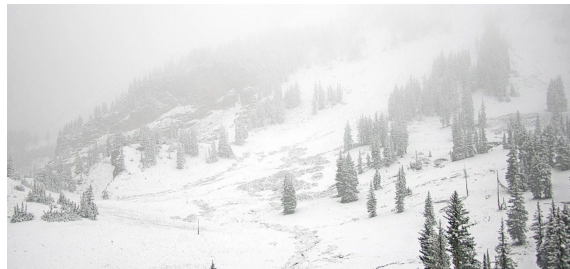




The 10th Annual WPC-HMT Winter Weather Experiment



12 November 2019 – 13 March 2020

Weather Prediction Center

College Park, MD

Findings and Results

Michael Bodner - NOAA/NWS/WPC, College Park, MD

Kirstin Harnos - CIRES: Cooperative Institute for Research in Environmental Sciences,
NOAA/NWS/WPC, College Park, MD

Benjamin Albright - Systems Research Group, NOAA/NWS/WPC, College Park, MD

Bill Lamberson - CIRES: Cooperative Institute for Research in Environmental Sciences,
NOAA/NWS/WPC, College Park, MD

James Correia Jr. - CIRES: Cooperative Institute for Research in Environmental Sciences,
NOAA/NWS/WPC, College Park, MD

Diana Stovern - CIRES: Cooperative Institute for Research in Environmental Sciences,
NOAA/NWS/WPC, College Park, MD

Updated: 5 June 2020

Table of Contents

Abstract	<u>2</u>
Transition Recommendations	<u>3</u>
Introduction and Background	<u>4</u>
<i>Science and Operations Objectives</i>	<u>5</u>
Data and Methods	<u>6</u>
<i>Data overview</i>	<u>6</u>
<i>Experiment Logistics and Participation</i>	<u>10</u>
<i>Forecast Exercise: Blending Methodology</i>	<u>11</u>
<i>Verification Exercise: Subjective and Objective</i>	<u>12</u>
Results	<u>13</u>
<i>Cases</i>	<u>13</u>
<i>Subjective Verification</i>	<u>16</u>
<i>WWE Experiment Blend, WPC Official, and NBMv3.2</i>	<u>16</u>
<i>Precipitation Type Verification: GFS and NAM12</i>	<u>19</u>
<i>Precipitation Type Verification: ECMWF and CMC Decision Tree</i>	<u>25</u>
<i>CAMs Verification</i>	<u>27</u>
<i>Western US Downscaling</i>	<u>29</u>
<i>Objective Verification</i>	<u>31</u>
<i>Model Explicit Snowfall</i>	<u>31</u>
<i>Performance Diagram</i>	<u>32</u>
Summary & Recommendations	<u>34</u>
Acknowledgements	<u>37</u>
References	<u>38</u>
Appendix A: WPC MODE Settings for Objective Verification	<u>40</u>

Abstract

The Hydrometeorology Testbed at the Weather Prediction Center (HMT-WPC) conducted the 10th Annual Winter Weather Experiment (WWE) from 12 November 2019 through 13 March 2020. The experiment brought together members of the operational forecasting, research, and academic communities to address winter weather forecast challenges. The 10th Annual WWE focused on the following science goals: (1) Explore the use of ensemble clustering to assess forecast sensitivity on snowfall footprints and identify precipitation type (p-type) challenges in the forecast process. (2) Utilize clusters to test and evaluate blending of snowfall solutions relative to current operational blending of ensemble means and precipitation type fields. (3) Explore downscaling methodologies in the NAM and GFS over western United States for the improvement of forecasts for higher terrain snowfall and snow liquid ratio (SLR). (4) Explore the utility of Convection Allowing Models (CAMs) in the Day 2 and Day 3 time period.

During forecast sessions, experiment participants collaborated to create a 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 snowfall forecasts were created throughout the entire season, the majority of them for Day 2. These blended snowfall forecasts as well as the individual experimental tools 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.

Transition Recommendations

Table 1. Transition metrics for selected experiment data sources 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					
Ensemble clustering		X	X		WPC
Precipitation Type Methods					
WPC Ensemble (EPTypE)		X			WPC
Change in Snow Depth (SNDP)	X¹	X²			WPC
Forecast Blending					
Manual blending of snowfall			X		WPC
CAMs Snowfall Guidance					
FV3-SAR			X		EMC
SSEF Ensemble Mean			X		CAPS
Western US Downscaling					
Downscaling Technique		X³	X⁴		UUtah
Totals	1	4	5	0	

For more detailed recommendations please see the [Summary and Recommendations](#) Section of this report.

¹ SNDP is already operationally available in AWIPS as part of the model post processing

² SNDP recommendation is to add to the WPC PWPF product

³ Transfer of code to the WPC WWD for the forecast process over the Western US

⁴ Continue development of technique, apply to CONUS, and provide comparison to day 2 and 3 CAM simulations

Introduction and Background

This year marked a milestone with the successful completion of the 10th annual Winter Weather Experiment (WWE) at the Hydrometeorology Testbed in the Weather Prediction Center (HMT-WPC). WWE serves the greater winter weather community as an important component in the research-to-operations (R2O) process. The experiment provided participants with the unique opportunity to interact with members of the operational forecasting, research, and academic communities to address winter weather forecast challenges in a mock-operational environment.

Participants in the 10th annual WWE tested and evaluated modeled snowfall solutions in real-time. This year there was an emphasis on the forecast process, mainly in the day 2 time frame which included the creation of a participant-generated snowfall forecast blend. As part of this blending forecast process, multiple techniques to achieve snowfall solutions were highlighted. These included cluster analysis of the global ensemble systems, post-processed microphysics techniques, and convective allowing model (CAM) solutions. Details on each of the tools used in this year's experiment can be found in the [10th Annual WWE science plan](#).

For the third year in a row, the 10th Annual WWE was conducted remotely utilizing the web-based distance communication software, GotoMeeting, paired with a teleconference to encourage and promote interactivity and engagement with the participants. Remote interactions began 12 November 2019 and ran until 13 March 2020. The team also hosted two in-house weeks at the National Center for Weather and Climate Prediction (NCWCP) in College Park, MD. The first week was 10-14 February 2020 and the second 2-6 March 2020. These weeks were instrumental in providing an environment for more in-depth discussion on the experimental tools as well as planning for future WWE.



Figure 1. Left: Dr. Louis Uccellini welcoming participants to the first in-house week (photo courtesy of Dr. David Novak). Right: Participants working through a verification activity during the second in-house week. (photo courtesy of Dr. James Correia, Jr.).

Science and Operations Objectives

The main science objectives of the 10th Annual Winter Weather Experiment were to:

- Explore the use of ensemble clustering to assess forecast sensitivity on snowfall footprints and identify precipitation type (p-type) challenges in the forecast process.
- Utilize clusters to test and evaluate blending of snowfall solutions relative to current operational blending of ensemble means and precipitation type fields.
- Explore downscaling methodologies in the NAM and GFS over western United States for the improvement of forecasts for higher terrain snowfall and snow liquid ratio (SLR).
- Explore the utility of CAMs in the Day 2 and Day 3 time period using the Stand-Alone Regional (SAR) configuration of the FV3.
- Enhance collaboration among NCEP centers, WFOs, and academic partners on improving spatial and quantitative snowfall forecasting.
- Use both event and season long verification to assess the utility and performance of experimental data sets.

Additionally, the main operations objective for this year's WWE was to run the forecast blending exercises in graphical forecast editor (GFE) on WPC-HMT's AWIPS2 system.

Data and Methods

Data Overview

In addition to the full multi-center suite of deterministic and ensemble guidance available to WPC forecasters, participants were asked to consider several different experimental datasets while preparing the snowfall forecasts.

Ensemble Clustering

Ensemble forecasts contain a wealth of information about the range of possible outcomes for a given weather event. More information is good until it becomes too much information for a forecaster to sift through in a timely manner. This is where fuzzy clustering comes in. It can extract a signal from the noise by picking out the few dominant forecast scenarios being predicted by the ensembles. The ensemble clusters are calculated from 90 members of the GEFS, ECENS, and CMCE as follows:

- The dominant patterns in the ensemble forecast are derived by calculating the first two Empirical Orthogonal Functions (EOFs) of the 500-hPa height field over the Continental United States (CONUS).
- In general, an EOF of 500-hPa heights will take on one of two patterns (or a combination of the two):
 1. A monopole centered on the ensemble mean location of a 500-hPa feature, such as a trough. This indicates uncertainty in the amplitude of the 500-hPa feature among the ensemble members.
 2. A dipole centered between the ensemble mean location of two 500-hPa features (e.g., a trough-ridge couplet). This indicates uncertainty in the location of these 500-hPa features among the ensemble members.
- A phase space of forecast scenarios is constructed from the two EOFs by using the principal components (PCs) of EOF1 as x-axis coordinates and the PCs of EOF2 as the y-axis coordinates. Each ensemble member is plotted on the phase space diagram using its principal components for EOF1 and EOF2.
- The ensemble members are then clustered (using the k-means algorithm) into 4 clusters based on where they fall in the phase space.

Precipitation Type Methodology

Multiple algorithmic methods of forecasting precipitation type were applied to the quantitative precipitation forecasts (QPF) of NCEP's NAM12 and GFS models to generate a snowfall forecast for the respective deterministic model. These methods provided an instantaneous estimate of which precipitation type species can be expected for a given grid point. A snow-to-liquid ratio (SLR) was then applied to the resultant frozen precipitation to derive a snowfall forecast.

The first precipitation type method applied to the forecast process was the WPC decision tree algorithm. This algorithm features a straightforward logic check of critical temperatures at 700, 850, 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 time stamp it is applied to.

The second method, which is 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 (Manikin 2005) to diagnose precipitation type at each grid point. Each of the methodologies that make up the NCEP dominant scheme evaluate the freezing level of the wet bulb temperature to assess hydrometeor changes. This multi-algorithmic method generates multiple precipitation types 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.

The third method assigns a precipitation type for each grid point using the percent of frozen precipitation (POFP) and rime factor parameterization from the model microphysics. The POFP parameter in both the NAM12 and GFS estimates the percentage of model QPF to be frozen in the lowest level of the model at a particular forecast time. Thus the fraction of frozen precipitation field in this lowest level of the model 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 rime factor parameter in the NAM12 is generated by the Ferrier-Aligo microphysics scheme. This is a variable density graupel parameter which estimates the ratio of the growth of snow by the liquid accretion plus vapor deposition divided by vapor deposition relationship. Thus the higher the liquid accretion, the higher the resultant rime factor. Higher rime factors present in the lowest sigma layer of the model are used to assess the potential for sleet at a given grid point due to the increased graupel (HMT-WPC 2015).

An exclusive rime factor parameter is not available through the GFDL microphysics scheme in the GFS, however a “rime factor proxy” can be derived using the snow and graupel mixing ratios in the lowest level of the model.

$$\text{RF} = \text{Graupel Mixing Ratio} / 1 + \text{Graupel Mixing ratio} + \text{Snow Mixing Ratio}$$

Exploration of the rime factor proxy the past two winters suggests that values greater than 0.7 align with increased presence of graupel and a higher likelihood of sleet as the instantaneous precipitation type (HMT-WPC 2019).

Determination of precipitation type using microphysics parameters:

	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

NAM12 and GFS Bucket Ensemble

For both the NAM12 and GFS, an ensemble averaging of all three precipitation type methodologies was computed every six hours to arrive at a weighted percentage of each precipitation species for the respective 6-hour period. The respective model QPF was then sequestered into each precipitation type bucket based on the weighted percentages. For example if the weighted percentage of snow was 70%, then 70% of the 6-hour QPF was partitioned as snow, then the remaining 30 % consisted of the other precipitation species. An SLR was applied to the QPF to derive snowfall amounts. For sleet, a 2:1 SLR was used.

In the NAM12, the Roeber SLR was used, then the Ferrier-Aligo rime factor parameter was applied to adjust the SLR in precipitation type transition areas. For the GFS, the Baxter climatological SLR was used, then adjusted using the rime factor proxy derived from the GFDL microphysics over precipitation type transition areas. Snowfall forecasts from the NAM12 and GFS precipitation type ensemble were made available in the AWIPS forecast experiment blender.

Individual snowfall forecasts using each of the aforementioned precipitation type methodologies, the respective model QPF, and 10:1 SLR were computed. The three snowfall solutions for each model were averaged to make a GFS and NAM12 average forecast. These average forecasts were not made available in the forecast blending exercise, but were evaluated in both the visual and objective seasonal verification.

Change of Snow Depth (SNDP)

A change in snow depth field was made available from the NAM12 and GFS. Both models use the Noah Land Surface scheme to generate a snowfall forecast based on estimated ice density and melting processes. Both solutions were made available in the AWIPS forecast blender, and evaluated over the course of the winter season.

A change in snow depth was also available in the FV3SAR, but was not available in the AWIPS blender. The model explicit snowfall from the FV3 was verified over the 2019-20 winter season.

ECMWF and CMC Decision Tree

Post processed snowfall forecasts were generated using the QPF from the ECMWF and CMC, the WPC decision tree method to assign a precipitation species, and the Baxter climatological SLR. The ECMWF forecasts were on a quarter degree grid, while the CMC a half degree grid. To increase sharpness and forecast resolution to be more in line with the quarter degree ECMWF and GFS, the QPF from the CMC was downscaled to 5km using PRISM QPF climatology. The downscaling was applied over the CONUS starting December 23, 2019 and the resultant 5km QPF was used to generate the CMC snowfall forecasts.

