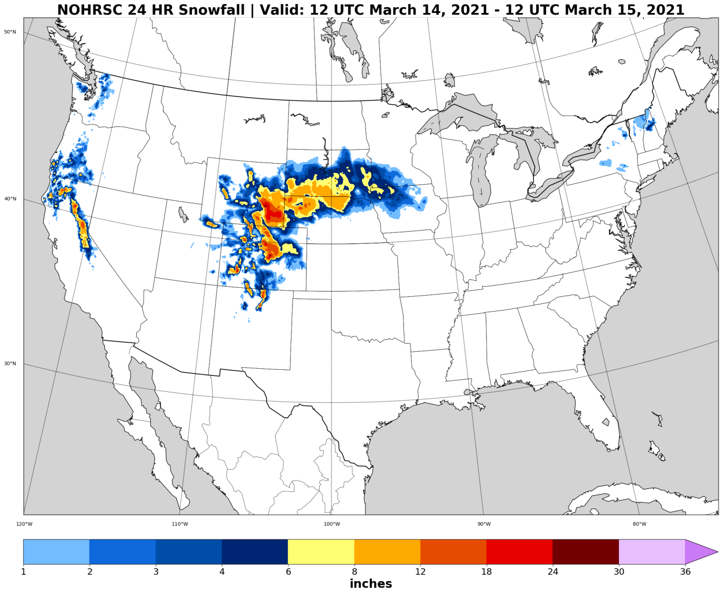
11th Annual Winter Weather Experiment:
Findings and Results


NOHRSC 24 HR Snowfall | Valid: 12 UTC March 14, 2021 - 12 UTC March 15, 2021

1 November 2020 - 15 March 2021

Weather Prediction Center
Hydrometeorology Testbed
College Park, MD

Kirstin Harnos - CIRES/CU Boulder, NOAA/NWS/WPC
James Correia Jr. - CIRES/CU Boulder, NOAA/NWS/WPC
Benjamin Albright - Systems Research Group, NOAA/NWS/WPC
Mike Bodner - NOAA/NWS/WPC
James Nelson - *NOAA/NWS/WPC*

Table of Contents

1. **Introduction** 2
2. **Data and Experiment Logistics** 3
   Experiment Logistics, Evaluation, and Participation 3
   Data Overview 9
3. **Experiment Findings and Results** 16
   Subjective Survey Results 17
      Eastern CONUS 17
      Western CONUS 22
   Seasonal Performance Diagrams 27
   Maximum Snowfall and Timing Product 30
   Highlighted Events 34
4. **Summary and Recommendations** 45
5. **Acknowledgements** 47
6. **References** 48

Appendix A: MODE Configuration 49

1. *Introduction*

In support of the ongoing mission to improve National Weather Service (NWS) products and services for winter weather, the Hydrometeorology Testbed (HMT) within the Weather Prediction Center (WPC) conducted the 11th annual Winter Weather Experiment (WWE) during the 2020-2021 winter season. The WWE provides an immersive collaborative research to operations (R2O) experience bringing together members of the forecasting, research, and academic communities to evaluate and discuss winter weather forecast challenges. As the WWE moved into its 11th year, it looked to explore the future of snowfall products and services by providing a large suite of experimental FV3[1] convective allowing models (CAMs) and operational simulations in the day 2 and day 3 time period. The suite of FV3 CAMs are more formally referred to as the FV3-limited area model (LAM) and will be designated as such for this report. These FV3-LAM configurations seek to directly inform developers of the new Rapid Refresh Forecast System (RRFS[2]) for wintertime snowfall performance. The RRFS is the new single-model, convective-allowing, ensemble-based data-assimilation and forecasting system aimed to advance high-resolution ensemble forecasting methods while supporting the Unified Forecast System (UFS[3]).

While the remote aspects of the WWE have been successful for the past few years, this year looked to adapt and expand in an attempt to be more flexible and garner more CONUS-wide participation. The primary avenue of this expansion is the use of a dynamic GIS based website[4]. Participants were asked to explore our experimental data through this interface paired with an online survey asking for subjective rankings and comments of the data both before and after an event occurred. Another component of interaction involved asking the participants to use an online drawing tool[5] to create polygons of snowfall footprint, maximum snowfall amount, and maximum precipitation rates within their defined snowfall footprint. The full details of these activities will be described in later sections, but this activity sought to gain insight into participant use of both CAM and probabilistic information for finding extremes in the snowfall and utility of the experiment data.

As with past remote years, the WWE had twice weekly (Tuesday and Wednesday) sessions beginning 17 November 2020 and ending 10 March 2021. The experiment also hosted two weeks of invited intensive evaluations where selected retrospective cases were evaluated more

[1] FV3 stands for finite volume cubed-sphere and is the new dynamical core planned for all future NWS modeling systems.

[2] RRFS web page: https://gsl.noaa.gov/focus-areas/unified_forecast_system/rrfs

[3] UFS web page: https://ufscommunity.org/

[4] WWE web page: https://origin.wpc.ncep.noaa.gov/hmt/wwe2021/11th_annual_wwe.php

[5] WWE Drawing tool: https://origin.wpc.ncep.noaa.gov/hmt/wwe2021/draw_Snowfall.html

rigorously. Overall, the WWE evaluated eight cases in a real time setting and 11 cases as retrospective evaluations. There are also objective seasonal evaluations beginning 1 November 2020 through 15 March 2021.

2. Data and Experiment Logistics

While the WWE season started out slowly, the experiment was able to capture and evaluate several impactful and large accumulation events. Observations from the National Operational Hydrologic Remote Sensing Center version 2 (NOHRSCv2[6]) snowfall accumulation can be seen in figure 1 for the WWE season. This seasonal image highlights the regions where WWE focused: the northeastern CONUS, the Central Plains into the Midwest, and the intermountain west especially over Sierra Nevada California and the eastern Rockies.

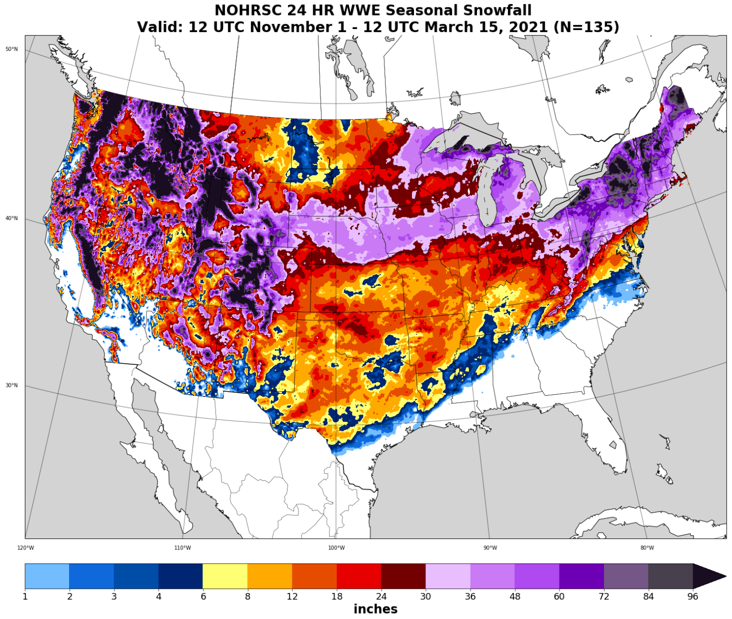

NOHRSC 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=135)
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
1
2
3
4
6
8
12
18
24
30
36
48
60
72
84
96
inches

Figure 1: WWE seasonal snowfall accumulation from NOHRSCv2.
The WWE season ran 1 November 2020 - 15 March 2021.

Experiment Logistics, Evaluation, and Participation

Following the previous years' success utilizing the remote format and the ongoing COVID-19 global pandemic, this year ran with twice weekly remote sessions over the full winter season beginning Tuesday, 17 November 2020 and ending Wednesday, 10 March 2021. Experiment participants and field representatives were asked to join on Tuesdays (10:30 am-12:00 pm EST) and Wednesdays (10:30 am-12:00 EST) throughout the season. This perspective allowed the WWE to subjectively evaluate events before they occurred as a 'pre-event' evaluation and then

[6] NOHRSCv2 snowfall analysis can be found here: https://www.nohrsc.noaa.gov/snowfall/

again after as a ‘post-event’ evaluation. The results of this pre- and post-event subjective evaluation will be discussed in later sections.

Throughout the WWE season, the participants evaluated 10 events during the Tuesday/Wednesday sessions and an additional 8 events during the intensive weeks. Several of these events were evaluated from the day 3 and day 2 perspective including Case 7, Case 11, Case 14, and Case 15. Table 1 lists the initialization and valid times for the cases as well as the region of focus and whether it was evaluated as a live event or as a retrospective case.

Table 1: List of cases that were evaluated in the WWE.

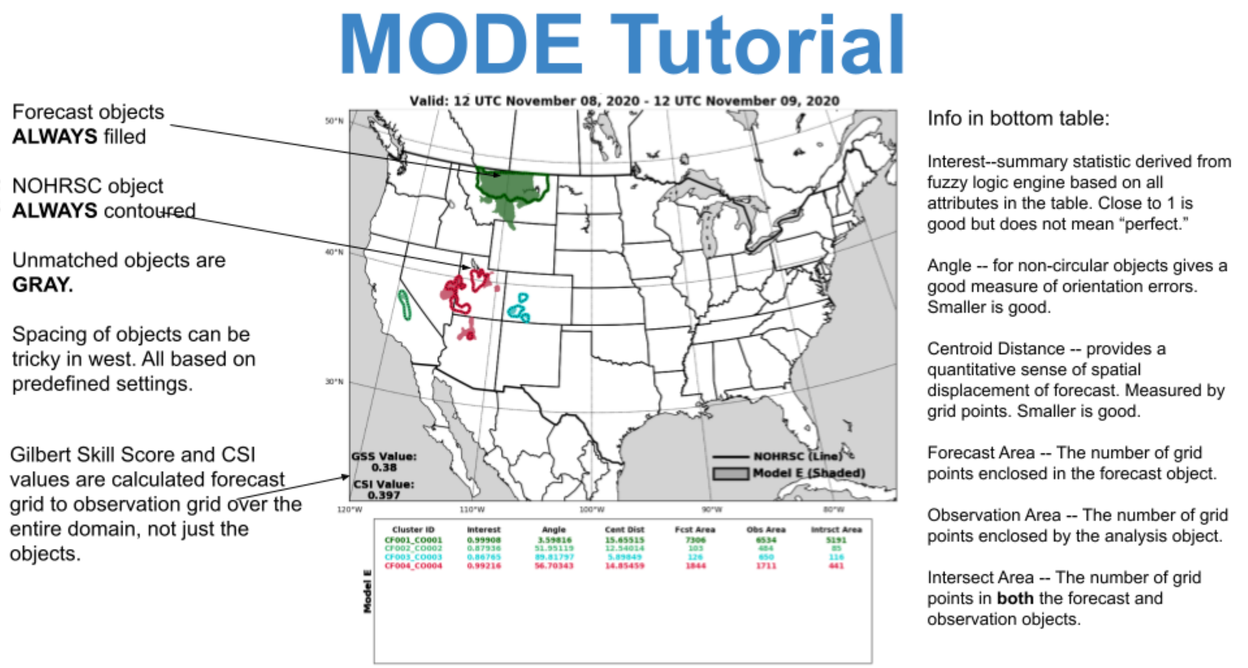
	Initialize Date (00z model run)	Valid End Date (24 hour period 12z - 12z)	Live or Retrospective	Region of Focus
Case 1	7 November	9 November	Retrospective	Montana/Western CONUS
Case 2	1 December	3 December	Live	Kansas/Oklahoma
Case 3	8 December	10 December	Live	Northeastern CONUS
Case 4	15 December	17 December	Live	Mid-Atlantic
Case 5	5 January	7 January	Live	Central CONUS
Case 6	10 December	12 December	Retrospective	Central CONUS/Midwest
Case 7	25 January 26 January	28 January	Retrospective Live	California/Western CONUS
Case 8	2 February	4 February	Live	Colorado/Great Plains
Case 9	9 February	11 February	Live	Central/Eastern CONUS
Case 10	23 February	25 February	Live	Colorado
Case 11	21 December 22 December	24 December	Retrospective	Upper Midwest
Case 12	23 December	25 December	Retrospective	Great Lakes
Case 13	29 November	1 December	Retrospective	Great Lakes
Case 14	29 January 30 January	1 February	Retrospective	Mid-Atlantic
Case 15	27 December 28 December	30 December	Retrospective	Midwest
Case 16	27 December	29 December	Retrospective	Western CONUS
Case 17	29 December	31 December	Retrospective	Texas
Case 18	9 January	11 January	Retrospective	Southern CONUS

During the Tuesday pre-event sessions participants were asked, through the use of surveys, to subjectively rank the experimental deterministic 24 hour snowfall totals, from best to worst, over the region of interest. Next, based on the experimental deterministic and probabilistic information, the participants were asked to draw polygon contours on a digital map for the snowfall footprint (typically the 1" contour), the highest accumulation amount based on their confidence in the forecasts, and mark where the largest precipitation rates will occur within their snowfall footprint. This activity, named the Maximum Snowfall and Timing Product (MSTP), is adapted from the Flash Flood and Intense Rainfall (FFaIR) experiment's Maximum Rainfall and Timing Product (MRTP). The survey questions included questions about the start time and maximum duration of snowfall inside the footprint. For more information about the MSTP or MRTP products, please visit the science and operations plan for either the 2020 FFaIR experiment or the 11th annual WWE.

The post-event evaluations focus on the prior week's pre-event activity. Participants were again asked via survey to provide a subjective ranking and comments of how well the experiment data performed from best to worst. However, with this evaluation, they had access to the NOHRSCv2 snowfall accumulations as well as Method for Object-Based Diagnostic Evaluation (MODE) statistics to aid in the ranking process[7]. MODE is part of the Model Evaluation Tools (MET) package[8] (Bullock 2016) and is the main objective verification system used by HMT to provide both event and seasonal statistics for evaluating the experimental datasets; see Appendix A for the specific MODE configuration used in WWE. The objective evaluation from MODE was computed on the 24 hour snowfall accumulations at the one, two, four, six, eight, and 12 inch thresholds. A slide that was presented during WWE and outlines what information is provided from the MODE maps can be found in figure 2.

[7] WWE MODE verification page can be found here: https://origin.wpc.ncep.noaa.gov/hmt/wwe2021/mode/wwemode.php

[8] Information on MET can be found at the Developmental Testbed Center website: https://dtcenter.org/community-code/model-evaluation-tools-met.



MODE Tutorial

Valid: 12 UTC November 08, 2020 - 12 UTC November 09, 2020

Forecast objects **ALWAYS** filled

NOHRSC object **ALWAYS** contoured

Unmatched objects are **GRAY.**

Spacing of objects can be tricky in west. All based on predefined settings.

Gilbert Skill Score and CSI values are calculated forecast grid to observation grid over the entire domain, not just the objects.

GSS Value: 0.38

CSI Value: 0.397

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.99908	3.59816	15.65515	7306	6534	5191
CF002_CO002	0.87936	51.95119	12.54014	103	484	85
CF003_CO003	0.86765	89.81797	5.89849	126	650	116
CF004_CO004	0.99216	56.70343	14.85459	1844	1711	441

Info in bottom table:

Interest--summary statistic derived from fuzzy logic engine based on all attributes in the table. Close to 1 is good but does not mean "perfect."

Angle -- for non-circular objects gives a good measure of orientation errors. Smaller is good.

Centroid Distance -- provides a quantitative sense of spatial displacement of forecast. Measured by grid points. Smaller is good.

Forecast Area -- The number of grid points enclosed in the forecast object.

Observation Area -- The number of grid points enclosed by the analysis object.

Intersect Area -- The number of grid points in **both** the forecast and observation objects.

Figure 2: MODE tutorial slide presented during WWE.

Along with the MODE maps showing the experimental and NOHRSCv2 spatial comparisons, participants were also given access to performance diagrams derived from the statistics output from MODE. Figure 3 provides the tutorial for how to interpret the diagrams which were computed at the same 24 hour snowfall accumulation thresholds as the spatial MODE maps. Within the subjective surveys, participants were asked if and how they utilized either the MODE maps or performance diagrams. The WWE team is interested in what type of objective information is useful and whether or not that information changes with storm type or location. The MSTPs were also discussed during these post-event sessions where participant drawings were verified in a comparison to NOHRSCv2 (for snowfall amounts) and Multi-Radar Multi-Sensor (MRMS[9]; for precipitation rate) data. Insights into both the subjective survey results and objective statistics can be found in later sections.

[9] MRMS can be found here: https://www.nssl.noaa.gov/projects/mrms/

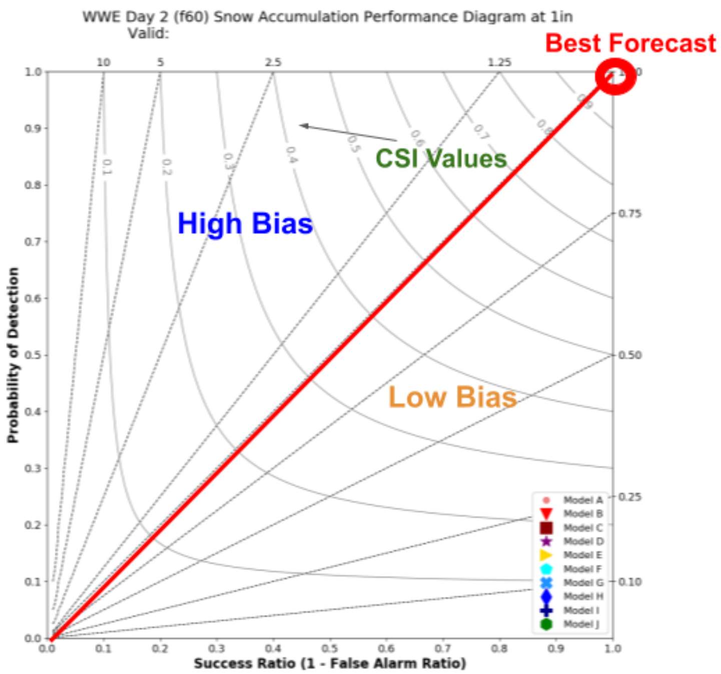


Figure 3: Roebber Performance diagram. X-axis represents the success ratio, y-axis represents the probability of detection. Dashed diagonal lines are bias values. Curved lines are CSI values. In general for a specified threshold, the closer a forecast is to the upper right corner, the better the forecast.

In addition to the weekly sessions, the WWE team hosted two intensive weeks where specific partners were invited to provide more directed evaluations for specific winter storm event types. The goal of these intensive weeks was to garner evaluations on retrospective cases that were not well represented within the weekly WWE sessions. The first ran 16-18 February 2021 and focused on Lake Effect cases (Table 1: Cases 11-14). The second intensive week ran 2-4 March 2021 and focused on cases over the Western CONUS (Table 1: Cases 15-18 and Case 7: retrospective). Each day of these weeks were structured to capture pre-event evaluation survey responses and MSTP drawings for two cases in the morning with the respective post-event evaluations occurring in the afternoon.

The final component of this year's WWE were the invited presentations. Throughout the WWE season, team members invited 12 scientists to present their work on current and future winter weather challenges. The specific presenters and their topics can be found in Table 2. The slides for each of these presentations have been archived on the WWE webpage. Feedback was positive from these presentations and will hopefully be incorporated into future WWEs.

