While the forecasts themselves were generally quite similar, participants tended to favor the HPC Autoensemble because of its inclusion of other model data.

Objectively, the HPC Autoensemble mean was among the best performing ensemble mean forecasts, although the difference is only statistically significant at the 4 in threshold (Fig. 9). Probabilistically, the HPC Autoensemble provided extremely reliable forecasts, with the forecast and observed probabilities almost identical at most thresholds (Fig. 10). This result is supported by the subjective verification, which found that the HPC Autoensemble was better able to capture the range of possible forecast solutions than the operational SREF (Fig. 7).

SREF Weighted Mean

The SREF Weighted Mean forecasts were also very similar to those from the operational SREF. In many cases, the magnitude of the differences was so small that they did not impact the forecast. In the few cases in which larger differences were observed, the SREF Weighted Mean provided slightly better forecast guidance than the operational SREF (Fig. 5a). The objective verification confirms that the forecasts from the SREF Weighted Mean and the operational SREF were very similar, with no statistically significant differences (Fig. 9). These results are consistent with those in Du and Zhou (2011), which notes that the improvement generated by the weighted mean decreases as the number of ensemble members increases.

NAM Rime Factor-Modified Snowfall Accumulations

Overall, the rime factor-modified snowfall accumulations provided slightly worse forecast guidance than the accumulations derived from the operational NAM using the Roebber Technique (Fig. 5a). This result is supported by the objective verification, which shows that the snowfall forecasts from the rime factor modification were slightly worse than those from the operational NAM, with the difference being statistically significant at the 8 in threshold (Fig. 11). The objective verification also reveals that both techniques have a pronounced high bias. While the impact of the rime factor modification was clear in some cases (Fig. 12), subjectively it was often difficult to determine whether the additional forecast details seen in the rime factor-modified accumulations were primarily the result of the rime factor modification technique or were instead driven by resolution differences between the Roebber Technique (40 km) and the modified snowfall accumulation (12 km). These differences in model resolution can result in differences in QPF, temperature, and other fields used to calculate the snowfall accumulation.

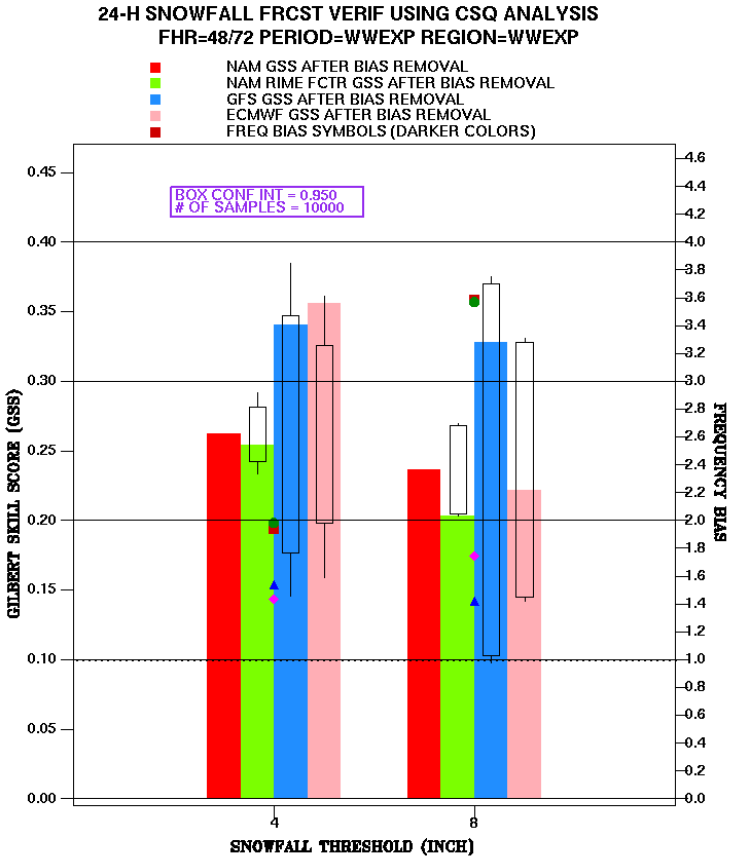

Figure 11. Bias corrected Gilbert Skill Score (GSS; shaded bars with error bars) and frequency bias (symbols) for 4-inch and 8-inch deterministic snowfall forecasts during the 2013 HMT-HPC Winter Weather Experiment.