Convective Allowing Models (CAMs)

The WWE was also provided two Stand Alone Regional versions of the FV3 (FV3-SAR). A beta version of the 3km FV3-SAR was run out to 60-hours and made available by EMC for the experiment. This data set allowed the WWE to assess the skill of Day 2 snowfall forecasting using a CAM solution. A post processed snowfall was computed using the snow water equivalent field in the model, and applying a 5km version of the Baxter climatological SLR. This version of the FV3-SAR featured the GFDL microphysics therefore the rime factor proxy, derived from the graupel and snow mixing ratios, was used to adjust the SLR over precipitation type transition areas. The FV3SAR was available in the AWIPS forecast blender.

The second CAM was a FV3-SAR ensemble from CAPS. The five member ensemble was run out to 84 hours with the simple ensemble mean provided to the WWE for verification activities. Membership included differing the Microphysics and LSM schemes within the FV3-SAR. The microphysics schemes included the GFDL, NSSL, and Thompson. LSM schemes included the NOAH and EDMF. Individual members as well as three ensemble means were available for real-time guidance via the [CAPS HMT Winter Products webpage](#).

Western United States Downscaling

This year's WWE began exploring the utility of downscaling techniques over the Western US for improved snowfall forecasting. The University of Utah provided the snowfall forecasts over the Western US for the verification activities with additional guidance available in real-time from [their website](#). The technique used during the experiment involves three major steps: (1) Identification of the transition zone and precipitation type, (2) Downscaling of precipitation to an 800-m grid based on high-resolution precipitation-altitude relationships derived from monthly PRISM analyses (see Lewis et al. 2017), and (3) Applying snow-to-liquid ratio algorithms based on historical relationships between snowfall density and atmospheric predictors such as wind and temperature (Alcott and Steenburgh 2010).

More complete details on each of these datasets can be found in the [10th Annual WWE science plan](#). Additional information on the ensemble clustering can be found in the [10th Annual WWE cluster documentation](#).

Experiment Logistics and Participation

The experiment was conducted weekly over the full winter season beginning Tuesday, November 12, 2019 and ending Friday, March 13th, 2020. This year marked the third year of remote WWE interactions with the team running the experiment out of 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 Tuesdays for the forecasting exercise (10:30 am-12:00 pm EST) and on Wednesdays for verification (10:30 am-12:00 pm EST) throughout the season. Overall, there were 11 weeks of remote WWE with specific participation in forecast sessions from NCEP centers, WFOs, research labs, and universities shown below.

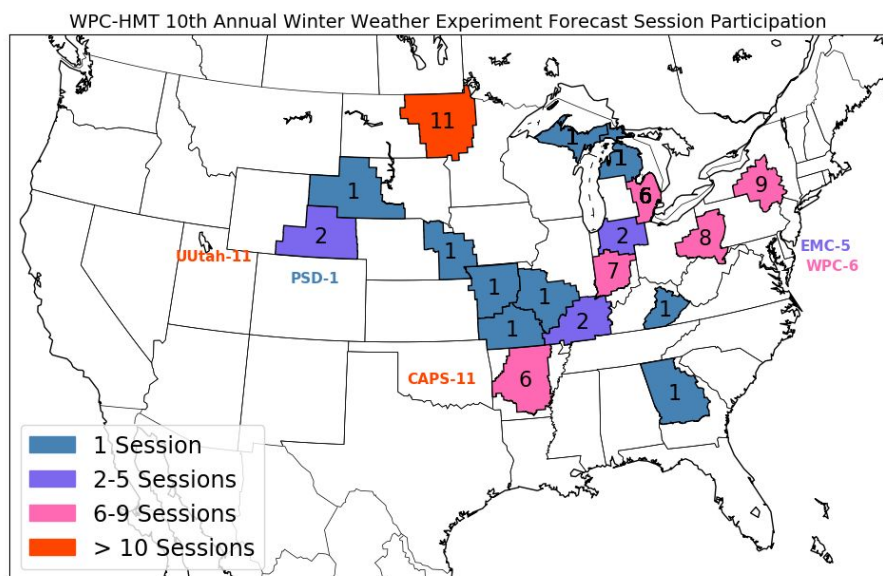


Figure 2. WWE participation in forecasting sessions by WFO (shaded) or other National Center, University Partner, or location (text) colored according to participation frequency.

In addition to the remote experiment, there were two weeks of residence experiments at NCWCP. The first residence week was 10-14 February 2020, and the second was 2-6 March 2020. During the weeks, each day was structured to allow for maximum discussion and interactions amongst the participants. The mornings began with a verification and forecast activity. These daily sessions allowed for multiple in-depth discussions on the experimental tools and datasets. While the winter weather pattern was not cooperative for forecasting live events, the groups were able to explore past events to create forecast blends.

Afternoons were structured around a ‘topic of the day’ with 2 invited presentations followed by discussion sessions. Topics included: CAM computing issues, non-snow winter precipitation, impact-based product development, downscaling techniques, and the NWS winter forecasting process. Some direct impacts of these presentations and discussion sessions included the addition of snow squalls to the [mPING](#) database, the beginning steps for verification of impact-based products, and suggestions for future WWEs. An archive of the invited presentations can be found within the [experiment website](#).

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 24-hour period and region of interest was selected based on snowfall potential. 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.

Blending of the snowfall forecast guidance was performed with a GFE procedure called the Scalar/Vector Grid Manipulation procedure (ScaVec) (Fig. 3). ScaVec was developed at WFO Green Bay and modified for operational and testbed purposes at WPC. ScaVec is an interactive blending tool that allows blending of forecast guidance by individual model weight. Slider bars associated with each model can be adjusted to add more or less weight, and as the weight is adjusted, the weighting associated with all other models dynamically update to reflect their contribution to the blend. The tool is flexible so that different blends can be applied to grids over varying time ranges and edit areas. In this experiment, forecast blends of snow amount were applied over the entire continental United States, varying between 6- and 24-hour periods. When different blends were applied to individual 6-hour grids within a 24-hr period, a summation tool within ScaVec was used to create the final 24-hour, 12Z-12Z grid used for verification.

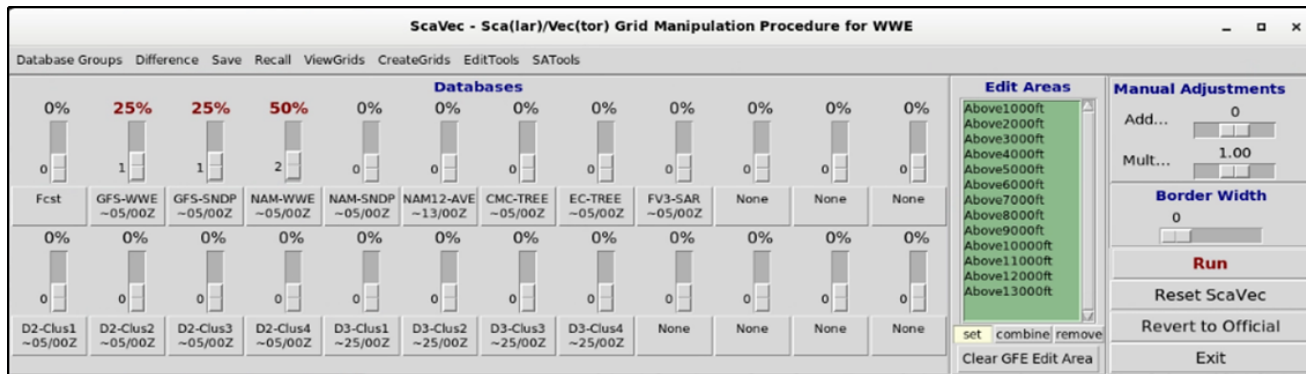


Figure 3. The Scalar/Vector Grid manipulation procedure used for the WWE. The buttons shown below each slider indicate the guidance used in the forecast blends.

Verification Exercise

Subjective

Experiment data were subjectively validated against 24-hour snowfall accumulation amounts available from the National Operational Hydrologic Remote Sensing Center version 2 (NOHRSCv2) dataset (Clark, 2017). Collection of the subjective verification scores and comments was completed through the use of Google surveys and Google slides. An example of the survey and corresponding slide is shown in the figure below. Participants were asked to visually provide a score between one and ten, where one is considered a low score and ten a perfect score, followed by written comments on the snowfall footprint, position, and magnitude. Verbal discussion was highly encouraged during the Wednesday verification sessions; however, the current remote verification format and lack of interesting cases made participation low for most weeks.

10th Annual WWE Forecast Evaluation

* Required

1. Experimental Blend Snowfall Forecasts

Using 6/24-hr NOHRSC, please evaluate the overall quality of the 6/24-hr forecast of accumulated snowfall.

Use the text box to enter a numerical rating from 1-10 where 0=N/A, 1=Very Poor, 10=Very Good (decimals allowed)

a. WWE Snowfall Forecast: Please score from 1 through 10: *

Your answer

b. Please provide commentary of forecast snowfall footprint, position, and magnitude of any banding

Your answer

Back
Next

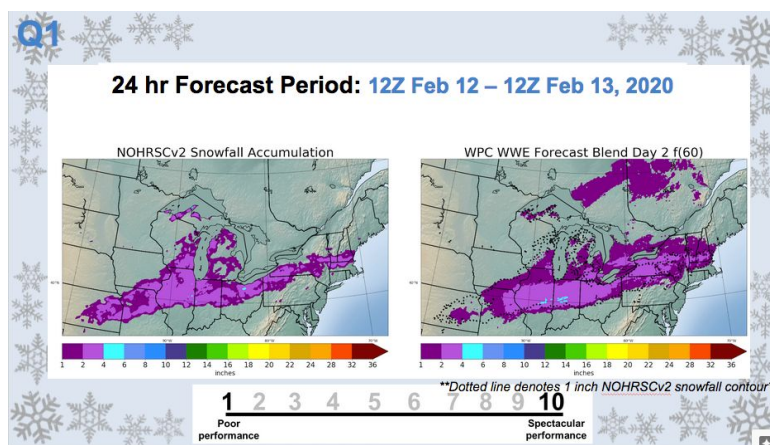


Figure 4. Sample of google survey and presentation used for visual subjective verification.

After each verification session, the subjective scores were plotted and displayed on the [10th Annual WWE Subjective Verification website](#). This provided the data partners and WWE participants a chance to see how the experimental tools were scoring in near real time. Scores over all the WWE verification sessions as well as word clouds highlighting the written comments for each tool will be presented in the results section later in this document.

Objective

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 Day 2 (48/60 hour) forecasts of 24-hour snowfall were thresholded by 1, 2, 4, 6, 8, and 12 inch amounts and forecast objects were matched to observed objects. Statistics derived from the MODE output assess similarities between objects such as spatial coverage, orientation, distance, etc. The details of the MODE configuration used for this analysis can be found in [Appendix A](#).

MODE also outputs contingency table statistics over the whole grid for each day at each threshold. Combining these daily statistics, seasonal verification covering November 19, 2019 to March 10, 2020 over the whole CONUS was done consisting of threat scores, frequency biases, etc. for Day 2 forecasts of 24-hour snowfall at thresholds of 1, 2, 4, 6, 8, and 12 inches over the whole CONUS. Additional configuration details are provided in [Appendix A](#). In addition to MODE, the Forecast Verification Software (FVS; Novak et al. 2014, their Appendix B) was also used to generate seasonal objective statistics over the whole CONUS as well as Eastern and Western CONUS domains. The verification was done for all forecast methodologies from the GFS, NAM12, FV3-SAR, CMC, and ECMWF that were available to forecasters, and computed for events aggregated over the period November 1, 2019 through March 15, 2020.

Results

Cases

A total of 19 forecast blends were created during the course of the experiment. Figure 5, below, shows the NOHRSCv2 accumulation over the cases used in the forecast exercises. It highlights the fact that there were three main regions of snow for the experiment this year: Intermountain West terrain snow, Central Plains to Upper Midwest cyclones, and Lake Effect.