Table 2: Invited presenters and titles from the 11th annual WWE

Date	Presenter	Title
Nov 18, 2020	Josh Kastman	Winter Storm Severity Index Development Work
Dec 2, 2020	Mike Erickson	The Weather Prediction Center’s Snowband Prototype Page
Dec 9, 2020	Keith Brewster	FV3-LAM Storm-Scale Ensemble Forecasts and Ensemble Consensus Products for the HMT Winter Weather Experiments
Dec 16, 2020	David Wright	Using Model-Based Lake Surface Conditions in the Unified Forecasting System to Improve Lake-Effect Snowfall Forecasts
Jan 6, 2021	Sarah Perfater	The NWS Winter Progra: Enabling Innovation to Achieve Consistent, Collaborated Products and Messaging
Jan 20, 2021	Jacob Radford	Verification and Visualization of HREF Snowband Forecasts
Jan 27, 2021	Massey Bartollini	Evaluating Stochastic Parameter Perturbations in Ensemble Forecasts of Mesoscale Winter Precipitation Events
Feb 3, 2021	Heather Reeves	Spectral Bin Classification of Hydrometeor Phase as a Means to Fulfill Emerging FAA Requirements
Feb 10, 2021	Kim Elmore	More mPING’ery
Feb 17, 2021	Phil Schumacher	Incorporating Probabilistic Information into Winter Storm Services
Feb 24, 2021	Julie Demuth	Development of CAM Ensemble-derived Winter Weather Timing Guidance for Forecasters and Partners
March 3, 2021	Mike Wessler	Multivariate Snow-to-Liquid Ratio Forecasts in the Weathern United States

One of the main goals of the experiment this year was to increase CONUS wide participation. Figure 4 maps out the participation numbers for this year. Overall, there was a dramatic increase in the number of participants over previous years. The team especially wants to acknowledge the frequency of participation from the Pueblo CO, Gaylord MI, Binghamton NY, Detroit MI, and Chanhassen MN weather forecast offices (WFOs). There was also an increase in regional level participants with the eastern region NWS headquarters (ERH) and the Mid-Atlantic River Forecast Center (MARFC) engaging in a few sessions. On average, each Tuesday/Wednesday session boasted around 10-15 partners which is a remarkable improvement over last year’s WWE which saw a few remote sessions with zero to one partner.

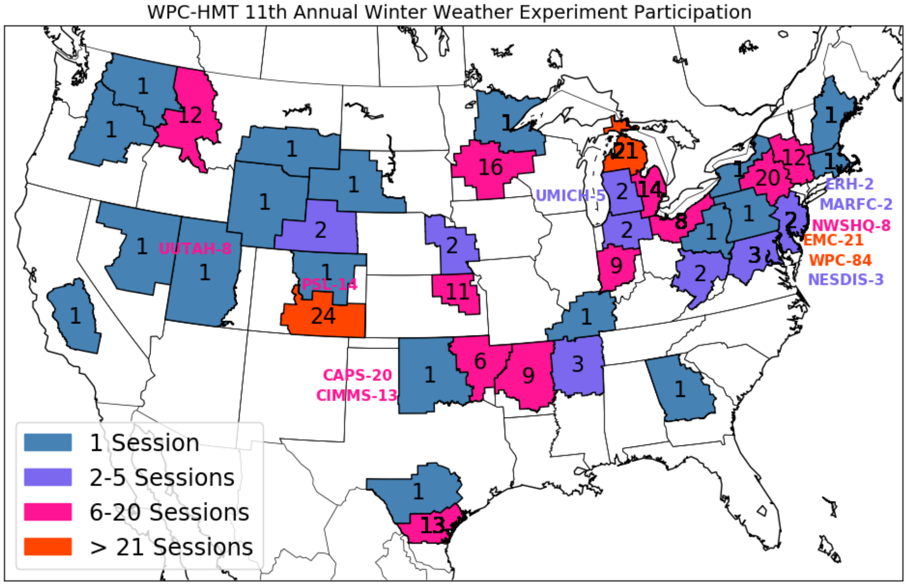


Figure 4: Participation map showing the number of sessions attended by WFO (outlined areas), National Labs, National NWS Centers, Universities, NWS headquarters (NWSHQ), and other partners.

Data Overview

WWE participants evaluated a suite of experimental data centered around the FV3-LAM which is summarized in Table 3. To limit any inherent biases to specific configurations, each model was presented anonymously as 'Model A', 'Model B', etc. Details for each configuration as well as its anonymous designation used in WWE can be found in the following subsections.

Table 3: Model guidance that was evaluated in the 11th Annual WWE

Model	WWE Designation	Provider	Resolution	Forecast Hours
FV3-LAM (3 members)	Models G - I	EMC	3km	60
FV3-LAM Ensemble (5 members)	Models B - F	CAPS	3km	84
GFSv15	Model J	Utah	2.5km	84
NBMv4.0 (deterministic & probabilistic)	Model A	MDL	2.5km	84
PWPF (probabilistic only)	n/a	WPC	5km	72

FV3-LAM Configurations

There were eight FV3-LAM configurations evaluated during WWE: three provided by the Environmental Modeling Center (EMC) and five provided by the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma. For both systems, only the 00UTC initialization was archived and evaluated.

All of EMC's simulations were configured with the same microphysics (Thompson; Thomson et al. 2008), planetary boundary layer (PBL; MYNN, Olson et al. 2019), surface layer (MYNN, Olson et al. 2019), and land surface model (LSM; NOAH, Tewari et al 2004). The control simulation (hereafter referred to as EMC FV3-LAM) was run on the operational CONUS High Resolution Rapid Refresh (HRRR) grid outlined in yellow in figure 5. The second configuration is similar to the control configuration but includes an hourly data assimilation cycle over a 6 hour period prior to issuance of a free forecast (EMC FV3-LAMDA). The third EMC configuration is branded as EMC FV3-LAMX which ran the configuration of the EMC FV3-LAM on the proposed RRFS grid (Fig 5: white box).

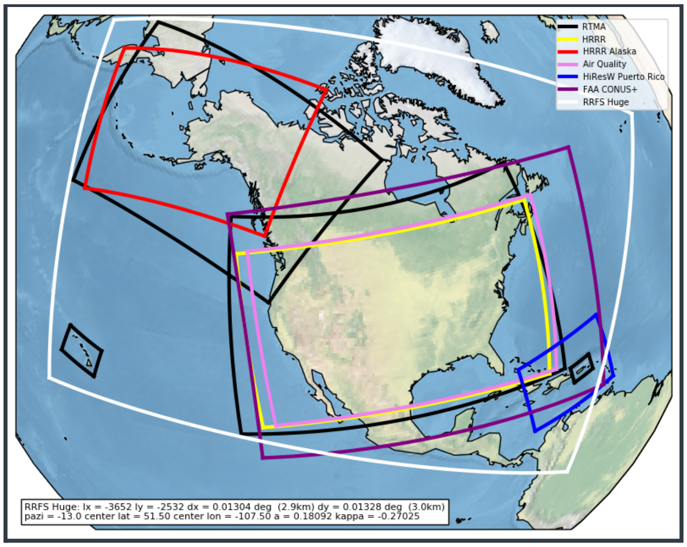

RTMA
HRRR
HRRR Alaska
Air Quality
HiResW Puerto Rico
FAA CONUS+
RRFS Huge
RRFS Huge: lx = -3652 ly = -2532 dx = 0.01304 deg (2.9km) dy = 0.01328 deg (3.0km)
pazi = -13.0 center lat = 51.50 center lon = -107.50 a = 0.18092 kappa = -0.27025

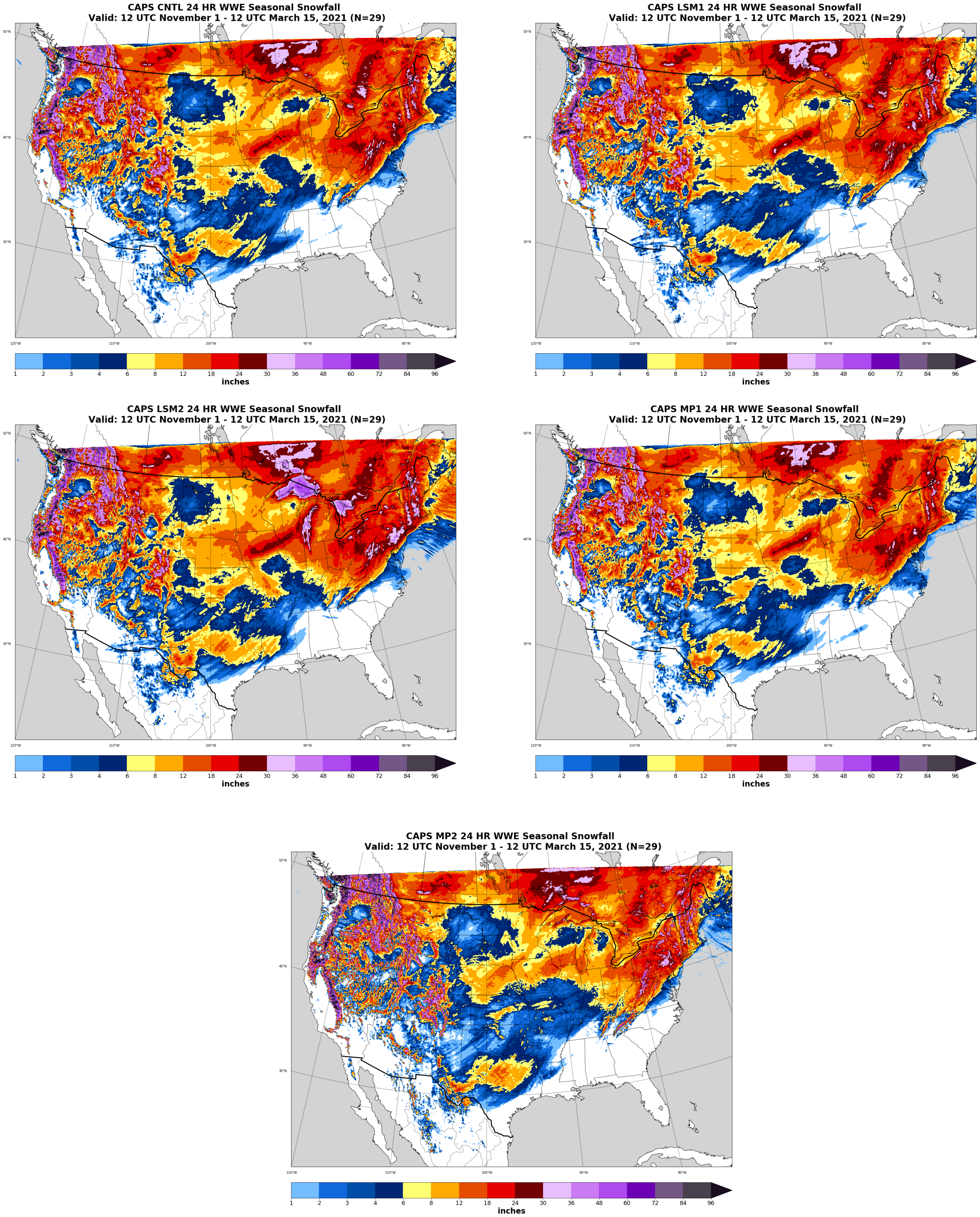
Figure 5: FV3-LAM domain grids.

The CAPS ensemble ran the same control configuration as the EMC FV3-LAM. It then modified the microphysics scheme (Thompson; NSSL, Mansell et al, 2010; or Ferrier-Aligo, Aligo et al. 2014), PBL (MYNN, K-EDMF, or TKE-EDMF), surface layer (MYNN or GFS), or LSM (NOAH, RUC, or NOAHMP) to provide the four additional ensemble members. WWE evaluated each of these members separately as it was determined in previous WWEs that this ensemble setup does not provide sufficient spread to evaluate the system ensemble averaging techniques. Naming conventions as well as configuration details for each of the CAPS configurations and EMC can be found in table 4.

Table 4: Summary of all FV3-LAM configurations evaluated by WWE

Model Name	WWE Designation	Microphysics	PBL	Surface Layer	LSM
EMC FV3-LAM EMC FV3-LAMDA EMC FV3-LAMX CAPS_CNTL	Model G Model H Model I Model B	Thompson	MYNN	MYNN	NOAH
CAPS_MP1	Model E	NSSL	MYNN	MYNN	NOAH
CAPS_MP2	Model F	Ferrier-Aligo	K-EDMF	GFS	NOAH
CAPS_LSM1	Model C	Thompson	MYNN	MYNN	RUC
CAPS_LSM2	Model D	Thompson	TKE-EDMF	GFS	NOAHMP

The CAPS ensemble was run weekly out to 84 hours for the Tuesday evaluations or specific retrospective dates as requested by the WWE team. In total, there were 29 events where the CAPS ensemble was run. The total 24 hour snowfall accumulations at day 2 over all of the CAPS ensemble member events can be seen in figure 6.



CAPS CNTL 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=29)

CAPS LSM1 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=29)

CAPS LSM2 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=29)

CAPS MP1 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=29)

CAPS MP2 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=29)

Figure 6: WWE case 24 hour snowfall accumulations for day 2 in inches for the five CAPS FV3-LAM configurations. CAPS_CNTL in upper left, CAPS_LSM1 in upper right, CAPS_LSM2 in middle left, CAPS_MP1 in middle right, CAPS_MP2 in lower panel.

For EMC, each system was run daily out to 60 hours. Due to some computing issues, there are a total of 10 (EMC FV3-LAM), 20 (EMC FV3-LAMDA), and 29 (EMC FV3-LAMX) missing days within the archive throughout the WWE season from 1 November 2020 - 15 March 2021. The total 24 hour snowfall accumulations from day 2 for the WWE season of the EMC configurations can be seen in figure 7.

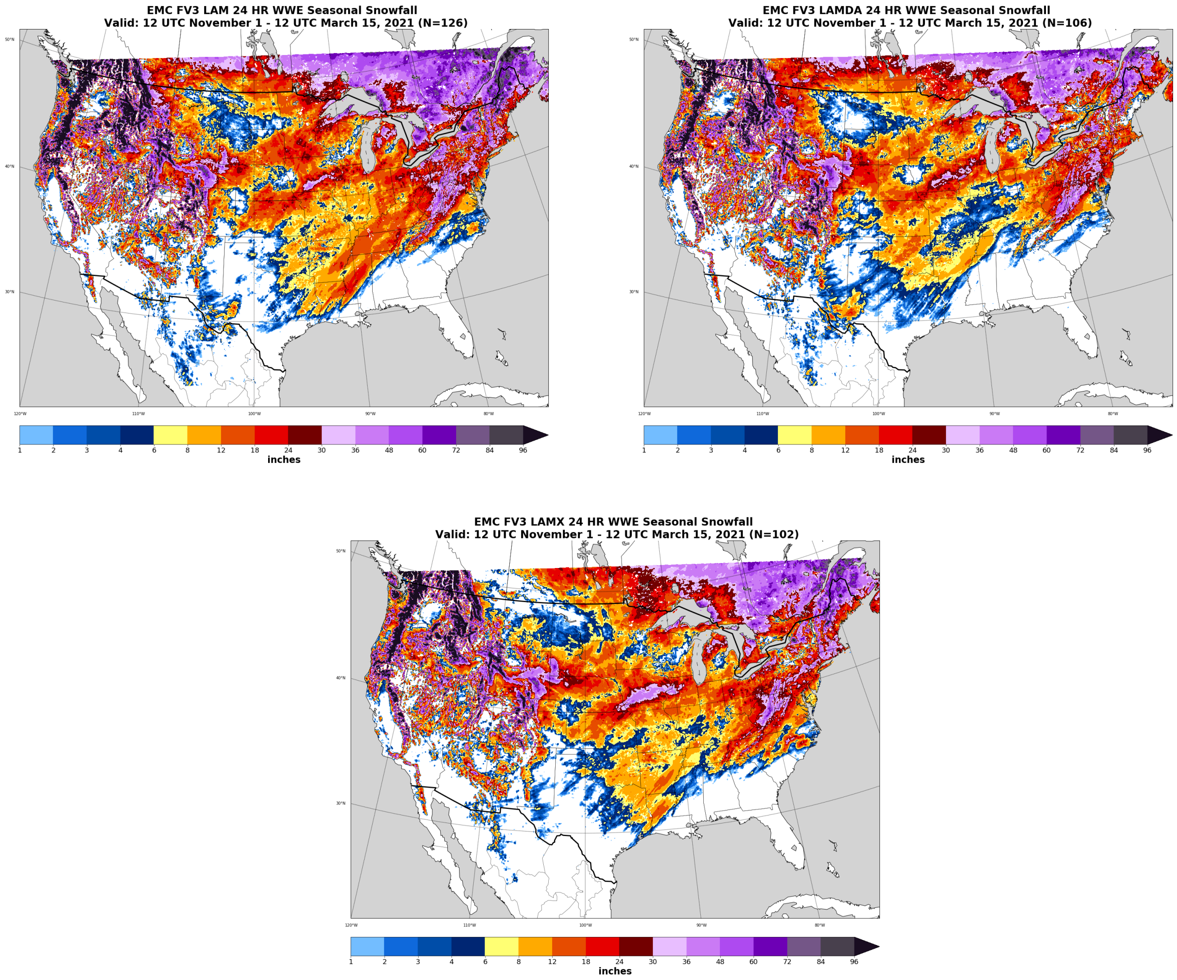

EMC FV3 LAM 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=126)
EMC FV3 LAMDA 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=106)
EMC FV3 LAMX 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=102)
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
1
2
3
4
6
8
12
18
24
30
36
48
60
72
84
96
inches

Figure 7: WWE seasonal 24 hour snowfall accumulation at day 2 in inches for the three EMC FV3-LAM configurations. EMC FV3-LAM in upper left panel, EMC FV3-LAMDA in upper right panel, and EMC FV3-LAMX in lower panel.

Details on the specific weekly cases and seasonal evaluations can be found in later sections. It should also be noted that the snow-to-liquid ratio (SLR) for all of the FV3-LAMs was set to 10:1. The WWE team agrees that this is not an ideal solution to SLR and introduces some error into the evaluation process; however, the focus of the experiment this year was on the utility of the CAMs in the day 2 and day 3 time periods. Future WWEs hope to include more dynamic SLRs as they are planned in future CAM and RRFS developments.

GFSv15 Downscaled from University of Utah

This year again featured gridded, high resolution snowfall forecasts over the western CONUS from the University of Utah. These downscaled snowfall amounts west of 100°W were evaluated alongside the experimental FV3-LAM data suite to gauge how well this methodology compares to CAMs. Below outlines the downscaling methods that were applied to the operational GFSv15. It should be noted that this is the same methodology that was used in last year's WWE.

1. Interpolating wet-bulb temperatures to an 800-m grid and determining precipitation type based on the profile of wet-bulb temperature.
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).
3. Applying snow-to-liquid ratio algorithms based on historical relationships based from Alta, UT.

A complete archive between 1 November 2020 - 15 March 2021 was collected for this dataset and the 24 hour snowfall accumulations at day 2 can be seen in figure 8 below.