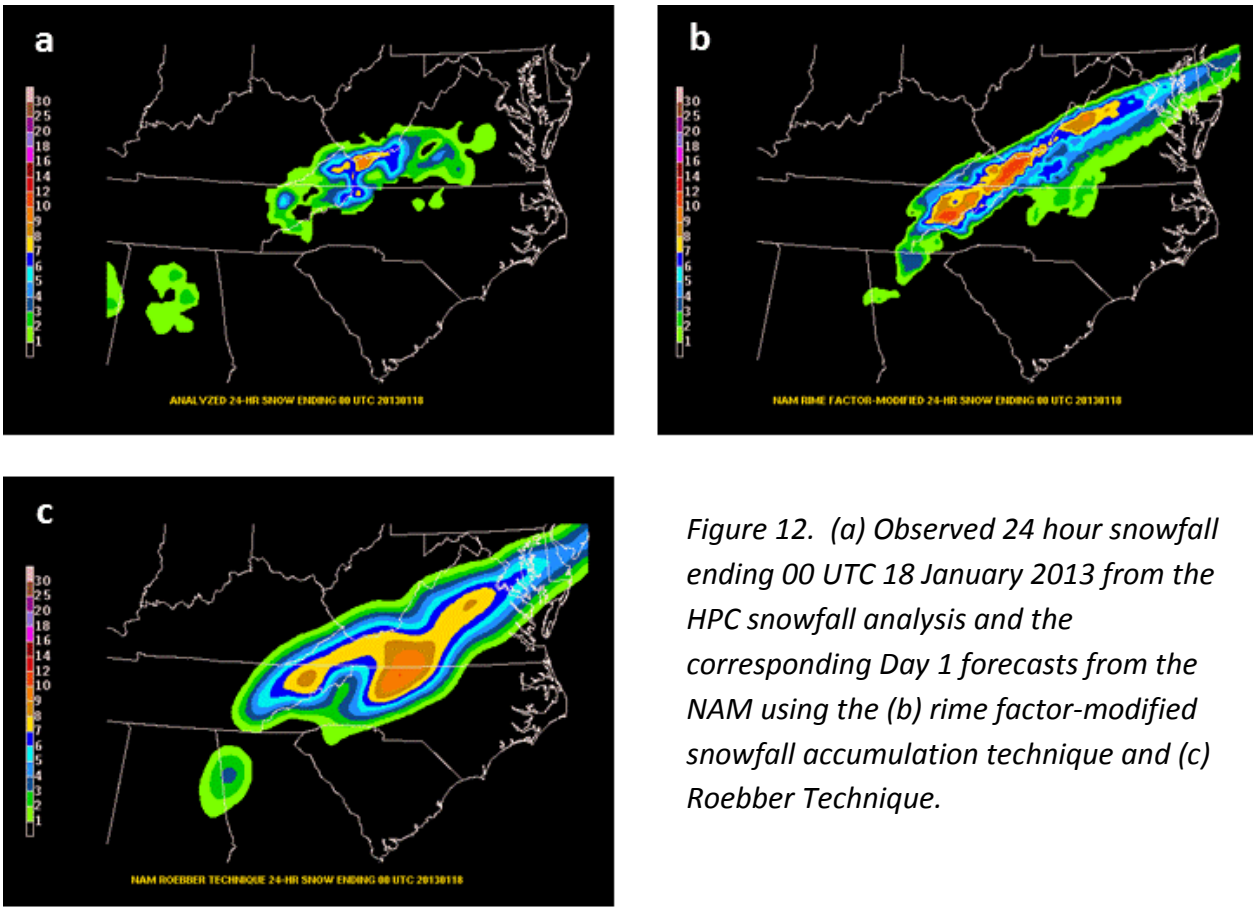

Figure 12. (a) Observed 24 hour snowfall ending 00 UTC 18 January 2013 from the HPC snowfall analysis and the corresponding Day 1 forecasts from the NAM using the (b) rime factor-modified snowfall accumulation technique and (c) Roebber Technique.

Despite the mixed performance of the snowfall accumulations, participants were encouraged by the potential of this new approach. In addition to the snowfall accumulations, participants also had access to the percent of frozen precipitation and rime factor fields that were used in the calculation of the modified snowfall accumulations (Appendix B). Many participants found that these fields provided useful forecast information, particularly in marginal events and events with the potential for mixed precipitation. Participants also noted that they would like to see this technique expanded to other models.

5. EXPERIMENTAL FORECAST TOOLS

In addition to the experimental guidance, two forecast tools were highlighted during the preparation of the Day 4-5 winter weather outlook forecasts. After using these tools throughout the week, participants were asked to provide feedback about the potential utility of the tools in an operational forecasting environment. The evaluation of both tools was made more difficult by the lack of well-defined low pressure systems throughout much of the experiment.

Ensemble Sensitivity Analysis

While participants generally found the concept of the ensemble sensitivity analysis tool intriguing, they raised a number of questions about how to effectively use this information in the forecast process. For example, while the goal of the tool is to help forecasters identify changes to the initial conditions that may ultimately impact the forecast outcome several days in the future, similar information is already available through careful model interrogation. In their limited exposure to the tool, it was unclear to participants whether the ensemble sensitivity analysis could provide them with additional information or help them identify alternative solutions more easily than the methods already in place operationally. In addition, although the tool can distinguish between sensitivities associated with the track and intensity of the cyclone, it was not clear how to best use this information. In one case, participants attempted to use the tool to identify the downstream impact of model differences 12 hours into the forecast period, but the differences were too small to make any definitive conclusions.