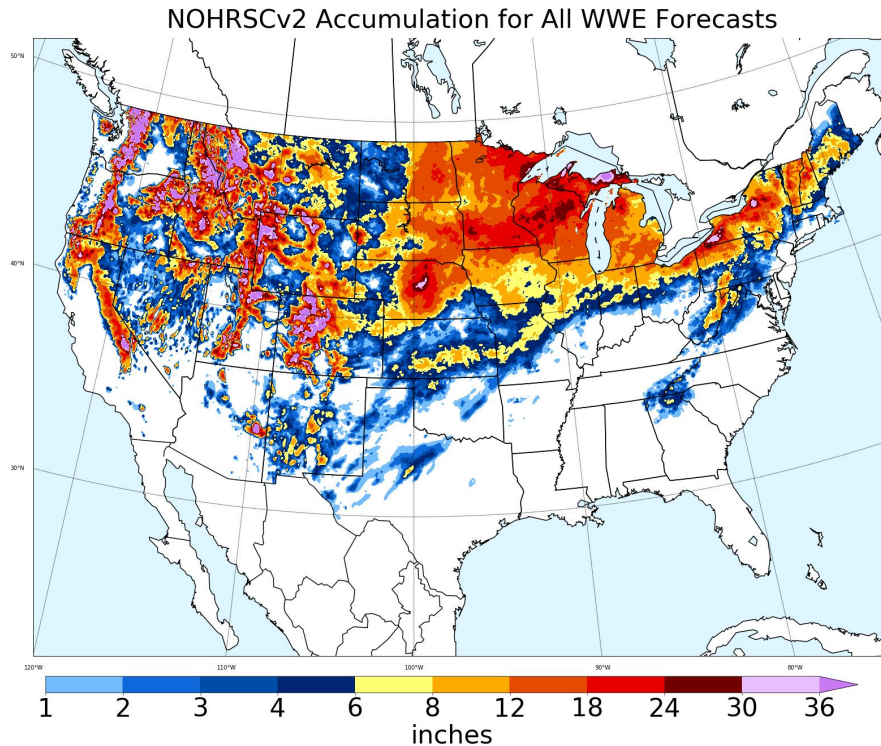


Figure 5. *NOHRSCv2 accumulation of snow over the 19 cases used in the forecasting exercises in the 10th Annual WWE.*

Details on the forecast date and domain for each of the cases are listed in the table below. With the exception of the forecasts on 10 December 2019 and 28 January 2020, all forecasts were for a day 2 12z - 12z time period. The 10 December 2019 case was a day 3 12z - 12z forecast. The 28 January 2020 case was a day 1 00Z - 12Z case. While the team attempted to focus each WWE forecasting session on a specific region, there were three weeks where there was no concentrated event to focus. Therefore, the forecasting session was completed over the entire CONUS for the 10 December 2019, 17 December 2019, and 7 January 2020 cases. A CONUS forecast provided an interesting experiment in that it forced participants to create a blend that captures all regions and snowfall types. Feedback on these CONUS sessions were not favorable. Participants found that they were more keen to provide guidance for a focused forecast region only. Based on this feedback, future WWEs will look to examine retrospective cases should the live forecast outlook be unfavorable for region specific exercises.

Table 2. List of all WWE experiment session 24-h cases.

Case	Forecast Date	Forecast Hour	Valid Date	Forecast Domain	WFO Participation
1	12 Nov 2019	F60	14 Nov 2019	Upper Michigan/Great Lakes	DTX, IWX, BGM, GFG, IND, PAH, MQT, CYS, FFC, EAX, CRH, OAX
2	19 Nov 2019	F60	21 Nov 2019	Intermountain West	DTX, IND, LZK, APX, SGF, CYS, FGF, PBZ
3	3 Dec 2019	F60	5 Dec 2019	Intermountain West	DTX, LZK, FGF, PBZ, BGM
4	10 Dec 2019	F84	13 Dec 2019	CONUS	DTX, IND, LZK, BGM, FGF, PBZ
5	17 Dec 2019	F60	19 Dec 2019	CONUS	DTX, PBZ, PAH, FGF
6	7 Jan 2020	F60	9 Jan 2020	CONUS	BGM, FGF, PBZ, LZK, JKL
7	21 Jan 2020	F60	23 Jan 2020	Midwest	DTX, IND, FGF, BGM, UNR
8	28 Jan 2020	F36	29 Jan 2020	Southern/Central Plains	IND, FGF, LSX, PBZ
9	4 Feb 2020	F60	6 Feb 2020	Southern/Central Plains	IND, LZK, IWX
10*	10 Feb 2020	F60	12 Feb 2020	Southwest	In house week
11*	11 Feb 2020	F60	13 Feb 2020	Midwest to New York	
12*	12 Feb 2020	F60	27 Nov 2019	Central Plains to Midwest	
13*	13 Feb 2020	F60	18 Jan 2020	Upper Midwest	
14	18 Feb 2020	F60	20 Feb 2020	Colorado/Kansas/Nebraska	*No WFO participation*
15	25 Feb 2020	F60	27 Feb 2020	Great Lakes to New England	BGM, FGF
16*	2 Mar 2020	F60	29 Dec 2019	Nebraska to Minnesota	In house week
17*	3 Mar 2020	F60	8 Jan 2020	Mid-Atlantic	
18*	4 Mar 2020	F60	9 Feb 2020	Montana to Minnesota	
19*	5 Mar 2020	F60	10 Feb 2020	Wisconsin to Michigan	
*In-house Week Sessions Retrospective case forecasts					

In addition to the lack of live winter weather cases, weekly participation was the other major issue with this year’s WWE. A combination of the lack of winter events and minimal advertising caused the team to struggle with gathering participants each week. Table 2, above, lists the WFO participants for each week. Out of the 122 offices within the NWS, the largest participation was 8 offices during the first week. In conversations at the in-house weeks,

participants gave suggestions for how to increase interest. These include reaching out to the Science and Operations Officers in the region where our experiment will be focusing that week, flexibility in when the forecasting exercises occur, and orientation presentations to each of the NWS regions. Going forward, these suggestions will be carefully considered to encourage greater participation from the NWS field offices. In addition, there will be greater communication with the outside partners and research community for their participation.

Experimental Forecasts and Tools

As stated earlier, there were multiple tools and techniques utilized in this year's WWE. In an effort to gain some insight from the subjective feedback and scoring, this next section will look at each dataset and method individually.

Subjective Forecast Verification

WWE Experiment Blend, WPC Official Forecast, NBMv3.2

There were three forecasts evaluated during the verification activities. These include the WWE experiment blend, the official WPC Winter Weather Desk (WWD) forecast, and the National Blend of Models (NBM) version 3.2. Inclusion of the WPC WWD forecast is due to the WWE's focus on evaluating blending of snowfall solutions.

The current snowfall forecast methodology used on the WPC WWD utilizes the official WPC quantitative precipitation forecast (QPF) prepared by the Days 1-3 WPC QPF desk forecasters. WPC WWD desk forecasters then modify the precipitation type and snow-to-liquid (SLR) ratios to generate the snowfall forecast. This current method often results in forecast inconsistencies between the liquid QPF and snowfall forecasts because the model solutions used to blend the QPF are not the same as the model precipitation type solutions selected. For example, if the QPF is heavily weighted towards the GFS, and the precipitation type forecast and SLR are NAM based, quantitative inconsistencies result between QPF and snowfall amounts, particularly on the periphery of the snowfall forecast near precipitation type transition areas.

The WWE team noted that the blending of snowfall solutions using the same blend as applied to the QPF would make for a more consistent forecast (i.e. internally consistent QPF, PTYPE, SLR, Snow, and Ice) than the current methodology of editing each parameter (QPF, PTYPE, SLR). The WWE team recommends further exploration is needed to increase internal consistency methodology.

Below in Figure 6 are the snowfall accumulations from the three forecasts evaluated in the WWE verification sessions. In general, the WWE experiment blend had broader coverage over the west and lower amounts over the Upper Midwest and Great Lakes. The WPC WWD and NBMv3.2 show similar coverage over the CONUS with the NBMv3.2 accumulation values slightly higher from the Central Plains to Wisconsin.

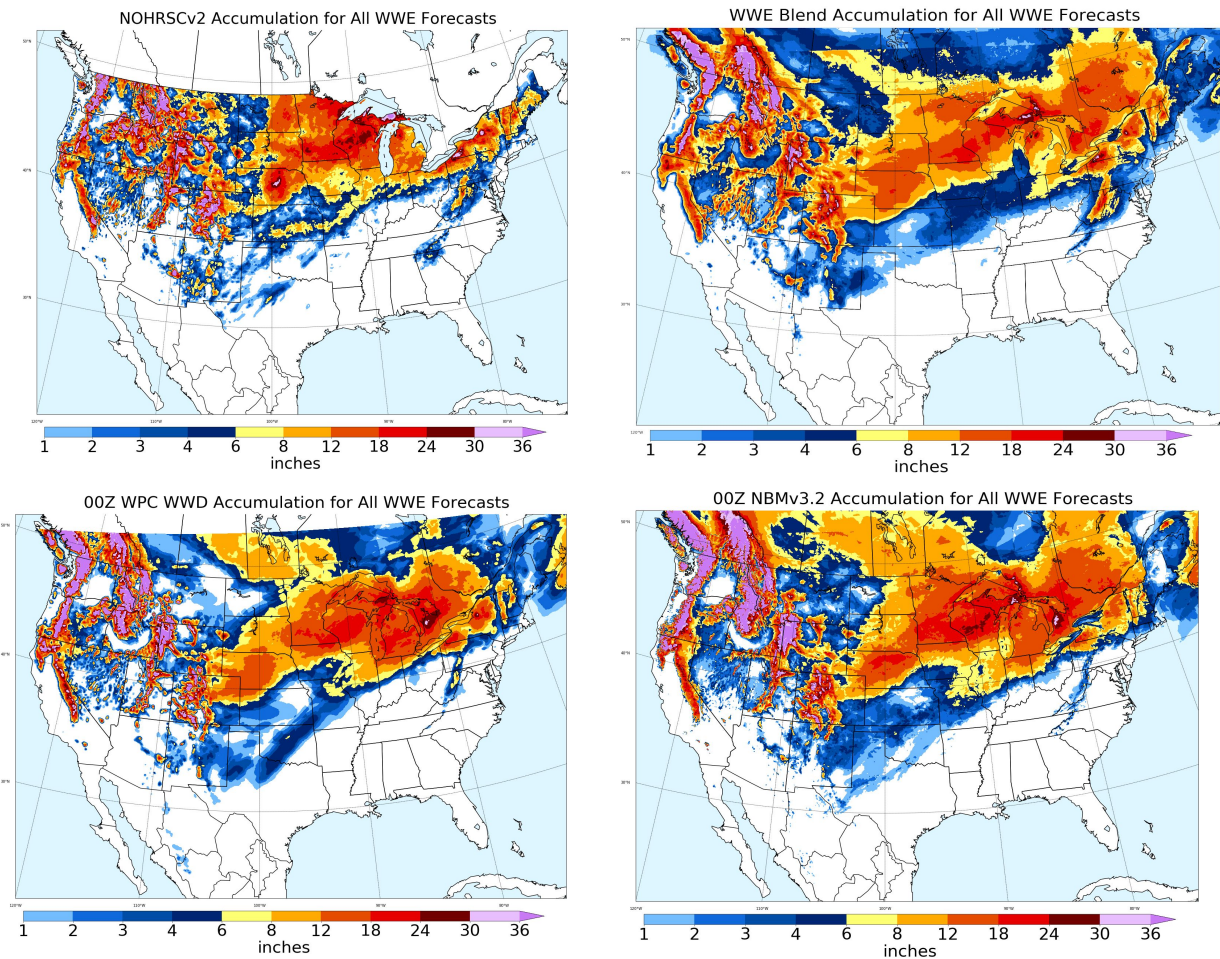


Figure 6. Total accumulated snowfall from all 19 events from the WWE experiment blend (upper right), the official WPC WWD (lower left), and the NBMv3.2 (lower right). NOHRSC observations in upper left for reference.

The general comments noted above were also consistent with the subjective comments throughout the verification process. Figure 7, shown below, are word clouds generated from the every google survey filled out during the sessions. The participants were asked about the footprint and amounts of each forecast, hence the reason for those words to appear large. However, for the WWE forecast, words like 'high' and 'missed' are associated with the broader footprint out west and the lower values in the Midwest, respectively. The NBMv3.2 comments showed that it was consistently underdone and missed the footprint for a lot of the cases. WPC WWD forecast had both over and under done cases, with comments highlighting 'missed' forecast amounts throughout the evaluation.

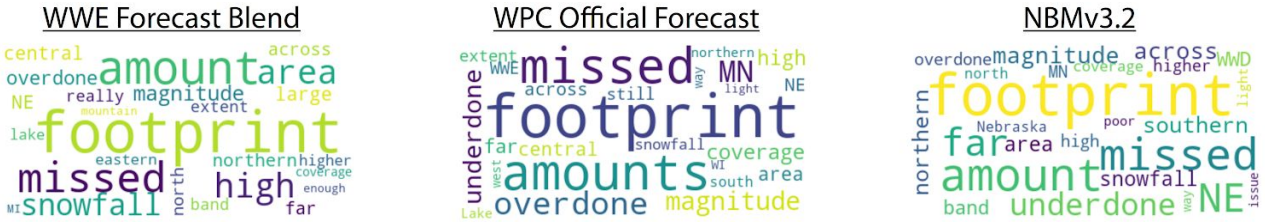


Figure 7. Word clouds based on the subjective evaluation comments.

For the subjective evaluation scoring (Figure 8), participants found the WWE experiment blend to be the best of the three forecasts with an average score of 5.12 over all cases, the WPC WWD next at 4.68, and finally the NBMv3.2 with 4.43. To compare, the highest score for the WWE occurred during case 13 (18 Jan 2020) with a value of 6.8 (WWD: 6.65, NBMv3.2: 5.69), WWD occurred during case 2 (19 Nov 2019) with a value of 7.14 (WWE: 5.07, NBMv3.2: 6.0), and NBMv3.2 occurred during case 19 (10 Feb 2020) with a value of 6.5 (WWE: 5.28, WPC: 5.72). The lowest score for both WWE and WPC WWD was case 1 (12 Nov 2019) with a score of 3.86 and 4.0 (NBMv3.2: 5.14) and the NBMv3.2 lowest score was case 17 (8 Jan 2020) with a value of 0.93 (WWE: 5.36, WPC: 2.11). As with all subjective scoring initiatives, there may be some bias in the low values at the beginning of the experiment as participants have not yet been involved in the process. It is also worth noting that there may be a slight positive bias in the retrospective cases in the WWE experiment blend, due to prior knowledge of the event influencing the creation of the blend. Participants were asked not to look up the case until after the forecasting exercise was completed, and those directly affected by the case excluded themselves from participation. Figures used in the verification sessions for each case can be found on the [experiment webpage](#).