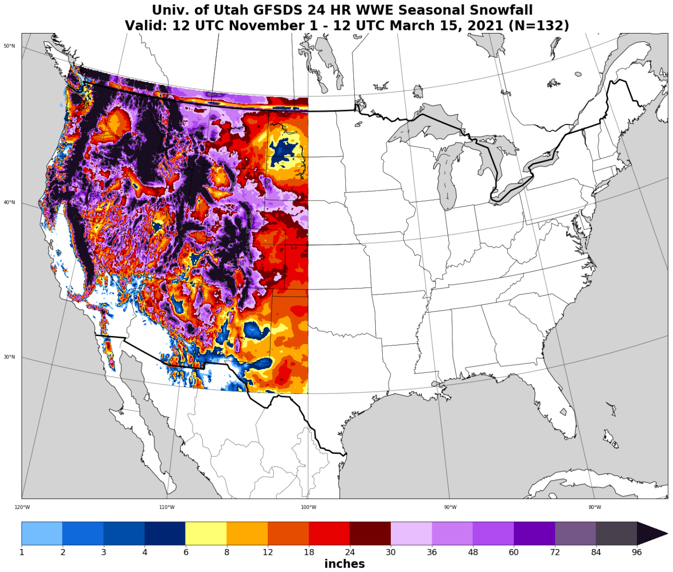

Univ. of Utah GFSDS 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=132)
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
1
2
3
4
6
8
12
18
24
30
36
48
60
72
84
96
inches

Figure 8: WWE seasonal 24 hour snowfall accumulation at day 2 in inches of the downscaled GFSv15 from the University of Utah. This is a complete archive from 1 November 2020 - 15 March 2021.

It should be noted that the science plan included downscaled GFSv15 snowfall using machine learned techniques; however, the Utah team was unable to get this updated algorithm working in time for this year's WWE. The updated technique will be included in next year's WWE science plans.

National Blend of Models version 4.0 (NBMv4.0)

The NBM is a "nationally consistent and skillful suite of calibrated forecast guidance based on a blend of both NWS and non-NWS numerical weather prediction model data and post-processed model guidance" (NBM Webpage[10]). It was created in an effort to help NWS meteorologists more efficiently create their forecasts by providing a consistent starting point for forecasters across the NWS. Version 4.0 of the NBM became operational in the fall of 2020. Within WWE, participants were asked to evaluate the deterministic NBM snowfall with the CAMs and use the 24 hour probabilities at various snowfall thresholds for the web drawing activity. The seasonal 24 hour snowfall accumulations on day 2 for NBMv4.0 can be seen below in figure 9.

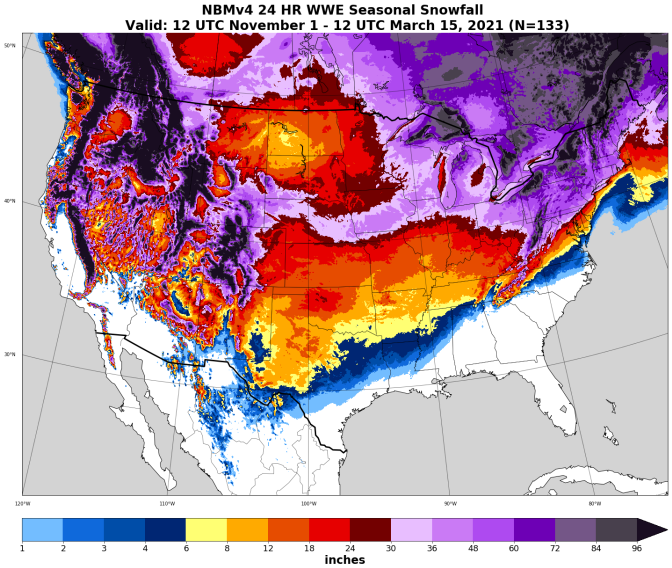

NBMv4 24 HR WWE Seasonal Snowfall
Valid: 12 UTC November 1 - 12 UTC March 15, 2021 (N=133)
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
1
2
3
4
6
8
12
18
24
30
36
48
60
72
84
96
inches

Figure 9: WWE seasonal 24 hour snowfall accumulation at day 2 in inches for the NBMv4.0. This is a near complete archive from 1 November 2020 - 15 March 2021

It should be noted that after the conclusion of the WWE season, the NBM team made the decision to revert the NBM's quantitative precipitation forecast (QPF) to the previous version (3.2) due to identified errors in the QPF.

Probabilistic Winter Precipitation Guidance (PWPF)

In addition to the NBMv4.0 probabilities, participants were also presented with the WPC Probabilistic Winter Precipitation Guidance (PWPF[11]) to help guide and inform their drawing

[10] NBMv4.0 web page: https://www.weather.gov/mdl/nbm_home

[11] PWPF web page: https://www.wpc.ncep.noaa.gov/pwpf/wwd_accum_probs.php

activity. The PWPF products are based on the deterministic WPC Winter Weather Desk (WWD) accumulation forecasts and are generated automatically using an ensemble of model forecasts along with the WWD forecasts. The automatic nature of this product generation allows an extensive set of displays of probabilities for snowfall or freezing rain exceeding a number of thresholds and accumulations of snowfall or freezing rain for various percentile levels. The percentile amounts and probabilities for 24-hour intervals are generated at six-hour increments through 72 hours.

3. Experiment Findings and Results

In addition to providing evaluations of the eight FV3-LAM configurations, the WWE focused on several science objectives this year including:

- The utility of the experimental CAM data in the day 2 and 3 time period.
- The process of conveying extreme snowfall information out of a deterministic forecast.
- Exploring the use of probabilistic data to inform forecasts and decision making.
- Comparing the experimental CAM data to downscaled efforts in Western CONUS for improvements of forecasts in higher terrain snowfall.
- Expand the interactive engagement of the remote aspects of the experiment.
- Enhance collaboration between NOAA, NCEP centers, WFOs, and academic partners on improving winter weather forecasting and messaging
- Use both event and season long verification to assess the performance of experimental data sets.

While the expanded engagement and enhanced collaboration objectives were already addressed in previous sections. The other objectives require more quantitative explorations. To assess the utility of the experimental CAMs in the day 2 and day 3 time period a combination of subjective survey results and objective performance diagrams can begin the analysis.

As stated earlier, participants evaluated cases from a pre- and post-event perspective. Specifically, they were asked to rank each deterministic solution from best to worst based on the footprint, timing, and amount of snowfall in the 24 hour forecast period. This gives the team the ability to determine which deterministic model subjectively appeared the best to the evaluators before the event took place compared to which model subjectively performed the best to the participants when compared to the observations. This method of evaluation, along with the anonymity of the model names, was viewed favorably by the people who took part in the experiment as it tested their own inherent model biases and made them really investigate what each model solution was showing.

Additionally, the format of this year's experiment allowed for further detailed evaluation by splitting events into Eastern and Western CONUS at day 2 and day 3. The following subsections show the results from the subjective surveys, the seasonal objective performance diagrams, and finally a few highlighted cases.

Subjective survey results

Eastern CONUS

The majority of cases and survey entries focused over the Eastern CONUS (east of 100°W) on day 2. Pre-event rankings are shown in Figure 10. There were a total of 145 survey entries for this event type with nine deterministic solutions. The CAPS_CNTL configuration was the best ranked deterministic solution with an average ranking of 3.52 over all survey entries and a selection of first nearly 20% of the time and second over 20%. Interestingly, the EMC FV3-LAM configuration which has a similar setup (Table 4) was not ranked as highly with an average of 4.48. The full order of the rankings (averages) are: CAPS_CNTL (3.52), CAPS_LSM1 (3.85), EMC FV3-LAM (4.48), CAPS_MP1 (4.50), CAPS_LSM2 (4.69), CAPS_MP2 (4.71), NBMv4.0 (5.30), EMC FV3-LAMX (5.42), EMC FV3-LAMDA (5.52). Overall, the percent selected was relatively evenly distributed for all the models. This points to a recurring comment throughout all of the WWE sessions that in general all of these models presented similar solutions. The ranking exercise was commented to be especially challenging in the pre-events because there was little other provided information beyond the experimental snowfall amounts and the quantitative precipitation forecast (QPF) amounts.

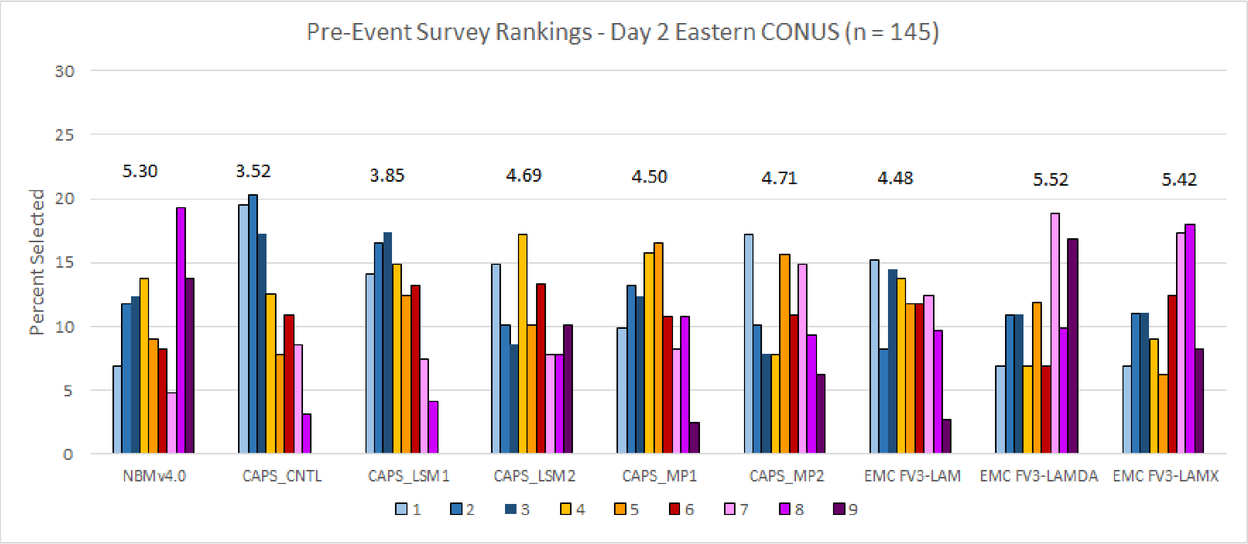


Figure 10: Pre-event survey results from Day 2 in the Eastern CONUS. There were a total of 145 survey entries for this ranking. Color bars represent the rank order with 1 being the best and 9 being the worst. Y-axis shows the percent that ranking was selected in the survey. Number above the bars indicates the average ranking value over all survey entries.

Post-event rankings for the day 2 Eastern CONUS cases are shown in Figure 11. There were a similar number of survey submissions to the pre-event evaluations with 143 total. The largest difference with this evaluation is that participants had access to the NOHRSCv2 observations as well as a suite of MODE statistics and performance diagrams. While on average the CAPS_CNTL configuration again has the best ranking with a value of 3.70, it did not have the highest percent selected as first or second. The CAPS_LSM2 solution was chosen almost 25% of the time as the best with CAPS_LSM1 selected almost 25% of the time as the second best. Another interesting outcome is the EMC FV3-LAMDA and EMC FV3-LAMX configurations ranked on average equally as poor with anywhere between 18 - 26% of the time as last or second to last. The full order of rankings (averages) from best to worst are as follows: CAPS_CNTL (3.70), CAPS_LSM1 (3.75), CAPS_LSM2 (4.32), NBMv4.0 (4.47), CAPS_MP1 (5.00), EMC FV3-LAM (5.26), CAPS_MP2 (5.97), EMC FV3-LAMX (6.01), and EMC FV3-LAMDA (6.02). The largest change in ranking comes from the EMC FV3-LAM which was third in the pre-event evaluation and falls to sixth in the post-event. Participants found that the post-event rankings were easier to achieve due to the extra available information. This is why the percent selected values are generally more skewed toward either better, middle, or lower rankings.

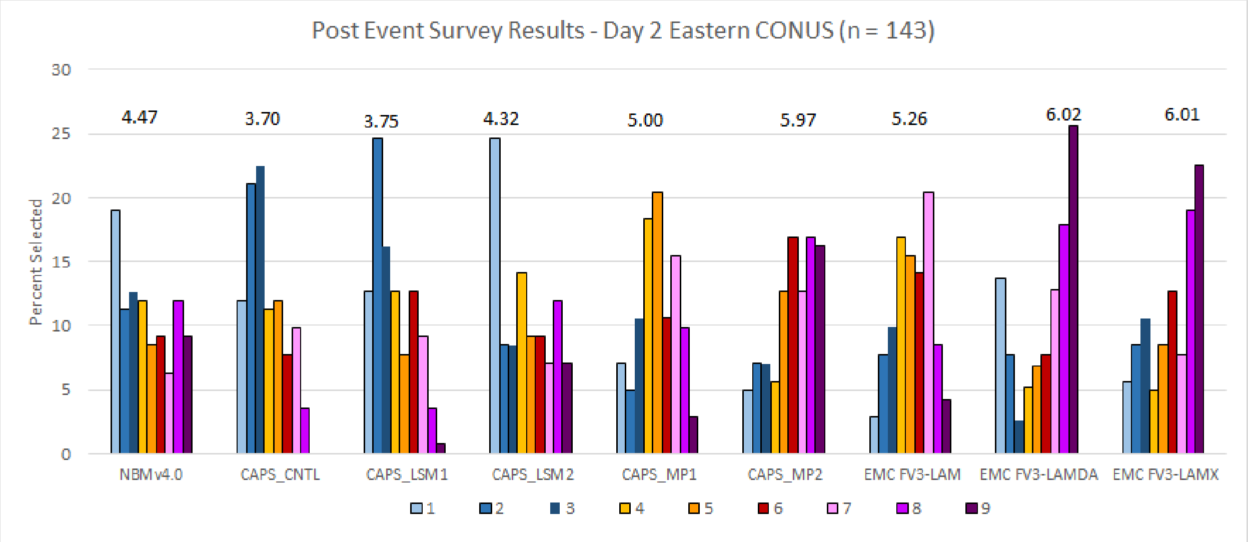


Figure 11: Post-event survey results from Day 2 in the Eastern CONUS. There were a total of 143 survey entries for this ranking. Color bars represent the rank order with 1 being the best and 9 being the worst. Y-axis shows the percent that ranking was selected in the survey. Number above the bars indicates the average ranking value over all survey entries.

Participants were also asked to provide written comments as to their rankings as summarized in Figure 12 as a word cloud. While many noted that for some cases the similar solutions made the evaluation difficult, the deciding factor for the rankings in the pre-event (figure 12; left) was snowfall amount followed by the footprint. There were also many comments on banding features and, when applicable, the capturing of any Lake Effect/Enhancement occurring. In summarizing the comments based on the post-event rankings (figure 12; right), participants shifted focus from the snowfall amount to the snowfall footprint. Remember that for these post-event rankings they had access to the NOHRSCv2 observations as well as objective MODE statistics and performance diagrams. This may be the reason the focus shifted since MODE is object oriented so it is easy to compare the footprint of the experiment data to the footprint of the observations. In addition, participants still commented frequently on the position of any banding structures.

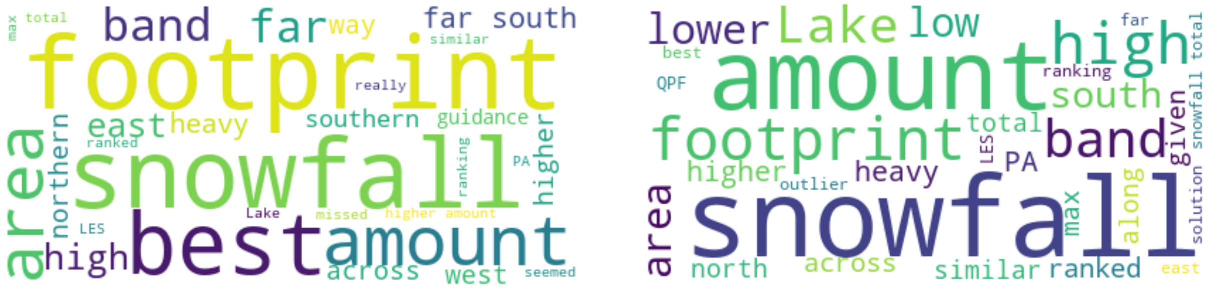


Figure 12: Word Cloud generated from Eastern CONUS, Day 2, Pre-Event (Left) and Post-Event (Right) Survey ranking comments.

As part of the WWE intensive weeks, the team asked participants to focus on several cases using day 3 data. From these cases there were a total of 39 survey entries for the Eastern CONUS. It should be noted that the EMC configurations were not run to day 3, so there were only six deterministic solutions to rank. The results of these rankings can be found in Figure 13. Differing from the day 2 results, the CAPS_CNTL is not the best ranked configuration on day 3. Both the CAPS LSM configurations had the highest percent selected first with near 30% for each. CAPS_LSM2 has the best ranking of all the models with an average of 2.74. NBMv4.0 was overwhelmingly the last ranked solution both on average with a value of 4.77 and in the percent selected last at over 50% of the time. Due to operational CAM availability, the NBMv4.0 relies heavily on global model solutions for day 3. **Participants commented that the details the CAMs provided at day 3 were a swaying factor for rating them higher than the NBMv4.0. This highlights a potential forecaster bias toward solutions that provide more details whether or not those details verify correctly.** Much of the discussion during these cases and comments within the survey reflected this fact for the NBMv4.0. The rankings (averages) for the models are as follows: CAPS_LSM2 (2.74), CAPS_LSM1 (2.90), CAPS_MP1 (3.28), CAPS_CNTL (3.28), CAPS_MP2 (4.03), and NBMv4.0 (4.77).

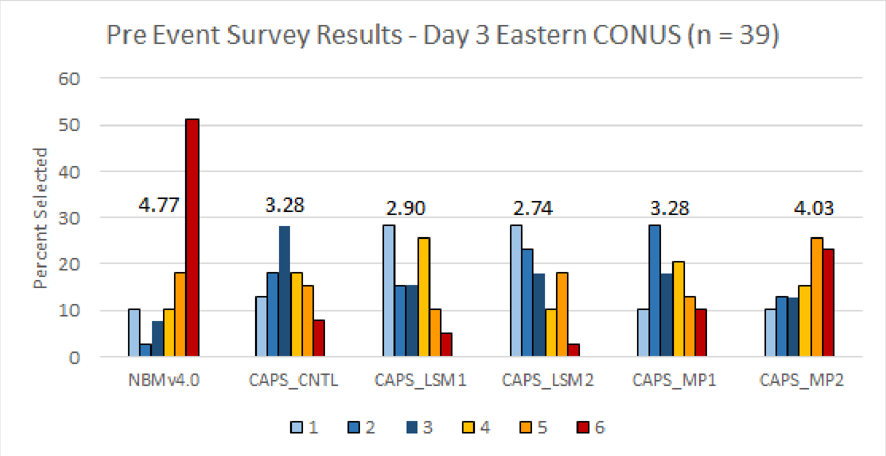


Figure 13: Pre-event survey results from Day 3 in the Eastern CONUS. There were a total of 39 survey entries for this ranking. Color bars represent the rank order with 1 being the best and 6 being the worst. Y-axis shows the percent that ranking was selected in the survey. Note, the EMC configurations are not run past day 2. Number above the bars indicates the average ranking value over all survey entries.

From a post-event perspective (Figure 14), the rankings see an interesting shift with respect to the NBMv4.0 which went from an average ranking of 4.77 to 3.66. Participants also were equally split at 29% for ranking the NBMv4.0 as either the best or worst model. Investigating the comments show this was due to the importance placed on the snowfall footprint by the individual participant. **Some found the utility in these models at day 3 comes solely from identifying where an event may occur. Others found at day 3 the utility should not only come from identifying where the event is going to occur (footprint) but also the potential for higher amounts and banding features.** CAPS_LSM2 is again the best ranked in terms of both the average ranking (2.40) and the percent selected as first over 35% of the time. CAPS_CNTL has a better ranking than the day 3 pre-event results due to having the highest percent selected as second 34% of the time. CAPS_MP2 has been consistently the worst ranked configuration for the CAPS ensemble. For this survey entry, it has the worst average ranking of 4.61 and the highest percent selected as last over 40% of the time. The full rankings (averages) for the day 3 post-event Eastern CONUS are as follows: CAPS_LSM2 (2.40), CAPS_CNTL (3.24), CAPS_MP1 (3.44), NBMv4.0 (3.66), CAPS_LSM1 (3.66), CAPS_MP2 (4.61).