While it was encouraging that different ensembles generally identified the same sensitive areas, participants often found the domain-based approach to the analysis limiting. There were several instances in which a system was too far off the northeast coast to be captured by one of the tool's fixed domains, yet still had the potential to impact the coast with any westward shift in its track. While this issue could have been somewhat mitigated by taking better advantage of the available floating domain, a better long term solution may be to increase the number of domains to better capture events across the entire country.

Another hurdle to widespread operational use is that despite its relatively straightforward premise, the concept of ensemble sensitivity analysis has proven to be difficult to explain. Some of this challenge may be mitigated by focusing training on the advantages and disadvantages of the tool and identifying specific ways to use this information in the forecast process.

ECMWF Extreme Forecast Index

Unfortunately, the lack of extreme events during the experiment period severely limited the use of the ECMWF EFI, making it difficult to evaluate its potential utility. With few opportunities for the EFI to provide value to the forecast process, participants rarely modified forecasts based on information provided by the EFI. In addition to the EFI, participants also had access to point-based cumulative distribution functions (CDFs) that provided additional context about the rarity of the event and trends over the last several runs of the model. Like the EFI,

while participants found this display interesting, it had limited utility during the experiment because of the lack of extreme winter weather events.

While the lack of extremes limited the utility of the EFI this winter, the concept appears to be sound and should be explored further. One limitation of the ECMWF EFI is that access to the data is extremely limited. Expanding this concept to other ensembles would allow for wider use of the data and create opportunities for further development. With this in mind, HMT-HPC has worked with the Earth Systems Research Laboratory (ESRL) to develop a prototype EFI product based on the second generation GEFS reforecast dataset (Hamill et al. 2013). Real-time experimental products are available at: http://www.esrl.noaa.gov/psd/forecasts/reforecast2/analogs/index.html.