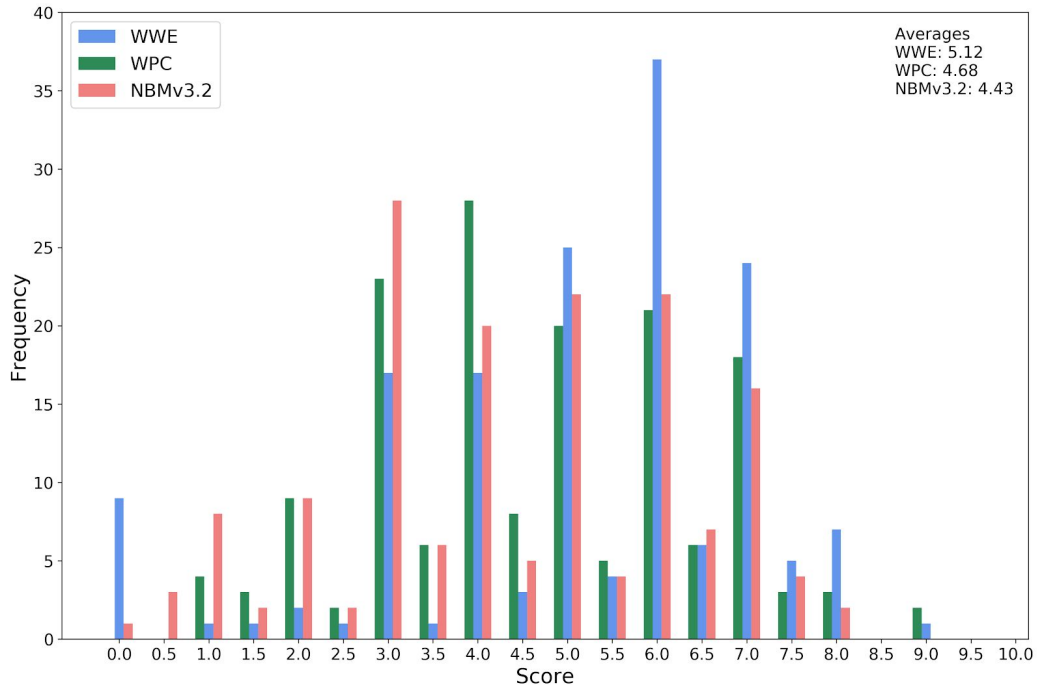


Figure 8. Subjective evaluation scores for the experimental WWE Blend, official WPC forecast, and NBMv3.2. Participants scored from 1 to 10, with 1 being the lowest score possible and 10 being the highest. 0 indicates a missing forecast. The mean score for each forecast labeled in upper right.

Precipitation Type Verification: GFS and NAM12

As stated earlier in the Data and Methods section, the WWE evaluated ensemble and averaged precipitation type techniques for the GFS and NAM12. Below are the maps showing the accumulated snowfall over all WWE forecast sessions for both of the precipitation type methods as well as SNDP (Figure 9). Overall, the SNDP accumulations were far less than the precipitation types. In fact, during the forecasting sessions, participants would generally lean towards blending the SNDP over the other solutions when the footprint looked good and the precipitation types looked overdone. Comparing the NAM12 to the GFS, the NAM12 was the snowier model throughout the WWE season, especially during the Central Plains to Midwest cyclone events. This was noted extensively in the forecast session discussions when the NAM12 would have higher QPF values in the cold sector of the system.

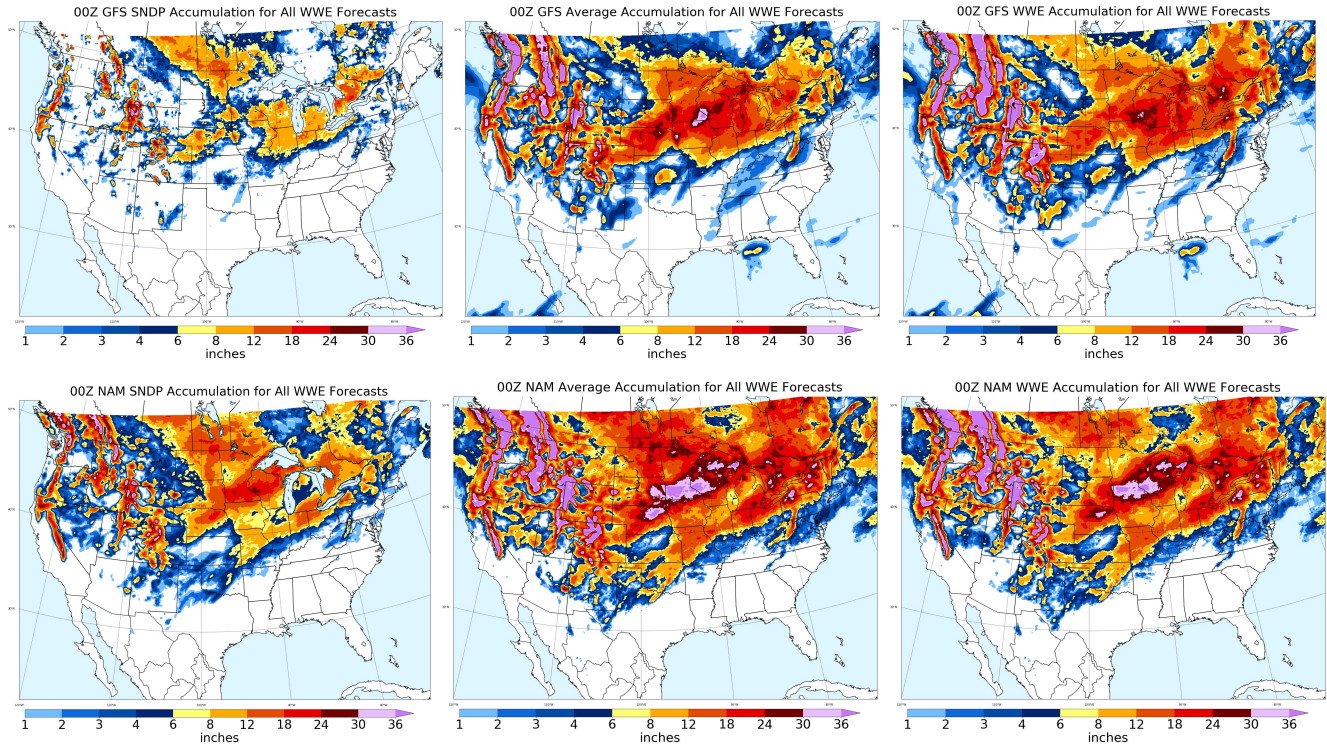


Figure 9. Total accumulated snowfall from all 19 events from the GFS SNDP (top left), the GFS Average (top center), the GFS WWE Ensemble (top right), the NAM SNDP (bottom left), the NAM Average (bottom center), the NAM WWE Ensemble (bottom right).

While only the WWE ensemble precipitation type was available in the forecast sessions, both the average and WWE ensemble were scored in the verification exercises. For the NAM12, participants commented frequently that there was little difference between both of the precipitation type methods, whereas the GFS was more likely to show differences between the methods. As an example, Figure 10 shows the verification images presented from case 19. The GFS ensemble method has both a larger footprint and higher amounts than the GFS average. In contrast, there are only slight differences in the amounts over west-central Wisconsin in the NAM12 products.

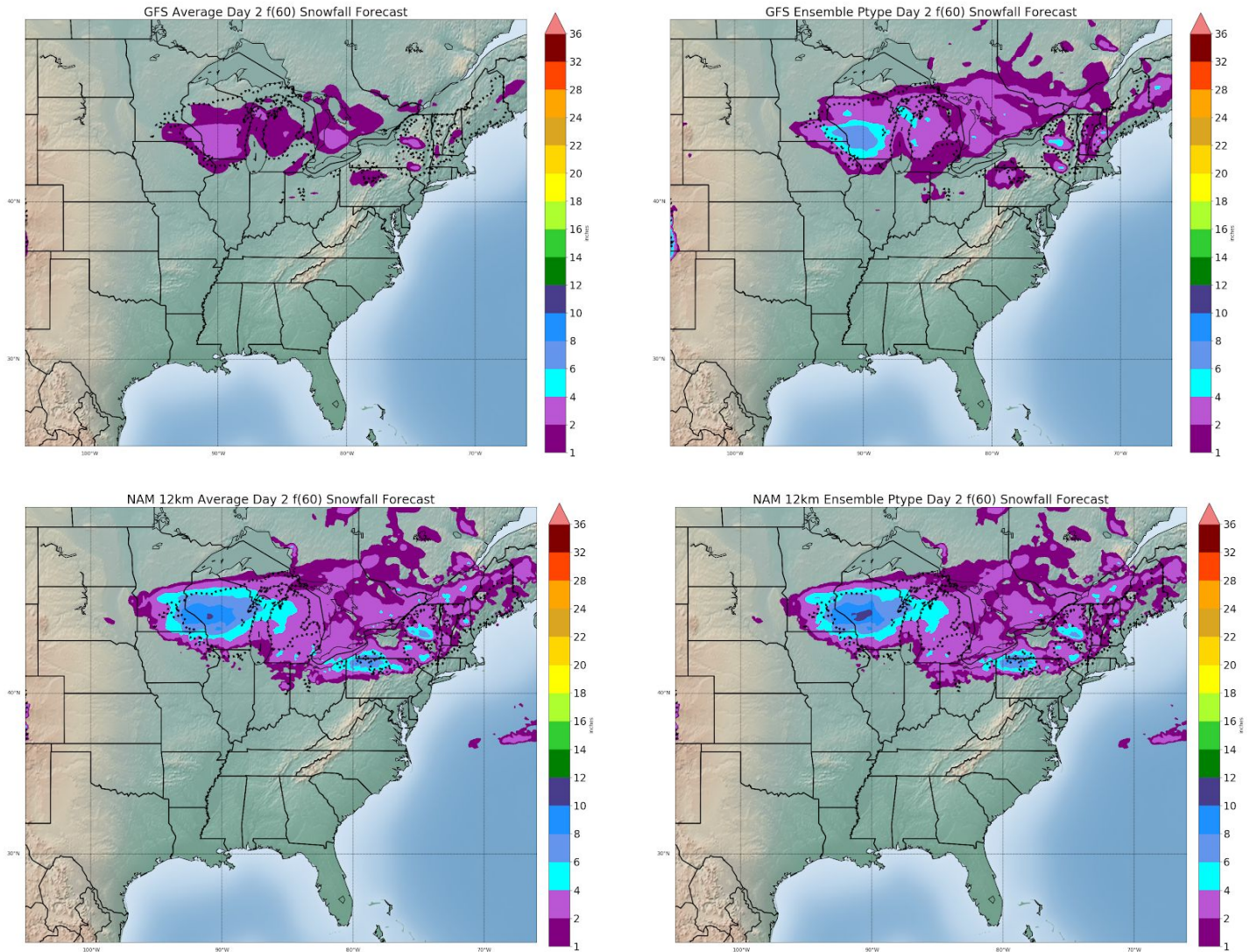


Figure 10. Case 19 (Valid 12Z - 12Z 9-10 February 2020) GFS (upper panels) and NAM12 (lower panels) verification images for the precipitation type methods. Average method values on the left. Ensemble precipitation type method values on the right.

The primary reason for these disparities in snowfall amounts is differences in SLR for each snowfall solution. The three GFS snowfall solutions averaged in the GFS AVE solution were computed using a 10:1 SLR, while the SLR used in the GFS WWE snowfall forecast was the 2.5km Baxter climatological SLR. The Baxter SLR contains ratios 25-50% higher than 10:1 for each grid point over the upper Midwest and Great Lakes region.

There were instances where both post-processed GFS snowfall solutions suggested higher amounts than the other model snowfall forecasts, and also forecasting of accumulating snowfall over areas where the precipitation type was clearly going to be rain. This was due to a cold bias in the GFDL microphysics in the GFS. Specifically the POFP parameter in the GFS, which is used in the experimental precipitation type algorithm at WPC, on several occasions

suggested percentages of 90-100 percent where the critical isobaric levels in the lower troposphere and the soundings were too warm for frozen precipitation. Thus a third of the post-processed snowfall forecast featured a precipitation type of snow which inflated the snowfall for those grid points.

Figure 11 shows a higher Day 3 snowfall forecast from Case 4 (12Z12 December - 12Z 13 December, 2019) over north central Wisconsin in the GFS WWE solution (lower right) than the GFS AVE (lower left). This is an example where the higher Baxter SLR is used in calculating the snowfall potential in the GFS WWE over the 10:1 SLR in the GFS AVE. Looking at Figure 11 along the Gulf Coast from New Orleans, LA to the western Florida Panhandle, accumulating snowfall was forecast in both the GFS WWE and GFS AVE solutions where the cold biased percent of frozen precipitation from the GFDL microphysics was used. Snowfall did not occur in this region as depicted by NOHRSC observations (upper left). No snowfall was correctly forecast by the model explicit GFS SNDP snowfall solution (upper right).