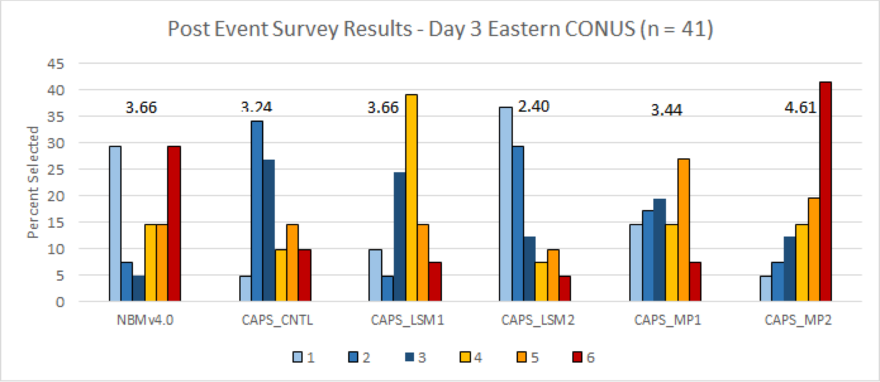


Figure 14: Post-event survey results from Day 3 in the Eastern CONUS. There were a total of 41 survey entries for this ranking. Color bars represent the rank order with 1 being the best and 6 being the worst. Y-axis shows the percent that ranking was selected in the survey. Note, the EMC configurations are not run past day 2. Number above the bars indicates the average ranking value over all survey entries.

Associated comments from these rankings focused more on snowfall amounts and any potential for banding. The lack of CAM data in the NBMv4.0 at day 3 also featured throughout the survey comments. Snowfall footprint was a predominant focus in the pre-event as seen in the word cloud in Figure 15 (left), the keywords used to describe the focus on footprint include placement, across, footprint, along, and area. When discussing the utility of CAM (or any deterministic model) at day 3 participants heavily commented that knowing the general location of snowfall was by far the most important factor followed by any indication of banding. This may account for the difference in NBMv4.0 rankings from pre- to post-event. Many participants may have gotten hung up on the details the CAM output shows, but at the longer lead times the synoptic pattern is dominant and possibly more important for messaging any potential future impacts. All of the CAMs are initialized with GFSv15 and as such are dictated by the same background synoptic initialization. The NBMv4.0 is not constrained to just a single synoptic source which seems to be an advantage for the utility our participants discovered through this exercise.

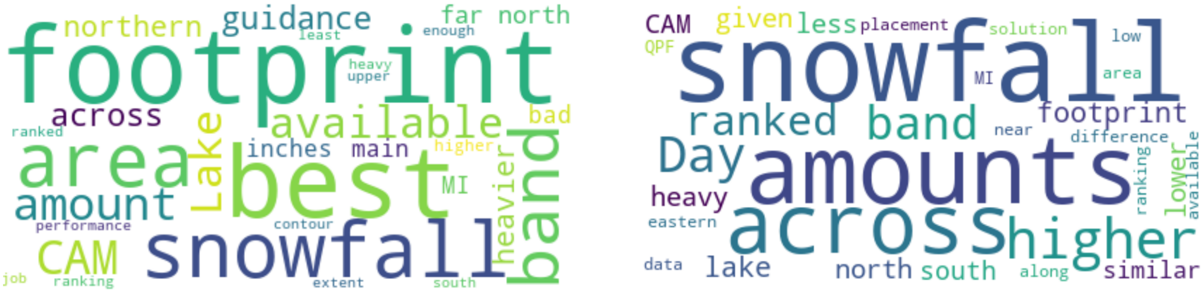


Figure 15: Word Cloud generated from Eastern CONUS, Day 3, Pre-Event (Left) and Post-Event (Right) Survey ranking comments.

Overall the subjective rankings over the Eastern CONUS favor members of the CAPS ensemble with the CAPS_CNTL member ranking first or second for three out of the four survey responses. For the EMC configurations, the FV3-LAM configuration was viewed more favorably than the FV3-LAMDA and FV3-LAMX solutions. However, all three configurations fell in the average rankings from the pre- to post-event evaluations once objective information became available. Conversely in the pre-event the NBMv4.0 was generally ranked low, however, when the post-event presented more objective statistics, the rankings improved for both day 2 and day 3. As expected, there is no clear subjective winner for deterministic experiment contributors in the Eastern CONUS. In fact, during the sessions participants often noted how close each solution footprint and amounts were to each other making this a challenging exercise.

Western CONUS

Following the same structure as the Eastern CONUS events, participants were asked to provide a best to worst ranking of the experiment models for Western CONUS events. Cases occurring west of 100°W allowed for the inclusion of the downscaled GFSv15 dataset from the University of Utah for evaluation. For the pre-event survey results on day 2, shown in figure 16, the results are quite different from the Eastern CONUS. While there were fewer events focused over the Western CONUS, there were still 97 entries submitted for the evaluation for the ten deterministic datasets. In this setup, CAPS continues to have the favored configurations. This time CAPS_LSM2 was ranked the highest with an average of 4.57. Also interesting to point out that the EMC ensemble members were ranked better than the majority of the CAPS ensemble which differs from the Eastern CONUS result. While the Utah dataset had the second lowest average ranking, it shows the highest percent selected as first at over 20% and worst at around 12%. This would indicate that participants were split on the downscaled data as they either ranked it as first or near the bottom. The full rankings (averages) for the pre-event evaluation on day 2 over the Western CONUS is as follows: CAPS_LSM2 (4.57), EMC FV3-LAMX (4.68), EMC FV3-LAM (4.73), EMC FV3-LAMDA (4.74), CAPS_MP2 (4.84), NBMv4.0 (4.97), CAPS_CNTL (5.07), CAPS_LSM1 (5.12), Utah (5.26), CAPS_MP1 (5.32). As with the Eastern CONUS the percent selected results are relatively evenly distributed with a few exceptions noted earlier.

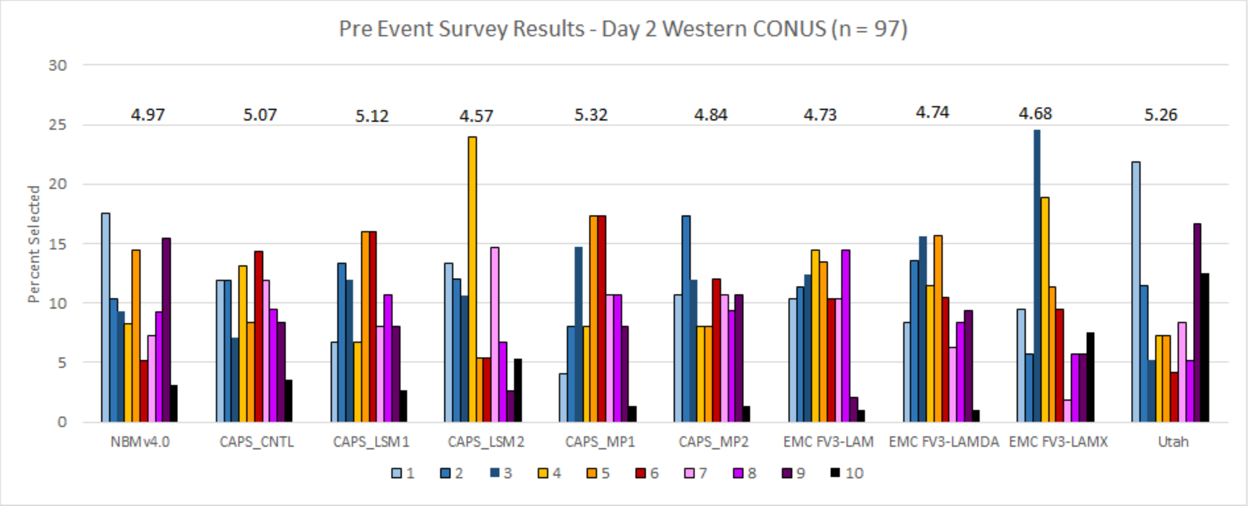


Figure 16: Pre-event survey results from Day 2 in the Western CONUS. There were a total of 97 survey entries for this ranking. Color bars represent the rank order with 1 being the best and 10 being the worst. Y-axis shows the percent that ranking was selected in the survey. Number above the bars indicates the average ranking value over all survey entries.

When the day 2 post-event survey results are examined from the 84 total entries (figure 17), it becomes clear that even though the average value isn't the best, participants were strong in selecting the Utah dataset as the best 35% of the time. As with the Eastern CONUS, the EMC configurations average ranking values fell quite drastically. For the pre-event EMC's ensemble were ranked second to fourth, here the FV3-LAMDA and FV3-LAMX are ninth and tenth respectively. Another similarity to the Eastern CONUS results is the NBMv4.0. Once participants had access to the objective statistics, the NBMv4.0 rankings increased. While the average value did not improve dramatically, it is tied for the best average ranking and has the highest percent selected as second at 25%. The CAPS_CNTL makes an appearance again as the highest average ranking tied with the NBMv4.0, although the percent selected shows it was generally selected as third or fourth place. In summary the rankings (averages) for the Western CONUS day 2 post event are as follows: NBMv4.0 (4.47), CAPS_CNTL (4.47), Utah (4.76), CAPS_LSM2 (5.13), CAPS_LSM1 (5.16), EMC FV3-LAM (5.27), CAPS_MP1 (5.53), EMC FV3-LAMDA (6.48), and EMC FV3-LAMX (6.69). Overall the percent selected distributions remain relatively even. Unlike the Eastern CONUS where the post-event results showed skewed distributions, again with the few noted exceptions.

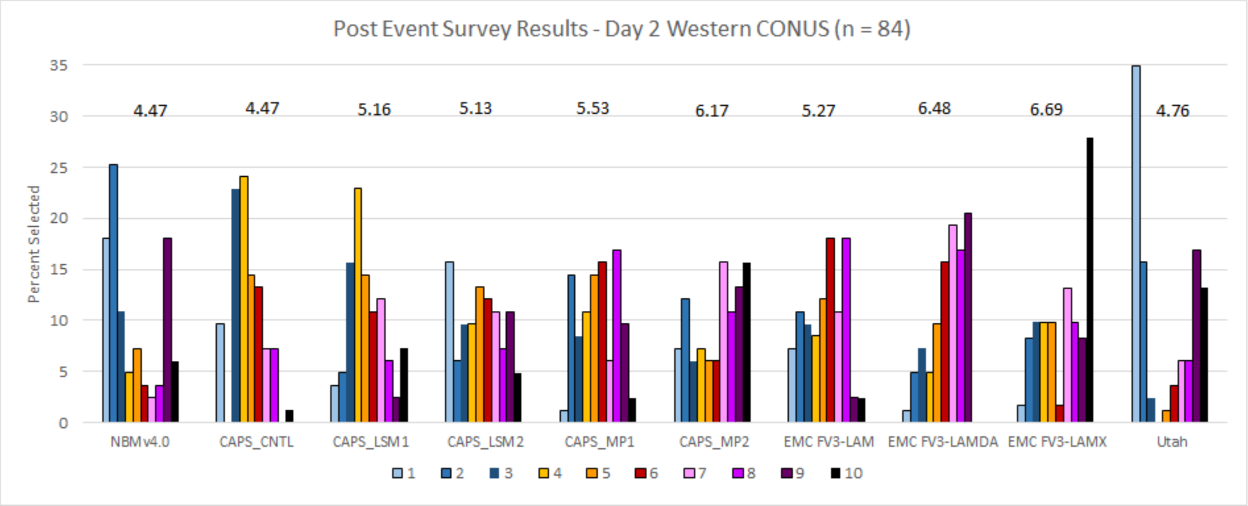


Figure 17: Post-event survey results from Day 2 in the Western CONUS. There were a total of 84 survey entries for this ranking. Color bars represent the rank order with 1 being the best and 10 being the worst. Y-axis shows the percent that ranking was selected in the survey. Number above the bars indicates the average ranking value over all survey entries.

It should be of no surprise that comments and discussion on Western CONUS cases focused on terrain and isolating which terrain features would receive the highest amounts. Within that vein, participants focused more on snowfall amounts than timing or footprint. This is reflected in the figure 20 word cloud where 'amount' features prominently. Within the sessions many noted that the MSTP activity for the Western CONUS became a 'draw the mountains' exercise and local knowledge of terrain became immensely important.

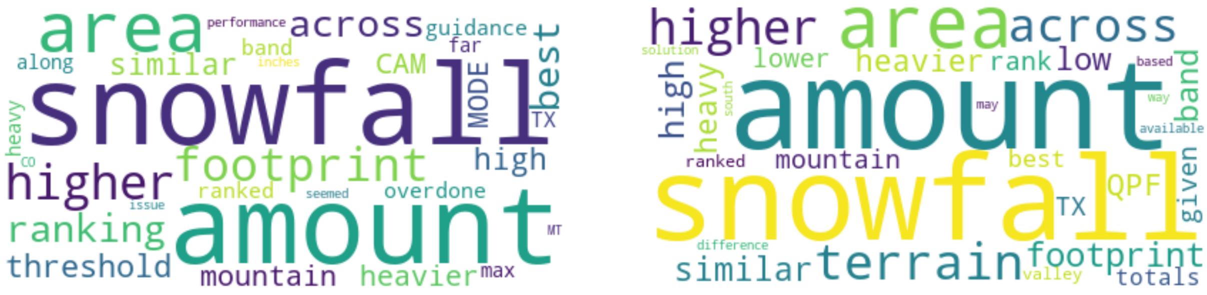


Figure 18: Word Cloud generated from Western CONUS, Day 2, Pre-Event (Left) and Post-Event (Right) Survey ranking comments

For the day 3 Western CONUS, there was only one case (Table 1; Case 7) that was examined during the WWE intensive weeks. This was due to both a lack of data and event availability. Some more details on this specific case can be found in a later section. For the survey results, there were 13 entries in both the pre- and post-event evaluations. As a reminder the EMC contributions only run through day 2 so there are only seven deterministic solutions to rank.

The average rankings for the pre-event results (figure 19) do not show any new insights as they are similar to both the day 3 Eastern CONUS (figure 13) and day 2 Western CONUS (figure 16). As with the Eastern CONUS, the NBMv4.0 is ranked the lowest in both the average ranking and percent selected last at over 45% of the time. The downscaled Utah dataset again has the highest percent selected first with a value of 38% despite being tied for the second worst average ranking. CAPS members show a mix of rankings with CAPS_LSM2 ranked best for this situation. Full rankings (averages) are as follows: CAPS_LSM2 (3.69), CAPS_LSM1 (3.85), CAPS_CNTL (3.92), CAPS_MP1 (4.23), Utah (4.38), CAPS_MP2 (4.38), and NBMv4.0 (4.92). Overall these rankings highlight the forecaster's focus on the utility of these models on day 3 is about snowfall amounts over the Western CONUS. It is well established in the West that if there is a system approaching, the terrain will see snowfall, so the utility comes from providing information on snow level and amounts.

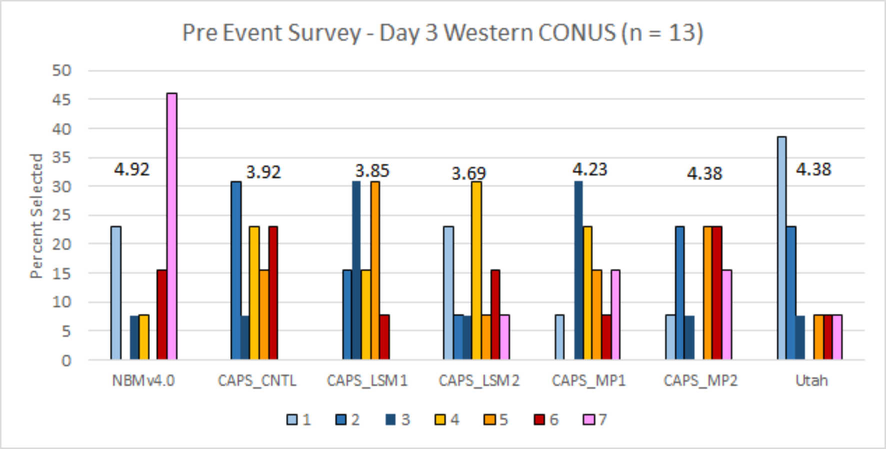


Figure 19: Pre-event survey results from Day 3 in the Western CONUS. There were a total of 13 survey entries for this ranking. Color bars represent the rank order with 1 being the best and 7 being the worst. Y-axis shows the percent that ranking was selected in the survey. Note, the EMC configurations are not run past day 2. Number above the bars indicates the average ranking value over all survey entries.

As seen with the day 3 cases over the Eastern CONUS, NBMv4.0's ranking improved dramatically from the pre-event to the post-event evaluations. Once participants have access to the objective statistics, they tend to rank the NBMv4.0 as second 39% of the time behind the Utah dataset which also drastically improved in average ranking. It remained the highest percentage selected first 48% of the time. CAPS_MP2 saw the largest fall in rankings as it was overwhelmingly (over 50%) selected as the worst for this case. As with the Eastern CONUS, these results show that there is still a high utility for global model information at day 3, with the high ranking of Utah and NBMv4.0, perhaps even more so than the CAM solutions. It also again shows the potential forecast bias toward the details the CAMs provide as once the objective information is available, their rankings skewed lower than the global model influenced solutions.

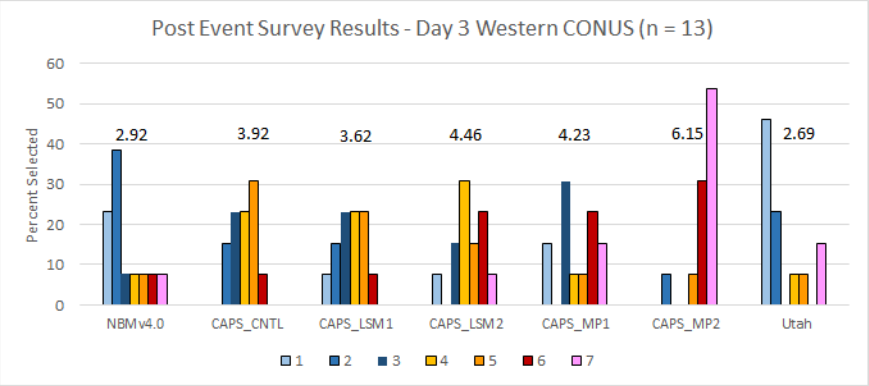


Figure 20: Post-event survey results from Day 3 in the Western CONUS. There were a total of 13 survey entries for this ranking. Color bars represent the rank order with 1 being the best and 7 being the worst. Y-axis shows the percent that ranking was selected in the survey. Note, the EMC configurations are not run past day 2. Number above the bars indicates the average ranking value over all survey entries.

This case was centered around the Sierra Nevada of California. As such, the comments reflected the focus on the snowfall amounts in the pre-event (figure 21; left) and the placement of the snowfall in the terrain in the post-event (figure 21; right). These comments and the associated discussion during the event were consistent with other day 3 and Western CONUS cases.

