The 2018-19 WPC-HMT Winter Weather Experiment

13 November 2018 – 15 March 2019
Weather Prediction Center
College Park, MD

Findings and Results

Michael Bodner *- NOAA/NWS/WPC, College Park, MD*
Sara Sienkiewicz *- I.M. Systems Group, NOAA/NWS/WPC, College Park, MD*
Benjamin Albright *- Systems Research Group, NOAA/NWS/WPC, College Park, MD*
Bill Lamberson *- I.M. Systems Group, NOAA/NWS/WPC, College Park, MD*

Updated: 15 May 2019

Table of Contents

Abstract

The Hydrometeorology Testbed at the Weather Prediction Center (WPC-HMT) conducted the 9th Annual Winter Weather Experiment (WWE) from 13 November 2018 through 15 March 2019. The experiment brought together members of the operational forecasting, research, and academic communities to address winter weather forecast challenges. The 2018-19 WWE focused on the following science goals: (1) the evaluation of precipitation-type algorithms applied to both the FV3-GFS (hereafter FV3) and 12-km NAM (hereafter NAM12) models during diverse weather events exhibiting precipitation-type transition zones, (2) the assessment of predictability of synoptic systems using ensemble cluster analysis, and (3) the evaluation of explicit microphysical data within convective allowing models (CAMs) for Great Lakes lake-effect snowfall and other mesoscale events.

During forecast sessions, experiment participants collaborated to create a 6-hour blended snowfall forecast using the real-time experimental snowfall guidance for Day 1, Day 2, or Day 3 utilizing several different precipitation type algorithms and ensemble cluster analysis for specific areas within the Continental United States. A total of 19 6-hour snowfall forecasts were created throughout the entire season. These blended snowfall forecasts as well as CAMs data were then subjectively evaluated through discussions and scoring during verification sessions by experiment participants. Snowfall forecasts were objectively evaluated for the entire winter season (November - March) in addition to the 19 experiment cases. Results informing recommendations are discussed for the application of precipitation type methodologies and the development of and training for ensemble clustering, forecast blending, and CAMs snowfall products.

Introduction and Background

In an effort to support improvements in both WPC and WFO winter weather forecasts, the Hydrometeorology Testbed at WPC (HMT-WPC) conducted the 9th Annual Winter Weather Experiment (WWE) from 13 November 2018 through 15 March 2019. The experiment brought together members of the operational forecasting, research, and academic communities to address winter weather forecast challenges.

The 2018-19 WWE provided an opportunity for participants to evaluate a suite of precipitation type algorithms in a real time setting to assign the best precipitation type for the thermal environment applied to forecast snowfall amounts. The 2018-19 WWE was conducted both remotely and on location in College Park, MD for the second year in a row utilizing the web-based distance communication software, Mikogo, paired with a teleconference to encourage and promote interactivity and engagement with the participants. More on the remote aspect of the experiment can be found in the Experiment Logistics and Participation section below.

Science and Operations Objectives

The main objectives of the 2018-19 Winter Weather Experiment were to:

- Explore the testing of multiple precipitation type methodologies to determine which methods best enhance the forecast process for accumulated snowfall.
- Explore the applications of ensemble clustering to the prediction of winter weather.
- Enhance collaboration among NCEP centers, WFOs, and NOAA research labs on winter weather forecast challenges including lake effect snow and mesoscale precipitation type transition zones.

Fast findings for science and operations objectives are summarized in Table 1. The concise reasoning behind each recommendation is found below the table with supporting evidence elaborated on in both the Results and Summary and Recommendations sections.

Table 1. *Transition metrics for select experimental guidance and techniques.*

Major Tests Conducted	Transitioned to Operations	Recommended Transition to Operations	Recommended for Further Development and Testing	Rejected for Further Testing	Funding Source
Predictability					
Fuzzy clustering		X	X		WPC
Precipitation Type Methods					
WPC Decision Tree (TREE)	X				WPC
NCEP-Dominant (NDOM)	X				WPC
Change in Snow Depth (SNDP)		X			WPC
Percent of Frozen Precipitation (POFP)			X		WPC
Bourgouin			X		WPC
Forecast Blending					
Manual blending of snowfall versus static blending (i.e., NBMv3.1)			X		WPC/ MDL
CAMs Snowfall Guidance					
1-km HRRRX			X		ESRL/ GSD/ GLERL
HRRRE			X		ESRL/ GSD
HREFv2.1			X		EMC
HREFv2.1 EAS Probabilities			X		EMC
Totals	2	2	8	0	

The following two products are transitioned to WPC operations: (1) WPC Decision Tree (TREE) and (2) NCEP-Dominant (NDOM) precipitation type methods. The following two products that are recommended for transition to operations are as follows: (1) fuzzy clustering, (2) change in snow depth (SNDP) precipitation type method. This could be done by providing web-based graphics or ingestion into AWIPS2 for use in GFE. The fuzzy clustering will still require assessment on a case-by-case basis so is also recommended for further development and testing.

The remaining products tested within the 2018-19 WWE are recommended for additional development and testing. This includes the two remaining precipitation type algorithms, (1) percent of frozen precipitation (POFP) and (2) Bourgouin (BOUR). For example, improvements to the intra-algorithm thresholds for available microphysical parameters would be a focus of development for the POFP precipitation type method and tuning the percentage of model QPF assigned to the snow precipitation type for certain environments would be a focus of development for the BOUR method. Further development and testing is recommended for the following five methods and products: (1) manual blending, (2) 1-km experimental HRRR (HRRRX), (3) HRRRE mean snowfall, (4) HREFv2.1 mean snowfall, (5) HREFv2.1 Ensemble Agreement Scale (EAS) snowfall probabilities. Manual blending would benefit from the creation of more forecast inputs from additional modeling systems (e.g., CAMs for Day 1, global Canadian deterministic model, etc.). The 1-km HRRRX had limited data availability for the 2018-19 season, so the observation of a low forecast bias compared to the 3-km HRRRX needs to be further assessed. The HREFv2.1 mean and probabilistic snowfall could be improved by not calculating snowfall from snow water equivalent, which aggregates all frozen species and doesn't isolate snow; this resulted in an over-forecast for many events analyzed in the WWE. For more detailed recommendations please see the Summary and Recommendations Section of this report.

Experiment Logistics and Participation

The experiment was conducted weekly over the full winter season beginning Tuesday, November 13, 2018 and ending Friday, March 15th, 2019. This year, the WWE was executed remotely from the WPC-OPC Collaboration Room at the NOAA Center for Weather and Climate Prediction (NCWCP) in College Park, MD. Experiment Participants and Field Representatives joined remotely on “Forecast Tuesdays” (10:30 am-12:00 pm EST) and “Verification Wednesdays” (10:30 am-12:00 pm EST) throughout the season. Experimental data and tools were explored and used to create experimental snowfall forecasts on Tuesdays. Forecast and data set verification were presented and discussed and subjective feedback was collected on Wednesdays. In addition to the remote experiment, there was one week of residence experiment at

NCWCP held from 11-15 March 2019. A total of 14 forecast and 15 verification sessions were conducted and participation for all 19 sessions is shown below.

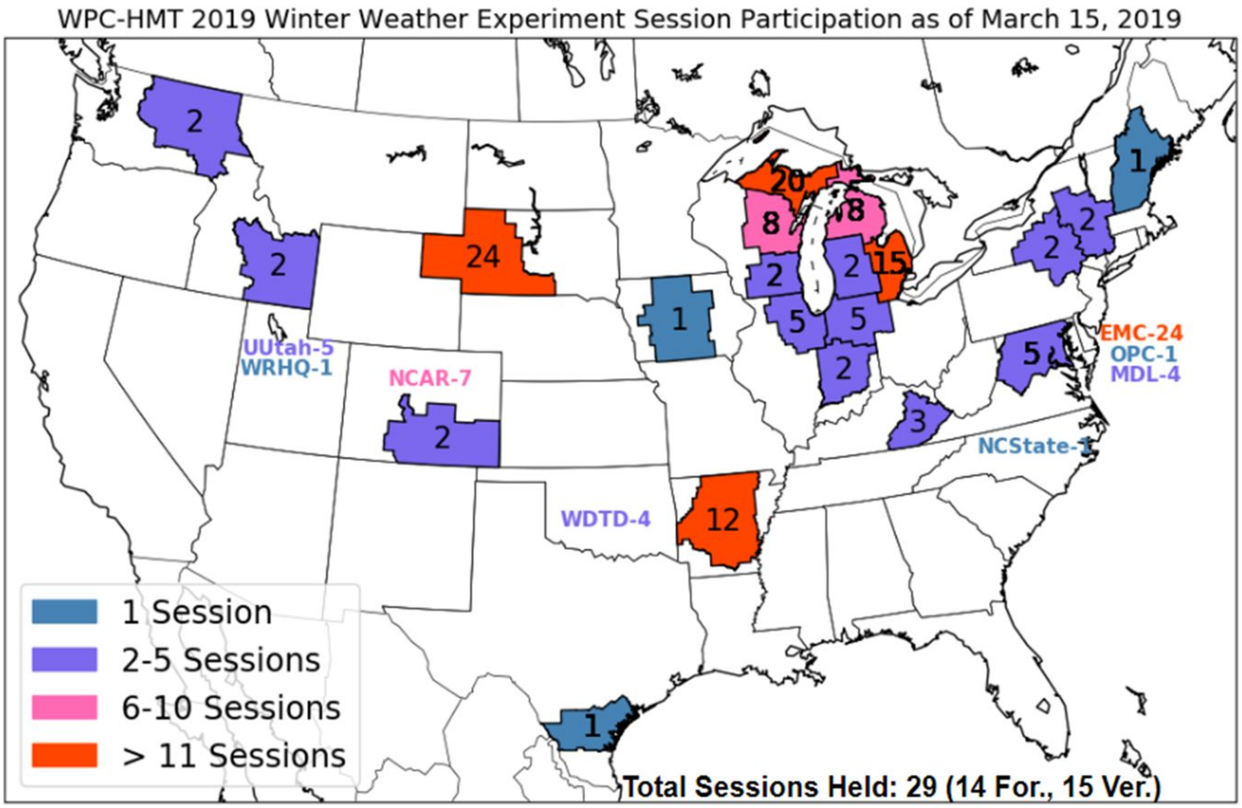
WPC-HMT 2019 Winter Weather Experiment Session Participation as of March 15, 2019
2
2
24
20
8
8
2
2
15
1
5
5
2
1
2
2
5
EMC-24
OPC-1
MDL-4
UUtah-5
WRHQ-1
NCAR-7
2
3
NCState-1
WDTD-4
12
1
1 Session
2-5 Sessions
6-10 Sessions
> 11 Sessions
Total Sessions Held: 29 (14 For., 15 Ver.)

Figure 1. *WWE total participation by WFO (shaded) or other National Center, University Partner, or location (text) colored according to participation.*

Each Tuesday, the experiment participants utilized a combination of operational and experimental precipitation type model guidance to create experimental 6-hour snowfall forecasts for selected regionalized domain(s), usually where precipitation type transition zones occurred. Forecast periods were also determined based on the potential for prediction challenges of heavy snowfall placement in which ensemble clustering guidance could also be incorporated. The regional forecasts covered either Day 1 (18 UTC - 12 UTC), Day 2 (12 UTC - 12 UTC), or Day 3 periods (12 UTC - 12 UTC).

Each Wednesday, participants subjectively evaluated the performance of both the experimental forecasts and the experimental model guidance over the previous week. Day 1 shorter-range guidance from the CAMS were evaluated for cases of mesoscale precipitation type transition zones and lake effect snowfall.

Data and Methods

Precipitation Type Datasets

Several algorithmic methodologies were explored in the 2018-19 WWE to assess hourly, instantaneous estimates of precipitation type. The methods were applied to both the 12-km NAM (NAM12) and FV3-GFS (FV3) deterministic model output to derive the various species of precipitation type. The precipitation type forecasts were then evaluated by verifying the performance of snowfall forecasts where each precipitation type method was used in combination with the respective model QPF and a standard snow to liquid ratio (SLR; 10:1 for snow, 2:1 for sleet) to produce the snowfall forecasts.

The first and most simple precipitation type method evaluated in the experiment was the **WPC decision tree algorithm** (“TREE”). This algorithm features a straightforward logic check of critical temperatures at 700 hPa, 850 hPa, and 925 hPa, and 2-meter temperatures for each grid point to ascertain the depth of warm and cold layers in the sounding. The decision tree algorithm generates a deterministic and instantaneous estimation of precipitation type for each model time stamp.

The second method tested and the most frequently used in NCEP model post-processing is an ensemble average of the Ramer, Bourgouin, and Baldwin precipitation type algorithms, better known as the **NCEP Dominant method** (“NDOM”). Each of the methodologies that make up the NCEP dominant scheme evaluate the freezing level of the wet bulb temperature to assess hydrometeor phase. This multi-algorithmic method generates multiple precipitation types at each grid point and at each forecast time stamp, and assigns the highest weighted type as a forecast for the grid point. In the event of a “tie” weighting, the assignment of freezing rain takes precedence over other species due to its high-impact.

The third method tested was the Percent of Frozen Precipitation method (“POFP”) which assigns a precipitation type for each grid point in the domain using the **fraction of frozen precipitation** and **rime factor parameterization** from the model microphysics. The fraction of frozen precipitation parameter in both the NAM12 and FV3 estimates the percentage of hydrometeors to be frozen in the lowest model level at a particular forecast time. Thus the fraction of frozen precipitation field in this lowest model level is assumed to have accounted for snowflakes that fall to the ground and those that melt then refreeze prior to reaching the surface. The fraction of frozen precipitation parameter is correlated directly with the QPF and the vertical thermal profile, thereby alleviating the need to empirically estimate warm layers in the forecast model soundings.

The POFP method is derived slightly differently for both the NAM12 and FV3 in this experiment. The Ferrier-Aligo microphysics scheme in the NAM12 features a variable density graupel parameter also known as the rime factor (“RF”). The rime factor is a ratio of the growth of snow by liquid accretion plus vapor deposition divided by vapor deposition. Thus the higher the liquid accretion, the higher the resultant rime factor. A rime factor for the lowest model level of the atmosphere is provided in the NAM12. Past evaluation of the rime factor parameter in precipitation type transition zones has demonstrated that rime factors over 10 typically align with increased graupel at the lowest model level (HMT-WPC 2015) and therefore present a higher likelihood of sleet being the instantaneous precipitation type at the surface.

A rime factor parameter is not available through the GFDL microphysics scheme in the FV3. Therefore a “rime factor proxy” (“RFP”) was derived using the snow and graupel mixing ratios in the lowest level of the model. Investigation of the rime factor proxy suggests that values greater than 0.7 align with increased presence of graupel and a higher likelihood of sleet as the instantaneous precipitation type at the surface.

RFP = Graupel Mixing Ratio / 1 + Graupel Mixing Ratio + Snow Mixing Ratio

Figure 2. *The equation for the rime factor proxy (“RFP”) used for the FV3 POFP precipitation type algorithm.*

POFP > 90% = Snow
POFP > 70% and < 90% = Sleet
POFP > 5% and < 70% = Rain/Snow Mix
POFP < 5% = Freezing Rain when 2-meter temperature ≤ 32F
POFP < 5% = Rain when 2-meter temperature > 32F
NAM12: POFP > 5% and < 90% with RF > 10 = Sleet
FV3: POFP > 5% and < 90% with RFP > 0.7 = Sleet

Figure 3. *Determination of precipitation type using microphysics parameters for the POFP precipitation type algorithm.*

A fourth precipitation type methodology, the **Bourgouin method** (“BOUR”), was tested using the FV3 only. The Bourgouin method estimates precipitation type by calculating the total energy available within a layer, then determines whether enough positive energy is available to initiate melting of the hydrometer or if enough negative energy is available in the environment below to allow for refreezing of a liquid hydrometeor (Bourgouin 2000). For each hour of FV3 output, the percentage of each of the following precipitation types was calculated: rain, snow, sleet, and freezing rain. To calculate the snowfall for an hour, the percentage of the snow precipitation type was multiplied by that hour’s QPF and a 10:1 SLR. The hourly data were then summed into 6-hourly snowfall amounts. Since sleet is it’s own precipitation type (with a 2:1 SLR applied to get accumulation), it is not included in the snowfall amounts.

The fifth methodology tested was snowfall forecasts evaluated from the **change in snow depth field** ("SNDP") in the NAM12 and FV3. Both models are coupled with the NOAH Land Surface Model (LSM) therefore no evaluation of precipitation type was needed with these snowfall forecasts and no SLR needed to be applied. Variations in snowfall accumulation and melting are determined by surface thermal fluxes and expansion of the range of potential snow/ice densities from the model microphysics. Increased snow densities result in decreased snowfall accumulations and vice versa.

In addition to the five precipitation type methods described above, an additional snowfall methodology called the Filter method was computed for both the FV3 and the NAM12 and assessed for the entire winter season (November - March). The **FV3 Filter** snowfall amounts were calculated with the FV3 QPF, fraction of frozen precipitation, and the Baxter climatological SLR (Baxter et al. 2005). The Baxter SLR was adjusted within environments likely exhibiting a precipitation type transition zone by applying a rime filter to reduce the SLRs where the FV3 microphysics indicated less than optimal snowflake aggregation. WPC has been applying this method in the NCEP NAM model for several years (i.e., **NAM Filter**), where a rime factor parameter featuring the ratio of snow growth by liquid accretion plus vapor deposition divided by the vapor deposition is used to modify a SLR computed by the Roebber neural network (Roebber et al. 2007). The rime factor is generated from the Ferrier-Aligo microphysics scheme, which is only available in the NAM. For the FV3, a rime factor proxy must be computed using the snow and graupel mixing ratios available in the GFDL microphysics. This rime filter is an instantaneous field calculated by dividing the graupel mixing ratio by the snow mixing ratio.