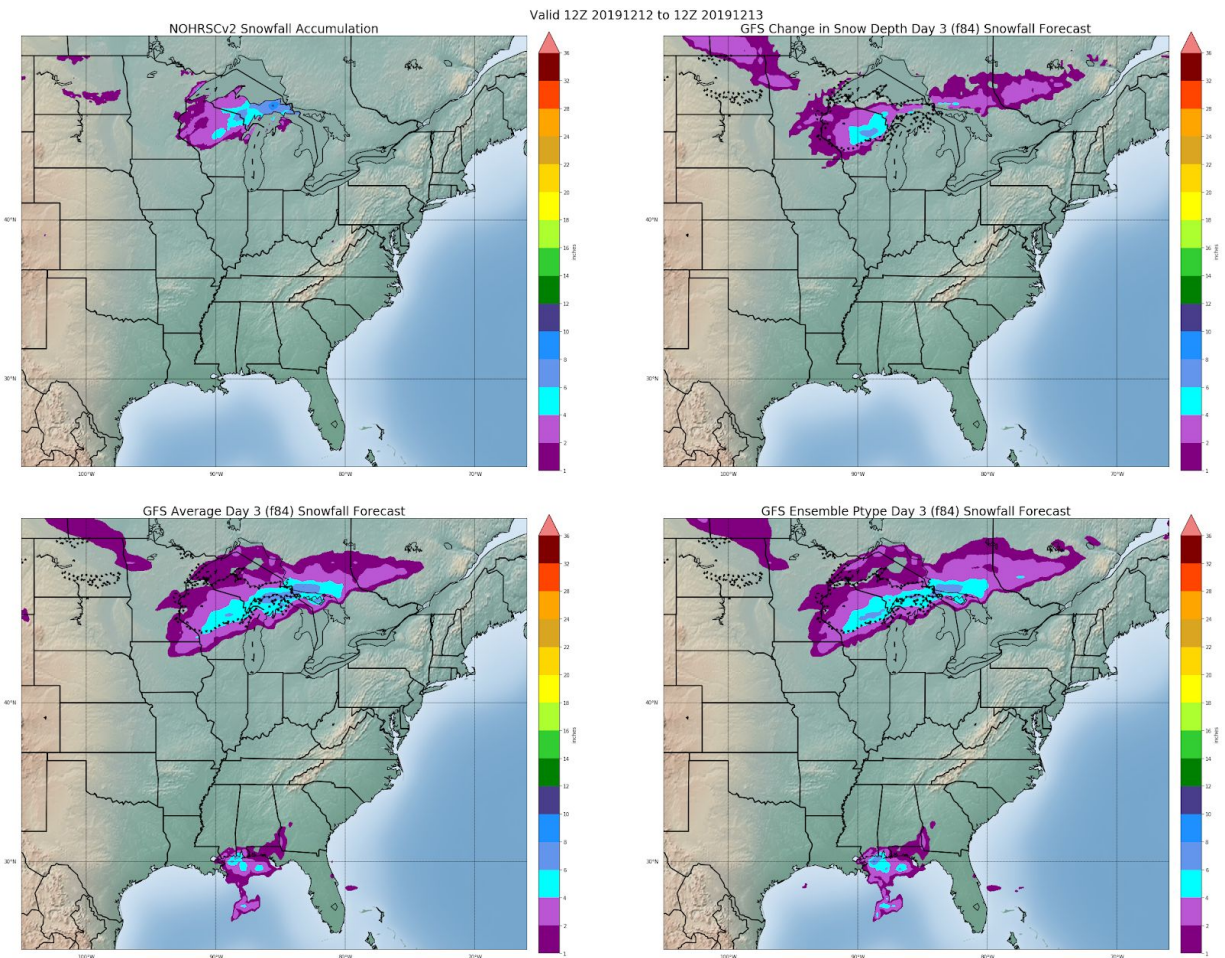


Figure 11. Case 4 (Valid 12Z - 12Z 12-13 December 2019) NOHRSC snowfall analysis (upper left), GFS SNDP snowfall forecast (upper right), GFS Average Snowfall (lower left), GFS Ensemble Ptype snowfall (lower right).

Figure 12 depicts an instantaneous plot of POFP from the GFS during the snowfall forecast period. Note the darker blue areas which represent 100 percent frozen precipitation where warm sector QPF was forecast over the northern Gulf of Mexico. This cold biased POFP resulted in an incorrect identification of snow as precipitation type, thereby predicting several inches of errant snowfall from the QPF along the Gulf Coast.

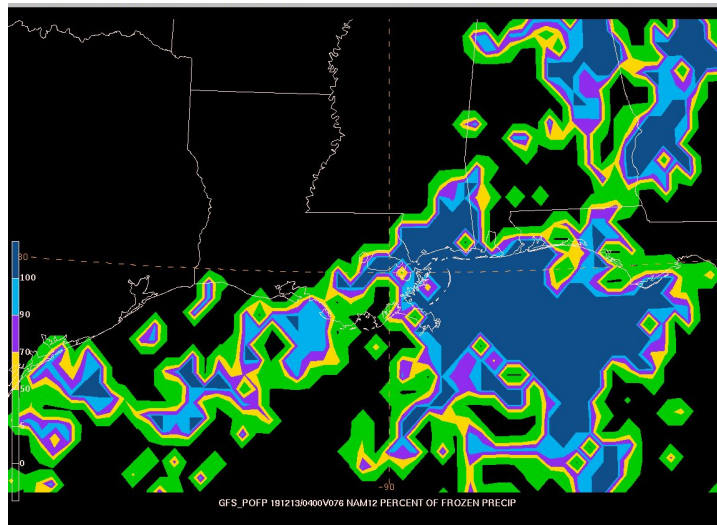
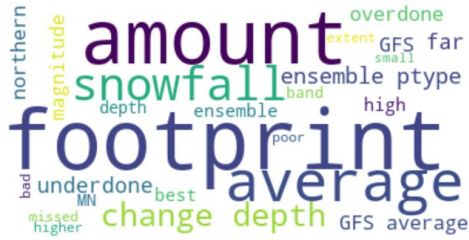


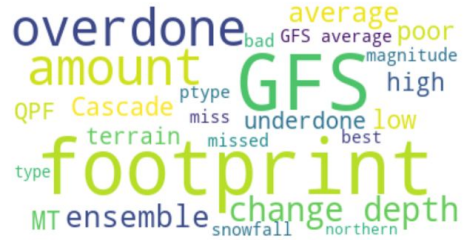
Figure 12. Instantaneous POFP at 0400 UTC 13 December 2019 (76-hr forecast)

From the word clouds below in Figure 13, participants were focused on the footprint of the forecast for the GFS in both the Eastern and Western US. In the Western US, the word ‘overdone’ also stands out, bringing to light a GFS wet-bias over the Intermountain West that was present throughout the sessions. For the NAM12, the snowfall amount was of bigger focus in the Eastern US, with participants noting the higher values seen in the previously shown maps. Over the Western US, the NAM12 SNDP was most commented on, as participants thought this method was generally better over the mountains.

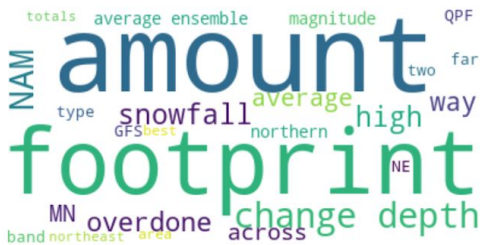
GFS Eastern US



GFS Western US



12km NAM Eastern US



12km NAM Western US

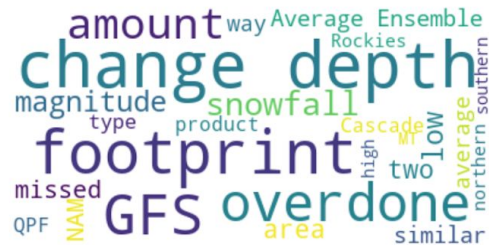


Figure 13. Word clouds based on the subjective evaluation comments.

For all of the techniques the subjective scores are found in Figure 14. The NAM12 SNDP over the Eastern US had the highest values of 4.69. In fact, all three techniques scored better over the Eastern US for the NAM12, with the lowest score of 3.98 in the Western US ensemble precipitation type. Based on all of the scores, the NAM12 scored higher over both the Eastern and Western US for the SNDP and average precipitation type. The GFS scored slightly higher for the ensemble method.

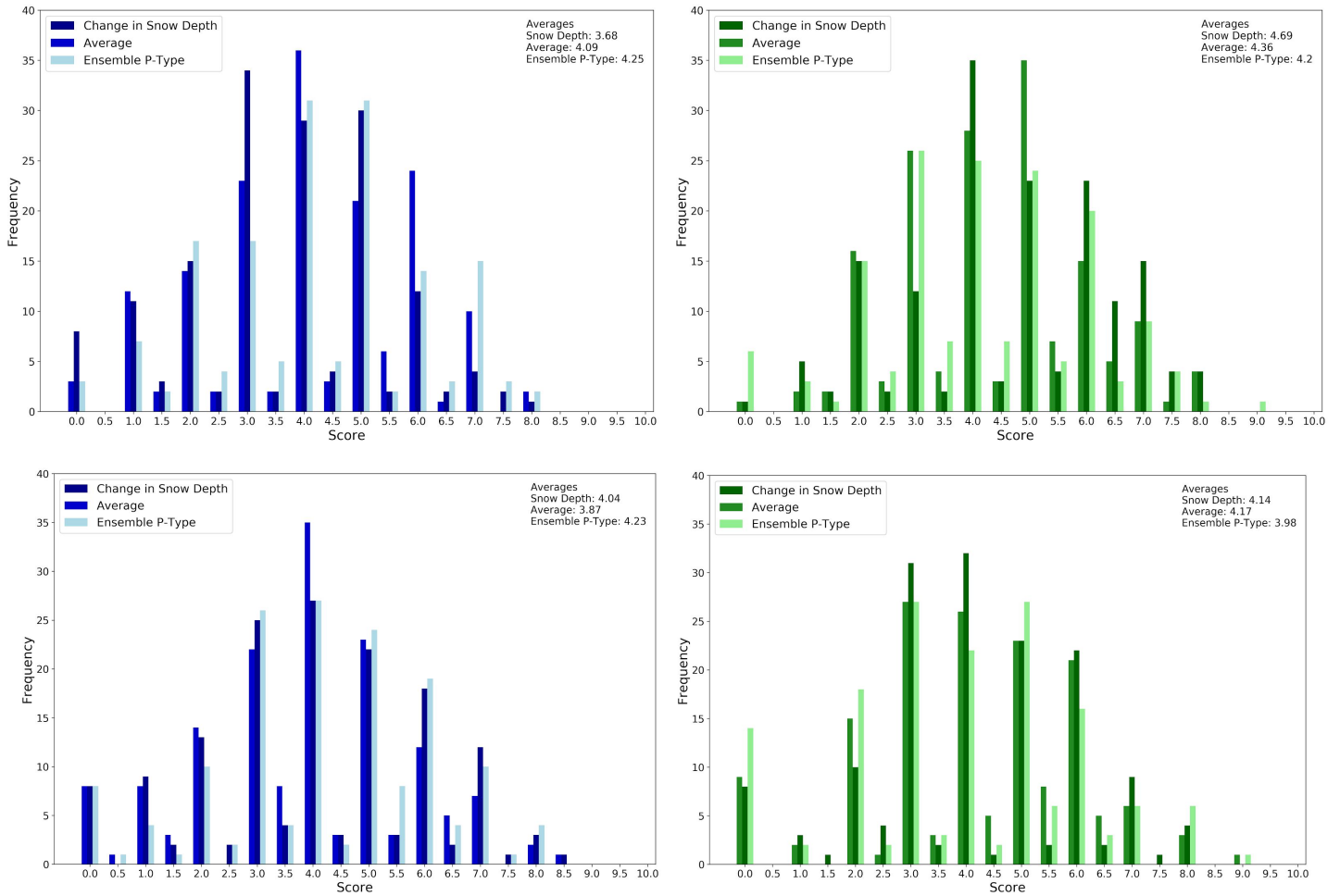


Figure 14. Subjective evaluation scores for the three precipitation type methods for the Eastern US (upper) and Western US (lower). GFS scores are the blue figures on the left. NAM12 are the green figures on the right. Participants scored from 1 to 10, with 1 being the lowest score possible and 10 being the highest. 0 indicates a missing forecast. The mean score for each forecast labeled in upper right.

Precipitation Type Verification: ECMWF and CMC Decision Tree

In an effort to provide participants with more options in blending forecast solutions, the WPC decision tree was applied to the ECMWF and CMC. In addition, beginning 23 December 2019 the CMC QPF was downscaled in hopes of improving the coarser resolution snowfall. The maps below show the snowfall accumulation for both of these models. Throughout the experiment period, the ECMWF was a popular choice for the forecast blend. In comparison to the NOHRSC, the ECMWF (Figure15) did a generally good job of capturing the amounts and area coverage for the three main forecasting regions. The CMC had too much snow accumulation over Minnesota and Wisconsin, which was an artifact of the statistical downscaling. However, the footprint was generally in line with what the observations showed, keeping in mind the coarser nature of this model.