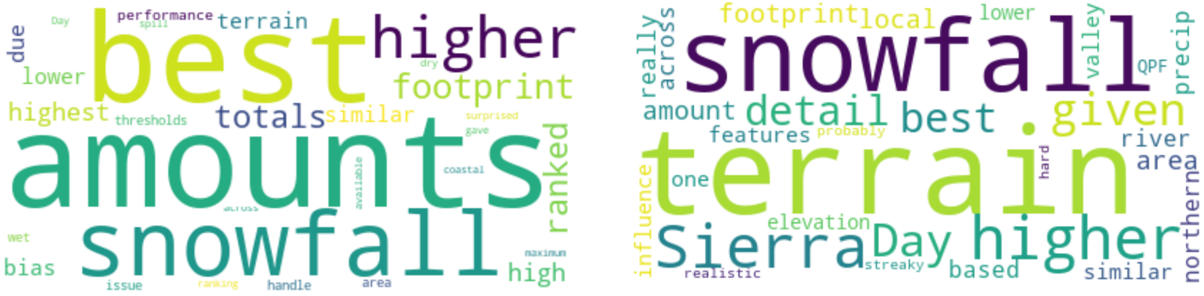


Figure 21: Word Cloud generated from Western CONUS, Day 3, Pre-Event (Left) and Post-Event (Right) Survey ranking comments

Overall the subjective evaluation over the Western CONUS showed some different preferences than the Eastern CONUS. When presented in the pre-event sessions, the EMC ensembles were ranked more favorably on day 2, however, once the objective information was available in the post-event, their rankings dropped significantly. Participants were also split on the downscaled GFSv15 from Utah with percent selections generally either first or last. This appears to be due to a high bias that will be shown in a later section. For both day 2 and day 3 the Utah data was overwhelmingly the preferred selection for first even if the average rankings did not reflect that fact. The CAPS ensemble seems to be the most consistent in terms of rankings for Western to Eastern CONUS and day 3 to day 2 with the CAPS_CNTL and the LSM configurations ranking relatively middle to high for everything. CAPS_MP1 and CAPS_MP2 are more mixed on their rankings but seem to generally fall in the middle to lower ranks. Similar to the Eastern CONUS, the NBMv4.0 rankings improved from the pre-event to post-event results. This again highlights the importance of the synoptic setup and the utility of these datasets lies in the ability to capture the general event pattern with hints at higher amounts over the terrain in the West.

One comment that was a frequent point of discussion throughout the experiment was how similar each of the models were to each other. It made ranking in the pre-event scenario difficult as participants felt there wasn't enough spread in solutions to really provide rankings. The activity was easier during the post-event evaluations thanks to the inclusion of the objective information. Exit survey comments stated that participants overall found this pre-/post-event format engaging and challenging and would be interested in seeing it continue. There was some debate on keeping the anonymity of the models since there shouldn't be too much forecaster bias if we are only testing configurations of experimental models. If future experiments include operational systems it may be worth keeping the anonymous feature, but most participants do not know the full details and biases of the individual experimental configurations. Finally, there was a lot of discussion and favorable comments about the availability of the objective metrics for the post-event. Future WWEs will continue to make the objective information available as it fits within the science objectives.

Seasonal Performance Diagrams

Performance diagrams of the experimental deterministic data over the entire WWE season support the subjective comments from the previous section. While the previous section provided evaluation over all of the WWE evaluated cases (Table 1), the performance diagrams here are completed for every available day starting 1 November 2020 through 15 March 2021. For a full detailing of missing days for each dataset please see the Data Overview section. Also, due to computational expense these diagrams were computed only on day 2 over the full CONUS and the Western CONUS as opposed to the east/west separation in the subjective evaluations.

For full CONUS, as mentioned extensively in the subjective comments, all the models are relatively similar in their position on the performance diagram seen in figure 22. This is mostly due to the fact that these are seasonal averages whereas individual cases may have greater spread between the models. Similar to previous years, as the threshold of snowfall accumulation increases from 1 inch (figure 22; upper left) to 12 inches (figure 22; lower right) CSI values decrease. This was true for all accumulation thresholds. For simplicity only the 1, 4, 8, and 12 inch thresholds are shown in the figures below. On the whole, there is a slight high bias in the models at the 1 inch threshold that disappears for the higher thresholds. In terms of individual models, the NBMv4.0 stands out as the model with the most favorable diagram position for all of the thresholds, which is reflective of the NBMv4.0s improvement in the rankings for the post-event evaluations. CAPS_MP2 consistently has a lower probability of detection (POD) and bias than the rest of the configurations. It was also ranked lower in most of the subjective evaluations. At the thresholds above 1 inch, the three EMC configurations have the highest POD and bias. The rest of the CAPS members were in the middle of the group with no stand out configurations.

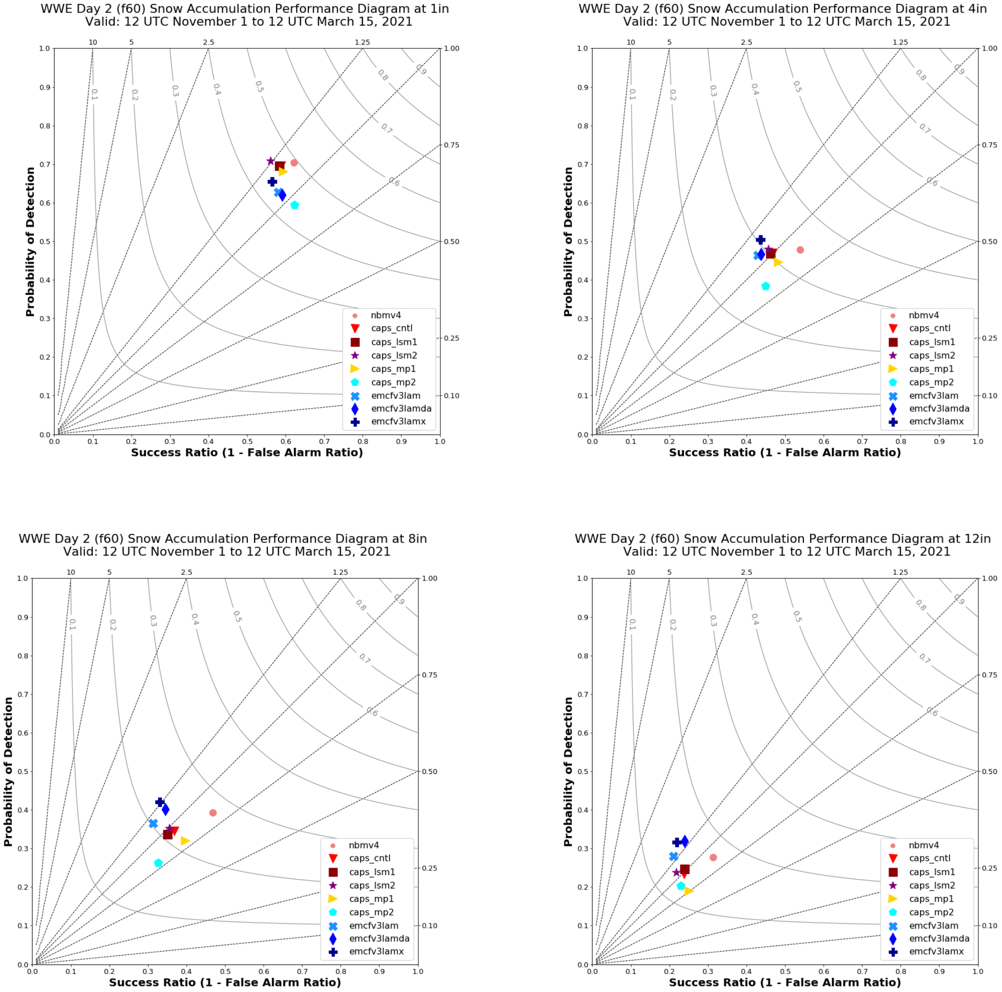


Figure 22: Seasonal, full CONUS, day 2 performance diagram at the 1 inch (upper left), 4 inch (upper right), 8 inch (lower left), and 12 inch (lower right) thresholds.

The seasonal performance diagrams for the Western CONUS domain (west of 100°W) shows a similar result to the full CONUS for the NBMv4.0 and CAM models (Figure 23). These diagrams also include the seasonal information from the downscaled GFSv15 data from the University of

Utah which stands out from the other models with a higher POD at all four thresholds. With the higher POD also comes a high bias which may account for participants ranking of either first or last for this dataset.

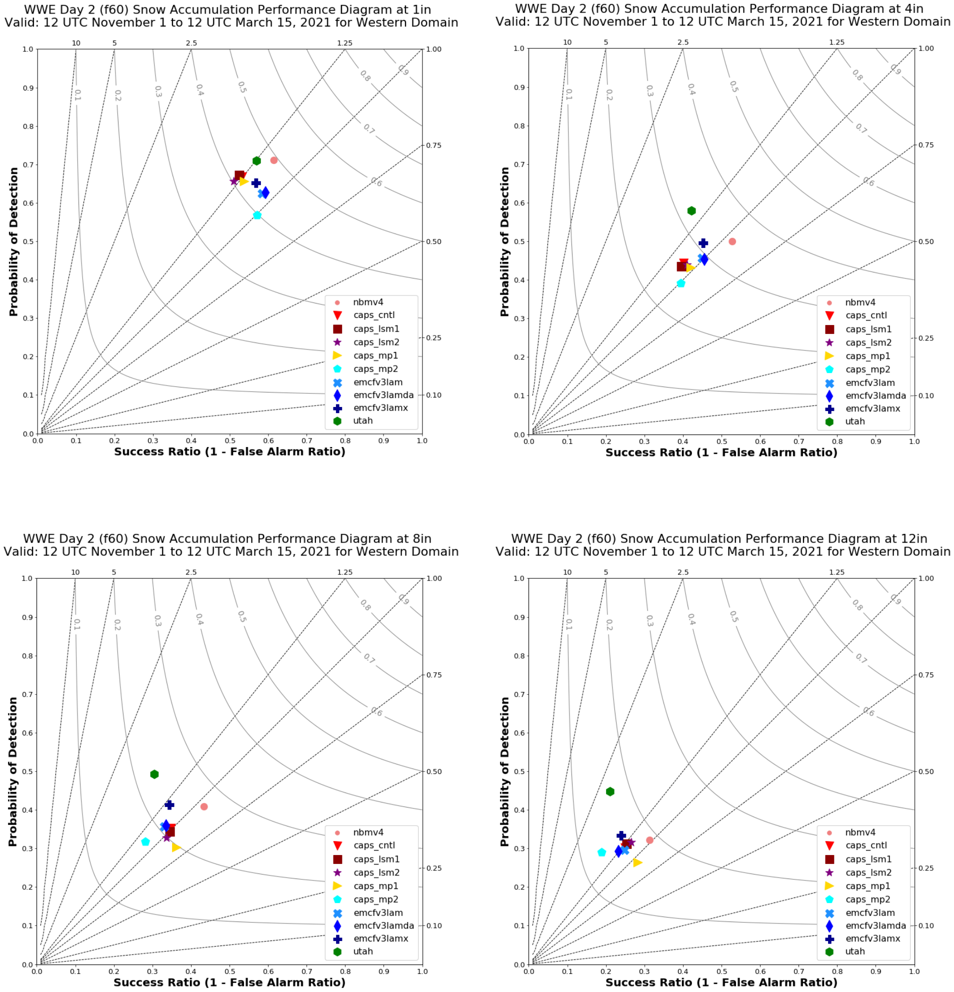


Figure 23: Seasonal, Western CONUS, day 2 performance diagram at the 1 inch (upper left), 4 inch (upper right), 8 inch (lower left), and 12 inch (lower right) thresholds.

Within the post-event survey, participants were asked to identify which objective statistics they used to inform their rankings. The team is then able to determine if there is a preference for specific objective information for day 2 or day 3 or Eastern or Western CONUS cases. Figure 24 shows the forecasters heavily favored the position on the performance diagram for everything. Bulk CONUS statistics like the Gilbert Skill Score (GSS) were used the least or not at all. The day 2 events also relied on the values within the MODE table (example in figure 2). The individual statistics provided on the performance diagram were selected less and were used on a case-by-case basis. Comments generally stated they viewed the performance diagrams as a quick summary of all the objective information and it was easy to use to inform the ranking exercise.

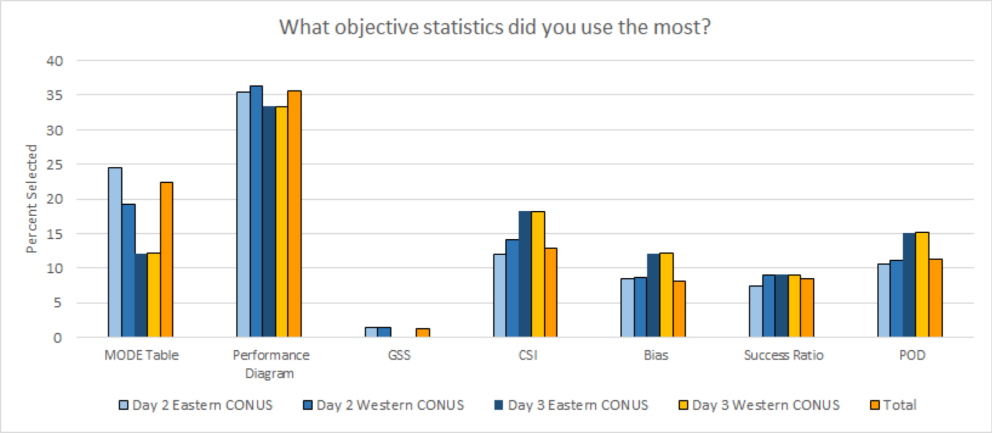


Figure 24: Subjective survey results from the post-event evaluations on which objective statistics participants use for day 2/day 3 and Easter/Western CONUS.

Maximum Snowfall and Timing Product

The MSTP product was intended to be a focusing mechanism for participants to investigate the model depictions of snowfall and then make decisions about the extent of the footprint and maximum snowfall contour. While we asked participants for timing and duration information there is no verification data oriented to complete an assessment of these forecasts. However, it was hoped that by collecting these data we could generate discussion on these elements during the evaluation process and furthermore spur research or observations in this area.

Performance diagrams of participant forecasts of trimmed domains for MSTP (on day 2 only) is shown in figure 25. For the majority of cases there was considerable variability in these footprint forecasts. On 15 of the 18 events participants, or a group thereof, achieved a POD above 0.5. In a couple of events, forecasts were biased significantly below 1 but for the most part were between 1 and 3. That is, participants were more likely to draw larger than observed footprint areas to ensure POD, but usually incurring a penalty in success ratio (SR). In situations where the snow field was large and continuous, these actions benefited the participants POD. Whereas in the mountain west, there were more SR penalties. While participants seemed to extract positive benefits from the available models, it was rather difficult at times to pick the maximum snow contour.

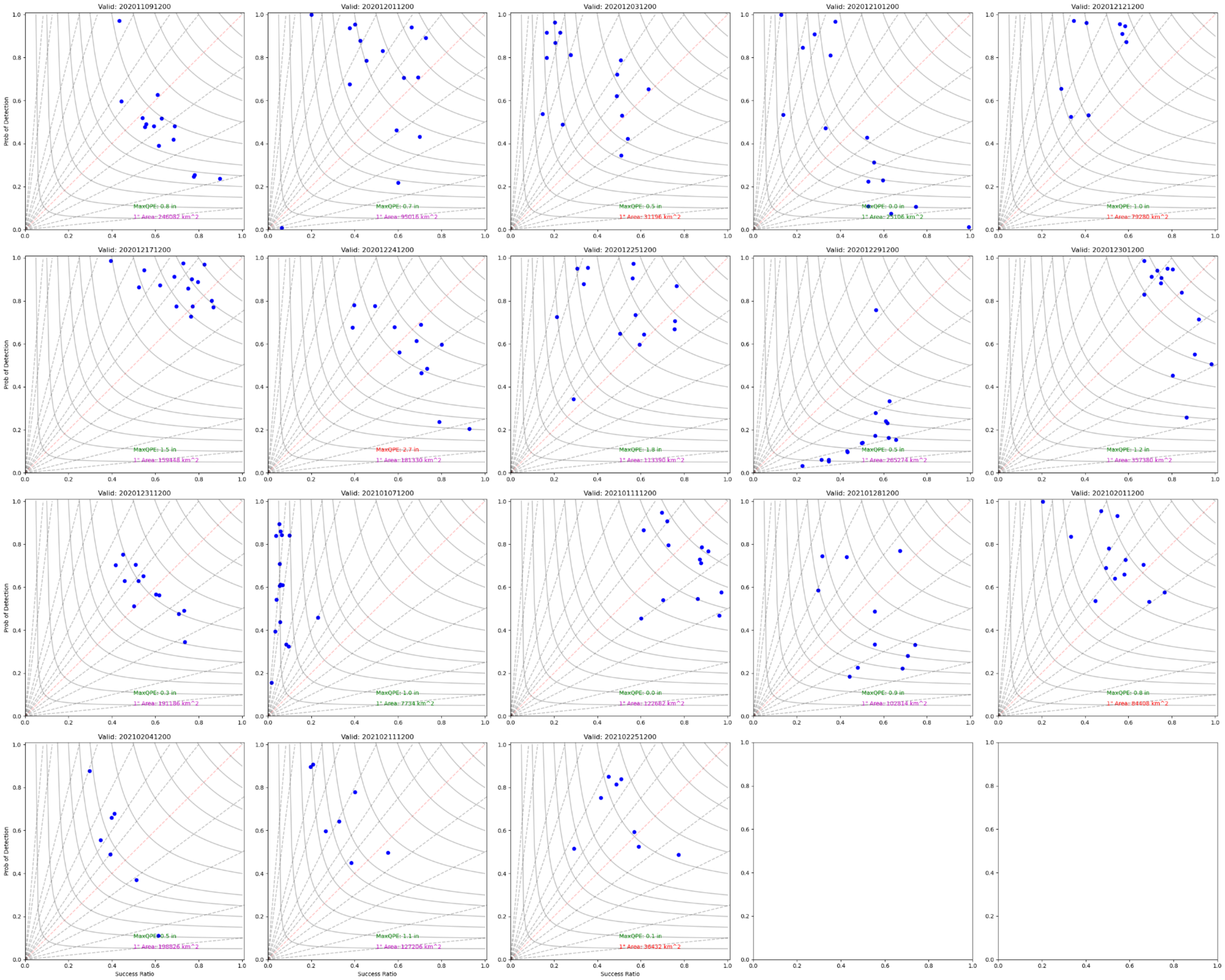


Figure 25: Performance diagrams for participant drawn MSTP for each of the 18 evaluated cases.

Histograms of participants maximum snow contour are shown in figure 26. From an aerial coverage point of view, we examined the NOHRSC snow data and chose a contour that best depicted where the most snow occurred for each event (the grey dashed line) and also depict the grid point maximum (dashed magenta line). Participants forecast maximum contours that generally agreed with the verification across most events. For some of the events, participants forecast max contours could be much higher - a possible reflection of increased confidence. This did not happen on 17 December where all but 1 participant was confident enough to forecast a max contour of over 24” of snow. Usually the participants' maximum snow contour was clustered toward the verification. In only a few events, where model forecasts were suggesting much more snow than occurred, did participants significantly overforecast the highest amounts (bounded by the grid point maximum).