Ensemble Clustering

The 2018-19 WWE employed ensemble clustering as a way to assess the predictability of and understand the possible forecast scenarios for winter weather events. The clustering used in the 2018-19 WWE was a version of fuzzy clustering and was developed based on the use and feedback of the clusters used during the 2017-18 WWE which were provided by Stony Brook University (SBU). The clusters were generated in the following 3 step process:

1. The first two Empirical Orthogonal Functions (EOFs) of the 500-hPa forecast from a 90-member superensemble combination of the 20 Canadian (CMCE), 50 ECMWF (ECENS), and 20 Global Ensemble Forecast System (GEFS) on the day of interest were calculated and interpreted. This differs from the SBU clusters which calculated the EOFs of the mean sea level pressure field. Calculating the EOFs of the 500-hPa geopotential height field is capable of capturing weaker synoptic systems. Figure 4 shows the EOFs for the day 3 forecast initialized at 0000 UTC on 30 October 2018.

This forecast period featured a strong trough over the Mississippi Valley with an attendant surface cyclone and heavy precipitation just downstream of the trough. The first EOF highlights that there are considerable differences between the ensemble members with respect to the east-west location of the trough. The second EOF highlights that there are also differences in the amplitude of the trough between the 90 ensemble members.

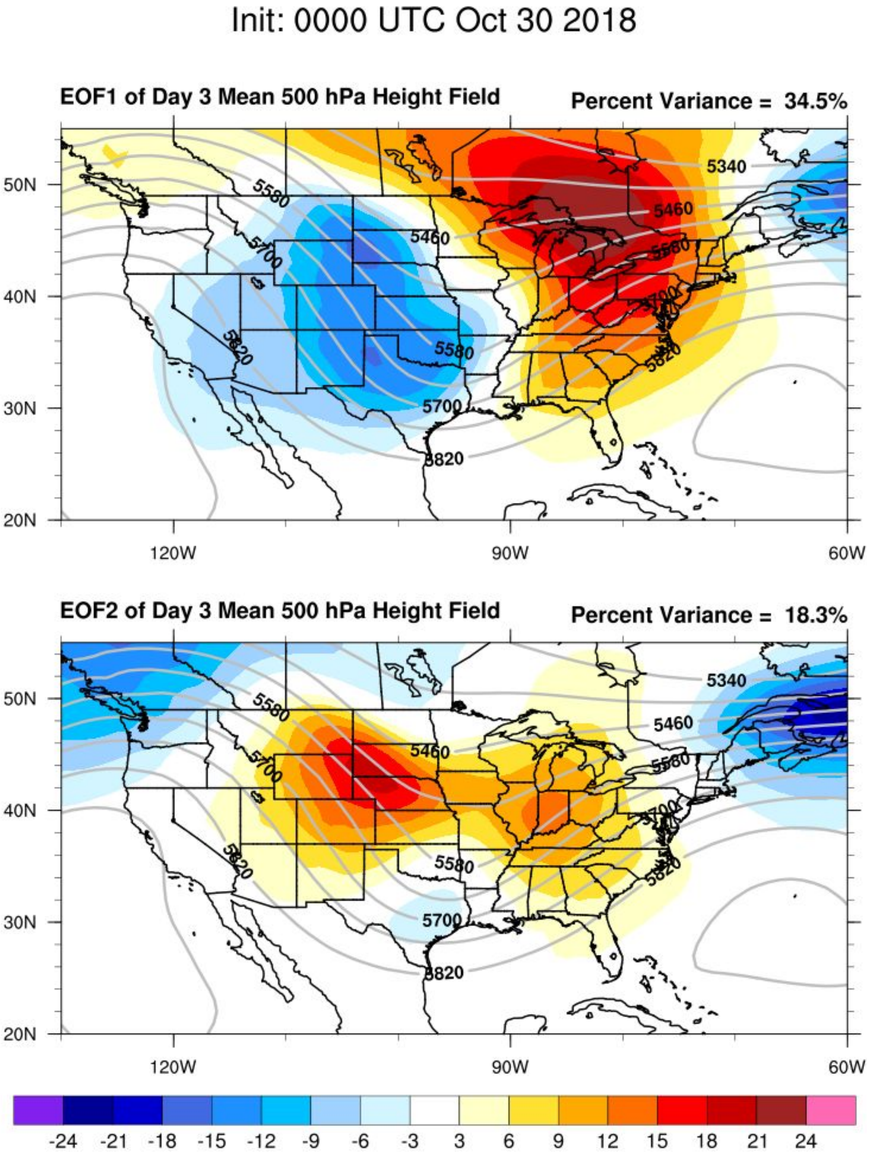

Figure 4. *Example of two EOF patterns (shaded) overlaid on the total 90-member ensemble mean 500 mb geopotential height field. The variance explained is provided in black text.*

2. The Principal Components (PCs) for the first two EOFs for each ensemble member were plotted on a phase space and a k-means clustering algorithm applied to group

ensemble members with similar forecasts into five distinct clusters (i.e., possible forecast scenarios). An example of the phase space and cluster membership for the day 3 forecast initialized at 0000 UTC on 30 October 2018 is shown in Figure 5.

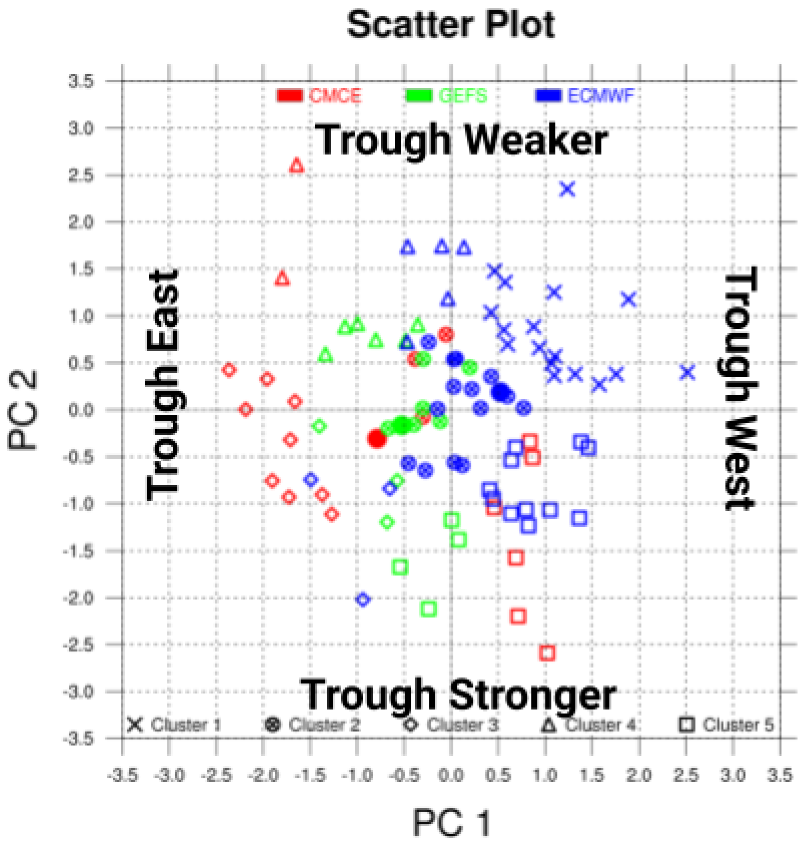

Figure 5. *The phase diagram showing each ensemble member colored by ensemble prediction system with the marker shape denoting cluster membership.*

3. Cluster mean forecasts of relevant parameters were viewed to further assess the event's predictiability and possible forecast scenarios. Figures 6a-c show differences in the location of the mid-level trough and its attendant surface cyclone and precipitation shield between the ensemble clusters. As expected, ensemble members that forecast the mid-level trough to be farther west (i.e., members in cluster 1) also forecast the attendant surface low and precipitation shield to be farther west. The members that forecast the mid-level trough to be farther east (i.e., the members in cluster 3) also forecast the attendant surface cyclone and precipitation shield to be farther east. Presenting ensemble forecasts in this manner gave WWE participants a better idea of the range of forecast outcomes for a given event.

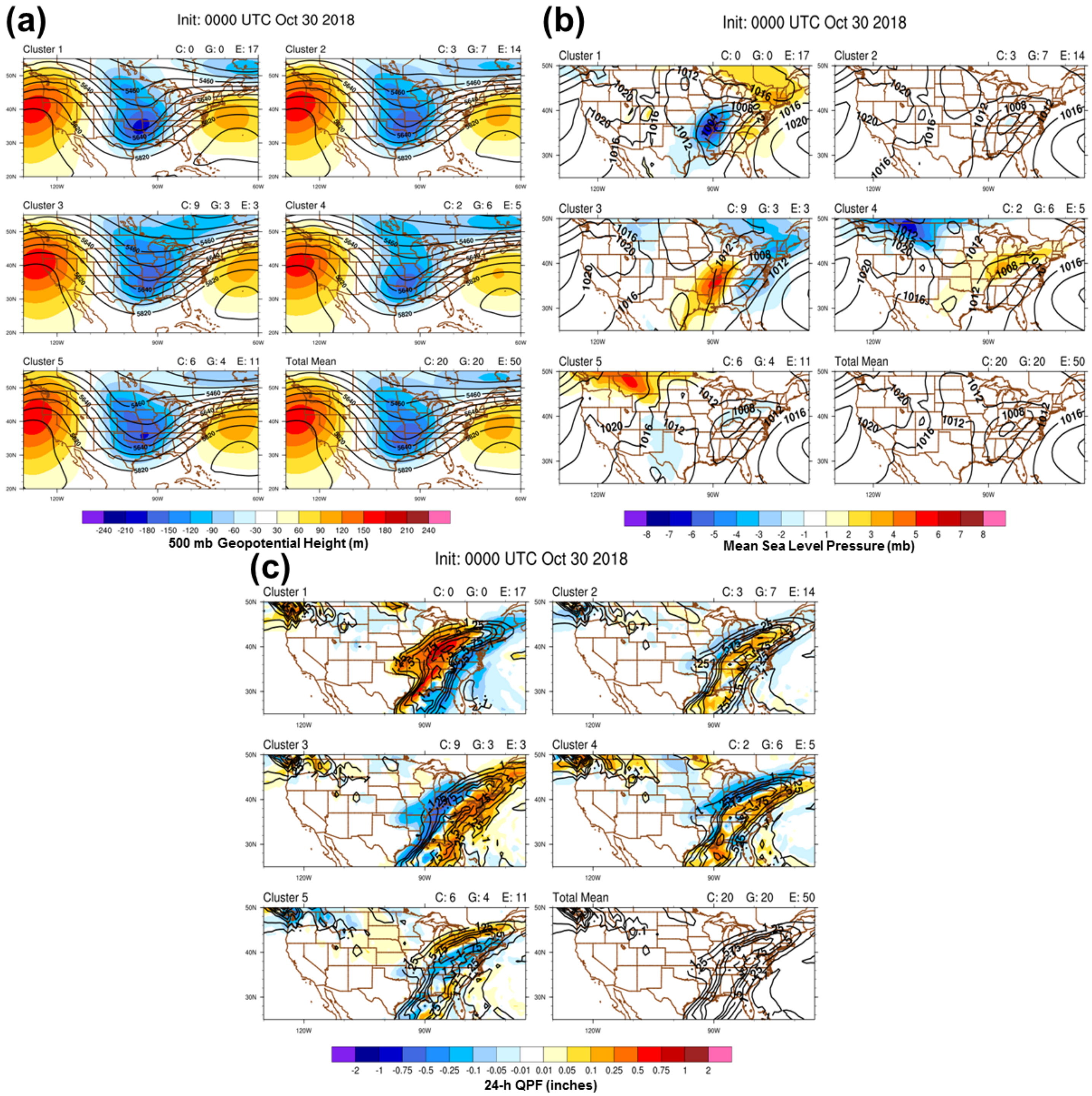

Figure 6. *Five cluster means and total 90-member ensemble mean of (a) 500 mb geopotential height with climatological anomalies shaded, (b) sea level pressure with differences from total ensemble mean shaded, (c) 24-h QPF with differences from the total ensemble mean shaded.*

For more information on fuzzy clustering please see Zheng et al. (2017) and references therein.

Forecast Exercise: Blending Methodology

To fully explore the utility of the experimental precipitation type methods and the synoptic scale predictability of the clusters to the snowfall forecasting process, a manual forecast blender was introduced to the WWE. Manual forecast blending of operational models and ensembles has been the cornerstone of the WPC forecast desks for the past decade and a half. Winter weather forecasters at WPC subjectively assign a weighted percentage to each model solution they wish to algebraically combine to generate a snowfall forecast. Although the experimental datasets were limited in comparison to the operational data sets available in real-time, the blending exercise nonetheless enabled experiment participants to dig deeper into the snowfall forecasting components of QPF, precipitation type data, and storm track variability.

At the start of each forecast exercise, a 6-hour period of interest was selected based on snowfall potential and the presence of a potential precipitation type transition during the forecast period. The selected 6-hour period focused on a small domain to allow for closer examination of the datasets. The WPC WWE team presented all of the experimental precipitation type and snowfall forecast guidance, and provided a thorough explanation of the cluster forecasts for the selected forecast period, as well as the ensemble sensitivity influencing the differences in storm track forecast and synoptic-scale features. The WPC WWE team then facilitated discussion to help participants select which model and cluster solution inputs to blend and how much weight to assign to each chosen input. Blends were run on NCEP Advanced Weather Interactive Processing System (NAWIPS) or Advanced Weather Interactive Processing System Graphical Forecast Editor (AWIPS-GFE) software then re-run with weight adjustments several times until the forecast was acceptable to all participants. The forecast was then saved and evaluated in the following week's verification session.

CAMs Datasets

Experimental CAMs guidance was provided by both GSD and EMC and are summarized in Appendix A and included: HRRR, HRRRE, HRRRX, and HREFv2.1. One featured piece of guidance was the experimental HRRR (i.e., HRRRX) with a 1-km nest within the parent 3-km domain. The FVCOM-ICE model run at GLERL provided lake ice concentration and lake surface temperature data used to initialize the HRRRX and was evaluated during the experiment for a few cases of lake effect snowfall and mesoscale snowfall events. It should be noted that there was limited data availability during the course of the experiment (18 cases total).

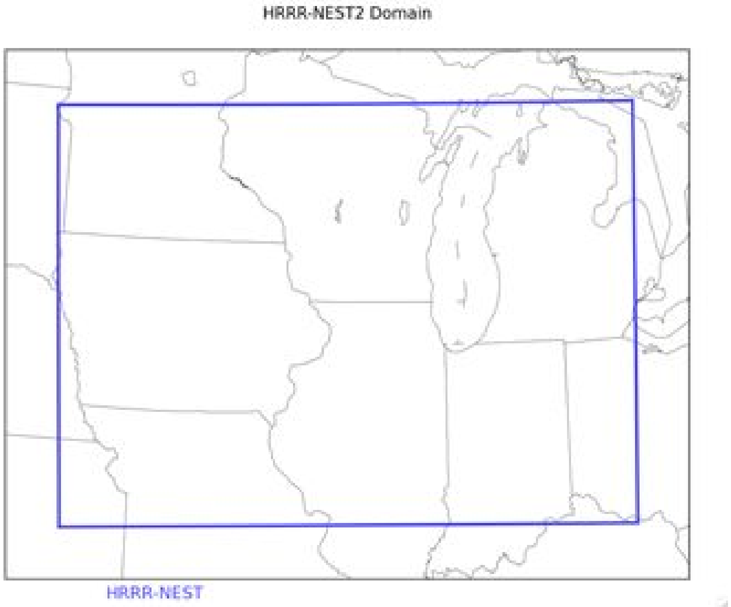

Figure 7. *The Experimental HRRR (HRRRX) 1-km domain (Image courtesy of Curtis Alexander).*

Verification

Experiment data was validated against 6-hour snowfall accumulation amounts available from the National Operational Hydrologic Remote Sensing Center version 2 (NOHRSCv2) dataset (Clark 2017). Visual evaluation with subjective scoring and discussion were conducted. 6-hour quantitative precipitation estimates (QPE) were also used to evaluate the 6-hour quantitative precipitation forecast (QPF) for all experimental blend inputs. Assessment of the large-scale pattern was done by comparing GFS 500 hPa geopotential height analysis over the United States with that of the blend inputs including ensemble clusters.

Objective verification of the experiment forecast blends and inputs was also conducted using the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2009) tool out of the Model Evaluation Tools (MET) software provided by the Developmental Testbed Center (DTC). The 6-hour snowfall forecast grids were thresholded by 1, 2, or 4 inch amounts and forecast objects were matched to observed objects, with statistics computed to assess their similarities (spatial coverage, orientation, distance, etc.). The details of the MODE configuration used for this analysis can be found in Appendix B.

Seasonal verification consisting of threat scores and frequency biases for 24-hour snowfall forecasts of the 1, 2, 4, 8, and 12 inch thresholds was computed for the Days 1,2, and 3 time periods using the Forecast Verification Software (FVS; Novak et al. 2014, their Appendix B). The verification was done for all forecast methodologies of the FV3 and NAM12, and computed for events aggregated over the period November 1, 2018 through March 23, 2019. MODE was also used to objectively score select CAMs guidance throughout the season; configuration details for the use with CAMs is provided in Appendix B.