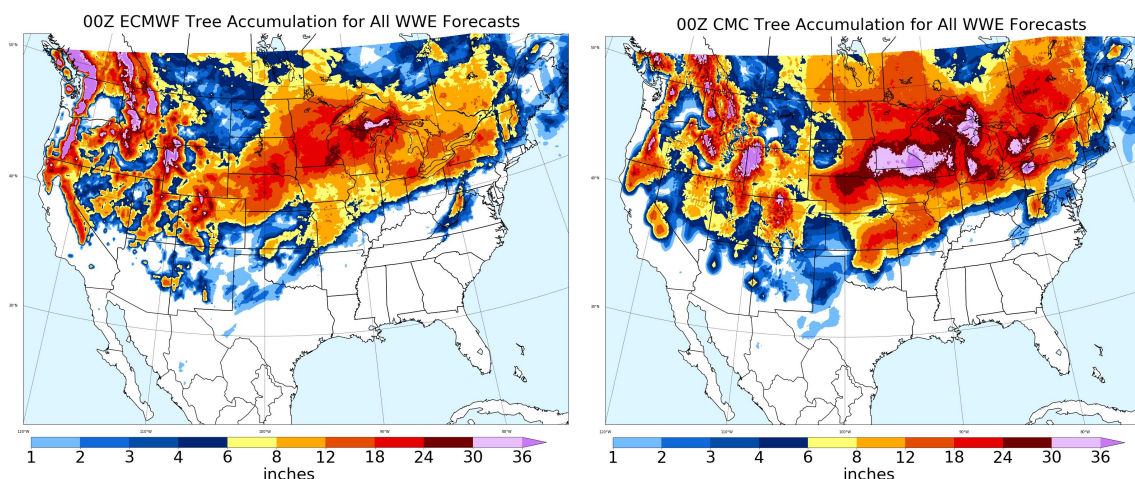


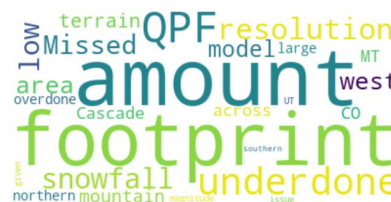
Figure 15. Total accumulated snowfall from all 19 events from the ECMWF Decision Tree (left) and CMC Decision Tree (right).

The word clouds for these models (Figure 16) are similar to the others in that the main focus of the participants was the footprint of the snowfall. Over the Western US, the amount is equally noted with the snowfall footprint due to the resolution issues over the Intermountain West terrain. For the Eastern US, the higher values in the CMC were noted with comments of ‘overdone’ and ‘high’ showing nearly as often as ‘amount.’ ECMWF word clouds also have ‘QPF’ as a large comment for the first time. With this model being heavily used within the blending exercise, it is unsurprising that participants felt the need to thoroughly comment on all components of the ECMWF forecast.

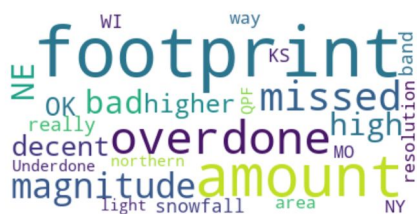
ECMWF Eastern US



ECMWF Western US



CMC Eastern US



CMC Western US

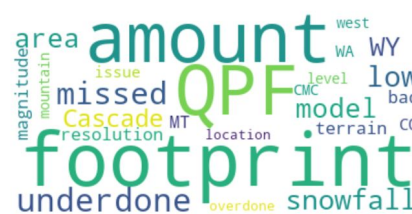


Figure 16. Word clouds based on the subjective evaluation comments.

Scoring for both of these models (Figure 17) is unsurprisingly consistent with what both the snowfall accumulation maps and word clouds show. The ECMWF was one of the highest overall scored models in Eastern US with a value of 4.68. The only higher score was the NAM12 S NDP over the Eastern US with a 4.69. Exactly the opposite is true for the CMC, as it had the lowest scores for both the Eastern and Western US out of all the experimental datasets with values of 3.43 and 2.78, respectively. The poor scores for the CMC can be attributed to the coarse resolution of the model, followed by the downscaling causing a sharp increase in snowfall amounts.

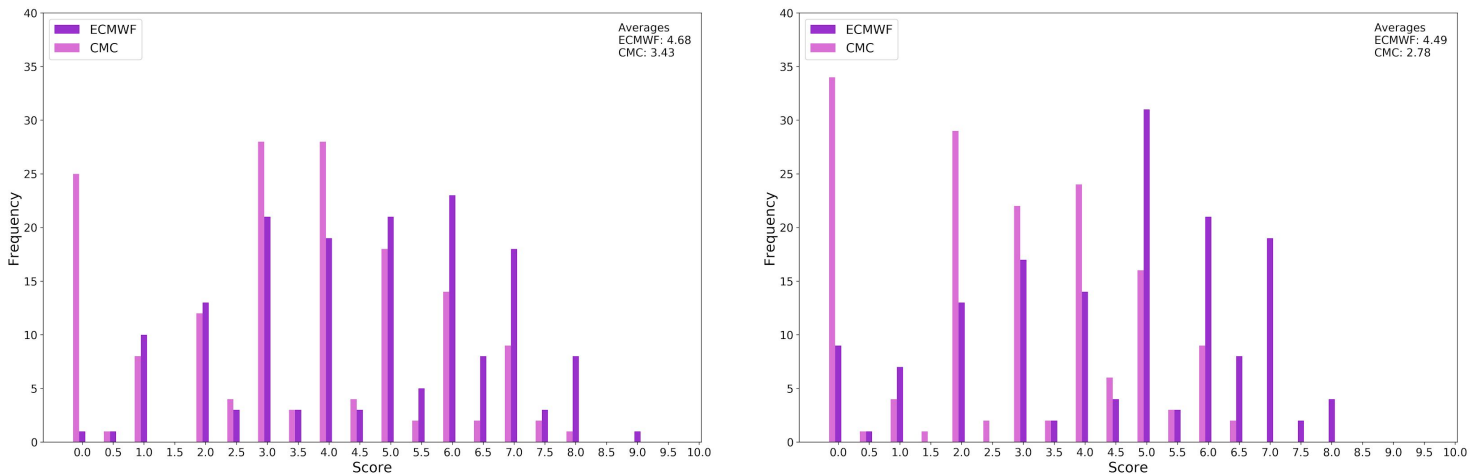


Figure 17. Subjective evaluation scores for the ECMWF and CMC Decision Tree for the Eastern US (left) and Western US (right). Participants scored from 1 to 10, with 1 being the lowest score possible and 10 being the highest. 0 indicates a missing forecast. The mean score for each forecast labeled in upper right.

CAMs Verification: FV3-SAR and SSEF Ensemble Mean

The FV3-SAR contributions to the WWE are shown in Figure 18. As stated earlier, the EMC FV3-SAR was available in both the forecasting and verification sessions. Due to the high resolution, it was a popular choice from participants in the forecast blend especially when the forecast activity was focused over the terrain. However, it was a bit overdone in the Intermountain West and over the Central Plains and Midwest storm tracks. The SSEF Ensemble mean from CAPS was only available for verification sessions. Because this is an averaged value, the values show smoother accumulations than the individual contributing members display. It should also be noted that as a direct result of the verification sessions, an error was found and corrected in the 2 meter temperatures. Future WWEs will hopefully examine the other ensemble averaging techniques from the SSEF Ensemble.

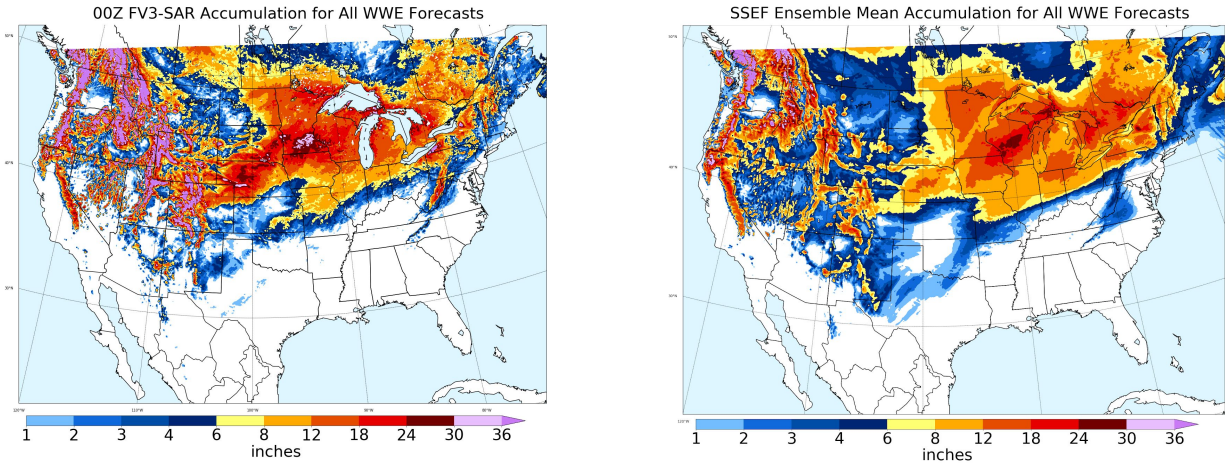


Figure 18. Total accumulated snowfall from all 19 events from the EMC FV3-SAR (left) and the SSEF Ensemble Mean (right).

Unlike the global models earlier, the high resolution models were only evaluated over the forecasted region. Therefore, there is no breakout of scoring for the Eastern and Western US. Word clouds (Figure 19) again reveal similar comments to what is mentioned in the previous sections. With the focus of the comments on the footprints of the snowfall. The FV3-SAR cloud also notes the overdone snowfall amounts for both lake effect and western terrain events. SSEF ensemble mean comments show an almost equal amount of ‘underdone’ and ‘overdone’ snowfall amounts. The word ‘shifted’ also appears for the first time in the comments, as participants noted the forecast tended to have a good footprint, although it was shifted in some way.

FV3-SAR

SSEF Ensemble Mean

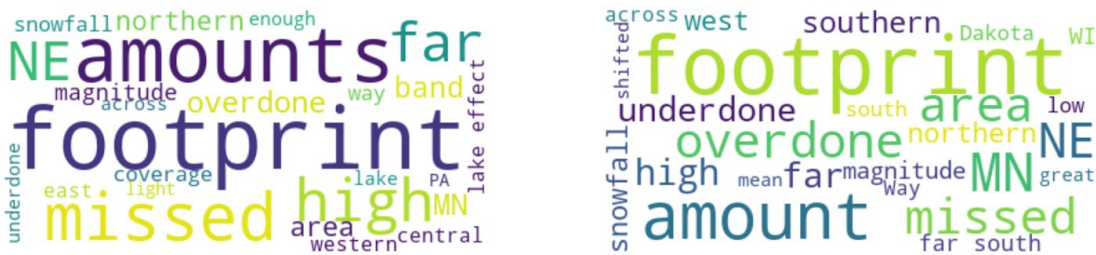


Figure 19. Word clouds based on the subjective evaluation comments.

Subjective scoring (Figure 20) for the CAM datasets are in line with the NAM12 and GFS scores with an average value of 4.06 for the EMC FV3-SAR and 4.13 for the SSEF Ensemble mean. These scores seem to be reflective of the high amounts in the western US terrain for the FV3-SAR and the smoothed nature of the SSEF Ensemble mean.

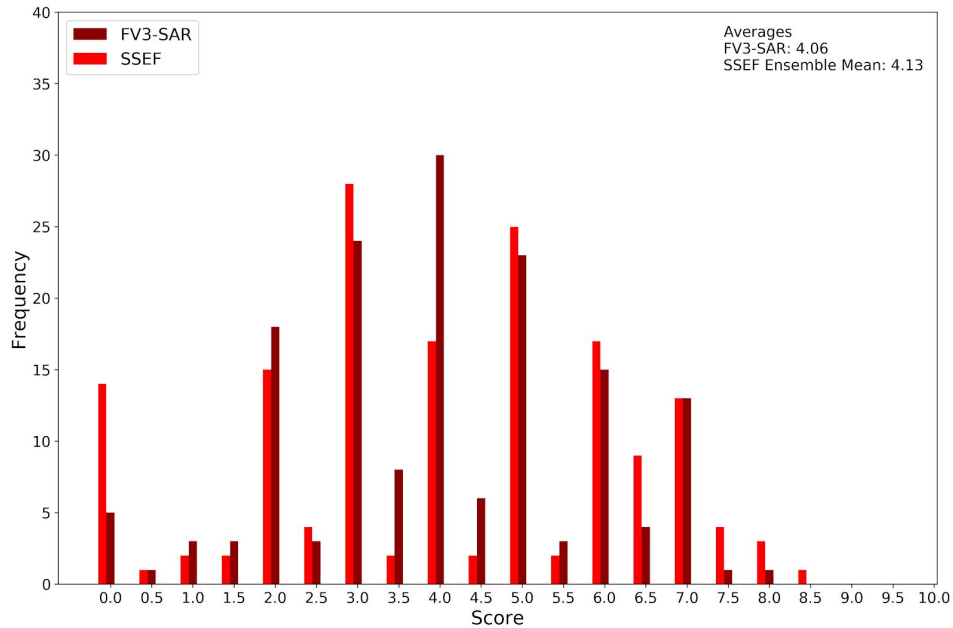


Figure 20. Subjective evaluation scores for the FV3-SAR and SSEF Ensemble Mean. Participants scored from 1 to 10, with 1 being the lowest score possible and 10 being the highest. 0 indicates a missing forecast. The mean score for each forecast labeled in upper right.