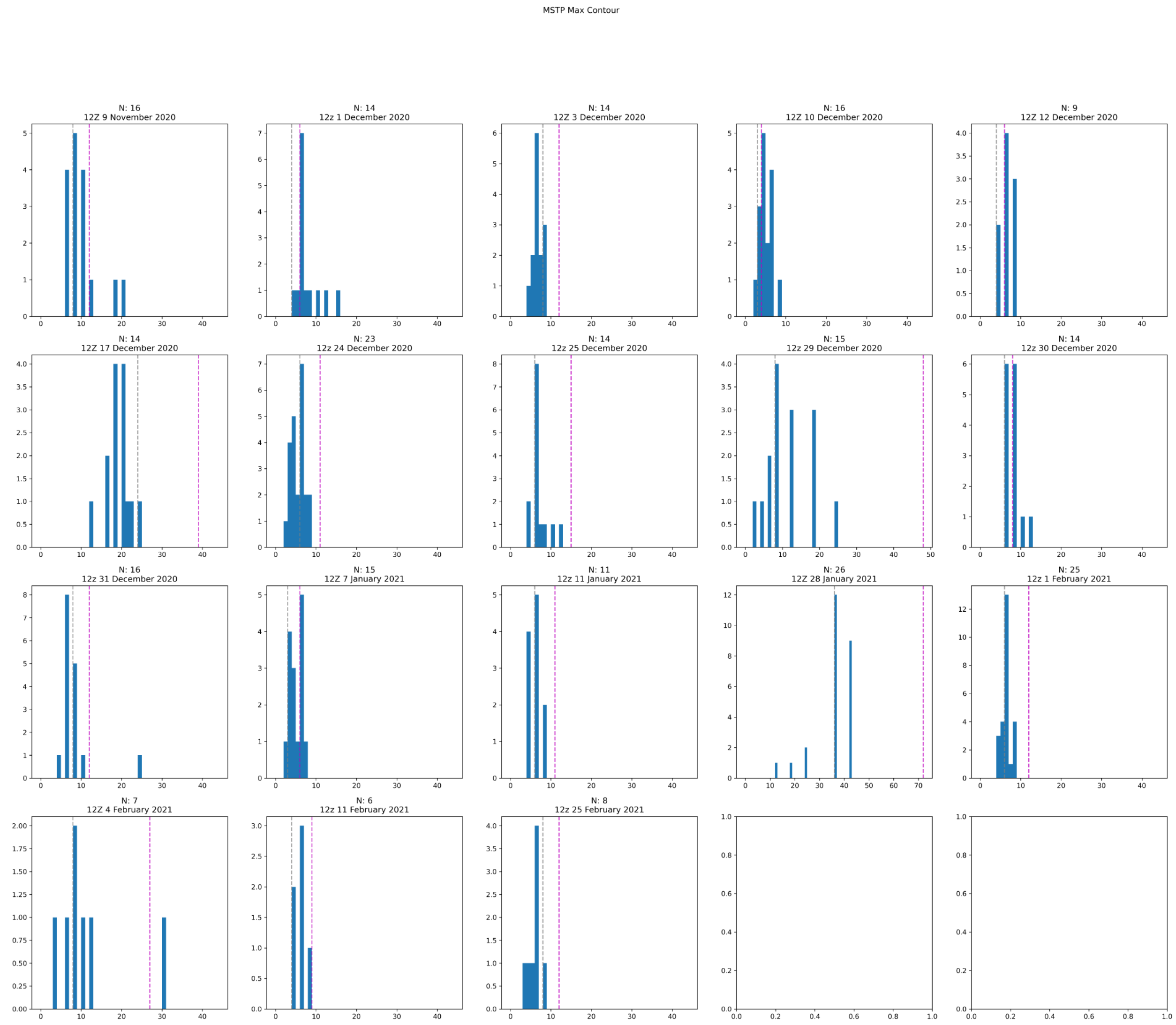


Figure 26: Histograms snowing participant maximum snowfall amounts for each of the 18 evaluated cases.

Overall, the following lessons learned were gleaned from the MSTP exercise.

- While we were able to provide multiple model depictions of 6 hourly and 24 hourly snow accumulation using a 10:1 SLR, WWE lacked detailed model data to ascertain why or why not a particular forecast evolution occurred. The bigger the event, such as the large snowstorms near Binghamton and the central California valley, drew more focus but still participants wondered if what any model predicted actually occurred. The lesson learned for organizers is to find ways to provide additional data (both observational and model) in regards to p-type which might then inform timing and duration evaluations.
- Footprint forecasts were generally quite good in the eastern US, a result that occurred for at least 2 reasons: participants drew large continuous areas (an effect of our simple

drawing software and ease of use) and our suite of model data generally contained good depictions of footprints in aggregate.

- The maximum contour exercise was intended to provoke notions of believability and/or trust in CAM guidance. For the most part, participants drew snowfall contours in areas where more snow occurred the majority of the time (not shown).

We encountered a few challenges related to the process of forecasting with an eye towards evidence based decision making. Participants tended to forecast based on the model evidence presented. In some ways the guidance influenced the participants to think about the event in general while still drawing what the models indicated. The participants had a difficult time accounting for the lower predictability of events at longer ranges and as a result could not account for changes to the synoptic pattern, since the majority of model guidance was based on the global GFSv15 (ie. deterministic solution spread was low). It is possible that as the UFS develops a CAM based ensemble system that longer range forecasts can use probabilistic snowfall to account for predictability issues in the longer ranges of 2-3 days. Therefore the WWE needs to further develop methods to assess how ensemble information might couple to some deterministic depictions to guide better snowfall forecasts.

The preliminary feedback from this exercise was positive. In some comments the participants emphasized that they liked the ability to think about the various guidance as opposed to doing grids or building blends with SLR. The time to reflect on what the CAMs show was also helpful as we discussed many aspects of why some CAMs have different representations of snow, possibly due to the individual microphysics used in the CAPS collection of deterministic forecasts. Another aspect that resonated with participants was trying to figure out how to make use of the CAMs (higher end amounts, mesoscale features not captured in global models, lake effect bands, the impact of lake ice or temperature feedbacks, etc) when they have various other first guess guidance like the NBM or WPC starting points.

This immersive forecasting activity was somewhat successful in helping us explore the CAM guidance in a realistic decision making environment with many deterministic, high resolution guidance. How to incorporate this type of guidance is complicated by the fact that predictability on the mesoscale is relatively low, especially with low precipitation amounts. In the future we hope to provide more data per model, fewer individual models, and more focus on precipitation type factors that influence the snowfall forecasts.

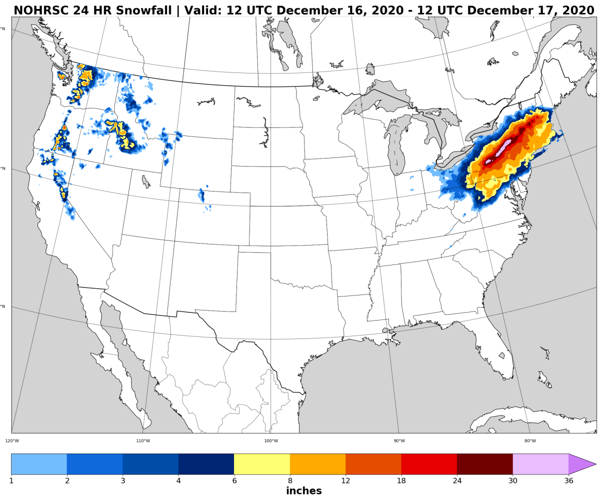
Future ideas emerged from the intensive sessions such as doing back to back day 3 - day 2 forecasts to focus on predictability issues whilst becoming more probabilistic as the RRFS continues in development. Likewise we could extend the forecasts to day 1 to continue the progression of incorporating CAM guidance from a consistency perspective. Lastly, we could continue to enhance WFO National Center perspective building by having participants face both WFO and National scale challenges and simulate collaboration amongst two groups (one focused on a WFO scale and another group focused on regional scale as we did this year).

Highlighted Events

While this year's WWE was able to capture and evaluate many impactful snowfall events, there are two cases that the team would like to highlight as they were record setting in terms of winter impacts and presented challenges for both the experimental data and participants. The first is Case 5 (Table 1) which was centered over the Northeastern CONUS. This system was evaluated live during our weekly WWE sessions. The final snowfall amounts were record breaking over much of central New York and Pennsylvania with Binghamton, NY measuring over 40 inches in 24 hours. The second is Case 18 (Table 1) which was centered over Texas, Louisiana, and Alabama. It was evaluated retrospectively as part of our second WWE intensive week. This system marked the beginning of the historic cold snap that caused widespread power outages throughout the region.

Case 5: Day 2 Northeast CONUS

The forecast challenges for this case were location and amount. As is common for Nor'easter type systems, placement of the low pressure center is vital for precipitation type and amount forecasting. While the final 24 hour snowfall amounts exceeded 30 inches in New York and Pennsylvania (figure 27), at day 2 the experiment data indicated the maximum precipitation would occur farther south and east in Southern New York, Pennsylvania, New Jersey, Connecticut, and Massachusetts.


NOHRSC 24 HR Snowfall | Valid: 12 UTC December 16, 2020 - 12 UTC December 17, 2020
1
2
3
4
6
8
12
18
24
30
36
inches

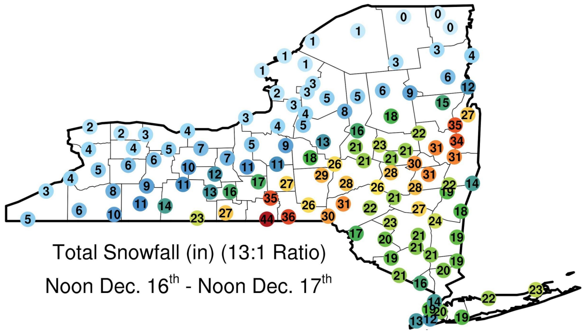
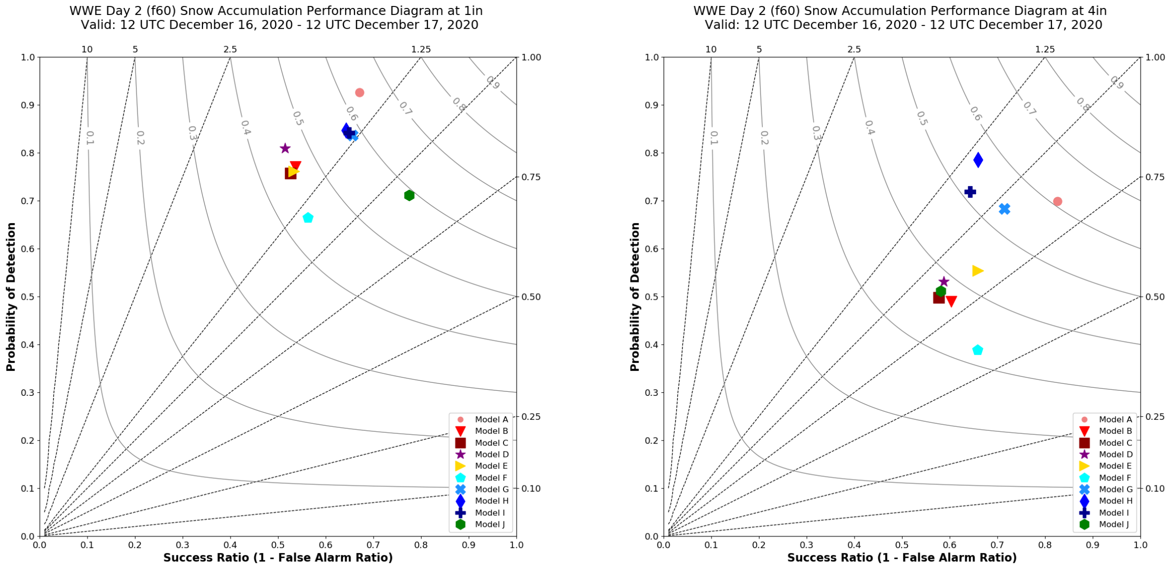

Total Snowfall (in) (13:1 Ratio)
Noon Dec. 16th - Noon Dec. 17th

Figure 27: NOHRSCv2 24 hour snowfall analysis for Case 5 (left). 24 hour snowfall for New York State (right). Image provided by NWS WFO Binghamton

Despite the misplacement of the precipitation maximum, the experiment data did a decent job of capturing the overall footprint. The performance diagrams for this case are shown in figure 28. At the 1 inch threshold all the models are in the upper right quadrant indicating a 'good' forecast. However, the positioning begins to degrade as the threshold increases with several configurations falling into the bottom left corner by 12 inches. While our evaluations did not extend above 12 inches, none of the experimental models even hinted at the record breaking amounts seen within this system. Discussion by the participants during the pre-event evaluation session did lead to the conclusion that somewhere in the forecasted footprint was going to get 'whacked' with a lot of snow, but even these discussions did not hint at a record event. When examining the performance diagrams, three models seem to stand out: the NBMv4.0 (Model A), EMC FV3-LAMDA (Model H), and CAPS_MP2 (Model F). The NBM's position remains consistent with the highest CSI values at the 1, 4, and 8 inch thresholds before falling off at the highest threshold of 12 inches. EMC FV3-LAMDA is also consistent in its high CSI values and seems to have the most favorable diagram position at the 12 inch threshold. Conversely, CAPS_MP2 has the lowest CSI and POD values for all four thresholds and falls into the lower left corner by 12 inches indicating it did not have snowfall accumulation values that high.



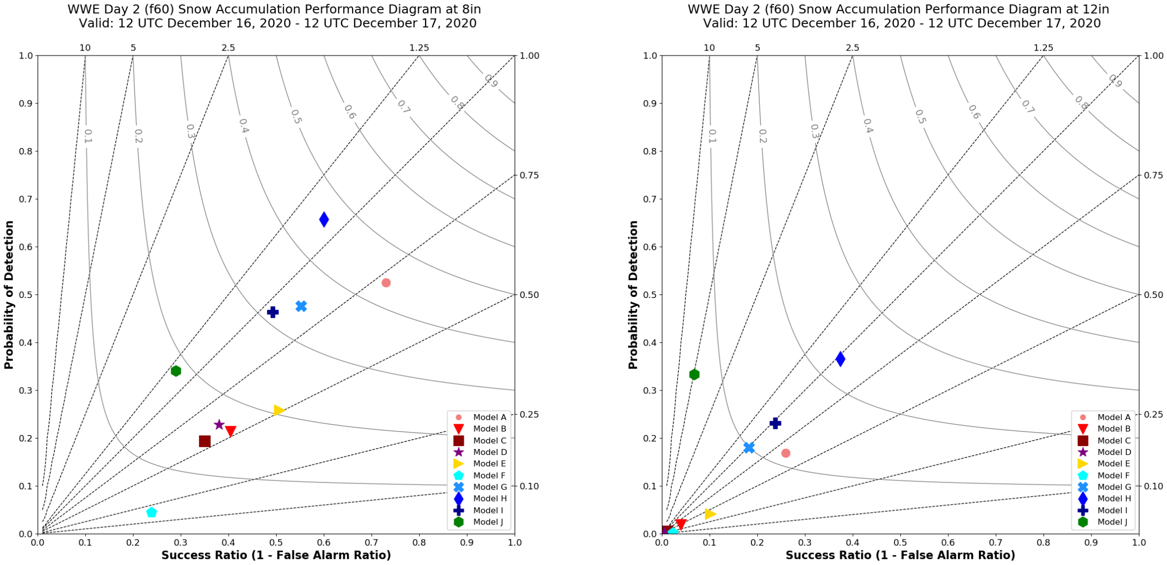
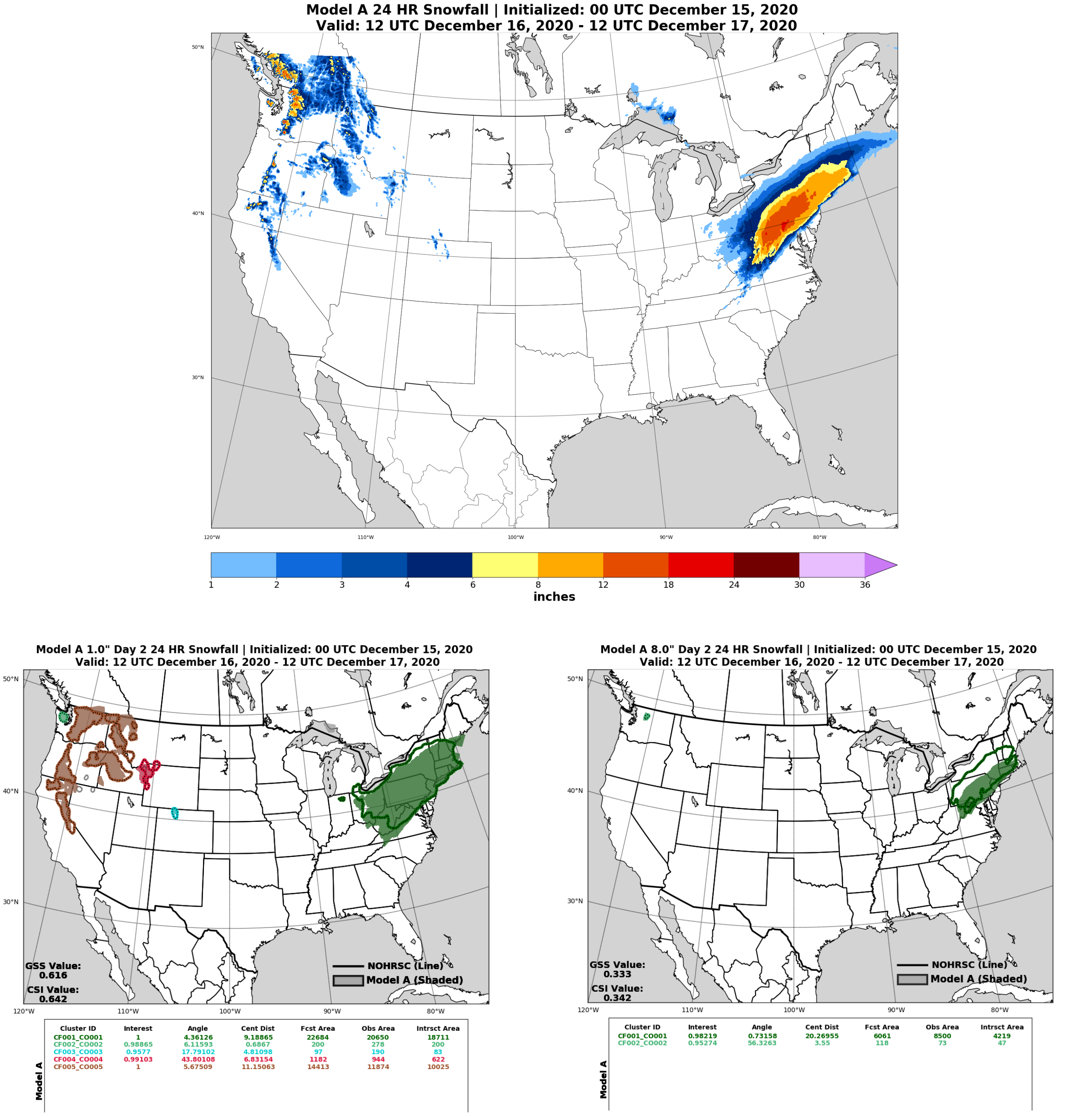


Figure 28: Performance diagrams for Case 5 at the 1 inch (upper left), 4 inch (upper right), 8 inch (lower left), and 12 inch (lower right) thresholds. Model A = NBMv4.0, Model B = CAPS_CNTL, Model C = CAPS_LSM1, Model D = CAPS_LSM2, Model E = CAPS_MP1, Model F = CAPS_MP2, Model G = EMC FV3-LAM, Model H = EMC FV3-LAMDA, Model I = EMC FV3-LAMX, Model J = Utah. Note Model J was not evaluated for this case since it occurred east of 100°W.