Results

Cases

A total of 19 6-hour forecast blends were created during the course of the experiment (Table 2). The event-specific regional domains spanned most of the northern half of the Contiguous US, with the most events occurring in the Middle and Upper Mississippi Valleys and the Northeast. The most common event types investigated were cases where mid-level warm air advection affected precipitation type out ahead of a shortwave trough (8 out of 19 cases) and more robust continental cyclones (5 out of 19 cases).

Table 2. *List of all 19 WWE experiment session 6-h cases.*

Case	Forecast Date	Forecast Hour	Valid Date	Forecast Domain	Event Type
1	6 Nov 2018	FH 66	18Z 8 Nov 2018	Central Plains	Shortwave Trough
2	13 Nov 2018	FH 60	12Z 15 Nov 2018	OH Valley	Continental Cyclone
3	27 Nov 2018	FH 42	18Z 28 Nov 2018	Northeast	Synoptic-enhanced Lake Effect
4	27 Nov 2018	FH 54	06Z 29 Nov 2018	Upper Miss. Valley	Shortwave Trough
5	4 Dec 2018	FH 72	00Z 07 Dec 2018	Northeast	Lake Effect
6	18 Dec 2018	FH 24	00Z 19 Dec 2018	Northern Rockies	Jet Streak
7	29 Jan 2019	FH 24	00Z 30 Jan 2019	Northeast	Arctic Front
8	2 Feb 2019	FH 54	06Z 7 Feb 2019	Plains/Upper Miss. Valley	Shortwave Trough
9	12 Feb 2019	FH 24	00Z 13 Feb 2019	Northeast	Shortwave Trough
10	19 Feb 2019	FH 30	06Z 20 Feb 2019	Central Plains	Jet Streak
11	19 Feb 2019	FH 42	18Z 20 Feb 2019	Mid-Atlantic	Shortwave Trough
12	26 Feb 2019	FH 54	06Z Feb 28 2019	Northeast	Jet Streak
13	26 Feb 2019	FH 84	12Z Mar 1 2019	Mid-Atlantic	Shortwave Trough
14	5 Mar 2019	FH 72	00Z 8 Mar 2019	Midwest	Shortwave Trough

15*	11 Mar 2019	FH 78	06Z 14 Mar 2019	Northern Plains	Mature Continental Cyclone
16*	12 Mar 2019	FH 54	06Z 14 Mar 2019	Northern Plains	Mature Continental Cyclone
17*	13 Mar 2019	FH 30	06Z 14 Mar 2019	Northern Plains	Mature Continental Cyclone
18*	13 Mar 2019	FH 36	12Z 14 Mar 2019	Northern Plains	Mature Continental Cyclone
19*	10 Jan 2019	FH 54	06Z 20 Jan 2019	Middle Miss. Valley	Shortwave Trough
*In-house Week Sessions					

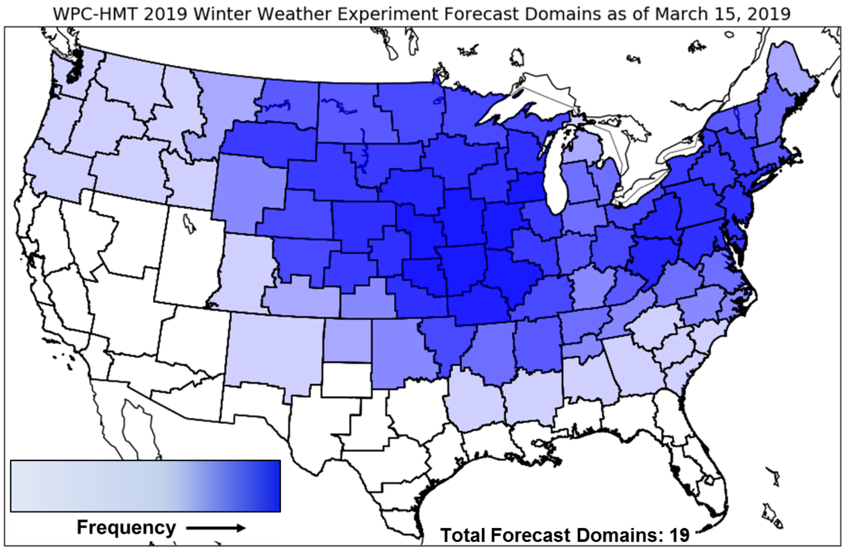
WPC-HMT 2019 Winter Weather Experiment Forecast Domains as of March 15, 2019
Frequency
Total Forecast Domains: 19

Figure 8. *Spatial WFO distribution of coverage within WWE experiment event domains.*

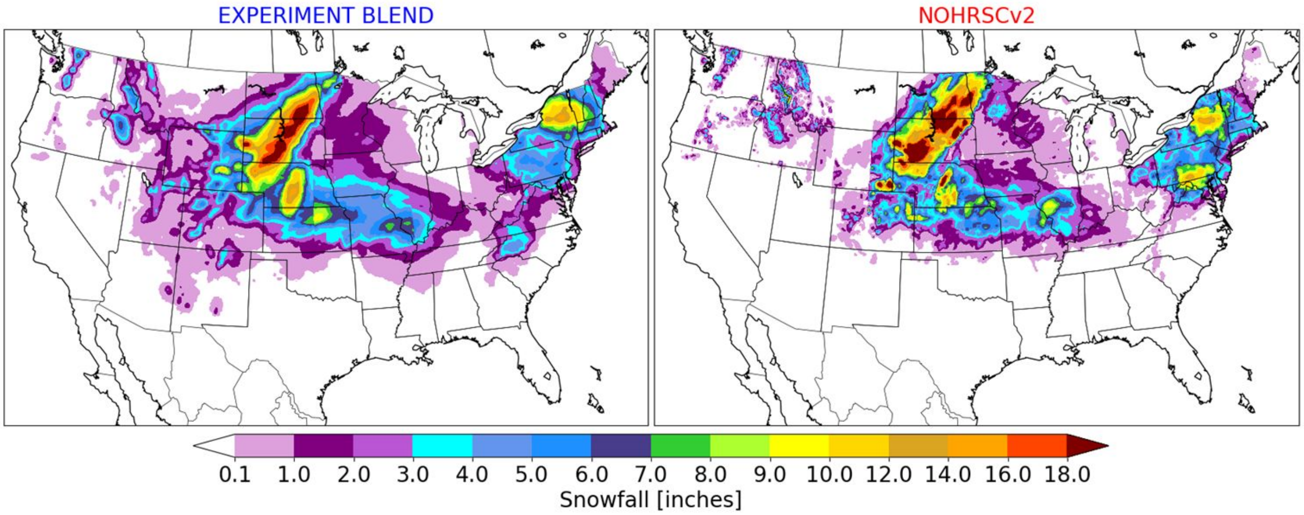

Figure 9. *Total aggregated snowfall from all 19 events from the experiment blend (left) and observed NOHRSCv2 amounts (right). The 13-14 March 2019 blizzard is prominent due to the fact that it comprised 4 out of all 19 events.*

Highlight: 12-14 March 2019

A notable strong continental cyclone occurred during the latter part of the in-house week which allowed participants to create forecasts for the same valid time for three consecutive experiment days for the periods valid between 06-12 UTC 14 March 2019. This large-scale event was well predicted by experimental guidance which allowed for in-depth discussion of the mesoscale details during the forecast sessions.

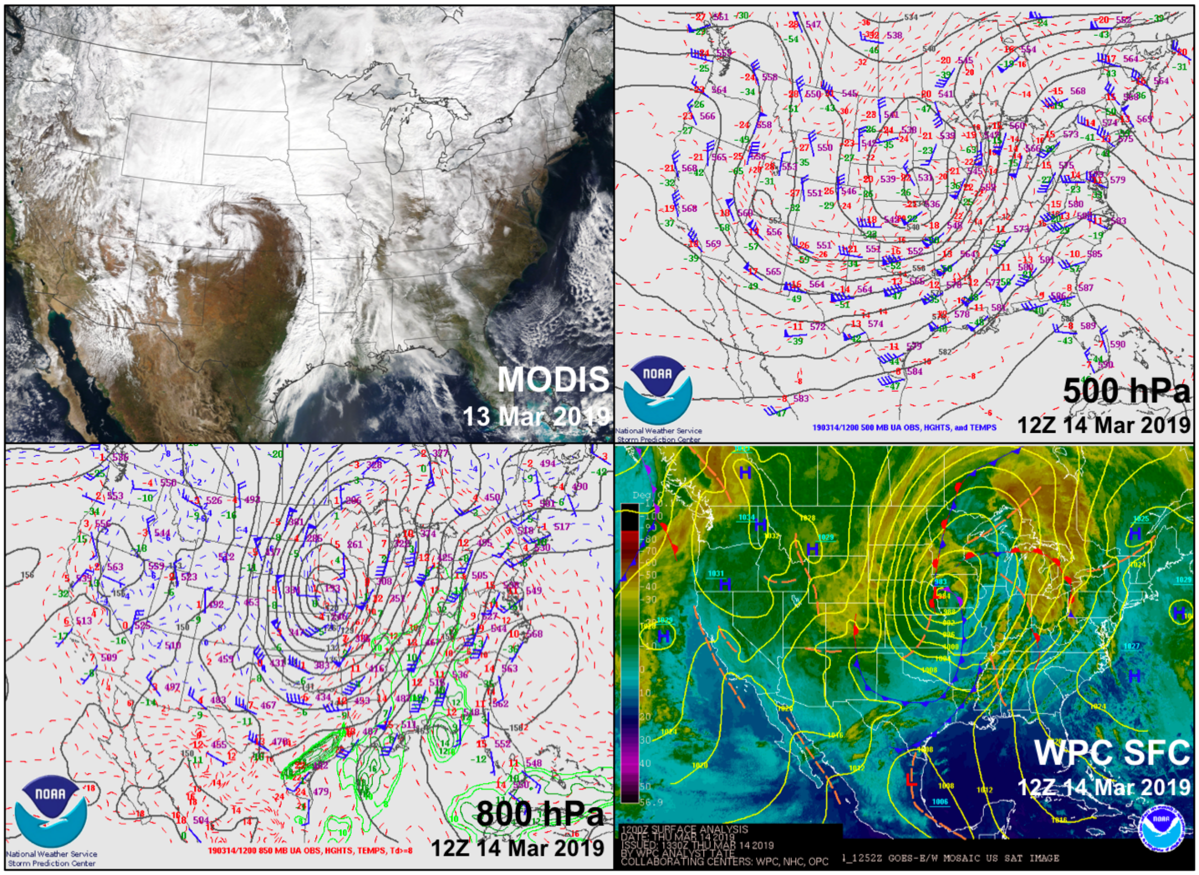

***Figure 10.** (upper-left) MODIS satellite imagery from 13 Mar 2019, (upper-right) 500-hPa, (lower-left) 800-hPa, (lower-right) WPC surface analysis analysis valid at 12Z 14 Mar 2019.*

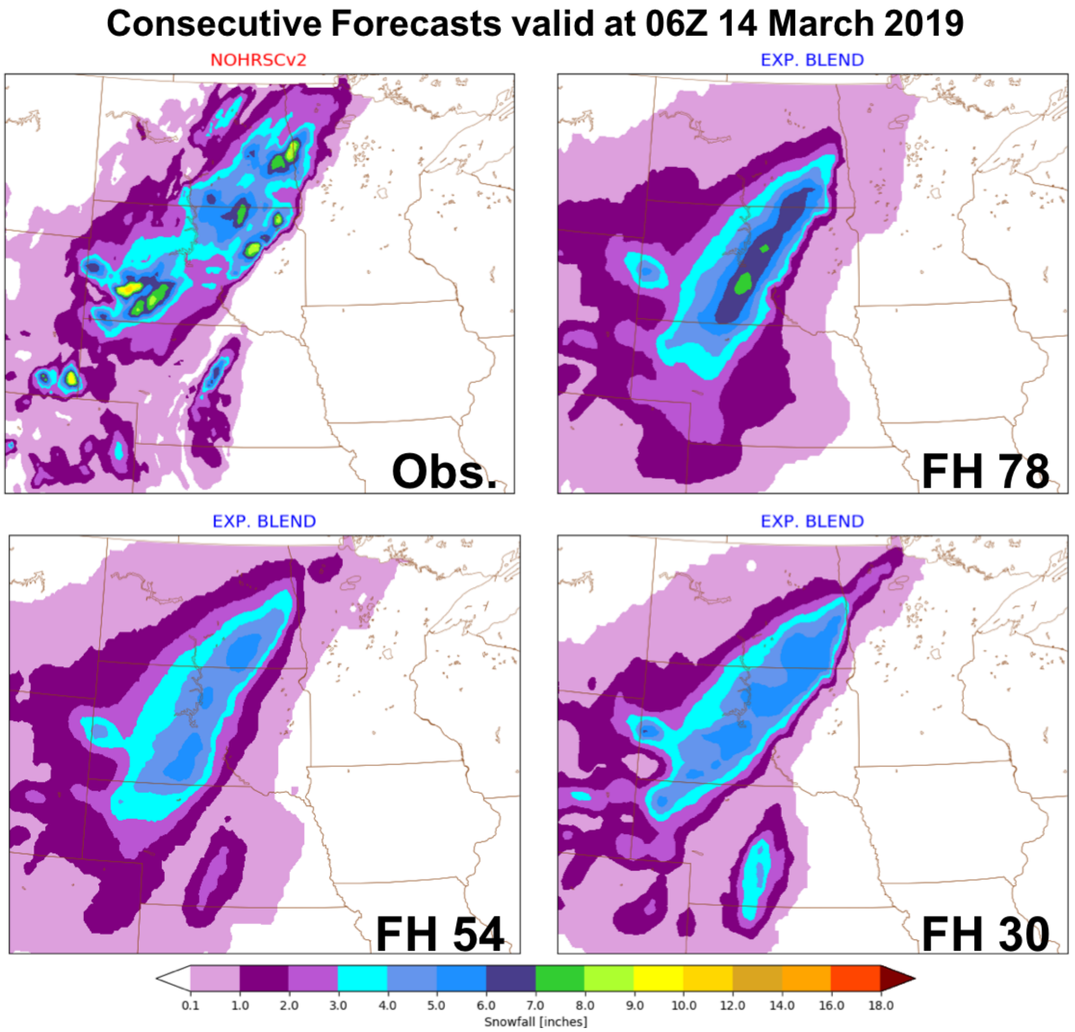

Figure 11. *(upper-left) NOHRSCv2 6-hour snowfall valid at 06Z 14 March 2019, experiment forecast blends with a lead time of 78 h (upper-right), 54 h (lower-left), and 30 h (lower-right).*

Experiment Forecast Verification

Experimental forecasts were verified using 6-hour NOHRSCv2. During verification sessions, subjective scores were collected and discussed. An example graphic is provided below that shows 6-hour snowfall from all blend inputs, the resultant experiment blend, the NBMv3.1 and the verifying NOHRSCv2. Additional graphics were examined that included spatial plots of the error or bias and how the experimental blend performed relative to the NBMv3.1.

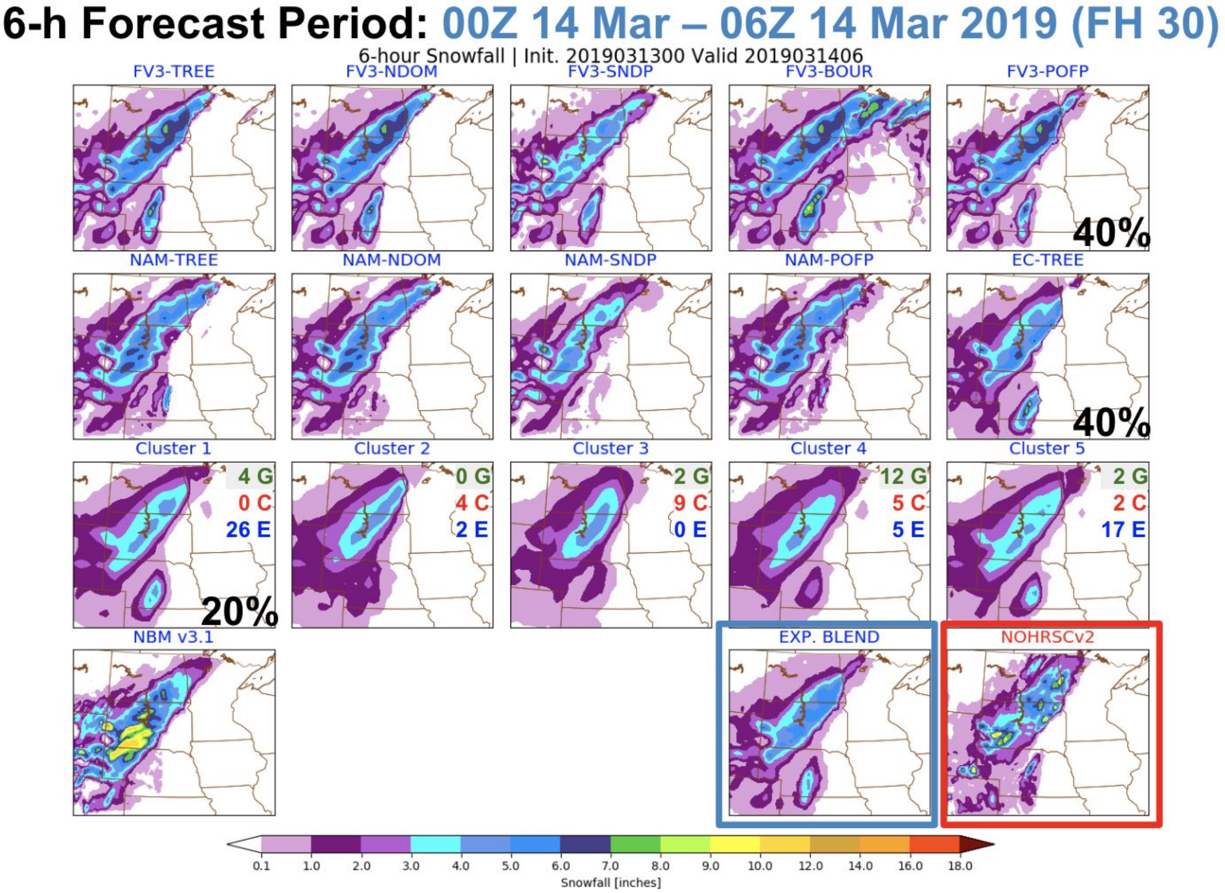

Figure 12. *Example of forecast inputs from precipitation type methods and clusters (with ensemble membership labeled), the NBM v3.1, experimental blend (blue box comprised of weights given in black text) that were subjectively evaluated with NOHRSCv2 snowfall (red box).*

According to participants, reasons for decreased scores included errors in the timing of the system, the thermal field (e.g., too much or too little warm air advection aloft), among others. Out of a scale of 1 to 10 with 10 being a perfect forecast, participants on average scored the experiment blend 5.23 and the distribution was skewed towards higher values. Participants tended to score higher versus lower because they analyzed the inputs into the blend and understood the limitations of the inputs as a barrier to creating a more accurate forecast blend.