Western US Downscaling: GFS and 12km NAM

The WWE data suite included the downscaling methodology applied to the GFS and NAM12 over the Western US from the University of Utah. Maps below (Figure 21) of the snowfall accumulation for the WWE forecast cases show this technique provided utility in gaining the terrain feature especially when comparing to the non-downscaled GFS and NAM12, as expected.

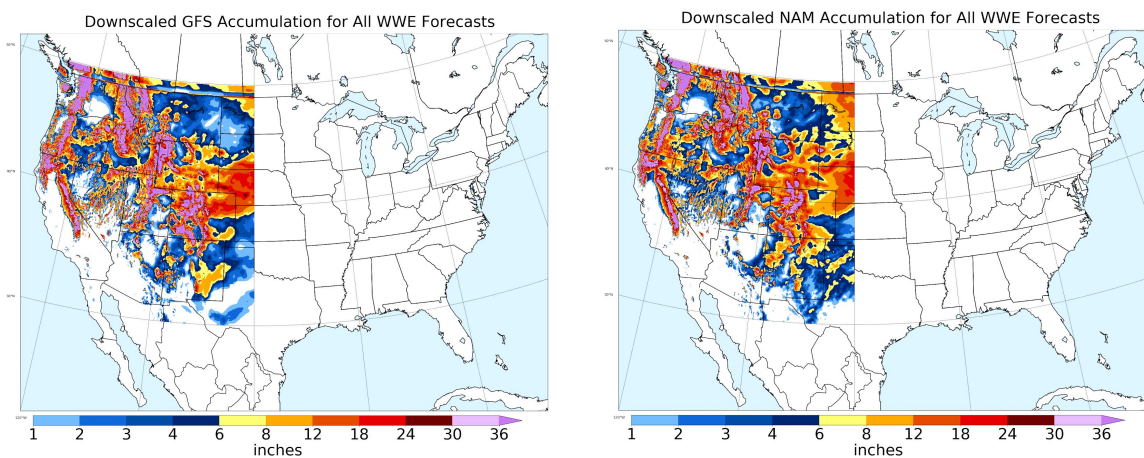


Figure 21. Total accumulated snowfall from all 19 events from the Downscaled GFS (left) and Downscaled 12km NAM (right).

Word clouds (Figure 22) for both the downscaled GFS and NAM12 reveal that participants found the downscaling technique to provide ‘overdone’ forecast snowfall amounts. This has been acknowledged by the University of Utah team as an artifact of the process. It should also be noted, however, that there are known issues with underreporting snowfall amounts in the NOHRSC dataset over the western US.

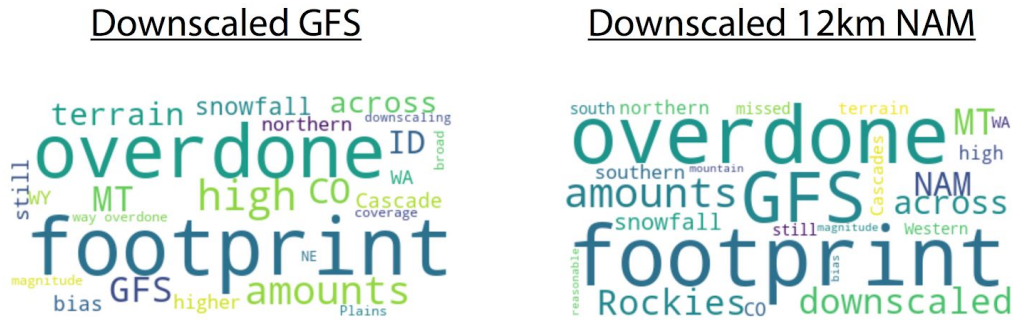


Figure 22. Word clouds based on the subjective evaluation comments.

Even though the word clouds indicate the downscaled forecasts were overdone, the subjective scores ranked the highest of any dataset throughout the WWE (Figure 23). Both the GFS and NAM12 scored at a 4.9. In fact, only the WWE experiment blend forecast scored higher with a 5.12. Participants repeatedly commented on the potential for this technique especially if a better verification dataset can be found. Discussions on this dataset often led to the request for transition into the operational process.

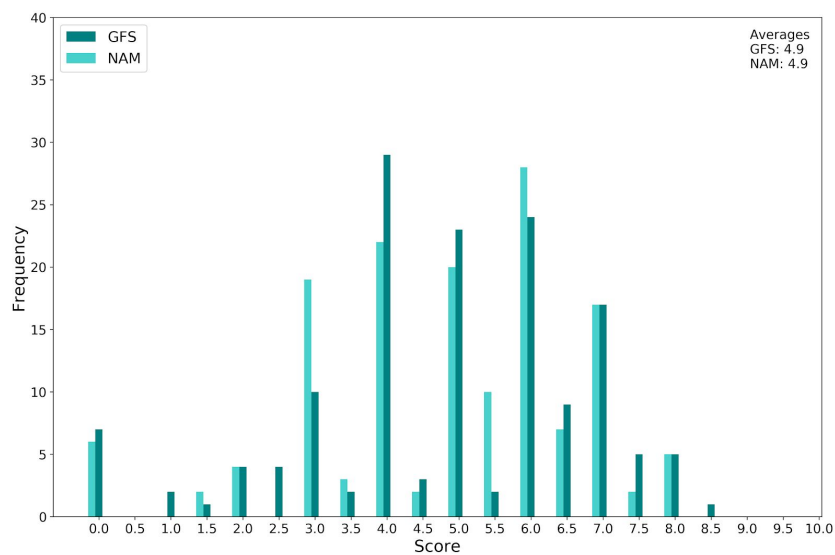


Figure 23. Subjective evaluation scores for the downscaled GFS and 12km NAM. Participants scored from 1 to 10, with 1 being the lowest score possible and 10 being the highest. 0 indicates a missing forecast. The mean score for each forecast labeled in upper right.

Objective Forecast Verification

Model Explicit Snowfall Forecasts

For the full WWE winter season (21 November 2019 - 10 March 2020), the SNDP from the NAM12 and GFS showed a low frequency bias for the 1, 4, 8, and 12 inch snowfall thresholds over both the eastern and western CONUS. Over the western CONUS (Figure 24), the magenta colored bar represents the GFS SNDP, and the gold colored bar, the NAM12 SNDP. This became more pronounced in higher end snowfall thresholds of 8 and 12 inches. The 3km FV3-SAR SNDP, depicted by the lime green colored bar scored a lower bias as well over the west (Figure 24) but it was closer to the benchmark 1.0 than the NAM12 or GFS solutions.

Across the eastern CONUS (Figure 25), the FV3-SAR SNDP (lime green bar) had a marginally low frequency bias at 1, 4, and 8 inches but performed quite well at 12 inches. The GFS SNDP (magenta bar) showed a lower frequency bias but not as low as over the western CONUS. The NAM12 SNDP (gold bar) showed a low frequency bias at all thresholds and was quite poor at the 8 and 12 inches snowfall thresholds.

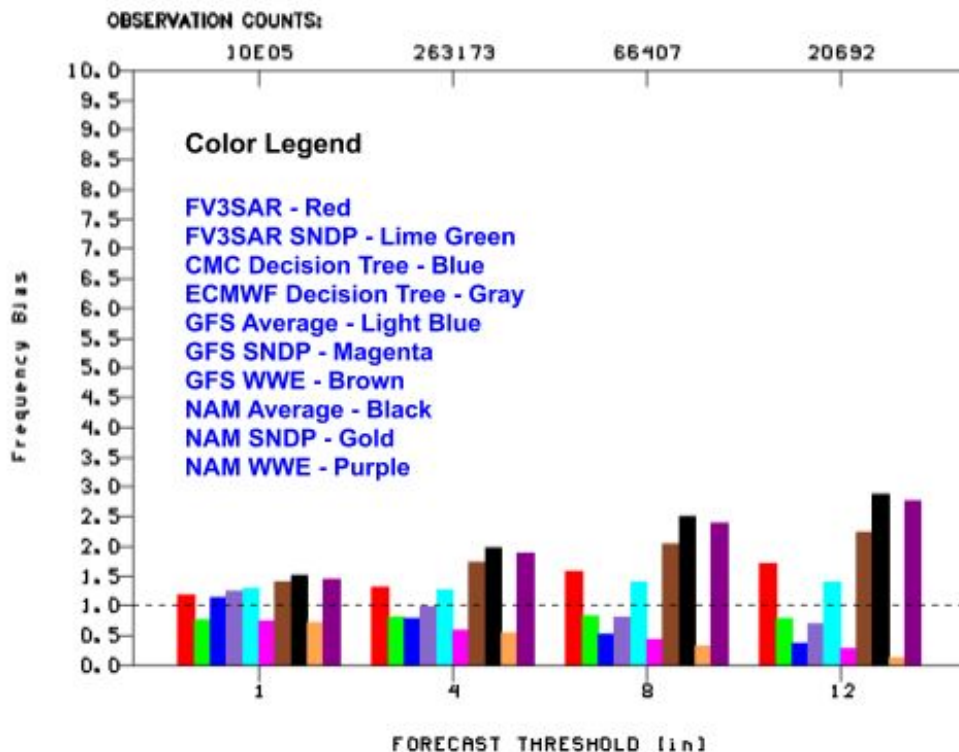


Figure 24. Western CONUS Frequency Biases

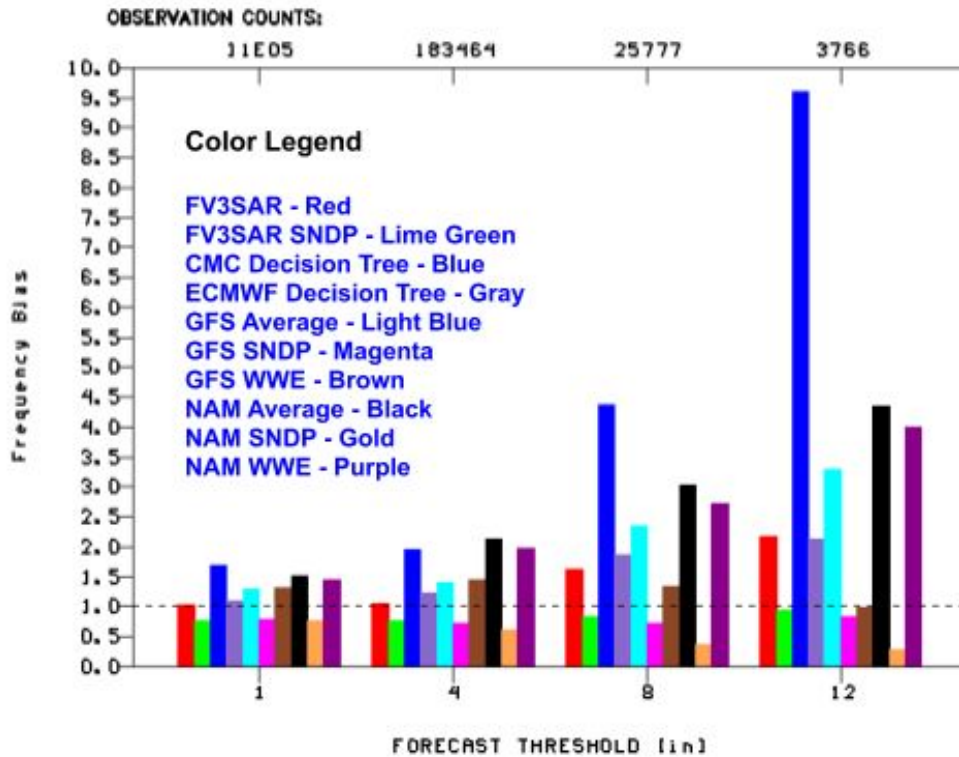


Figure 25. Eastern CONUS Frequency Biases

Performance Diagram

Seasonal verification from MODE for all WWE blender options at the 4 inch snowfall threshold for the CONUS are shown in Figure 26. Both NAM12 post processed forecasts (salmon/purple) showed similar skill with the highest probability of detection in the experiment, while the NAM12 SNDP (orange) verified with the lowest probability of detection. The probability of detection scoring of both GFS post processed forecasts (brown/light blue) showed more difference than the NAM12 values with slightly lower values. GFS SNDP (pink) is similar in value to the NAM12 SNDP. At multiple snowfall thresholds, both model solutions featured similar probability of detection with the precipitation type ensemble method (NAM12_WWE and GFS_WWE) and the three solution averaging method (NAM12_AVE and GFS_AVE) in the aggregate seasonal verification. It was noted in the visual verification sessions that there was occasionally a large disparity between these two solutions; an example of this can be found in the previous section.

The post processed ECMWF snowfall forecast using the WPC decision tree to identify frozen precipitation scored the highest in both the seasonal objective verification and subjective visual verification. The ECMWF solution is plotted in purple (Figure 26) and was the best performing solution in both threat score and frequency bias. In contrast, the CMC solution (blue) shows the lower end performance. There was a pronounced high bias to the CMC snowfall forecasts, particularly at the higher end thresholds, over the eastern CONUS (dark

blue bars in Figure 25). Conversely there was a low bias in CMC snowfall forecasts over the western CONUS (dark blue bars Figure 24). Finally, the FV3-SAR showed similar performance diagram values to the ECMWF (Figure 26; red). Also, similar to the NAM12 and GFS, there is a difference between the FV3-SAR and the FV3-SAR SNDP (light green) with the SNDP showing lower probability of detection.