Looking at more details for the NBMv4.0, figure 29 shows the 24 hour snowfall accumulation for the forecast period in the upper panel with the MODE map and table for the 1 inch threshold in the lower left and 8 inch threshold in the lower right. Comparing the 24 hour snowfall accumulation map to the NOHRSCv2 in figure 27 one can see the positioning of the snowfall footprint looks decent but there is a clear under forecast of the snowfall amounts. When matching the forecast objects from NBMv4.0 to the NOHRSCv2 observations the 1 inch contours (figure 29; lower left) match quite well. This is also reflected in the MODE table values below the map. However, when the threshold is increased to 8 inches the displacement of the forecast object to the southeast can be seen.

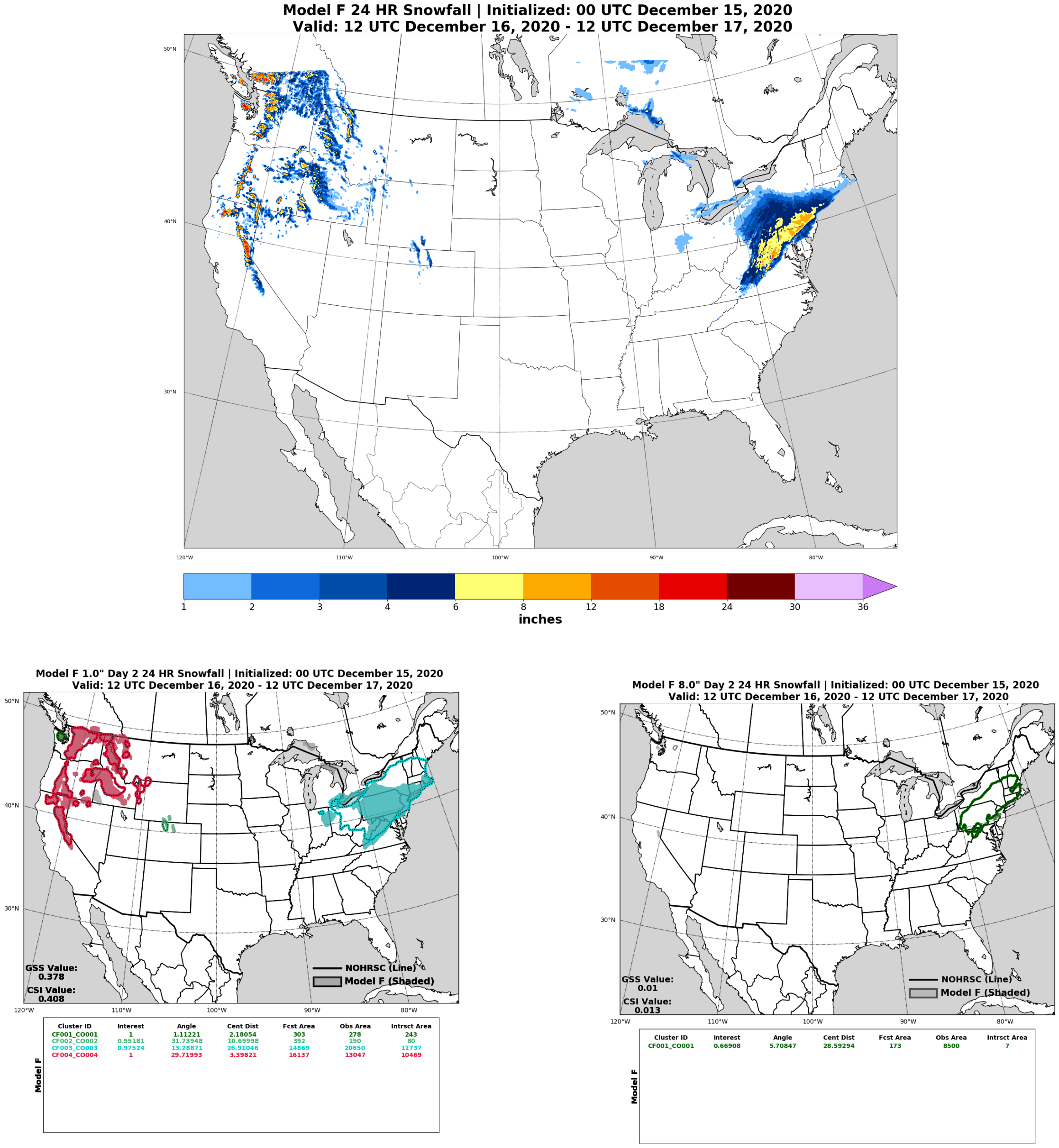

Model A 24 HR Snowfall | Initialized: 00 UTC December 15, 2020
Valid: 12 UTC December 16, 2020 - 12 UTC December 17, 2020
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
1
2
3
4
6
8
12
18
24
30
36
inches
Model A 1.0" Day 2 24 HR Snowfall | Initialized: 00 UTC December 15, 2020
Valid: 12 UTC December 16, 2020 - 12 UTC December 17, 2020
GSS Value:
0.616
CSI Value:
0.642
NOHRSC (Line)
Model A (Shaded)
Model A
Model A 8.0" Day 2 24 HR Snowfall | Initialized: 00 UTC December 15, 2020
Valid: 12 UTC December 16, 2020 - 12 UTC December 17, 2020
GSS Value:
0.333
CSI Value:
0.342
NOHRSC (Line)
Model A (Shaded)
Model A

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	1	4.36126	9.18865	22684	20650	18711
CF002_CO002	0.98865	6.11593	0.6867	200	278	200
CF003_CO003	0.9577	17.79102	4.81098	97	190	83
CF004_CO004	0.99103	43.80108	6.83154	1182	944	622
CF005_CO005	1	5.67509	11.15063	14413	11874	10025

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98219	0.73158	20.26955	6061	8500	4219
CF002_CO002	0.95274	56.3263	3.55	118	73	47

Figure 29: Case 5 24 hour snowfall accumulation for the NBMv4.0 (upper) with the MODE verification information at the 1 inch (lower left) and 8 inch (lower right) thresholds.

Looking at the details of CAPS_MP2 in figure 30, which has the lowest position on the performance diagram, both the lack of footprint and amounts can be seen. While the maximum snowfall amounts were shifted to the southeast, the 1 inch footprint (figure 30; lower left) does not extend far enough north or west. Also, the amounts were too low as there are only small 8 inch forecasted objects within the MODE map and table. Discussion of this configuration's performance was limited but a potential issue with how the precipitation type is assigned is a possible explanation.



Model F 24 HR Snowfall | Initialized: 00 UTC December 15, 2020
Valid: 12 UTC December 16, 2020 - 12 UTC December 17, 2020

Model F 1.0" Day 2 24 HR Snowfall | Initialized: 00 UTC December 15, 2020
Valid: 12 UTC December 16, 2020 - 12 UTC December 17, 2020

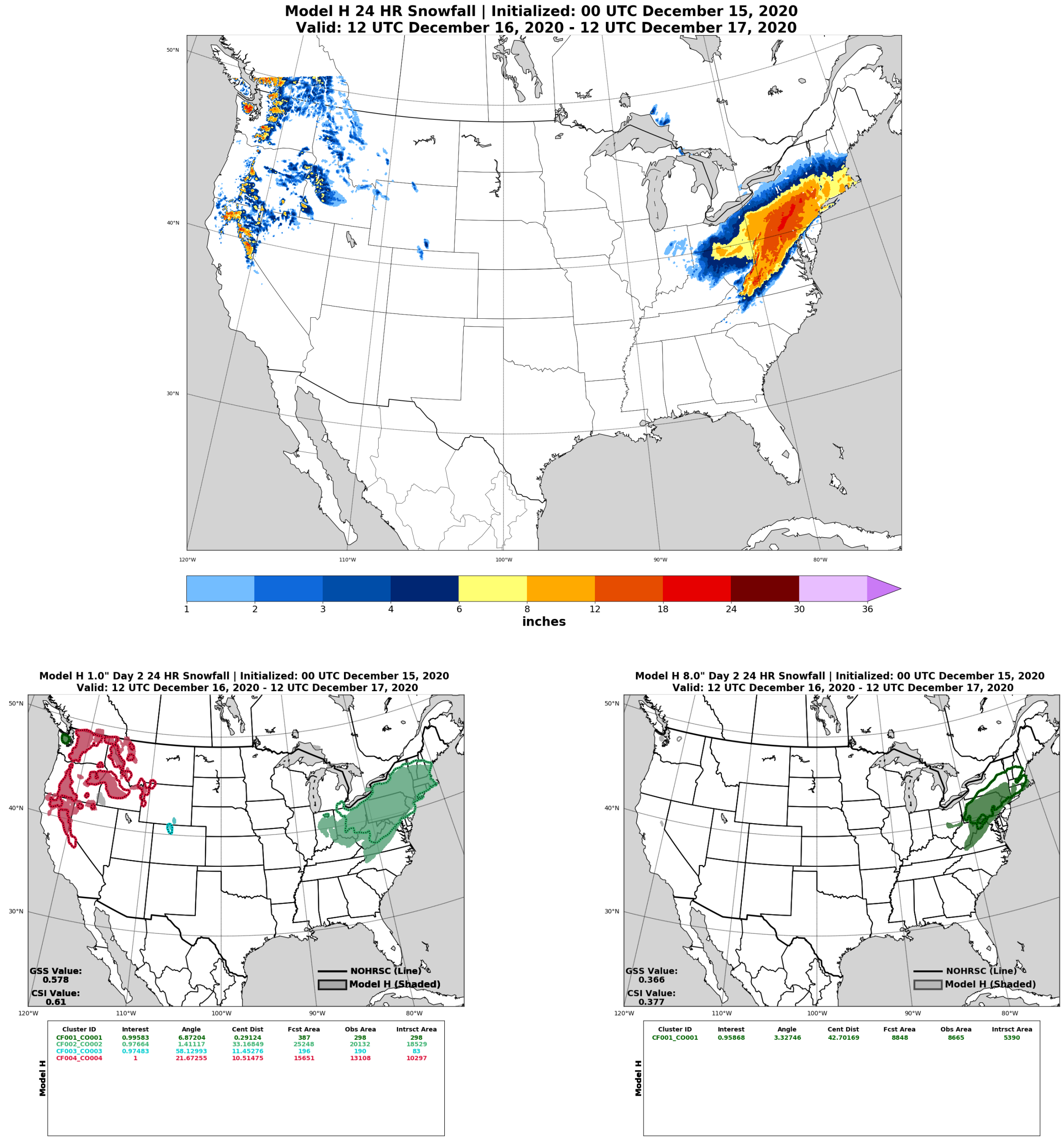
Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	1	1.11221	2.18054	303	278	243
CF002_CO002	0.95181	31.73948	10.69998	392	190	80
CF003_CO003	0.97524	13.28871	26.91046	14869	20650	11737
CF004_CO004	1	29.71993	3.39821	16137	13047	10469

Model F 8.0" Day 2 24 HR Snowfall | Initialized: 00 UTC December 15, 2020
Valid: 12 UTC December 16, 2020 - 12 UTC December 17, 2020

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.66908	5.70847	28.59294	173	8500	7

Figure 30: Case 5 24 hour snowfall accumulation for the CAPS_MP2(upper) with the MODE verification information at the 1 inch (lower left) and 8 inch (lower right) thresholds.

Based on the performance diagram, arguably the best performing deterministic model for this case was EMC FV3-LAMDA. This is most likely due to the high POD found due to the larger footprint that continues throughout all snowfall amounts. Figure 31 shows these with the 1 inch contour (lower left) extending too far into Indiana, West Virginia, Virginia, and North Carolina. The 8 inch map also shows the footprint extending too far south into West Virginia and Virginia.


Model H 24 HR Snowfall | Initialized: 00 UTC December 15, 2020
Valid: 12 UTC December 16, 2020 - 12 UTC December 17, 2020
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
1
2
3
4
6
8
12
18
24
30
36
inches
Model H 1.0" Day 2 24 HR Snowfall | Initialized: 00 UTC December 15, 2020
Valid: 12 UTC December 16, 2020 - 12 UTC December 17, 2020
GSS Value:
0.578
CSI Value:
0.61
NOHRSC (Line)
Model H (Shaded)
Model H
Model H 8.0" Day 2 24 HR Snowfall | Initialized: 00 UTC December 15, 2020
Valid: 12 UTC December 16, 2020 - 12 UTC December 17, 2020
GSS Value:
0.366
CSI Value:
0.377
NOHRSC (Line)
Model H (Shaded)
Model H

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.99583	6.87204	0.29124	387	298	298
CF002_CO002	0.97664	1.41117	33.16849	25248	20132	18529
CF003_CO003	0.97483	58.12993	11.45276	196	190	83
CF004_CO004	1	21.67255	10.51475	15651	13108	10297

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.95868	3.32746	42.70169	8848	8665	5390

Figure 31: Case 5 24 hour snowfall accumulation for the EMC FV3-LAMDA(upper) with the MODE verification information at the 1 inch (lower left) and 8 inch (lower right) thresholds.

Given the positioning of the experimental datasets, it is unsurprising that the vast majority of participants MSTP reflected the maximum snowfall too far to the southeast. Figure 32 shows one participant example where the 1 inch footprint outline (pink line) looks good, but the maximum amount (green line) is shifted.

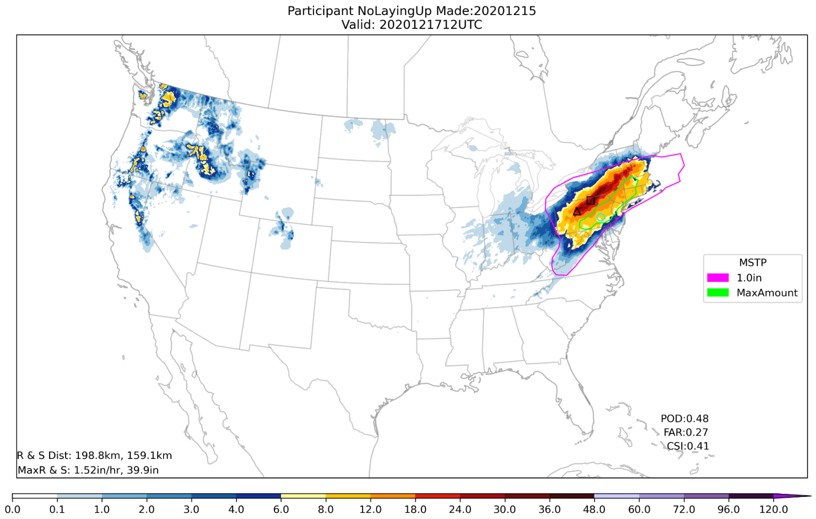


Figure 32: Participant MSTP for Case 5. Pink line is the participant drawn 1 inch footprint. Green line is the participant drawn maximum amount contour.

In the post-event evaluation discussion participants spent time discussing the impacts of this event and how to message a potential record setting snowfall amount to partners. For some participants, their partners do not care about specific amounts above the minimum 'we have to plow this' level. Others noted that this event specifically was an issue due to roof snow load on the temporary COVID19 treatment centers. While all agreed that at day 2, the messaging would not reflect a potential record breaking event, there needed to be information on where the maximum amounts may fall. From a probabilistic perspective, this case exceeded the 90 percentile and was truly a challenging event to forecast and message for local WFO forecasters.

Case 18: Texas/Southern CONUS

This case marked the beginning of the historic Texas power outages due to the winter storm and following record cold. While it did not line up with the live weekly WWE evaluations, the significant impacts of this event warranted its inclusion in the intensive sessions. From a winter storm perspective, the footprint spans a large swath of the southern CONUS (figure 33) while the 24 hour snowfall accumulation amounts are relatively low. There are isolated regions with 6 to 8 inches, but the vast majority of the region only received 1 to 3 inches. Despite the relatively low snowfall accumulation amounts, the experiment data showed a large spread of solutions.

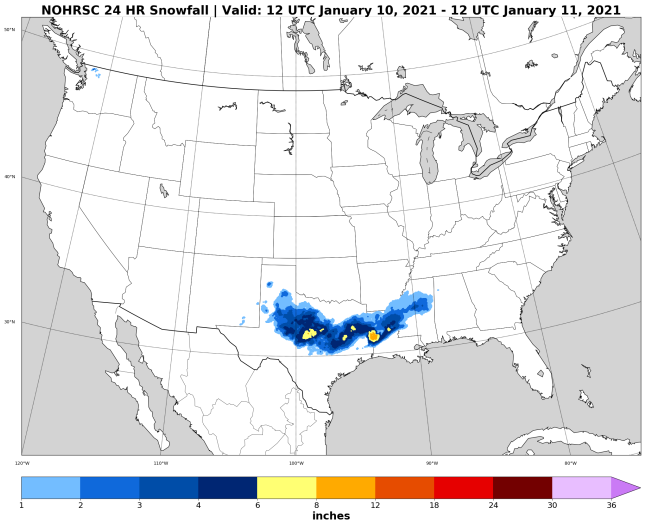

NOHRSC 24 HR Snowfall | Valid: 12 UTC January 10, 2021 - 12 UTC January 11, 2021
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
1
2
3
4
6
8
12
18
24
30
36
inches

Figure 33: NOHRSCv2 24 hour snowfall analysis for Case 18

Performance diagrams for the 1 and 4 inch thresholds are shown in figure 34. At both thresholds the experiment data varies drastically on the POD. The NBMv4.0 (Model A) stands out with the lowest POD and CSI values for this case while EMC FV3-LAMX had the highest POD. Discussions on this case led participants to note that models either had issues with the entire snowfall footprint or over forecasted the higher amounts.