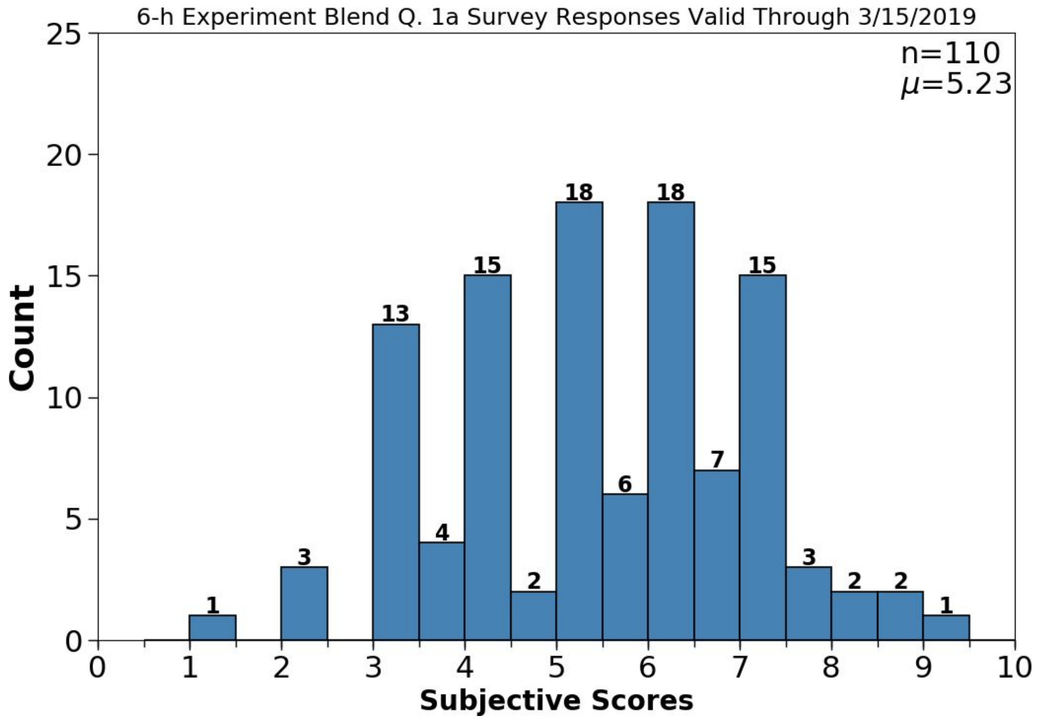

Figure 13. *Experiment blend subjective evaluation scores. Participants scored the blend from 1 to 10, with 1 being the lowest score possible and 10 being the highest. The total count (n) and mean score (μ) are labeled as well as the total counts for each score value in increments in 0.5.*

Statistics were calculated from the 19 cases examined from all of the experimental blend inputs and the experimental forecast blend itself. It was found that the FV3-BOUR exhibited a robust high bias (over-forecasting of snowfall), both the NAM and FV3 change in snow depth precipitation type (SNDP) exhibited a notable low bias (under-forecasting of snowfall), and the clusters exhibited a low bias on average (under-forecasting of snowfall, likely due to the coarse grid spacing of global ensemble members). The NBMv3.1 exhibited more over-forecasting than the experimental blend, likely due to the over-forecasting of its underlying QPF which was noted during several verification sessions. In general, the experimental blend often exhibited a higher threat score and thus performed better than individual precipitation type inputs.

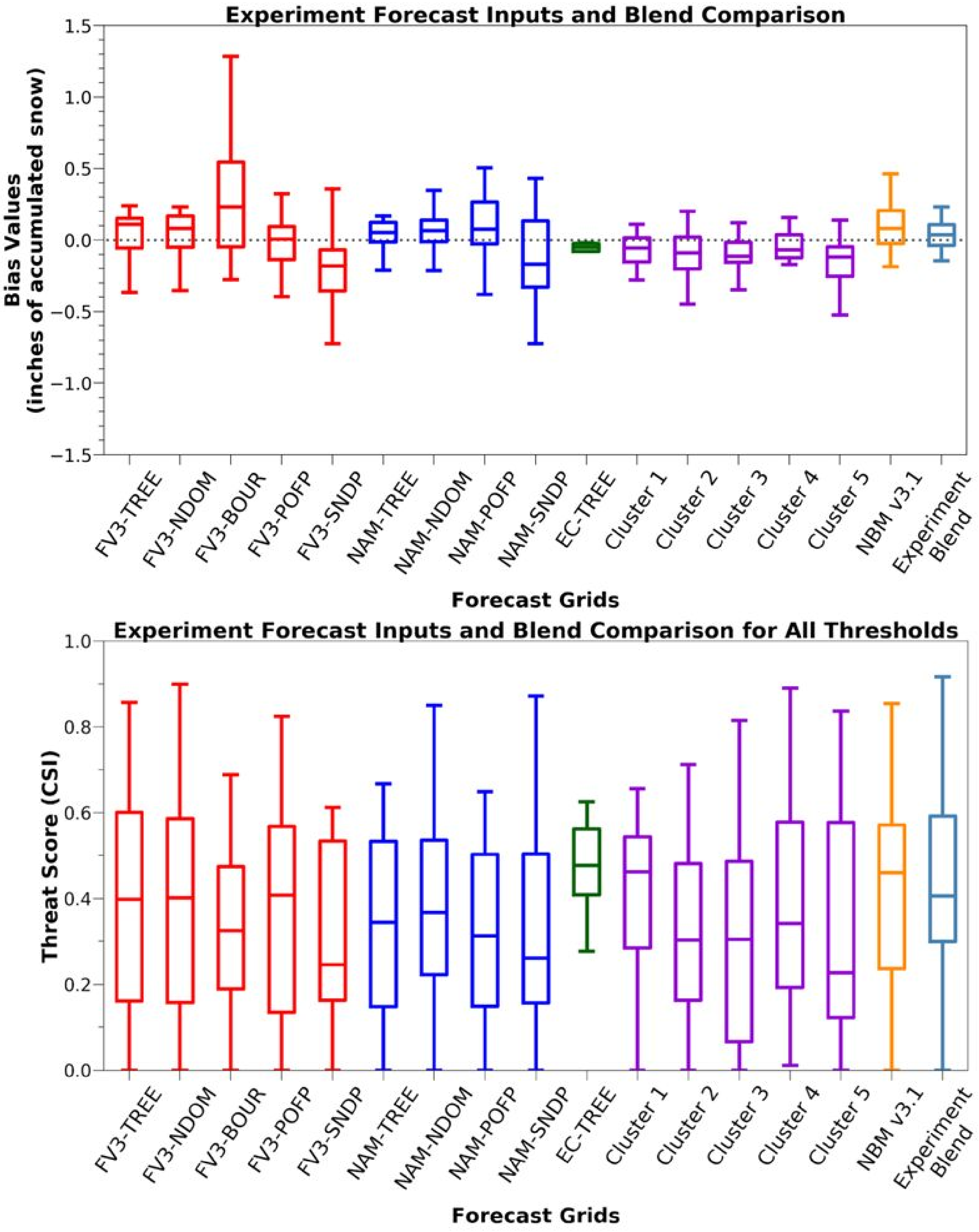

Figure 14. *Distribution of bias (top panel) and CSI/Threat score (bottom panel) values for each precipitation type experimental guidance for the FV3 (red), NAM-12 (blue), ECMWF (green), clusters (purple), NBM v3.1 (orange), and experimental forecast blend (light blue). The bar within the rectangle indicates the median of the distribution. The bounds of the box reflect the 75th (top) and 25th (bottom) percentiles of the values. The whiskers extend to the lowest and highest values. A bias of 0 is indicated by the gray dashed line.*

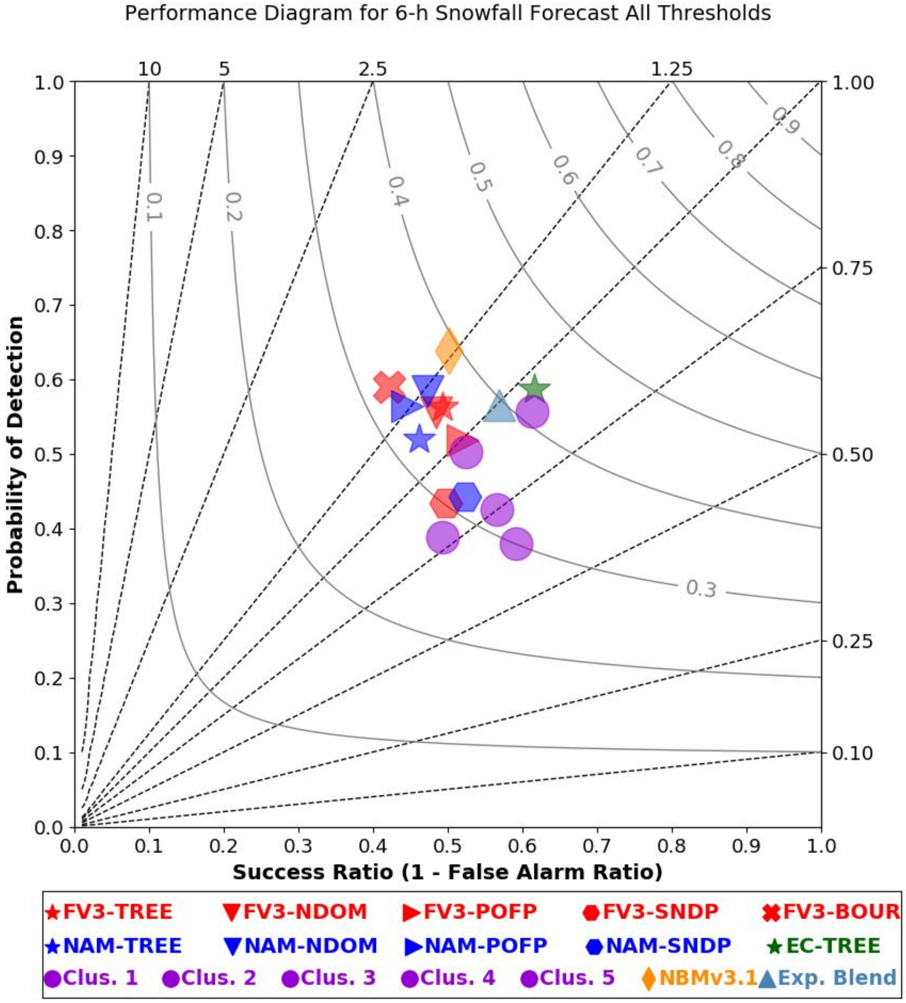

Figure 15. *Performance diagram for 6-h forecasts experiment cases of all objects >1", >2", >4" of accumulated snowfall with each marker color denoting the model and each marker type denoting the precipitation type method. Other markers include the 5 clusters (purple circles), NBM v3.1 (orange diamond), and experimental forecast blend (light blue triangle). The dashed lines in the plot indicate bias and the curved lines indicate CSI/Threat Score.*

MODE analysis of forecast objects, which accounts for shape and position errors, was conducted for all of the precipitation type inputs and blends. For forecasts that occurred at the > 1" threshold, the majority were displaced to the south of the observed object (i.e., indicating a southward bias). More NAM forecasts had a northward bias than FV3, which could perhaps be related to a cold bias resulting in a southward bias in FV3.

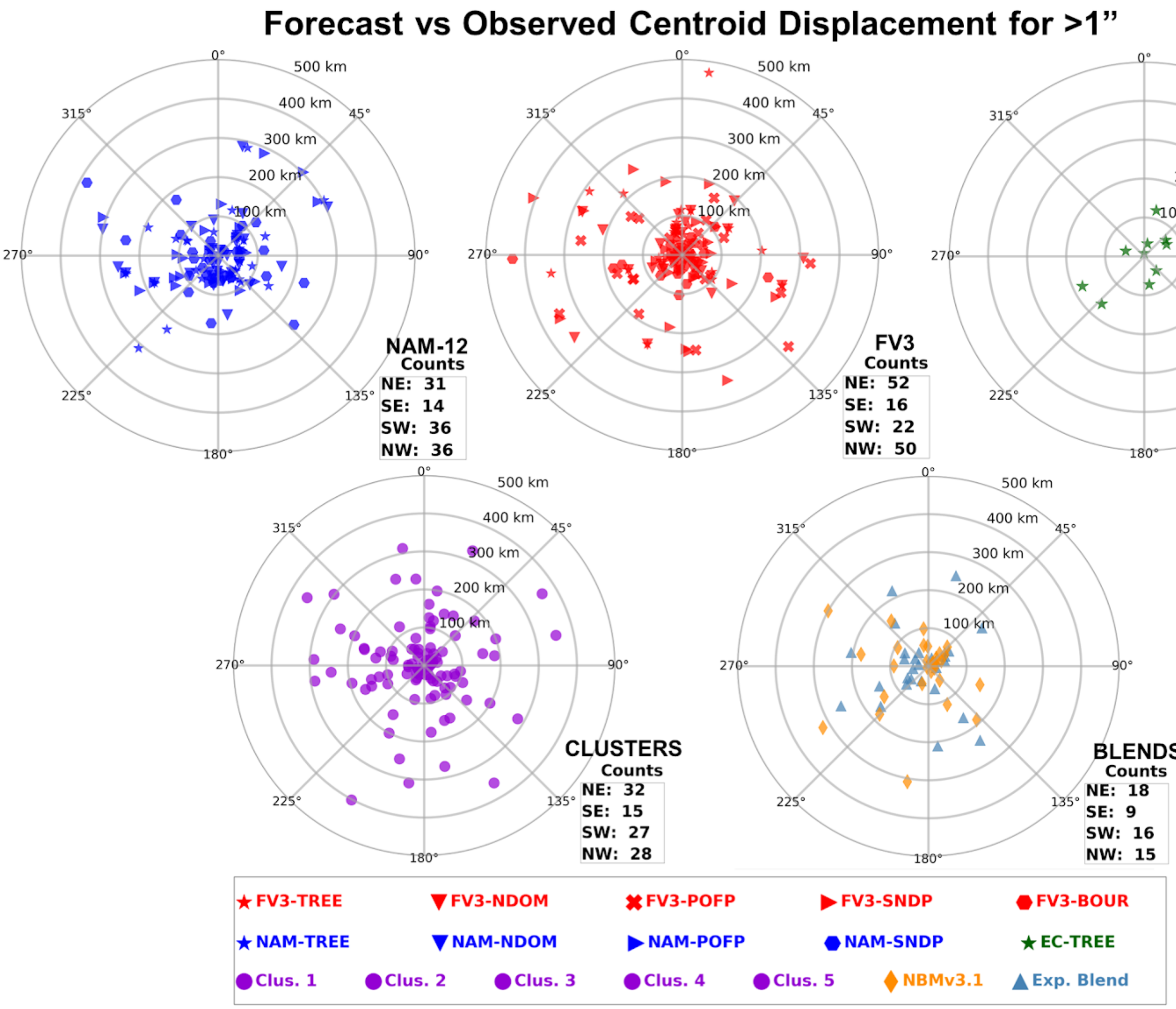

Figure 16. *Distribution of the position of the observed object relative to the forecast object (assumed to be at the origin) for 6-hour snowfall objects > 1 inch. Each marker color denotes the model and each marker type denotes the precipitation type method according to the legend at the bottom. Other markers include the 5 clusters (purple circles), NBM v3.1 (orange diamond), and experimental forecast blend (light blue triangle). The radial axis is the centroid distance in kilometers. The angular axis is the angle from true north of the observation object relative to the forecast object (e.g., 45° indicates that the observed object occurred to the NE of the forecast object, so the forecast had a SW bias).*

Season-long Verification: FV3 and 12-km NAM

Seasonal verification results for the FV3 showed similar threat scores among the WPC decision tree, NCEP dominant, and fraction of frozen precipitation methods for estimating precipitation type when applied to the FV3 precipitation forecast with a 10:1 SLR. Similar threat scores were also noted with the 24-hour change in snow depth forecast from the FV3, and the FV3 Filter which is a derived snowfall using the Baxter climatological SLR (Baxter et al. 2005), the FV3 QPF, and the fraction of frozen precipitation to estimate frozen precipitation reaching the ground. The rime factor proxy, derived from graupel and snow mixing ratios, is used to adjust the SLR in precipitation type transition zones.