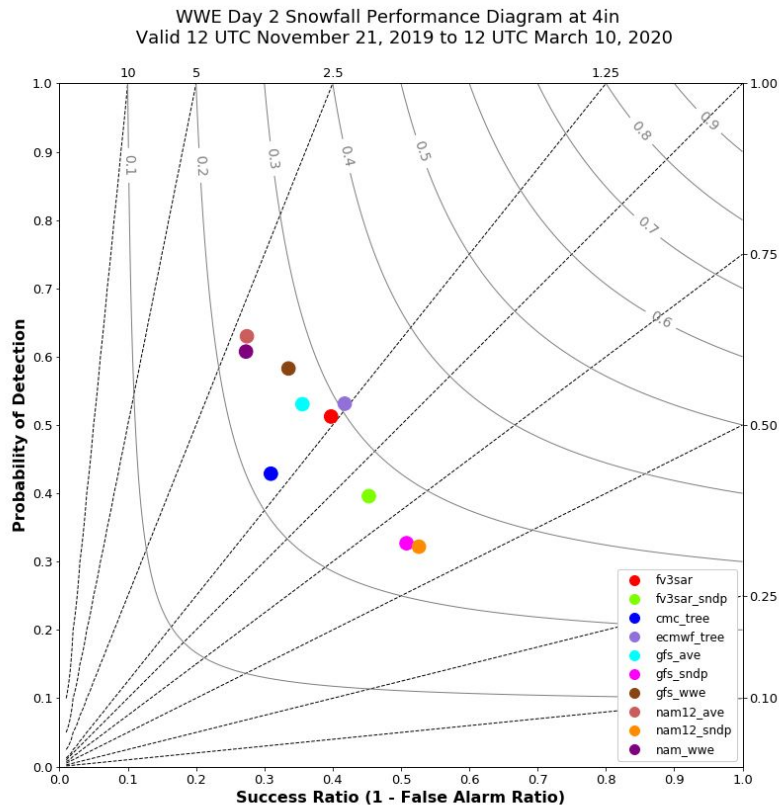


Figure 26. Performance diagram of WWE blender options at the 4 inch threshold. Calculated over the entire WWE winter season 21 November 2019 - 10 March 2020. Dashed lines indicate bias and curved lines indicate CSI/threat score.

While daily data is not available over the entire WWE season, the SSEF ensemble mean MODE performance diagram was still calculated over the 18 available case days. Figure 27 shows the diagram with the colored dots indicating different snowfall accumulation thresholds. Not surprisingly, probability of detection and threat score decrease with increasing snowfall amount, noting that there is probably a sample size issue with the larger accumulations.

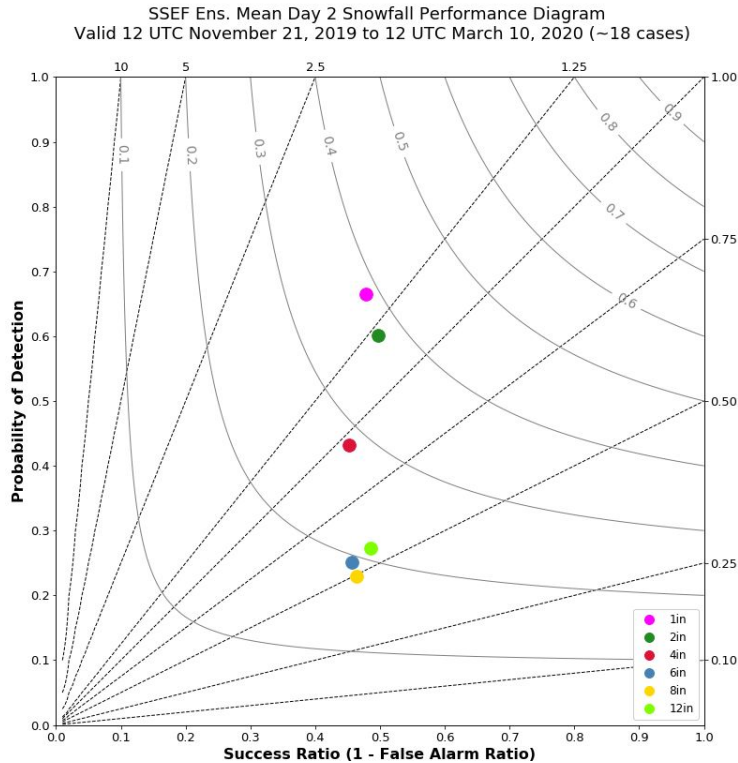


Figure 27. Performance diagram of SSEF Ensemble Mean. Colors indicate different snowfall thresholds. Calculated over 18 of the WWE exercise cases. Dashed lines indicate bias and curved lines indicate CSI/threat score.

Summary & Recommendations

This section provides more context and details to the Transition Recommendations table at the beginning of the report. Each dataset or method is summarized with specific recommendations from the WWE team as to its future development or implementation into NWS operations. Generally speaking, most of the datasets and techniques are recommended for continued testing and development. However, once a dataset or technique has been recommended for transition into operations, WWE team members will work with WPC developers to determine the best approach for implementation.

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 4 distinct cluster mean solutions. The main change from the prior winter's WWE was breaking the 90 ensemble members into 4 clusters instead of 5. The cluster mean snowfall forecasts were assessed along with the other experimental blend inputs during the forecast verification exercises.

Experiment participants provided mixed feedback about the use of ensemble clustering in forecasting winter events. Feedback about the clusters for cases that contained robust cyclones was positive. However, feedback for cases that were driven by weak, developing shortwaves, were mesoscale in nature, or heavily dependent on terrain was much more mixed and occasionally negative. Even if the clusters were not used in the forecast blend, most participants agreed that the clusters were useful for situational awareness. They provided an overview of the synoptic variability in the forecast large-scale pattern. They also assisted in defining the bounds of the possible forecast outcomes as well as identifying outlier guidance for certain events. Forecasters found this information helpful even if they struggled a bit at first with understanding how the clusters were derived and how they should be applied. **Recommendation: A common request was for more information and training on using ensemble clustering in the forecast process.**

There were two common criticisms of the clusters. The first was that they were too coarse and often underdid the magnitude of snowfall events, which precluded their use in the day's forecast blend. This was especially true for events that were mesoscale in nature or had a strong terrain component. The cluster forecasts are based on 0.5° ensemble data, the highest resolution ensemble data WPC has available. Midway through the season, downscaling was applied to the cluster snowfall forecasts and that did partially correct the issue of amounts being too low, especially in areas of higher terrain. However, coarse global ensembles will always struggle with resolving mesoscale features and there is no remedy for this. In these situations, CAMs or CAM ensembles should be used but they presently don't have enough members to make a technique like clustering useful. Forecasters will have to be made aware that the clusters can be a valuable tool but they are not valuable in all situations. There are certain situations (e.g., lake effect snow) where global ensemble forecasts are not generally of use and therefore the clusters will not be of use.

The second common criticism of the clusters was that there was often little discernible difference in the cluster snowfall forecasts. The aerial footprint of the forecasts provided some value but the quantitative range of snowfall amounts were limited. For example, several events were noted where the snowfall verification was outside the range of possible forecast scenarios as depicted by the four clusters. A potential explanation for this is that the underlying clustering methodology relies on ensemble differences in the forecast 500-hPa pattern over the CONUS which does not always translate to QPF differences in the Day 2 and Day 3 time periods. Future work on regional clustering directly on the ensemble forecasts of snowfall amounts of QPF may improve snowfall forecast dispersion. **Recommendation: Development and testing of refined cluster calculation techniques.**

Precipitation Type Methodologies

Based on the recommendations from the 2018-2019 WWE, this year looked to refine and adjust the precipitation type methodologies. For both the NAM12 and GFS, an ensemble

averaging of all three precipitation type methodologies were computed every six hours to arrive at a weighted percentage of each precipitation species for the respective 6-hour period. This ensemble average was evaluated for both the forecast and verification sessions. A simple average of the three precipitation type methodologies was also evaluated during the verification sessions. Feedback from participants on both the precipitation type methodologies and the averaging techniques was mixed. However, based on the discussion sessions, there was interest, especially from the WPC WWD, in the adjusted SLR processing that is incorporated into the precipitation transition zones. **Recommendation: WPC developers will work to incorporate the microphysics derived methods for adjusting SLR in the WPC WWD forecast process.** SNDP is already available to operational forecasters as it is part of the model post processes. However, there is an opportunity to add this information to the probabilistic winter weather forecast (PWPF) product. **Recommendation: WPC developers will work to add NAM12 and GFS SNDP to the PWPF.**

Forecast Blending

The blending exercise was well received by WWE participants. This year's WWE expanded the blending exercise to create 24 hour blends either from 6 hourly blends or as a direct 24 hour blend. In general, the forecast blends provided more accurate forecasts compared with individual model solutions. Blending of clusters was viewed favorably, and it was noted that they provide a good first guess field and forecast footprint. **Recommendation: Future WWEs will investigate blend creation further with more engagement from participants.** This could be achieved via the use of a dynamic website and possible submission of multiple participant created blend solutions.

CAMs

FV3-SAR was a popular member for the forecast blending exercise. However, the use of the GFDL microphysics was a noted issue by both EMC and experiment participants. EMC has announced that the FV3-SAR will be switching to the Thompson microphysics. There was also participant feedback wanting to include the FV3-SAR SNDP into the blending options. Based on the performance diagram, this should be considered for future WWEs. **Recommendation: Incorporate FV3-SAR with Thompson microphysics and FV3-SAR SNDP into future forecast blending exercises.** Another consideration for future WWEs is the utility of running CAMs out to 84 hours. While the team recognizes this is an extreme computing request, this is an important question that should be addressed within the testbed environment.

The SSEF ensemble was a valuable resource during the forecast discussions, even though it was not available as a blending option. Noting that this ensemble was run out to 84 hours and was a pivotal step towards evaluating the utility of day 3 CAMs solutions. Discussion with CAPs has indicated that future WWEs should look to evaluate their other ensemble

averaging techniques: the probability match mean (PM) and the local probability match mean (LPM). **Recommendation: Evaluate PM and LPM techniques and incorporate SSEF ensemble mean as an option in the AWIPS forecast blend exercise.**

Western US Downscaling

This was the first WWE evaluating downscaling, although this data was only available for the verification sessions. While the experiment only visually verified the GFS and NAM12 downscaled solutions, University of Utah also applied their downscaling to ensemble systems. Participants continually noted the positive contributions this downscaling would provide for snowfall forecasting, especially from WPC WWD. Due to data compatibility issues, objective verification was not able to be completed. The University of Utah has also indicated some refinement to their techniques. As such future WWEs will provide further testing and verification for the western US downscaling. **Recommendation: Continue evaluation of downscaling techniques by adding the eastern CONUS into the AWIPS forecast blend exercise.**

In addition to the previous recommendation, the WWE team also recommends the **transfer of code to WPC to use over intermountain west in WPC WWD forecasts.** The downscaling and mountain SLR methodologies in the code can be applied to all models not just NAM12 and GFS. This could entail early season testing in 2020-21 then implementation later in the season. As part of the future of WWE, the inclusion of this downscaled technique will help evaluate the utility of day 3 CAM solutions. This sets up the future science question of whether CAMs out to 84 hours is worthwhile or is downscaling of coarser models a viable alternative.

Acknowledgements

The WPC-HMT would like to sincerely thank WPC's Dave Novak, Kathy Gilbert, Jim Nelson, and Mark Klein for their commitment and diligence to enable this experiment to be successfully executed. We thank Chris Bailey for his AWIPS support and WPC DTB staff who assisted with verification. Alan Robson for his tech support and getting the USB capture device connected for successful AWIPS screen sharing. Josh Kastman for providing feedback on this report. Andrea Ray for her contributions to the evaluation process. 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.

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/tin15-05_bigrsc_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.
- 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.
- 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 2020, https://www.wpc.ncep.noaa.gov/hmt/WWE2015_final_report.pdf.

WPC-HMT, 2019: The 2019 HMT-WPC winter weather experiment: Final report. Accessed 4 May 2020, 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: WPC MODE Settings for Objective Verification

MODE Configuration for Experiment Cases

MODE was used to objectively analyze Day 2, 24 hour snowfall forecast objects from the following experimental snowfall precipitation type inputs over the whole season (November 19, 2019 to March 10, 2020: CMC Tree, ECMWF Tree, GFS Average, GFS Change in Snow Depth (SNDP), GFS Winter Weather Ensemble (WWE), NAM12 Average, NAM 12 SNDP, NAM12 WWE, FV3 SAR, FV3 SAR SNDP, NBMv3.2, WPC Winter Weather Desk (WWD). MODE was also used to analyze the mean snowfall from the SSEF for the 19 forecasts it issued over the whole experiment. 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, 6, 8, 12 inches of 24-hour snowfall	1, 2, 4, 6, 8, 12 inches of 24-hour snowfall
Convolution Radius	5 grid squares	5 grid squares
Area threshold	≥ 50 grid squares	≥ 50 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. A value of $T \geq 0.6$ 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

- 60 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 5 grid squares used
- Double thresholding technique applied