6. FORECAST CONFIDENCE

As part of the experimental forecast process, participants were asked to rate their overall confidence in their forecasts as above average, average, or below average. Figure 13 summarizes the team's forecast confidence. Forecast confidence tended to be average on Day 1, with results almost evenly split between average and below average on Day 2. Of the 23 forecasts issued during the experiment, participants only indicated above average confidence in one forecast. The overall level of forecast confidence is lower than last year's experiment, and is likely related to the abundance of fast-moving clipper systems during this year's experiment period.

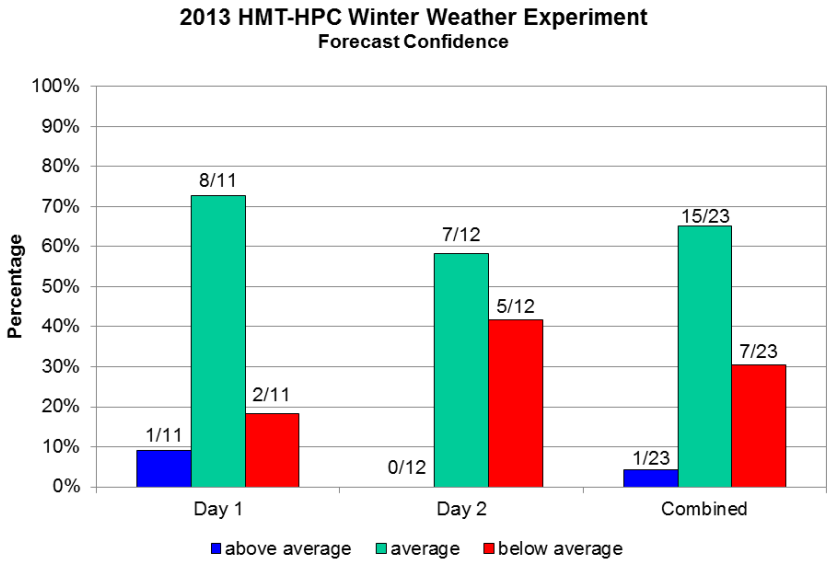

Figure 13. Overall forecast confidence as rated by participants in the 2013 HMT-HPC Winter Weather Experiment.

7. DECISION SUPPORT COMPONENT

Preparing a public forecast graphic and providing a corresponding mock decision support briefing exposed participants to the challenge of communicating complex meteorological information to a decision maker with limited meteorological knowledge. Participants found the orientation provided by the WxEM team to be extremely valuable. In particular, the orientation included several tips for conducting effective decision support briefings including the importance of:

- Understanding the needs of the audience
- Communicating forecast information clearly and concisely. This includes information about what specifically the hazard is as well as details about the expected timing, location, and duration of the event as well as any information about the history of the event and any additional information about forecaster confidence.
- Organizing the briefing so that the "bottom line" appears first.
- Expressing forecast confidence.

Based on the briefing advice in the orientation, participants generally provided mock briefings that avoided complex meteorological terminology, and the addition of the forecast graphic aided in the ability to clearly communicate the most important forecast information (the "bottom line").

While the addition of a graphic helped structure the mock briefings, participants often used either their experimental forecasts or a previous public forecast graphic as a starting point. Although this worked well in some cases, their approach generally didn't change in response to the current weather situation, which may have resulted in graphics that were less than ideal. Many participants appreciated the structure provided by the event-type classification scheme, but found that it could prove difficult to use when the classifications didn't adequately describe the anticipated threat. Ultimately, most groups combined the approaches into a more relaxed classification scheme based on the hazards posed by the current event.