Despite the similarity of the threat scores, there were noted disparities in frequency biases among the forecasts at the 8 and 12 inch thresholds. Over the CONUS domain, the change in snow depth forecasts presented a lower frequency bias for these heavier thresholds, while the filter method showed a higher frequency bias.

A snowfall forecast using the Bourgouin precipitation type applied to the FV3 QPF was also evaluated over the course of the 2018-19 winter season. Threat scores were generated with the FVS for selected snowfall thresholds from the Bourgouin based snowfall forecasts, and the results showed lower scores than the other FV3 methodologies. Since the WPC decision tree methodology is used operationally at WPC, the TREE forecasts were chosen as the control, and the other experimental methods were selected for comparison. A null hypothesis (Hamill 1999) that scores are identical is assumed, and the FVS uses resampling to compute confidence intervals (Brill 2017). The FVS uses an archive of forecast, hit, and observation fraction statistics. Using the filtered collection of this archive record, the FVS randomly swaps the membership of the archive between the control and comparison forecasts, and recomputes the scores for both with each iteration. The resultant histogram of score differences is used to compute the confidence intervals. The confidence intervals provide a measure of the likelihood of the null hypothesis to be true. So, if one bar is outside the 95% confidence interval range, it is unlikely that the scores are the same thereby the differences more significant. These forecasts also showed a much higher frequency bias for all thresholds on Days 1-3. The high frequency biases coupled with the lower threat scores can likely be attributed to over-forecasting of snowfall in areas where the other methods suggested no snowfall. This over-forecasting is a result of applying FV3 QPF into the 6-hour snowfall bin despite only a low probability of snow being the precipitation type. The graphic below plotted from FVS software shows the Day 2 seasonal results from the FV3 forecast methods for the domain east 105°W.

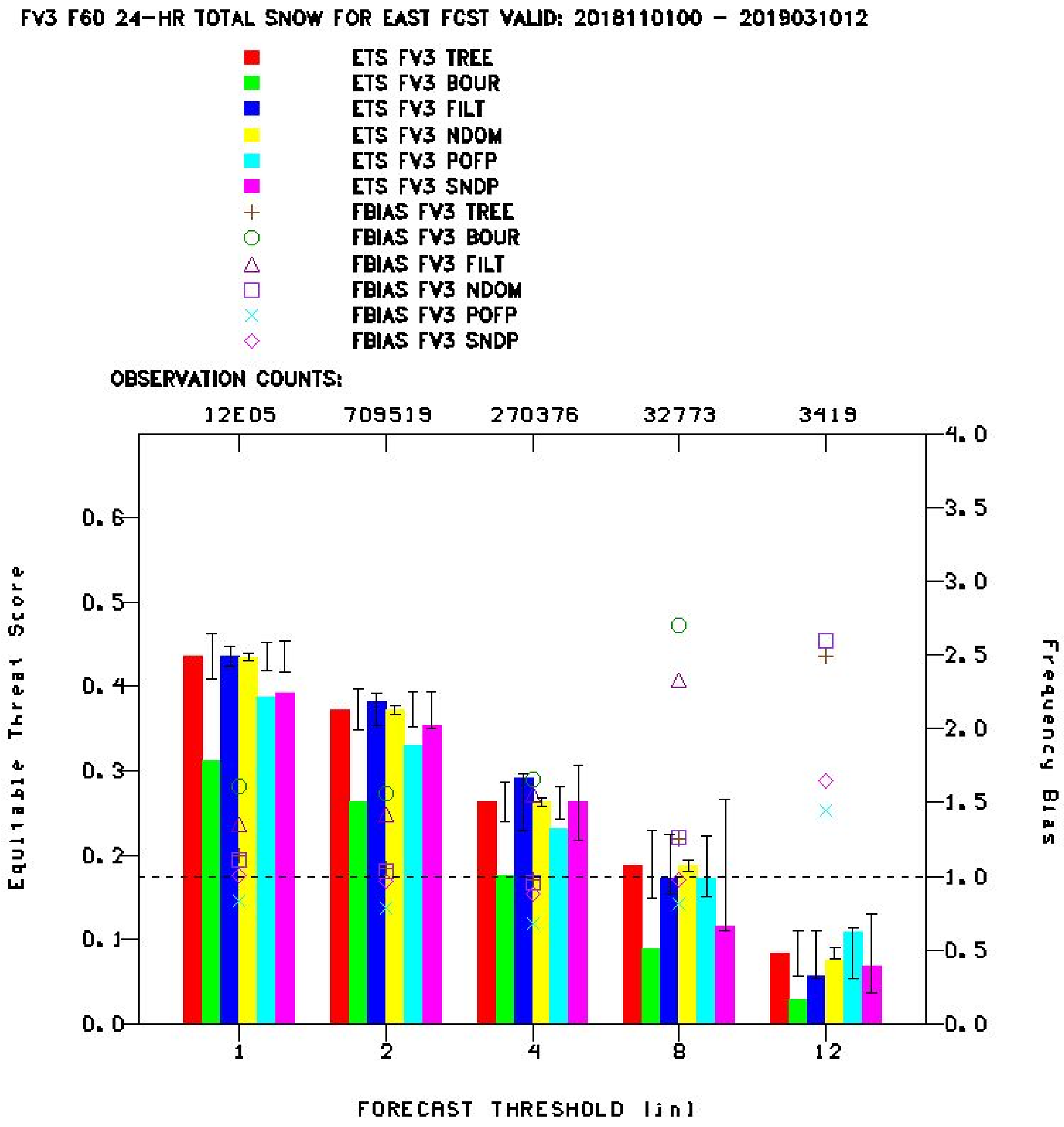

Figure 17. *Color bars represent the threat score for each FV3 snowfall forecast, at Day 2, for each threshold. The red bars (left most bar at each threshold) represents the FV3 snowfall forecast using the WPC decision tree to estimate precipitation type and statistical significance is tested against this solution since this method is used on the WPC Winter Weather Desk. Bourgouin is in light green, Filter in dark blue, NCEP Dominant in yellow, POFP in light blue, change in snow depth in magenta. Frequency biases are plotted using the shapes in the legend. Statistical significance in threat score differences is indicated by whisker bars completely outside or inside an individual color bar.*

Seasonal results from the NAM12 show similar threat scores among the suite of snowfall forecasts using the NAM12 QPF. No forecasts using the Bourgouin energy layers were available for the NAM12. The 24-hour change in snow depth forecast performed well at the higher thresholds, and showed a slightly higher frequency bias than was the case in the FV3. The graphic below depicts NAM12 threat scores and frequency biases for Day 2 over the domain east of 105°W.

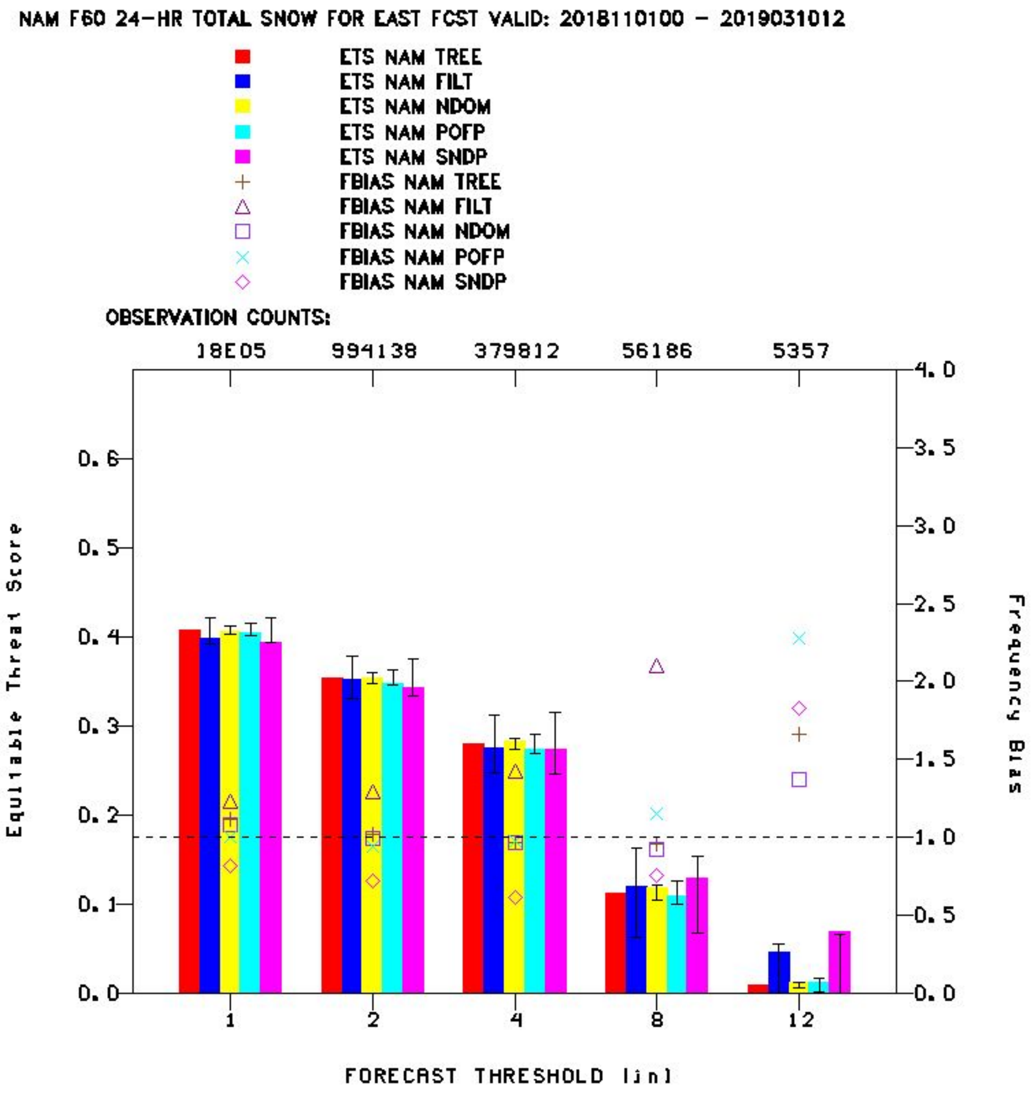

Figure 18. *Color bars represent the threat score for each NAM12 snowfall forecast, at Day 2, for each threshold. The red bars (left most bar at each threshold) represents the NAM12 snowfall forecast using the WPC decision tree to estimate precipitation type and statistical significance is tested against this solution as this method is used on the WPC Winter Weather Desk. FIlter is dark blue, NCEP Dominant in yellow, POFP in light blue, and change in snow depth in magenta. Frequency biases are plotted using the shapes in the legend. Statistical significance in threat score differences is indicated by whisker bars completely outside or inside an individual color bar.*

Highlight: FV3-GFS Evaluation for 15-16 November 2018 Event

Testing of the GFDL microphysics scheme in the FV3 was paramount in this year's WWE. Several cases that fell outside of the scheduled experiment times were analyzed including one from 15-16 November 2018. We found that for identifying precipitation type, the GFDL microphysics was an improvement over the Carr microphysics in the operational GFS. In spite of the numerous cold biases sited in the FV3, we found that in scenarios where the FV3 correctly forecast QPF, the precipitation type forecasts in the FV3 were on the whole quite good.

The precipitation type was well forecast using FV3 microphysics in an early season snow event where 6 inches or more of snow fell west of I-95 along the Philadelphia to New York City corridor in the six-hour period 1800 UTC November 15, 2018 to 0000 UTC November 16, 2018. The fraction of frozen precipitation field in the FV3 forecast values of ≥ 90% through this area, and the resultant precipitation type algorithm suggested snow over this area with sleet or a mixture of rain and snow along and to the east of I-95.

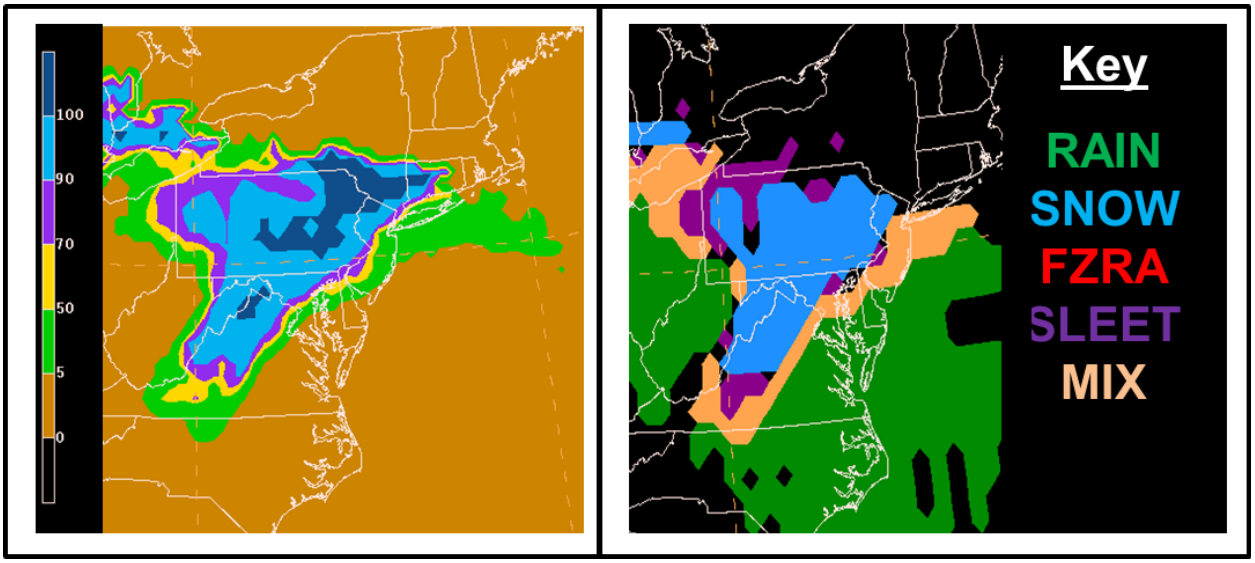

Figure 19. *Left image shows fraction of frozen precipitation field over northern Mid-Atlantic region. Dark blue represents POFP values of 100%, while light blue, 90-99%. Right image depicts instantaneous precipitation type at 1800 UTC November 15, 2018.*

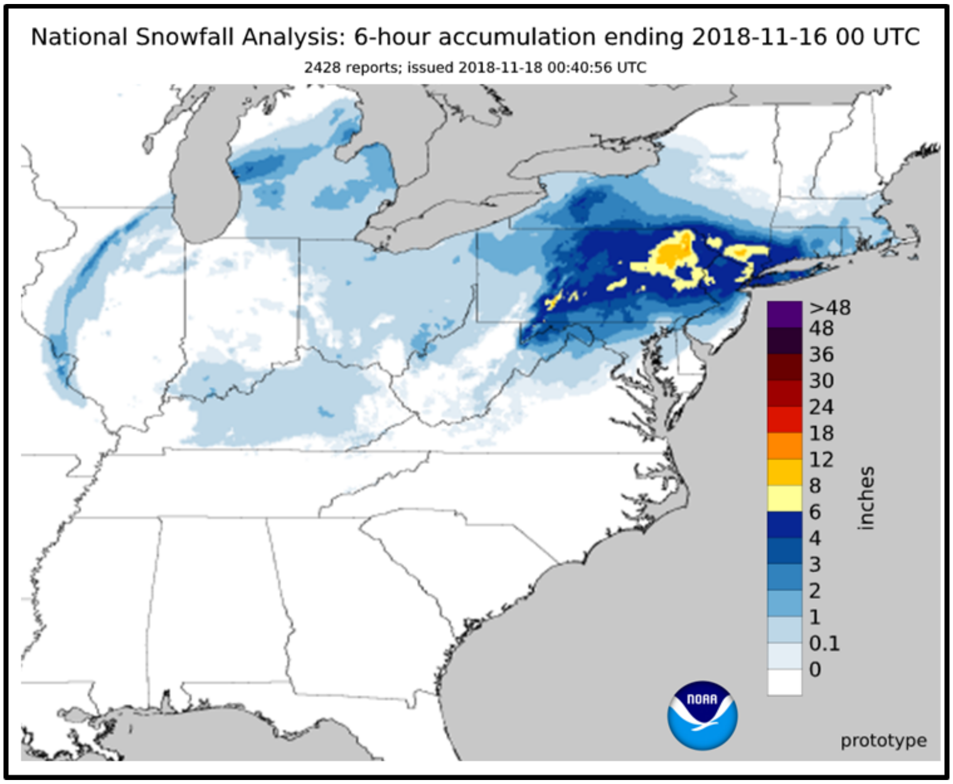
National Snowfall Analysis: 6-hour accumulation ending 2018-11-16 00 UTC
2428 reports; issued 2018-11-18 00:40:56 UTC
>48
48
36
30
24
18
12
8
6
4
3
2
1
0.1
0
inches
NOAA
prototype

Figure 20. *NOHRC 6-hour snowfall analysis ending 0000 UTC November 16, 2019.*

The lowest level of the FV3 graupel mixing ratio and snow mixing ratio were utilized to generate a rime factor proxy to identify sleet. This FV3 rime factor proxy was also used to modify the Baxter climatological SLR. The Baxter SLR was tapered down incrementally as values of the rime factor proxy increased. WPC employs a similar method to adjust the Roebber SLR using the actual rime factor produced by the Ferrier-Aligo microphysics in the NAM. This method for adjusting SLRs using the GFDL microphysics in the FV3 was tested over the course of the 2018-19 cold season by evaluating the Filter snowfall methodology cited in the seasonal verification discussion.