Compared to last year, the WxEM team found that the mock briefings were improved, with a greater emphasis on the bottom line and reduced use of complex meteorological terminology. In general, they recommended that more information be included on the graphics, particularly about snow and/or freezing rain amounts and event timing. While these details were generally covered as part of the verbal briefing, the WxEM team stressed the importance of including this information on the graphic as well since graphics are often viewed without the corresponding briefing.

It is important to note that during the experiment participants were asked to produce the public forecast graphic using NMAP and were limited to a single graphical image. This presented a number of limitations, as the graphics could quickly become difficult to read in complex situations and the black NMAP background limits the use of many colors. Future experiments may need to explore the possibility of preparing a complete briefing package in order to address some of the WxEM team's recommendations.

While the joint decision support briefings conducted in collaboration with AWT were a good opportunity to explore the potential for future cross-testbed collaborations, the divergent goals of the two experiments made true collaboration a challenge. For example, while the HMT-HPC experiment was focused on quantifying and communicating uncertainty in winter weather forecasts over the next several days, the AWT experiment was primarily focused on winter weather hazards within the next 24 hours. In addition, while HMT-HPC was focused on improving communication of this information to users with limited meteorological knowledge, aviation customers tend to be more weather-savvy, reducing the communication challenges. The difference in both the forecast period of focus and the target audience made it difficult to develop a joint activity that added value to both experiments. In the future, cross-testbed collaboration can be improved by focusing on specific areas where experiment goals and forecast time frames overlap.

8. DAY 4-5 OUTLOOK FORECASTS

The experimental Day 4-5 winter weather outlook forecasts issued during the experiment revealed that forecasts at these time ranges are a realistic operational goal. Winter weather forecasts are often dependent on small details in both the temperature and moisture fields that can be difficult for models and ensembles to resolve at longer forecast lead times. Despite these challenges, participants were generally able to issue forecasts that correctly highlighted areas expected to receive winter weather, and these forecasts generally improved from Day 5 to Day 4 (Fig. 14).

During the experiment, the winter weather outlook forecasts were broadly defined as indicating a "threat for winter weather". This allowed both for an exploration of the model guidance during these time periods and also generated discussion about how these forecasts should be defined operationally. One of the challenges is determining criteria that apply nationally. For example, while 1 inch of snow in the southern United States is considered a major event, the same is not true farther north. Another challenge is determining what the product should look like. While the operational Day 1-3 forecasts distinguish between snowfall and freezing rain, this distinction may not be appropriate at longer forecast lead times. Both

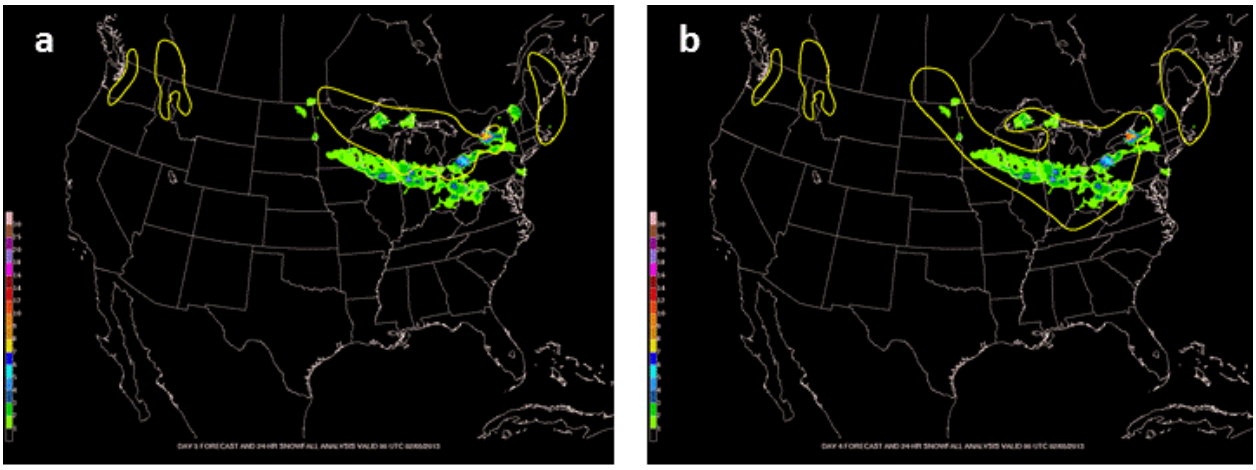

Figure 14. (a) Day 5 and (b) Day 4 experimental winter weather outlook forecasts valid 00Z 5 February 2013. The observed snowfall is shaded.