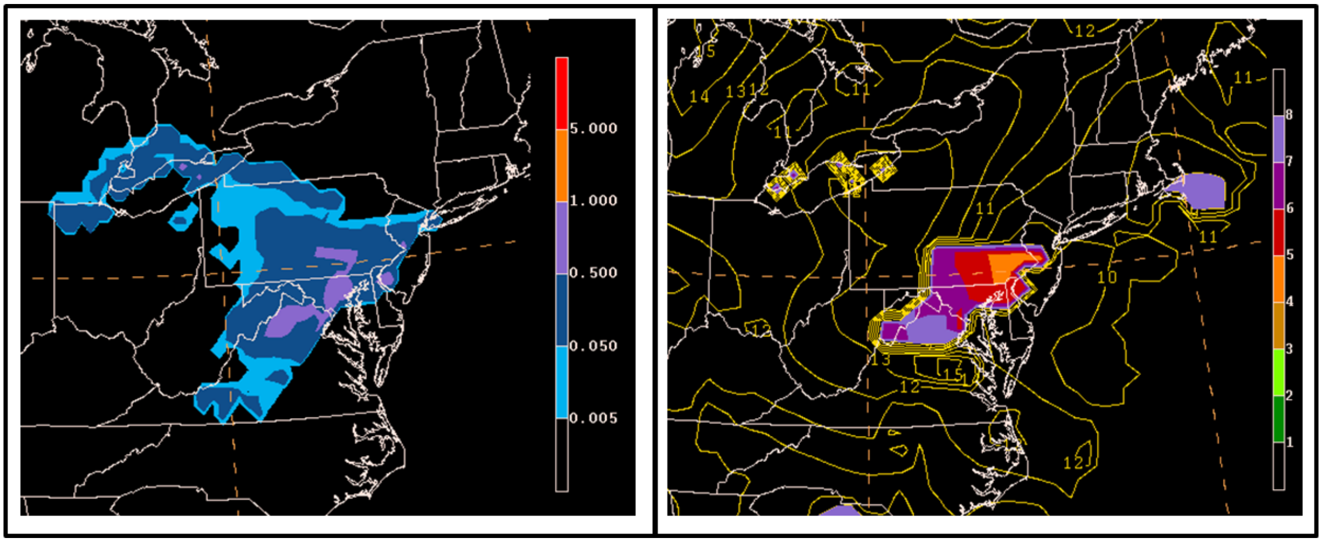

Figure 21. *(Left) Values of the rime factor proxy in the lowest level of the FV3. RF values of 0.5 (purple) and 0.05 (dark blue) initiated substantial reductions in the Baxter SLR. (Right) SLR values reduced to 5:1 and 4:1 over Southeastern PA. Note that SLRs were not reduced over Northeast PA where the heaviest 6-hour snowfall amounts were observed.*

CAMs Verification

Deterministic snowfall data from CAMs were evaluated for a number of mesoscale snowfall events and lake effect snowfall events. Experiment participants were asked to subjectively score 6 or 24 hour snowfall forecasts from both the operational HRRR and the experimental HRRR (HRRRX). On average, the HRRR scored higher than the HRRRX with a mean of 4.58 and 4.34, respectively. On a case-by-case basis, the forecast footprint of accumulated snowfall (i.e., where snow fell versus where it didn't) were rarely very different. The differences were in the magnitude of snowfall between the HRRR and HRRRX. Using data from the whole season (note: HRRRX data availability was very limited compared to HRRR), it was found that the 3-km HRRRX often over-forecasted amounts compared to the HRRR which often under-forecasted amounts at higher snowfall thresholds (> 4 inches).

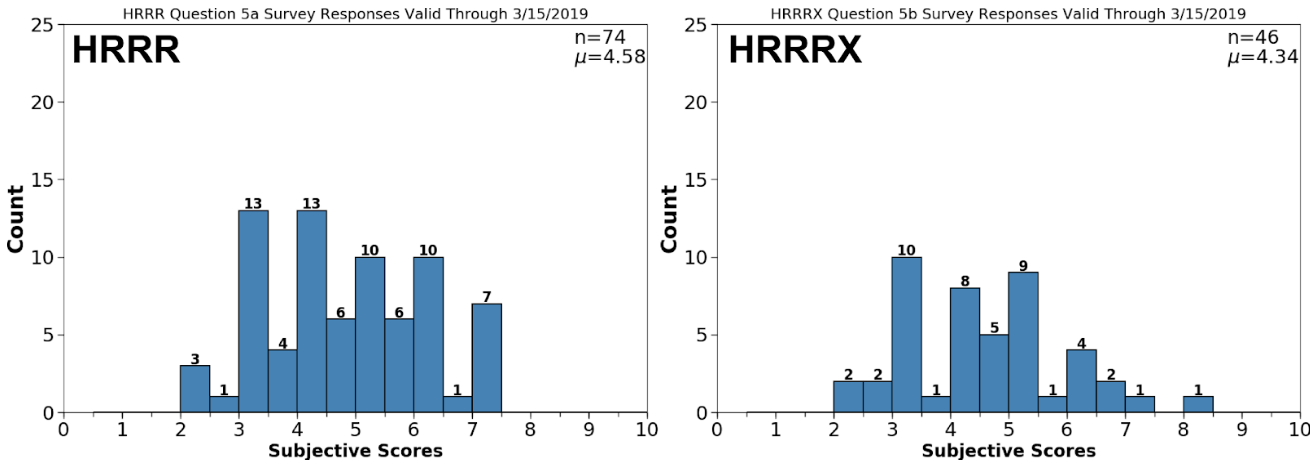

Figure 22. *Subjective evaluation scores for the Operational HRRR (left) and HRRRX (right) snowfall forecasts. Participants scored from 1 to 10, with 1 being the lowest score possible and 10 being the highest. The total count (n) and mean score (μ) are labeled as well as the total counts for each score value in increments in 0.5.*

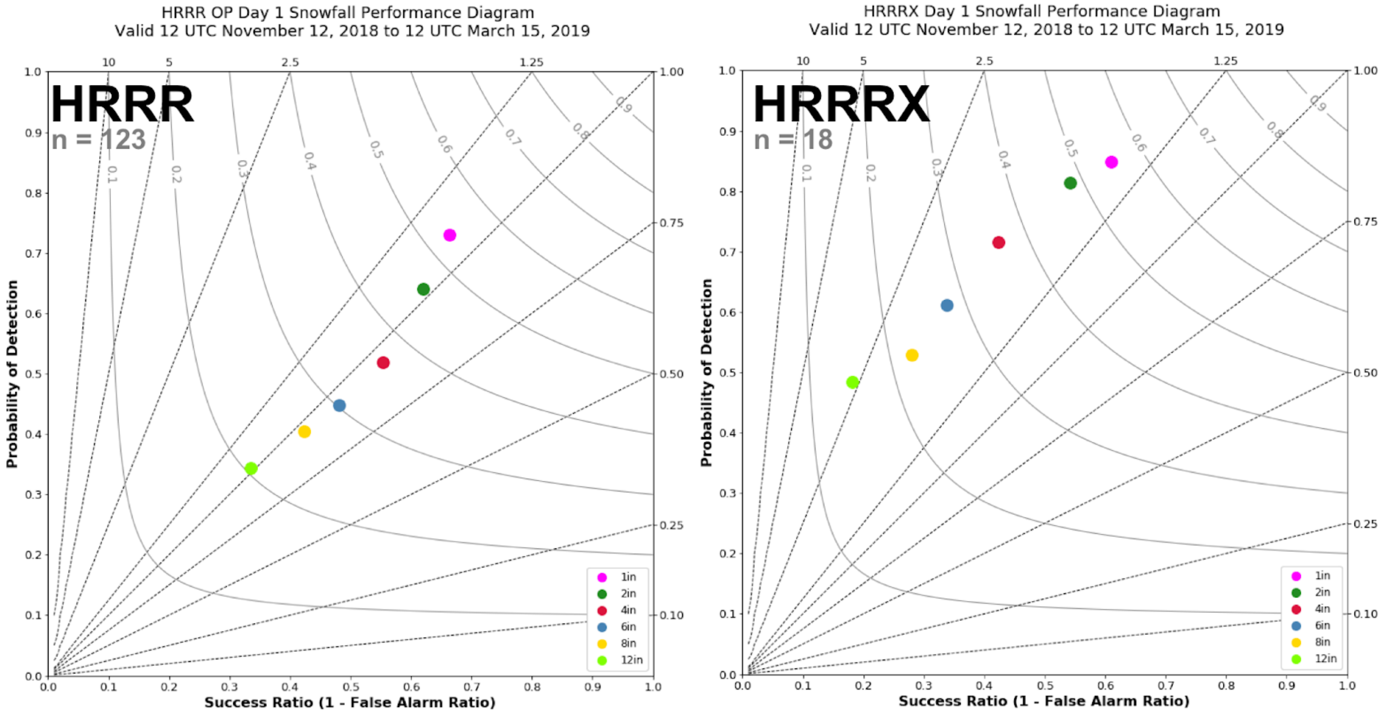

Figure 23. *Performance diagrams for each 24-h snowfall forecast threshold indicated by marker color for the HRRR (left) and the HRRRX (right) for all cases available for the season labeled for each model. The dashed lines in the plot indicate bias and the curved lines indicate CSI/Threat Score.*

The HRRRX-ICICLE 1-km nest was compared to the parent 3-km HRRRX domain for a selection of retrospective cases. In general, the 1-km nest was found to perform as well as or worse than the parent 3-km domain according to subjective evaluation by experiment participants. Feedback included an apparent under-forecasting of the snowfall field by the 1-km nest, which was likely due to the breaking up individual lake effect snowband circulations. Instead of having a larger band off of a lake producing measurable snowfall as in the 3-km domain, the 1-km domain may have produced more individual bands producing less accumulated snowfall at a single point that could look more realistic in fields like simulated reflectivity but less so accumulated snowfall.

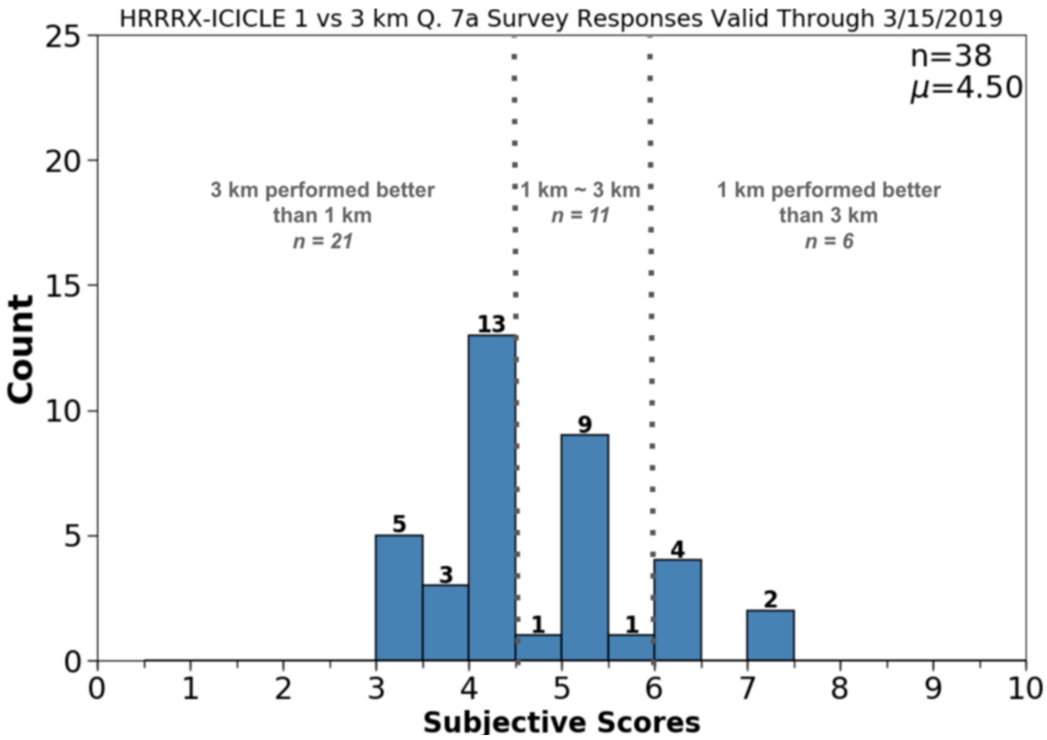

Figure 24. *Subjective evaluation scores comparing the 1-km HRRRX nest with the 3-km HRRRX snowfall forecasts. Participants scored from 1 to 10, with < 5 indicating a preference for the 3-km over the 1-km product, and > 5 indicating a preference for the 1-km over the 3-km product. The total count (n) and mean score (μ) are labeled as well as the total counts for each score value in increments in 0.5.*

Ensemble mean and probabilistic snowfall data from CAMs ensembles were evaluated for mesoscale snowfall events and lake effect snowfall events. The ensemble mean snowfall data for 6-hour or 24-hour amounts from the HRRRE scored lower on average than the parallel version of the HREF (HREFv2.1) with an average subjective score of 4.69 and 5.14, respectively. Qualitative feedback from experiment participants stated that the HRRRE produced an under-dispersed solution and that for some events failed to capture the uncertainty in timing and placement that the HREFv2.1 could, likely because the HREFv2.1 is comprised of multiple models instead of a single core.

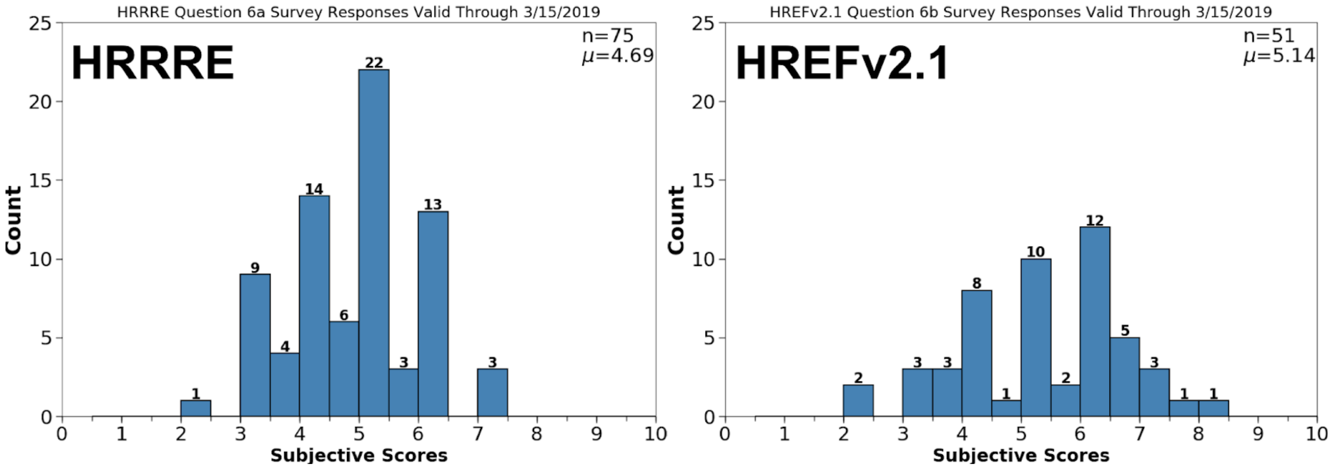

Figure 25. *Subjective evaluation scores for the HRRRE (left) and HREFv2.1 (right) ensemble mean snowfall forecasts. Participants scored from 1 to 10, with 1 being the lowest score possible and 10 being the highest. The total count (n) and mean score (μ) are labeled as well as the total counts for each score value in increments in 0.5.*

Evaluation of the neighborhood probability fields applied to the HRRRE and HREFv2.1, as well as ensemble agreement scale (EAS) probability fields applied to the HREFv2.1 were also conducted. Qualitative feedback that we received from the experiment participants was that the HREFv2.1 probabilities were preferred over probabilities from the HRRRE because the presentation of the probability field seemed under-dispersive in the HRRRE compared to the HREFv2.1. Gradients in the HRRRE appeared too tightly-packed together which could convey confidence in a solution but in reality was likely related to the lack of diversity of outcomes. The season-long analysis of the probabilistic data determined that the HREFv2.1 neighborhood probabilities performed well near snowfall footprint edges (Prob. ~10%) and for the highest probabilities (Prob. >90%), and certainly better in general than the under-dispersed HRRRE.

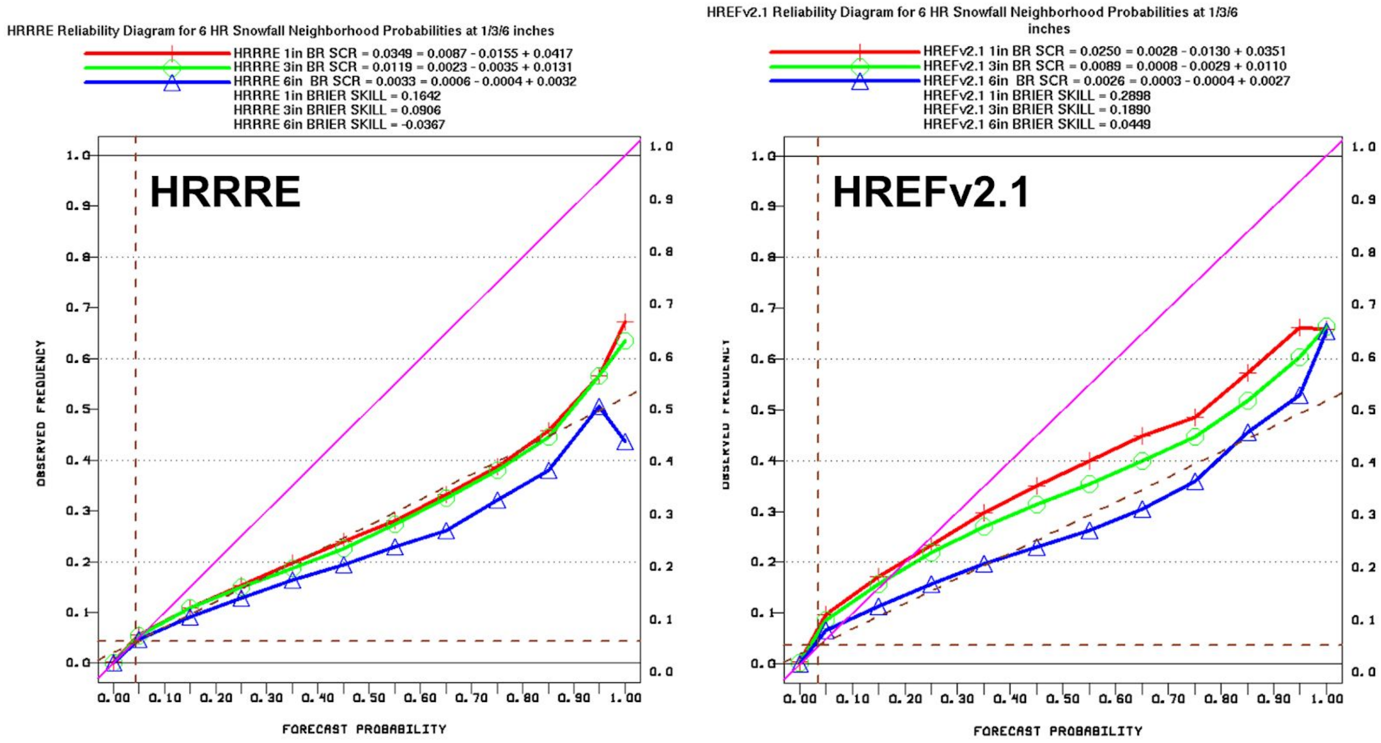

Figure 26. *Reliability diagrams for the HRRRE (left) and HREFv2.1 (right). Colors and markers represent the 6-hour neighborhood probability snowfall threshold used as 1 (red), 3 (green) or 6 (blue) inches.*

Evaluation of the MODE results applied to the CAMs data at the 2 inch threshold shows that the operational HRRR exhibited the most neutral, unbiased forecast. The NBMv3.1 which included CAMs at the Day 1 time-frame was found to have a slight high bias but exhibited a higher threat score. The HREFv2.1 and the WPC Winter Weather Desk forecast both exhibited the largest over-forecasting bias and as such lower threat scores.

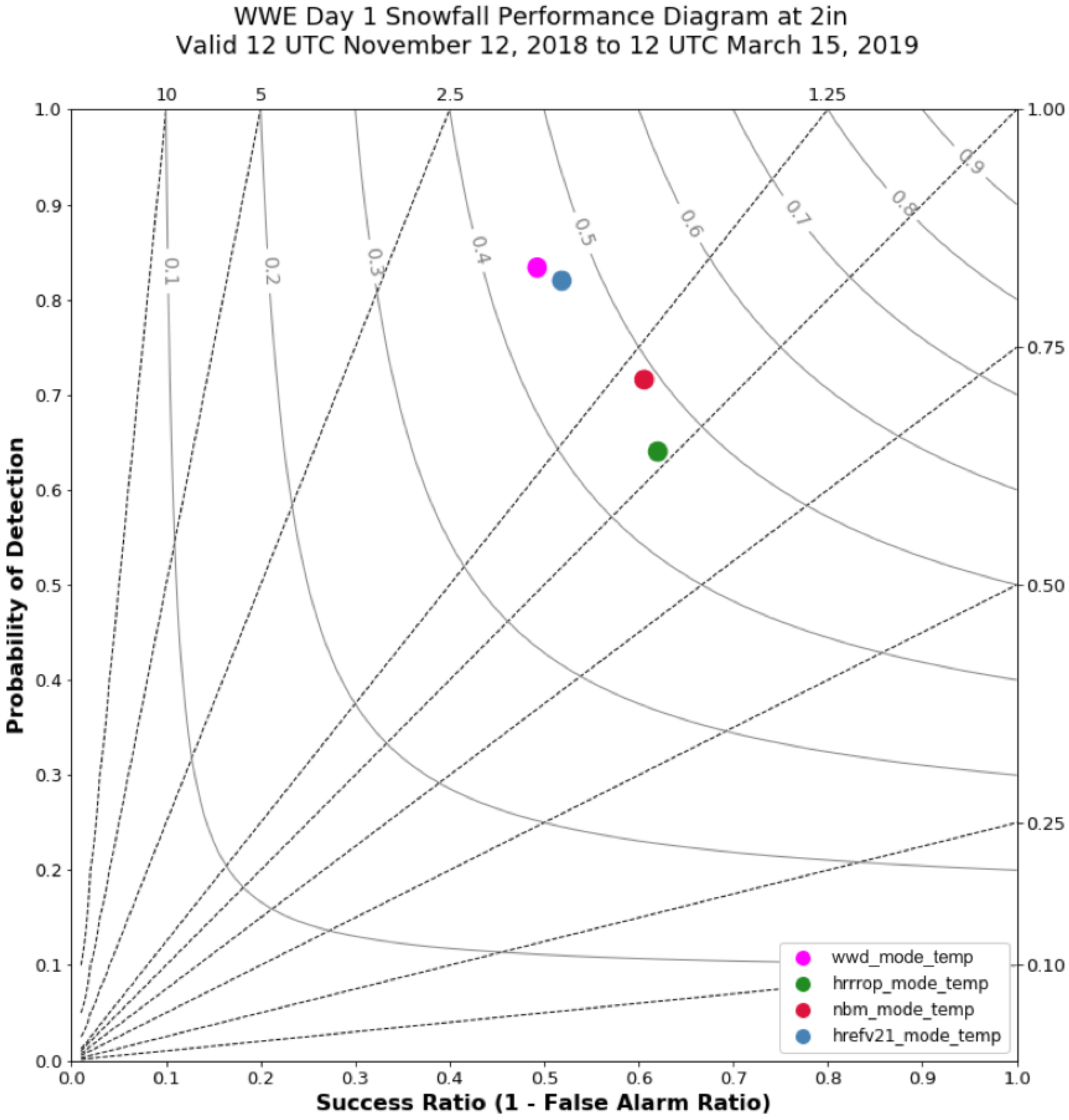

Figure 27. *Performance diagram for season-long statistics of 24-h forecasts of objects >2" of accumulated snowfall with each marker color denoting the model used. The dashed lines in the plot indicate bias and the curved lines indicate CSI/Threat Score.*

Summary & Recommendations

Precipitation Type Methodologies

The individual precipitation type methods were evaluated via snowfall analysis which entailed close examination of both the model solution QPF and the thermal environment. If the model QPF performed well, then differences in snowfall could be attributed to differences in precipitation type methodologies. For a majority of cases, the forecasts were similar and there was little forecast difference among the various precipitation type methodologies.

The FV3 snowfall forecast generated by the Bourgouin precipitation type methodology frequently resulted in an over-forecast of snowfall amounts and forecasts of a much larger footprint than observed in the NOHRSC analysis. It is hypothesized that this is due to the fact that at each hour, more than one precipitation type can be prescribed and the percentage of QPF determined to be frozen for any particular hour is assigned to the snow bucket then in turn summed to 6-hour amounts. For example, if the precipitation types within an hour are determined to be 20% snow and 80% rain, in reality there wouldn’t be any snowfall accumulation but this calculation allows for 20% of the QPF to be added to the total accumulated snowfall with a 10:1 SLR applied. This over-forecasting by the energy layer method was apparent in both the visual and objective verification. **Recommendation: BOUR be refined by applying logic to only accumulate snow under certain conditions and applying a dynamic SLR such as a modified Baxter climatological SLR or Roebber neural network-derived SLR.**

It was noted that in some cases, the environmental profile methods (WPC Decision Tree and NCEP Dominant precipitation type) performed better by producing lower snowfall amounts than the model explicit methods (Fraction of Frozen Precipitation and Change in Snow Depth). These instances may be attributed to over-forecasting of near-surface shallow cold pools by the implicit methods. Conversely, over-forecasting the depth of the warm layer by the profile methods may be responsible for under-forecasting snowfall amounts for some cases.

Comparing the two base models, the FV3 was often the colder solution and the NAM12 the warmer. This may be a result of differences in storm track or the model cold bias identified by the National Weather Service field-wide review of the FV3 model and the GFDL microphysics. It was noted on several occasions that the FV3 Fraction of Frozen Precipitation (POFP) solution forecast snow over sleet too often perhaps as a result of the model cold bias and the tuning of the post-processing of graupel mixing ratio to snow ratio (rime factor proxy). **Recommendation: WPC developers will adjust**

the rime proxy decision thresholds within the post-processing of the POFP algorithm for winter 2019-20.

Many forecasters expressed a preference for explicit precipitation type data over implicit, however both model explicit and post-processed snowfall guidance have utility. A key reason for this preference is that the implicit methods rely on coarse temporal resolution and instantaneous parameters to generate snowfall forecasts. There was a sentiment favoring further development of precipitation type and SLR post-processing which could then trend the preference closer to model explicit methods. **Recommendation: Future work to incorporate more model microphysics variables and tighten the temporal resolution of the implicit calculations are needed.**

Forecast Blending

The blending exercise was well received by WWE participants. In general, the forecast blends provided more accurate forecasts compared with individual model solutions. One recommended change to the forecast exercise was increasing the time period of the forecast from 6 hours to 24 hours because the 6-hour intervals were too restrictive. The 6-hour interval was chosen for the experiment because the shorter interval allowed for close examination of the evolution of precipitation type in snowfall forecasts. **In future WWE blending exercises, WPC can make a 24-hour blending option available in the AWIPS2 environment.**

Blending of clusters was viewed favorably, and it was noted that they provide a good first guess field and forecast footprint. Snowfall amounts, QPF, as well as basic mass fields were made available to forecasters. Snowfall from the clusters can be blended in 6-hour or 24-hour intervals, though 6-hour intervals only were used in the experiment blend exercise.

Several forecasters noted a preference for high resolution models to be included in the blend options. **High resolution model data can be added for Day 1 in future experiments.** It was also suggested that selected weighting percentages used by forecasters be saved over an entire season then compared with the fixed weights of the NBM with the goal of improving post-processing in the NBM. **The selected weights can be made available and shared with MDL.**

Concerns were noted regarding the lack of bias information available for each snowfall solution available in the blend, as well as physical background information on each methodology. **Seasonal biases were computed for the 2018-19 season however WPC will investigate a method to tally situational biases to better aid the forecaster. Training materials can also be developed with the aid of WDTD.**

The utilization of VLAB was suggested for hosting experiment imagery, videos, documents and training materials. VLAB could provide a platform for participants to access the data to use on their own. **WPC will investigate adding datasets as well as training and application materials to VLAB for forecaster use**.

A final concern was noted indicating that the blending exercise was not compatible with forecast builder procedures in the WFO forecast process. This was mentioned because any combination of blending forecast models results in changes to temperatures, winds, sky cover, and PoP as opposed to precipitation type and snowfall only. **WPC recommends more science sharing from the field via collaboration calls or VLAB to better streamline forecast coordination prior to adjustments to NBM output.**

Ensemble Clustering

Ensemble clustering was used during the experiment forecast exercise to ascertain predictability concerns and how those may impact the surface snowfall forecasts. Fuzzy clustering was applied to the 500 hPa geopotential heights of the 90 members of the global ensembles to produce 5 distinct cluster mean solutions. This method improved upon the fuzzy clustering employed during the 2017-18 WWE that produced regional clusters derived from the mean sea level pressure field which was more ambiguous and proved more challenging to interpret during the forecast exercise. The cluster mean snowfall forecasts were assessed along with the other experimental blend inputs during the forecast verification exercises.

Experiment participants provided predominantly positive feedback about the use of ensemble clustering in forecasting winter events, especially in cases with more robust cyclones rather than weak developing shortwaves. The availability of cluster data via a website was determined to be useful to provide an overview of the synoptic variability in the forecast large-scale pattern. The data could be used to define bounds of forecast outcomes as well as filter outlier guidance for certain events. **Recommendation: A common request was for more information and training on using ensemble clustering in the forecast process.**

There were a few suggestions about how to improve the use of ensemble clustering. The global ensemble member data analyzed was 1.0° data for most of the experiment before 0.5° data became available in March. Participants felt that the 6-hour cluster mean snowfall fields from the 1.0° data were too coarse to provide anything but the base footprint of the experimental blend snowfall forecast. In an effort to try to improve the usability of clusters, one idea was to decrease the number of clusters calculated, such as trying 4 clusters instead of 5 because it was often observed that one of the clusters would often have a very small membership (~3-5 members) compared to the others. Another suggestion was that fuzzy clusters could be improved by incorporating a time component to track the evolution of the uncertainty over time. For example, at what forecast hour do the clusters emerge from all ensemble members being clustered on the analysis at forecast hour 00? One would expect 1 cluster at forecast hour 06 could then become 2 clusters by forecast hour 24, then 3, 4, 5 etc. instead of always having to force the data to fit into 5 clusters at all forecast hours. **Recommendation: Development and testing of refined cluster calculation techniques.**

CAMs Data

Deterministic snowfall data were evaluated from the operational HRRR, experimental HRRR (HRRRX), and 1-km HRRRX nest (HRRRX-ICICLE). A randomly selected member of the HRRRE was also evaluated as a deterministic model. Feedback was generally positive about the CAMS snowfall output for a variety of mesoscale snowfall events and lake effect snowfall events. In the HRRR, an issue was raised that an overly simplistic relationship based on temperature alone doesn't accurately represent the impacts of wind or a varying vertical temperature profile on SLR which should affect snowfall output. When comparing the 1-km HRRRX-ICICLE domain to the 3-km HRRRX domain it was found that there was little difference between 24-hour snowfall forecasts. For a few cases where there were larger differences between the two, the 3-km HRRRX snowfall forecasts showed embedded higher amounts than the 1-km HRRRX-ICICLE forecast, suggesting that it could be resolving more dynamical structures such as individual snowband circulations which could be useful to know in the forecast process. The main feedback from the HRRRE member is that the snowfall field appeared smoothed compared to the output from the ensemble mean, so more insight is needed into how the fields are post-processed and made available to the field. **Recommendations: (1) More development is needed within the HRRR microphysical scheme to account for additional physical processes that could affect SLR and snowfall amounts. (2) More evaluation is required to assess the benefits of 1-km over 3-km output for mesoscale snowfall events versus the computational costs.**

CAMS ensemble snowfall data were evaluated from the operational HRRRE and the parallel HREF (HREFv2.1). Ensemble mean snowfall was examined for mesoscale snowfall events and lake effect events and general feedback was related to the predictability more so than the microphysics. For example, in some cases amounts were too high in the HRRRE because a particular feature such as a snowband was resolved but with little variability in timing or position which could convey too high of confidence in the product. Participants had some criticism for the ensemble mean snowfall from the HREFv2.1 because snowfall was derived from the Baxter climatological SLR applied to the snow water equivalent (SWE) field and was misleading in mixed-precipitation events where amounts were over-forecast by >10” in some areas. **Recommendation: Develop an alternative snowfall post-processing method from the HREFv2.1 that performs better during mixed precipitation events rather than applying a climatological SLR to the SWE field.**

Probabilistic snowfall data from the HRRRE and HREFv2.1 were also evaluated in the experiment. Data were presented as either neighborhood probabilities where a static 40 km filter was applied at every point or ensemble agreement scale (EAS) probabilities which had dynamic weighting at each point that depended on the agreement among the ensemble members near that point.

In general, the HREFv2.1 was thought to perform better than the HRRRE in the probabilistic phase space. For many cases, the HRRRE neighborhood probabilities were noted as being under-dispersive and over-confident. For a few events, there was considerable forecaster-to-forecaster variability in preference. For example, some found HRRRE probabilities useful at highlighting expected snowfall locations, especially at higher thresholds while others found that the HRRRE probabilities greatly overestimate the area of snowfall and smooth out critical terrain features or land/water interfaces.

Comparing the two probability methods applied to the HREFv2.1, participants thought that the EAS probabilities seemed to alleviate some of the concerns about terrain and water boundaries compared to the neighborhood probabilities by sharpening the gradients around those static features. However, looking at cases away from terrain or water bodies, the EAS method removed some detail compared to the neighborhood method. The EAS method does not provide useful guidance for higher threshold amounts (i.e., > 6” snow), as probabilities become very low (<5%) across the entire domain; therefore, the EAS method tends to have more utility at lower-end accumulation thresholds while the neighborhood probabilities are better for higher-end accumulations. **Recommendations: (1) More testing is needed of the neighborhood and EAS probabilities applied to diverse cases involving static features (i.e., lakes, mountains) and dynamic features (i.e., snowbands). (2) Testing and training are necessary to understand use cases where probabilistic snowfall forecast information is more appropriate than deterministic in the Day 1 timescale.**

Acknowledgements

The WPC-HMT would like to sincerely thank WPC's Kathy Gilbert, Dave Novak, Jim Nelson and Mark Klein for their commitment and diligence to enabling this experiment to be successfully executed. We thank Chris Bailey for his AWIPS support and WPC DTB staff who assisted with verification. We would also like to recognize Geoff Manikin for fostering deep collaboration with EMC, not only for data and subject expertise but with precise scheduling of EMC participants.

This experiment would also not have been possible without the commitment of our dedicated NWS SOO Advisory Group: David Beachler (WFO Marquette, MI), Greg Mann (WFO Detroit, MI), Eric Lenning (WFO Chicago, IL), Christopher Buonanno (WFO Little Rock, AR), and Steven Zubrick (WFO Sterling, VA).