issues will require further exploration as HPC continues to explore expanding its operational winter weather product suite.

9. SUMMARY AND OPERATIONAL IMPACTS

The third annual HMT-HPC Winter Weather Experiment was conducted January 15 – February 15, 2013. In addition to exploring the use of ensemble systems to better quantify and communicate uncertainty in winter weather forecasts, this year's experiment also focused on using information provided by model microphysics schemes to improve snowfall forecasts. Over the course of the five week experiment, 23 participants issued experimental probabilistic forecasts of exceeding 2 in, 4 in, and 8 in of snow over a 24 hour period. In addition to the experimental forecasts, participants evaluated the available experimental guidance, produced a public forecast graphic, participated in a mock decision support briefing, and issued experimental Day 4-5 winter weather outlook forecasts.

The experiment highlighted the importance of ensemble guidance for winter weather forecasts and revealed the potential benefits of using microphysics information to derive snowfall amounts rather than relying solely on snow-to-liquid ratio algorithms. A number of the experiment findings are directly relevant to operational winter weather forecasters:

- Of the ensembles tested during the experiment, **the operational SREF and the HPC Autoensemble provided the best forecast guidance.** While the AFWA ensembles provided some useful forecast details, the high bias in their snowfall amounts requires them to be used with caution.
- Although improvements in the NAM snowfall accumulations were limited, **the rime factor-modified snowfall accumulation technique and the underlying percentage of**

frozen precipitation and rime factor parameters appear promising. HPC is working with EMC to expand this technique to other models.

- **Day 4-5 winter weather outlook forecasts are a realistic operational goal.** Participants routinely provided valuable forecast guidance at these time ranges using current operational model guidance.
- While creating timely gridded snowfall analyses is challenging, **the HPC snowfall analysis needs improvement.** There are a number of potentially useful datasets available that are not currently included in the analysis.
- Ensemble sensitivity analysis raises a number of intriguing possibilities, **but it is still unclear how to best use this information in an operational forecasting environment.**
- **Communicating complex forecast information to a decision maker with limited meteorological knowledge can be improved with practice.** Clear briefings and graphics should include information about what the weather hazard is and details about the timing, duration, and location of the expected event.

The HMT-HPC Winter Weather Experiment provided a unique opportunity to bring the forecasting, research, and model development communities together to explore the challenges associated with winter weather forecasting. The experiment identified several potential ways to improve and expand current winter weather forecasts and snowfall verification which will continue to be explored.

ACKNOWLEDGEMENTS

The HMT-HPC Winter Weather Experiment would not be possible without the dedication of a host of individuals including Faye Barthold, Mike Bodner, Tom Workoff, Wallace Hogsett, Dave Novak, Dan Petersen, Rich Bann, Mike Musher, Rich Otto, and Brian Hurley. Keith Brill (HPC) provided the objective verification results. Becky Cosgrove (NCEP Central Operations; NCO), Justin Cooke (NCO), and Scott Rentschler (AFWA) were instrumental in providing the AFWA data. Brad Ferrier (EMC) and Eric Aligo (EMC) developed the rime factor-modified snowfall accumulation technique, while Eric Rogers (EMC) provided access to the necessary data. The WxEM team of Jessica Losego (University of North Carolina), Burrell Montz (East Carolina University), and Ken Galluppi (Arizona State University) provided valuable feedback throughout the decision support component.