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, https://doi.org/10.1175/WAF856.1.

Benjamin, S., J. M. Brown, and T. G. Smirnova, 2016: Explicit precipitation-type diagnosis from a model using a mixed-phase bulk cloud-precipitation microphysics parameterizations. *Wea. Forecasting*, **31**, 609–619, https://doi.org/10.1175/WAF-D-15-0136.1.

Bourgouin, P., 2000: A method to determine precipitation types. *Wea. Forecasting*, **15**, 583–592,https://journals.ametsoc.org/doi/full/10.1175/1520-0434%282000%29015%3C0583%3AAMTDPT%3E2.0.CO%3B2.

Brill, K., 2017: Resampling technique used by the forecast verification system (FVS). Personal communication.

Clark, E.P., 2017: Updated NWS Technical Implementation Notice 15-05. Accessed 8 April 2019, https://www.weather.gov/media/notification/tins/tin15-05bigrsc_snowfall_aaa.pdf.

Davis, C.A., B.G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1

Hamill, T., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea Forecasting*, **14**, 155-167.

Novak, D.R., C. Bailey, K.F. Brill, P. Burke, W.A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting*, **29**, 489–504, https://doi.org/10.1175/WAF-D-13-00066.1.

Roebber, P.J., M.R. Butt, S.J. Reinke, and T.J. Grafenauer, 2007: Real-Time forecasting of snowfall using a neural network. *Wea. Forecasting*, **22**, 676–684, https://doi.org/10.1175/WAF1000.1.

WPC-HMT, 2015: The 2015 HMT-WPC winter weather experiment: Final report. Accessed 8 April 2019, https://www.wpc.ncep.noaa.gov/hmt/WWE2015_final_report.pdf.

Zheng, M., E.K.M. Chang, B.A. Colle, Y. Luo, Y, Zhu, 2017: Applying fuzzy clustering to a multi-model ensemble for U.S. east coast winter storms: scenario identification and forecast verification. *Wea Forecasting*, **32**, 881-903, https://doi.org/10.1175/WAF-D-16-0112.1

Appendix A: Featured Guidance and Tools for Experimental Forecasts

Table A1. *Featured Day 1-3 guidance for the 2018-2019 HMT-WPC Winter Weather Experiment.*

Provider	Model	Resolution	Forecast Hours	Notes
EMC	NAM	12 km	60	12 km NAM CONUS *Deterministic Snowfall and precipitation type Guidance using 5 precipitation type algorithms*
EMC	FV3-GFS	13 km	10 days	3D hydrostatic dynamical core; vertically Lagrangian; GFS analyses initialization/physics; Noah LSM *Deterministic Snowfall and precipitation type Guidance using 5 precipitation type algorithms, Model implicit SLR*
ECMWF	ECMWF	27 km	10 days	*Deterministic snowfall and precipitation type guidance using WPC Decision Tree Algorithm*
MDL	NBMv3.1	2.5 km	Hourly out 36 hrs 3-hrly to Day 8 6-hrly Days 8-10	Runs every hour with 28 different deterministic and ensemble systems Weighted Conditional *Deterministic snowfall; Probabilistic precipitation type Guidance*
GLERL / ESRL / GSD	HRRRX w/ FVCOM-Ice	3km and 1 km nest	18 (all hours) 36 (00/06/12/18Z)	Experimental HRRR run with FVCOM-Ice and Heat Flux over Great Lakes (nest over Lake Michigan) *Deterministic Snowfall for CONUS (3 km) and Great Lakes domain (1 km)*
ESRL/ GSD	HRRR Ensemble (HRRRE)	3 km	18 hours at 12Z, 15Z, 18Z, 21Z 36 hours at 00Z	9 HRRR members, Sub-CONUS domain, stochastic *Deterministic and Probabilistic Snowfall and precipitation type Guidance*
EMC	HREF-p (HREFv2.1)	3 km	36 hours	Hydrostatic Multiscale Ensemble with 10 members including HRRR which produces probabilistic winter output *Probabilistic Ptype Guidance*

FV3-GFS

The FV3-GFS is expected to replace the GFS in June 2019. The development code in the model was initially frozen in September 2018, and it's this version of FV3-GFS that was evaluated in WWE 2019. The model physics, dynamical core, and stochastic package are the same as the GFS, however The Zhao-Carr microphysics scheme is replaced with GFDL microphysics which was a key motivation for testing winter weather forecasts derived from the FV3-GFS. Specifically, the ratio of graupel ratio to snow ratio in the lowest model level was tested as a proxy for sleet forecasting and SLR adjustment, and used similarly to the rime factor parameter from Ferrier-Aligo microphysics package in the NAM. The Noah Land Surface Model (LSM; Ek et al. 2003) is used to calculate snow depth.

During the course of the WWE, a couple different bugs were discovered in this frozen FV3 version related to the NOAH LSM snowfall accumulation and the radiation scheme. The version of the FV3 tested in the WWE contained these bugs but was found to not impact experimental results for the following reasons: (1) The experiment used data only through Day 3 (FH 84) so did not see the extreme effects of the cold bias related to the radiation driver that became more severe by Day 5, (2) The snowfall grids were generated mostly off of the QPF + WPC post-processing + fixed SLR.

National Blend of Models Version 3.1 (NBMv3.1)

The NBMv3.1 runs every hour with 28 different deterministic and ensemble systems. For the CONUS, typically 4 to 6 new model runs update each hour, with up to 7 or 8 (~50% new) every four cycles. The NBMv3.1 uses TOD (Time of Day) concept, rather than the "model" cycle. Therefore, a 12Z run of NBM v3.1 does not contain a single 12Z model run. It is a data cutoff time (newest models are from 10-11Z in this example; several 00Z to 06Z models included). NBM v3.1 is run at the top of each hour and available 50-60 minutes later. Models included are found in Table A2 and discussed in SCN 18-78.

Table A2. *Data Dependencies for the NBMv3.1.*

Global Models	Mesoscale Models
ECMWF deterministic - .25-degree	HRRR - 3 km
ECMWF ensemble - 1-degree	NAM Low-Res - 12 km
GFS - 0.117 degree	NAM High-Res - 3 km
GEFS mean + members - 0.5 degree	HIRESW (NMMB and ARW cores) - 3 km
CMC deterministic - 0.25 degree	Canadian regional deterministic model - 10km
CMC ensemble mean + members - 0.5 degree	Canadian regional ensemble - 15km (for precipitation products only)
NAVGEM deterministic - .50-degree	RAP - 12 km
FNMOC - 1 degree	SREF - 40 km
GMOS - 2.5 km	GLMP - 2.5 km

EKDMOS - 2.5 km	**URMA (CONUS, AK, HI, PR) - 2.5 km**
CCPA (used in NBM Precip SQM) - ~ 13 km	
NBM 2.5 Precipitation Stochastic Quantile Mapping	

EMC/ESRL Experimental Deterministic High-Resolution Rapid Refresh (HRRRX) and Ensemble HRRRE

The Experimental HRRR (HRRRX) in the WWE Experiment is a WRF-ARW-based 3 km model initialized with the latest 3-D radar reflectivity using a digital filter initialization (radar-DFI) technique (via the parent 13 km RAP) and is updated hourly. HRRRX contains numerous model changes including an update to WRF-ARW version 3.9.

The HRRRE is an experimental convective-allowing ensemble analysis and forecasting system run at NOAA/ESRL/GSD. It is being developed and tested for three main reasons: (1) improving 0-12 h high-resolution forecasts through ensemble-based, multi-scale data assimilation, (2) testing ensemble-design concepts for 0-48 h forecasts produced with a single model, and (3) providing a foundation for experimental, on-demand, very-high-resolution applications such as Warn-on-Forecast.

- **Specific Information (information in red was altered during the season)**
  - **Model**
    - WRF-ARW version 3.8+, combining elements of versions 3.8 and 3.9 plus other GSD-specific features
    - Configuration identical to experimental HRRR (https://rapidrefresh.noaa.gov/hrrr/), except that domain covers central and eastern US only (55% of HRRR domain), and a standard vertical coordinate is used instead of a hybrid coordinate
  - **Data-Assimilation Ensemble**
    - 36 members
      - 3-km horizontal grid spacing
      - Initial ensemble-mean atmospheric state from RAP analysis
      - Atmospheric spatial perturbations from members 1-36 of GDAS ensemble
      - Initial ensemble mean of land-surface state from HRRR
      - Random soil-moisture perturbations added to each member at initial time
      - Random perturbations to MU, U, V, T, and QVAPOR added to boundary conditions of each member
    - Initialization at 0300 and 1500 UTC, followed by hourly cycling for 9 h to 1200 and 0000 UTC, respectively

- **Hourly Data Assimilation**
  - Observations
    - NCEP bufr conventional observations, as in HRRR (http://www.emc.ncep.noaa.gov/mmb/data_processing/data_processing/)
    - MRMS gridded radar reflectivity observations, thinned in horizontal and vertical directions
  - Gridpoint Statistical Interpolation (GSI) for observation preprocessing and calculation of ensemble priors
  - DART ensemble adjustment Kalman filter (EAKF) for assimilation
    - Analysis variables: U, V, T, QVAPOR, PH, MU, QCLOUD, QRAIN, QICE, QSNOW, QGRAUP
    - Gaspari-Cohn compact pseudo-gaussian for localization
    - Horizontal localization radius (full radius, where weight reaches zero) 300 km and 18 km for conventional and radar observations, respectively
    - Vertical localization radius (full radius, where weight reaches zero) 8 km and 6 km for conventional and radar observations, respectively
    - GSI adjustments applied to each member individually after EAKF
      - Soil adjustment, as in HRRR
      - Cloud clearing based on satellite observations, as in HRRR
  - Relaxation to prior spread (inflation factor 1.2) after assimilation each hour
- **Ensemble Forecasts**
  - 9-member, 48-h forecast initialized from first 9 members of data-assimilation ensemble at 1200 and 0000 UTC
  - 3-km horizontal grid spacing
  - Random perturbations to MU, U, V, T, and QVAPOR added to boundary conditions of each member
  - Post processing: An ensemble post-processing system is applied to the nine HRRRE forecast members to produce all-season weather hazard probabilities including heavy rainfall as is done with the time-lagged HRRR. For the 2017-18 Winter Weather Experiment, HRRR-E probabilities are the fraction of members that exceed a given threshold, or predict a given precipitation type, at a point. The final probability field (100*(n/total)) is smoothed using a Gaussian filter of width 25 km.

HRRR/HRRRE Precipitation Type Algorithm

HRRR and HRRRE both use a microphysics-based algorithm to predict instantaneous precipitation type (snow, sleet, rain and/or freezing rain). Output is in the form of categorical (0 or 1) grids for each precipitation type, stored in GRIB2 fields CSNOW, CICEP, CRAIN, and CFRZR. Multiple precipitation types may be forecast simultaneously, up to 3, with the only exception being that the algorithm will predict either rain OR freezing rain. The algorithm uses 2-m temperature, 3-D hydrometeor mixing ratios and fall rates to provide a first guess of precipitation type reaching the ground. A flow-chart of the algorithm is shown below in Figure A1, with further details available in Benjamin et al. (2016).

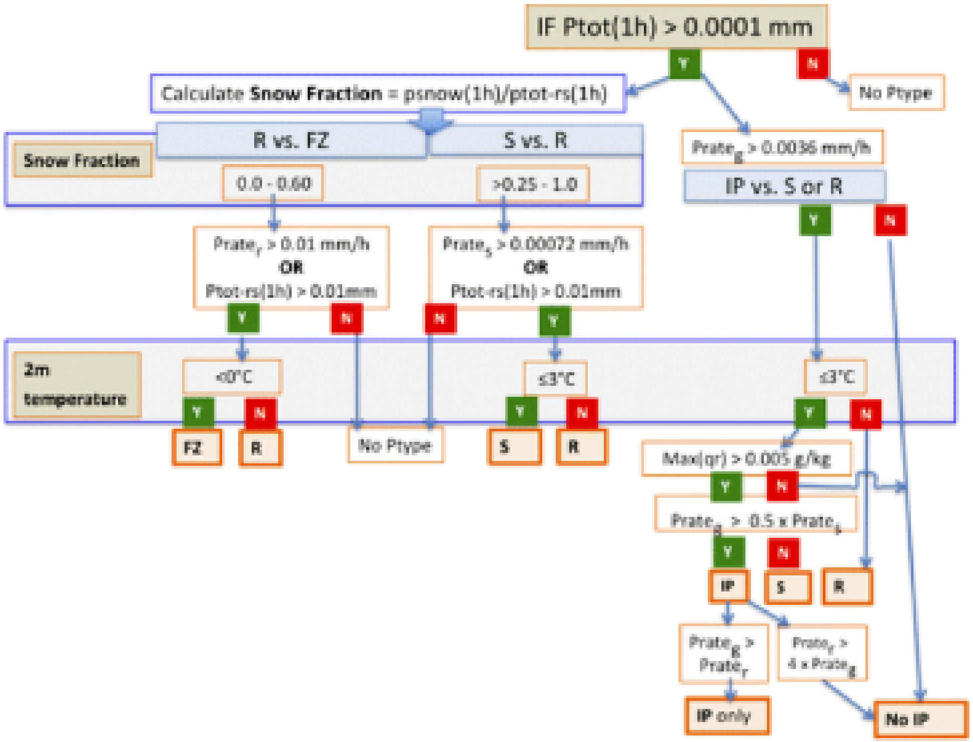

Figure A1. *Flowchart showing the HRRR/HRRRE microphysics-based algorithm to predict instantaneous precipitation type.*

HRRR/HRRR-E Variable-Density Snowfall Algorithm

HRRR and HRRRE both use a variable-density snowfall accumulation algorithm. The output of this algorithm, contained in the “ASNOW” GRIB2 field, is the total depth of new snowfall and graupel/sleet accumulation. Note: in the Thompson microphysics scheme and this algorithm, falling sleet is considered graupel.

At every model time step, the algorithm first determines snow-water-equivalent and graupel-water-equivalent precipitation rates using the mixing ratios and fall rates of these hydrometeors at the lowest model level. Next, some melted water equivalent is subtracted from these totals if the land surface temperature is above 0°C. Snow and graupel density functions are then applied to determine the rate of increase of total snow+graupel depth. The product of this accumulation rate (m/s) and the model time step (20 s) yields a snowfall+graupel depth accumulation that is added to a running total depth. At every hour, HRRR and HRRRE output this running total as ASNOW.

Snow-to-liquid and graupel-to-liquid ratios are linear functions of the temperature at the lowest model level (typically ~8 m), given by the following equations:

Eq. 1. *snow density (kg/m^3) = min(250, 1000/max(4.179, (13.*tanh((274.15- (x+ 273.15))/3))))*

Eq. 2. *graupel density (kg/m^3) = min(50., 1000/max(2,(3.5*tanh((274.15 -(x+273.15))/3))))*

The minimum and maximum densities possible from the snow equation are 76 and 250 kg/m^3, respectively, equal to snow-to-liquid ratios of 13:1 and 4:1.

GLERL Lake-effect Snowfall Forecasts

The Great Lakes Environmental Research Laboratory (GLERL) have developed a coupled lake hydrodynamic-ice system that can provide short term forecasts of lake surface temperatures, lake ice extent, evaporation rates, and heat flux from the Finite Volume Coastal Ocean Model (FVCOM-Ice). This data was used within the 3-km HRRRX and an experimental 1-km HRRRX nest (Fig. 7) to test whether lake effect bands are more realistically simulated given their sensitivity to lake surface parameters. FVCOM-Ice was run at GLERL twice daily at 00 UTC and 12 UTC for use in the HRRRX. The 1-km HRRRX nest was run twice a day at 03 UTC and 15 UTC and is run out for 24 h.

Parallel HREF (HREFv2.1)

A parallel version of HREF (HREFv2.1) was available for the winter season. This 3-km 10-member ensemble has membership from the two most recent runs of the following models: HRW-ARW, HRW-NMMB, HRW-ARW2, NAM nest, & HRRR.

Appendix B: WPC MODE Settings for Objective Verification

MODE Configuration for Experiment Cases

MODE was used to objectively analyze forecast objects from all 10 experimental snowfall precipitation type inputs, the 5 cluster inputs, the experiment blend, and the NBM v3.1 for each of the 19 6-hour cases. All data were interpolated onto a common 0.1° x 0.1° grid. Objects were identified based on the criteria in Table B1.

Table B1. *Metrics used in MODE to identify snowfall forecast and observed object pairs.*

	Forecast	**NOHRSCv2**
Threshold	1, 2, & 4 inches of 6-hour snowfall	1, 2, & 4 inches of 6-hour snowfall
Convolution Radius	2 grid squares	2 grid squares
Area threshold	≥ 10 grid squares	≥ 5 grid squares

Forecast objects were paired with observation objects using the Intensity Formula, *T*. *T* was proportional to the distance of the centroids, the distance from the boundary object edges, and the ratio of the area of each object. *T* did not consider any differences in orientation angle. A value of *T* ≥ 0.6 was required in order for a forecast object to be paired with an observation object.

MODE Configuration for Season-long CAMs Analysis

WPC MODE Settings for Objective Verification

- 36 hour model snowfall accumulation forecasts verified against 24 hour NOHRSCv2 snowfall accumulations
- 00Z cycles valid from 12Z to 12Z used
- Both snowfall accumulation forecasts and NOHRSCv2 snowfall accumulations re-gridded to a common 5km lat/lon grid
- Thresholds investigated varied.

MODE

- Grid stats harvested from daily MODE CTS. Daily MODE CTS were aggregated over the whole season and statistics calculated from the aggregated stats.
- Circular convolution radius of 3 grid squares used
- Double thresholding technique applied