REFERENCES

Baxter, M. A., C. E. Graves, and J. T. Moore, 2005: A climatology of snow-to-liquid ratio for the contiguous United States. *Wea. Forecasting*, **20**, 729-744.

Chang, E. K. M., M. Zheng, and K. Raeder, 2013: Medium range ensemble sensitivity analysis of two extreme Pacific extratropical cyclones. *Mon. Wea. Rev.*, **141**, 211-231.

Colle, B. A., and E. K. M. Chang, 2011: Predictability of high-impact weather during the cool season over the eastern U.S: From model assessment to the role of the forecaster. *CSTAR Semi-Annual Report*, April 2011.

Du, J. and B. Zhou, 2011: A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. *Mon. Wea. Rev.*, **139**, 3284-3303.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Jr., Y. Zhu, and W. Lapenta: NOAA's second generation global medium range ensemble reforecast data set. In press *Bul. Amer. Meteor. Soc.*

Hou, D., M. Charles, Y. Lou, Z. Toth, Y. Zhu, R. Krzysztofowicz, Y. Lin, P. Xie, D. -J. Seo, M. Pena, and B. Cui, 2013: Climatology-Calibrated Precipitation Analysis at fine scales: Statistical adjustment of Stage IV towards CPC gauge based analysis. In press *J. Hydrometeor*.

Lalaurette, F., 2003: Early detection of abnormal weather patterns using a probabilistic extreme

forecast index. *Quart. J. Roy. Meteor. Soc.*, **129**, 3037-3057.

Manikin, G. S., 2005: An overview of precipitation type forecasting using NAM and SREF data. *21st Conference on Weather Analysis and Forecasting/17th Conference on Numerical Weather Prediction*, Washington, D.C., 8A.6.

Roebber, P. J., S. L. Bruening, D. M. Schultz, and J. V. Cortinas, 2003: Improving snowfall forecasts by diagnosing snow density. *Wea. Forecasting*., **18**, 264-287.

Rooney, J. F., Jr., 1967: The urban snow hazard in the United States: An appraisal of disruption. *Geogr. Rev.*, **57**, 538-559.

Zsoter, E., 2006: Recent developments in extreme weather forecasting. *ECMWF Newsletter*, No. 107, ECMWF, Reading, United Kingdom, 8-17.

APPENDIX A

Participants

Week	HPC Forecaster	NCEP/WFO	Research/Academia/ Private Sector	EMC
Jan 15 – 18	Dan Petersen	Douglas Schneider (MRX) Paul Vukits (OPC)		Jun Du
Jan 22 – 25	Rich Bann	Seth Binau (ILN) Steve Lack (AWC)	Brian Colle (SBU)	Eric Aligo
Jan 28 – Feb 1	Mike Musher	Mary Wister (PDT)	Brian Kolts (First Energy)	Brad Ferrier
Feb 4 – 8	Rich Otto	Paul Frisbie (GJT)	Gary Lackmann (NCSU) Paul Stokols (NWSHQ)	Matt Pyle
Feb 11 – 15	Brian Hurley	Mike Eckert (AWC)	Stefan Cecelski (UMD) Jason Levit (NWSHQ)	Geoff Manikin

APPENDIX B

Experimental Model Guidance and Forecast Tools

Rime Factor-Modified Snowfall Accumulations

The rime factor-modified snowfall accumulation technique uses information from the NAM's microphysics scheme to modify the initial snow-to-liquid ratio provided by the Roebber Technique. The modification is applied hourly through the first 36 hours, then every three hours from 36-84 hours. Since the Roebber Technique only provides a snow-to-liquid ratio value every 6 hours, the snow-to-liquid ratio is linearly extrapolated in order to obtain approximate hourly/three hourly values. The extrapolated snow-to-liquid ratio values are then modified by the rime factor (RF), which indicates the amount of riming on an individual ice particle due to a combination of riming and liquid water accretion as follows:

Rime Factor	Modified SLR
1 < RF < 2 (fluffy snow)	$SLR_{RF} = SLR_{Roebber}$
2 < RF < 5 (rimed snow)	$SLR_{RF} = \frac{SLR_{Roebber}}{2}$
5 < RF < 20 (graupel)	$SLR_{RF} = \frac{SLR_{Roebber}}{4}$
RF > 20 (sleet)	$SLR_{RF} = \frac{SLR_{Roebber}}{6}$

The rime factor-modified snowfall accumulation is then calculated as:

$$Snowfall = (QPF) \times (POFP) \times (SLR_{RF})$$

where *POFP* indicates the percentage of precipitation that is frozen when it reaches the ground.

SREF Weighted Mean

The SREF Weighted Mean is calculated by taking the difference between the individual ensemble member solutions and the ensemble mean solution for 16 different variables including sea level pressure, 500 hPa, 700 hPa, and 850 hPa geopotential height, temperature, relative humidity, and the *u* and *v* components of the wind speed. Based on the overall difference, the ensemble members are then ranked from smallest difference (best member) to largest difference (worst member), and members with smaller differences are assigned more weight in the calculation of the ensemble mean. The weight assigned to the best member is

typically around 9%, while the weight assigned to the worst member is fixed at 0.1%. The differences are recalculated at each forecast hour (Du and Zhou 2011).

Ensemble Sensitivity

Ensemble sensitivity analysis was available for the GEFS, ECENS, Canadian ensemble (CMCE), North American Ensemble Forecast System (NAEFS), and SREF, This tool uses empirical orthogonal functions (EOFs) to identify the dominant patterns (system strength, location, etc.) associated with variations in initial conditions (Colle and Chang, 2011; Chang et al. 2013). Real time access to the tool is available at: http://dendrite.somas.stonybrook.edu/CSTAR/Ensemble_Sensitivity/EnSense_Main.html.

ECMWF Extreme Forecast Index

The Extreme Forecast Index (EFI) compares the current forecast probabilities from the ECMWF ensemble to probabilities derived from the model climatology to determine how climatologically rare an event is in the model history. The climatological probabilities used for this comparison are derived by rerunning a 5-member ensemble once a week for 20 years for the 30 day period (±15 days) surrounding the forecast valid date. The EFI is a measure of the difference between the forecast cumulative distribution and the climatological distribution for an event. Values range between -1 and +1, with values between 0.5 and 0.8 (-0.5 and -0.8) indicating an "unusual" event relative to model climatology and values greater than 0.8 (less than -0.8) indicating an "extreme" event. EFI is calculated for total precipitation, maximum 10 m wind gust, and 2 m temperature (Zsoter 2006; Lalaurette 2003).

APPENDIX C

Daily Schedule

A brief orientation session will be held at 8:00am on the first day of each week to explain the motivation and organization of the experiment as well as the data being evaluated. The WxEM team will provide a separate orientation for the decision support component at 12:30pm.

8:30am – 11:00am Determine forecast area and time period (Day 1 or Day 2)

Draw contours for probability (slight – 10%, moderate – 40%, high – 70%) of exceeding 2 in, 4 in, and 8 in of snow during the 24 hour period (00 – 00 UTC), substituting a 12 in snowfall threshold or 0.01 in, 0.10 in, or 0.25 in freezing rain thresholds as appropriate. Forecasts are based on 00 UTC guidance.

Write forecast confidence discussion

11:00am – 11:30am HPC-CPC map discussion

11:30am – 12:30pm Lunch

12:30pm – 1:30pm Prepare public forecast graphic and provide mock decision support briefing

1:30pm – 2:30pm Day 4-5 winter weather outlook forecast

2:30pm – 4:00pm Subjective verification using the HPC snowfall analysis to evaluate the performance of the experimental guidance as it relates to accumulations and indication of forecast uncertainty

4:00pm – 4:30pm Group discussion