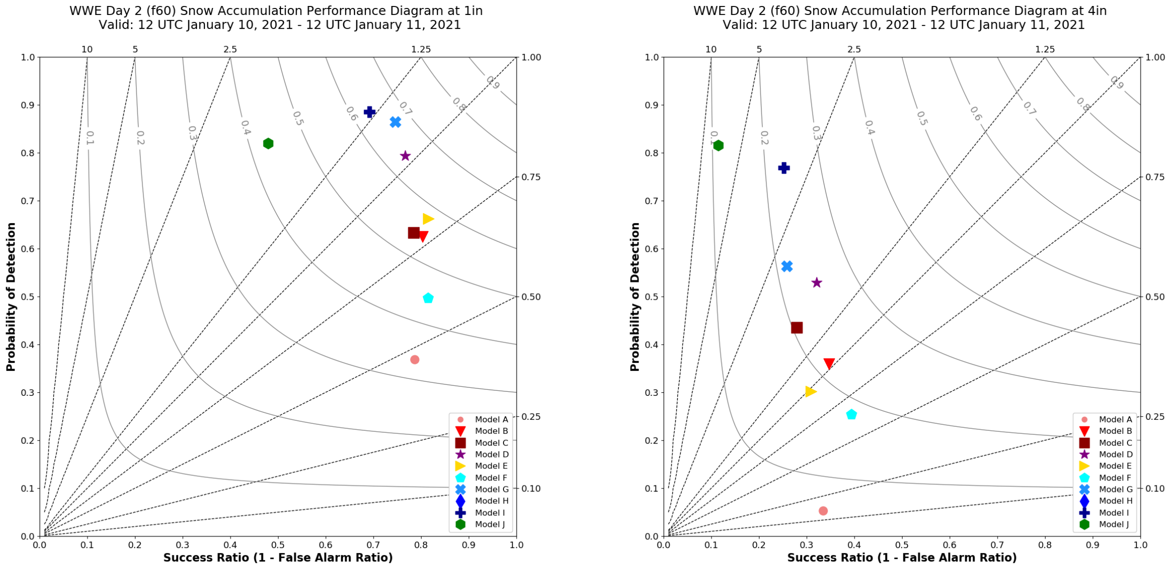
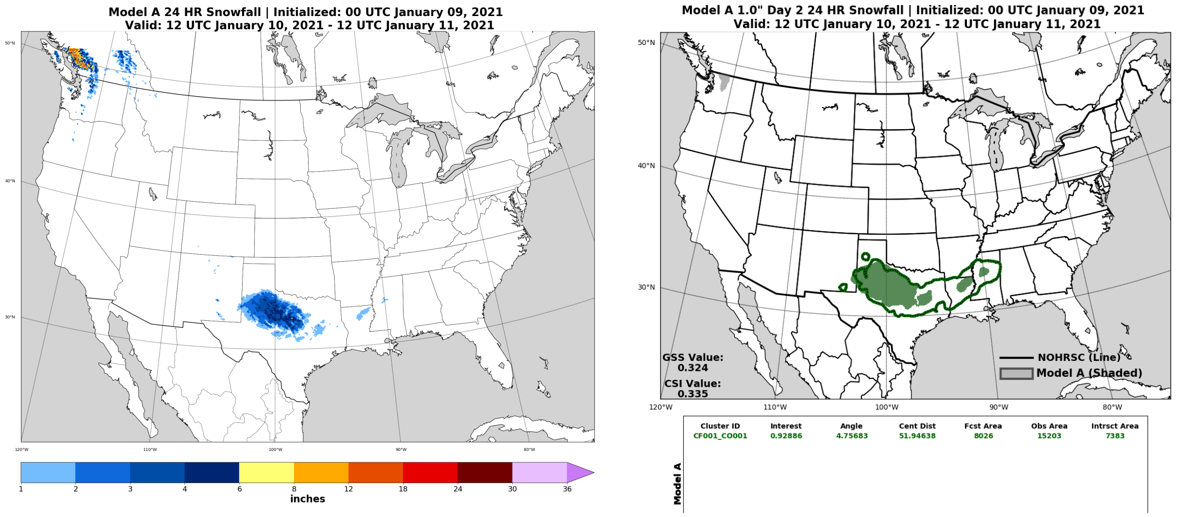


Figure 34: Performance diagrams for Case 18 at the 1 inch (left) and 4 inch (right) thresholds. Model A = NBMv4.0, Model B = CAPS_CNTL, Model C = CAPS_LSM1, Model D = CAPS_LSM2, Model E = CAPS_MP1, Model F = CAPS_MP2, Model G = EMC FV3-LAM, Model H = EMC FV3-LAMDA, Model I = EMC FV3-LAMX, Model J = Utah. Note Model J was not evaluated for this case since the majority occurred east of 100°W.

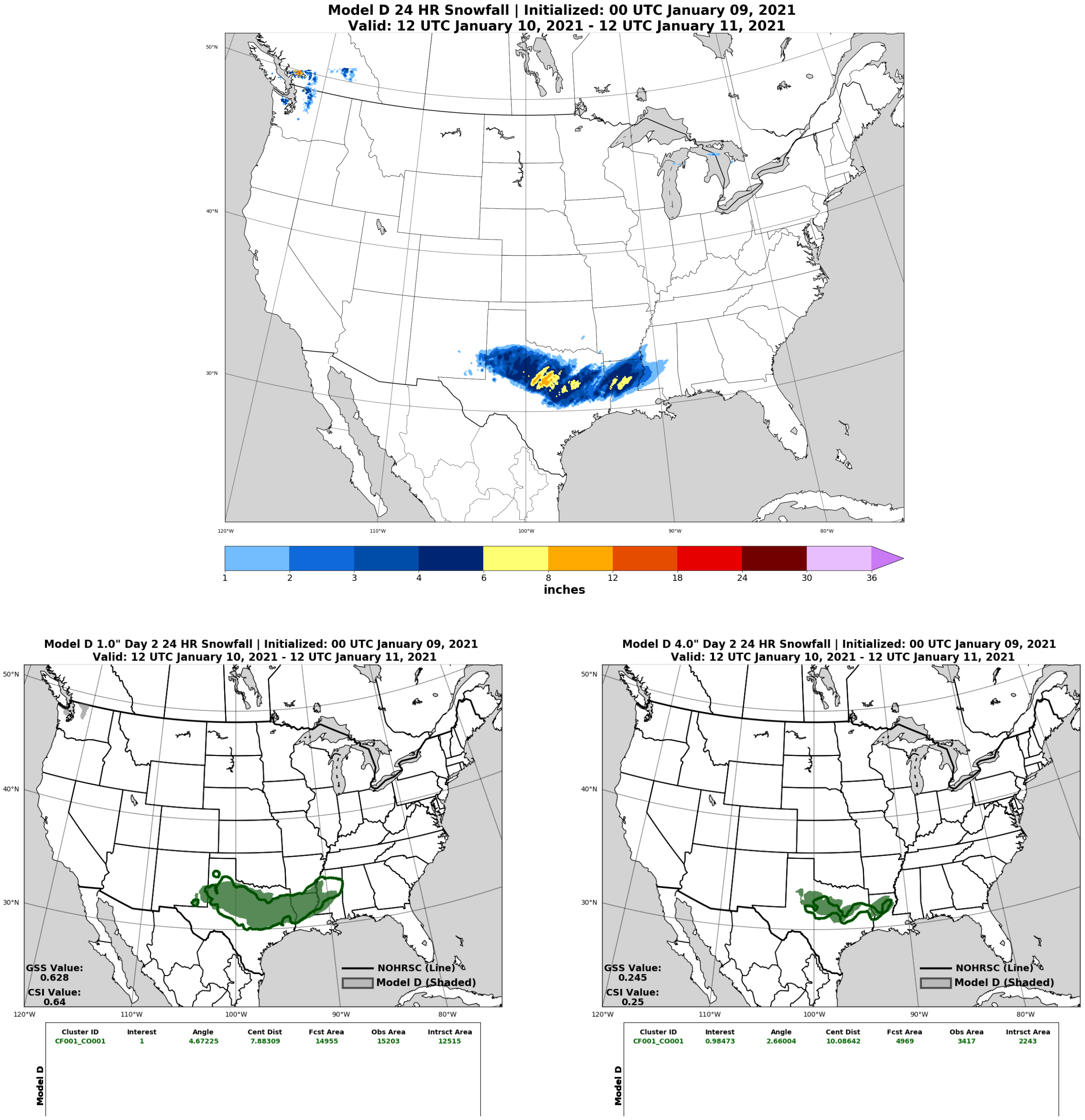
As shown in the performance diagrams, the NBMv4.0 has the lowest POD. The reason for this can be seen in figure 35 where there is only a small area of snowfall accumulation in central Texas (left panel). Discussion on the possible reasons for the poor performance of the NBMv4.0 for this case was focused on precipitation type. Participants found it likely the NBMv4.0 contributors were too warm creating sleet, freezing rain, or mixed precipitation scenarios which would lead to an incorrect snowfall footprint.


Model A 24 HR Snowfall | Initialized: 00 UTC January 09, 2021
Valid: 12 UTC January 10, 2021 - 12 UTC January 11, 2021
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
1
2
3
4
6
8
12
18
24
30
36
inches
Model A 1.0" Day 2 24 HR Snowfall | Initialized: 00 UTC January 09, 2021
Valid: 12 UTC January 10, 2021 - 12 UTC January 11, 2021
50°N
40°N
30°N
GSS Value:
0.324
CSI Value:
0.335
NOHRSC (Line)
Model A (Shaded)
120°W
110°W
100°W
90°W
80°W
Model A

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.92886	4.75683	51.94638	8026	15203	7383

Figure 35: Case 18 24 hour snowfall accumulation for the NBMv4.0 (left) with the MODE verification information at the 1 inch (right).

One of the best performing models for this case was the CAPS_LSM2. Based on the performance diagram at both thresholds it had one of the highest CSI values with minimal bias and was located closest to the upper right corner of the diagram. Looking at the details of this model figure 36 shows how well it predicted the 1 inch footprint (lower left panel). At the higher 4 inch threshold (lower right panel), this model displaced the maximum areas slightly but was able to correctly hint at pockets of higher accumulations.

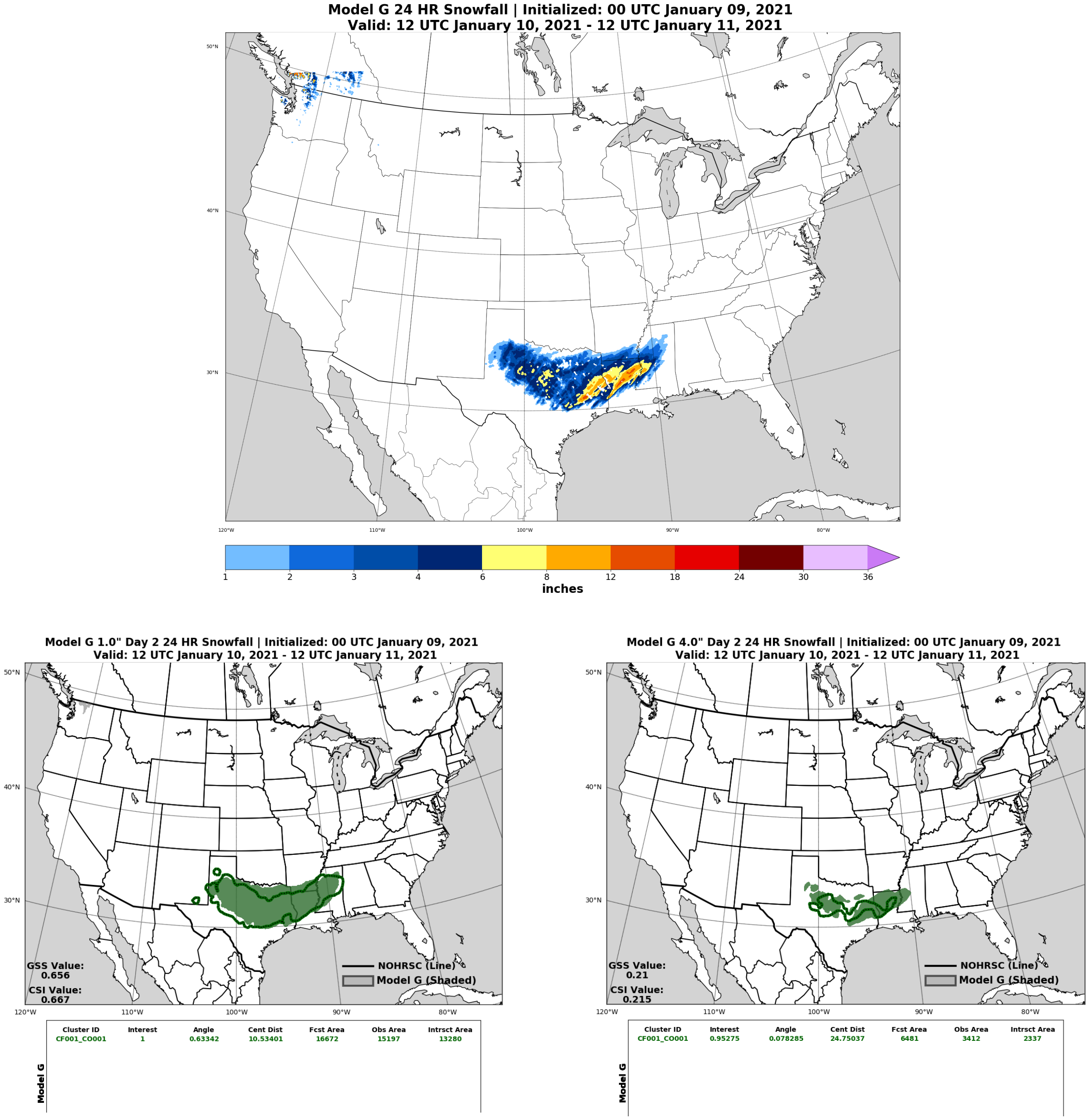

Model D 24 HR Snowfall | Initialized: 00 UTC January 09, 2021
Valid: 12 UTC January 10, 2021 - 12 UTC January 11, 2021
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
1
2
3
4
6
8
12
18
24
30
36
inches
Model D 1.0" Day 2 24 HR Snowfall | Initialized: 00 UTC January 09, 2021
Valid: 12 UTC January 10, 2021 - 12 UTC January 11, 2021
GSS Value: 0.628
CSI Value: 0.64
NOHRSC (Line)
Model D (Shaded)
Model D
Model D 4.0" Day 2 24 HR Snowfall | Initialized: 00 UTC January 09, 2021
Valid: 12 UTC January 10, 2021 - 12 UTC January 11, 2021
GSS Value: 0.245
CSI Value: 0.25
NOHRSC (Line)
Model D (Shaded)
Model D

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	1	4.67225	7.88309	14955	15203	12515

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.98473	2.66004	10.08642	4969	3417	2243

Figure 36: Case 18 24 hour snowfall accumulation for the CAPS_LSM2(upper) with the MODE verification information at the 1 inch (lower left) and 4 inch (lower right) thresholds.

Finally, the EMC FV3-LAM model also stood out as it did a nice job capturing the 1 inch footprint (figure 37; lower left), but was over zealous in the amounts at 4 inches (figure 37; lower right). This CAM showed higher snowfall accumulations throughout northern Louisiana as opposed to isolated in the western portion of the state. Some of this over forecast could be attributed to the WWE's use of 10:1 SLR, but comments for this case focused more on the footprint signal than the total amounts.


Model G 24 HR Snowfall | Initialized: 00 UTC January 09, 2021
Valid: 12 UTC January 10, 2021 - 12 UTC January 11, 2021
50°N
40°N
30°N
120°W
110°W
100°W
90°W
80°W
1
2
3
4
6
8
12
18
24
30
36
inches
Model G 1.0" Day 2 24 HR Snowfall | Initialized: 00 UTC January 09, 2021
Valid: 12 UTC January 10, 2021 - 12 UTC January 11, 2021
GSS Value:
0.656
CSI Value:
0.667
NOHRSC (Line)
Model G (Shaded)
Model G
Model G 4.0" Day 2 24 HR Snowfall | Initialized: 00 UTC January 09, 2021
Valid: 12 UTC January 10, 2021 - 12 UTC January 11, 2021
GSS Value:
0.21
CSI Value:
0.215
NOHRSC (Line)
Model G (Shaded)
Model G

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	1	0.63342	10.53401	16672	15197	13280

Cluster ID	Interest	Angle	Cent Dist	Fcst Area	Obs Area	Intrsct Area
CF001_CO001	0.95275	0.078285	24.75037	6481	3412	2337

Figure 37: Case 18 24 hour snowfall accumulation for EMC FV3-LAM(upper) with the MODE verification information at the 1 inch (lower left) and 4 inch (lower right) thresholds.

Based on the experiment data, participants generally followed the CAM guidance over the NBMv4.0 when drawing their MSTP forecasts. Figure 38 (left) is one example where the participant followed guidance that placed the footprint and maximum snowfall amount areas too far north. The ensemble (figure 38; right) shows the majority of forecasts drew their footprint over central Texas with fewer extending all the way into Louisiana and Alabama.

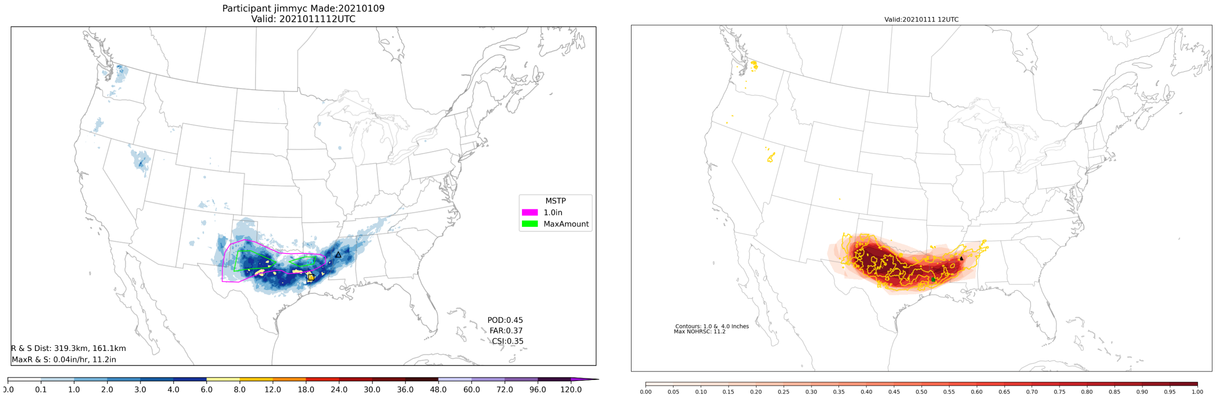


Figure 38: Left is a participant MSTP for Case 18. Pink line is the participant drawn 1 inch footprint. Green line is the participant drawn maximum amount contour. Right is the ensemble generated from all the participant footprint MSTP polygons. Red shading indicates the fraction of MSTPs drawn over the area. Yellow contours represent NOHRSCv2 snowfall accumulations at the 1 and 4 inch thresholds.

This case was arguably the most impactful of the 2020-21 winter season in terms of power outages, life, and property loss. The winter storm preceded a days long cold snap that ultimately caused widespread issues across Texas and the southern CONUS. While the total accumulated snowfall amounts may not be enough to disrupt life in some regions of the CONUS, the fact that even in day 2 there was a large spread in solutions made this system difficult to message especially in an area not accustomed to these types of weather conditions.

4. Summary and Recommendations

The 11th annual WWE was conducted over the 2020-2021 winter season with weekly evaluations of the experimental data on Tuesdays and Wednesdays. This year participants were asked to evaluate the models from a pre-event and post-event perspective. Experiment data was heavily focused on FV3-LAM configurations in preparation for the RRFS which is currently set for deployment in Fall of 2023. Science objectives for this year's WWE were mainly focused on the utility of these CAM configurations in the day 2 and day 3 time frame for snowfall. Participant comments show the utility comes from identifying potential areas of snow banding and higher snowfall amounts. They become especially useful in low confidence-high impact events when the global models are having a difficult time with the set up. These CAMs provide more detailed insight into possible scenarios. One recurring comment from participants related to keeping the synoptic scale issues in mind when evaluating these types of model solutions. While forecasters can be awed by the details the CAM models produce, they can still be inaccurate and need to be taken within context to the global models and synoptic setup.

Another aspect of the experiment was to increase participation and expand engagement with forecasters. Increased advertisement and early orientations proved useful in that our weekly attendance was greatly increased over the previous year. Using the new interactive website and MSTP drawing activity, the WWE team was able to actively engage participants with the experiment data that ultimately ended with their creation of a snowfall forecast. Feedback on the format and activity were positive with many people noting how it made them actually think about what each data set was showing and really assess what metrics they deem useful for making and evaluating a forecast. For each experiment dataset, the following bullet points will summarize the team's thoughts and recommendations with each being categorized as 'recommended for transition', 'recommended for further development', or 'rejected for further testing'. Table 5 also summarizes the recommendations.

- **EMC** provided three FV3-LAM configurations: A control FV3-LAM, the FV3-LAMDA that provides hourly data assimilation for the first six hours, and an experimental FV3-LAMX that runs on the proposed RRFS domain. Performance diagrams show all three tend to have a higher bias than other configurations. Subjective ranking also tended to have these configurations in the middle to bottom of the rankings mostly due to position on the performance diagrams for each individual case. Due to these results the team **recommends further testing and development** on each of the FV3-LAM configurations to refine which is most suitable for snowfall events.

- **CAPS** provided five FV3-LAM configurations: the CAPS_CNTL which had a similar setup to the EMC FV3-LAM, two configurations that tested LSMs CAPS_LSM1 and CAPS_LSM2, and two configurations that tested microphysics CAPS_MP1 and CAPS_MP2. The CAPS_CNTL and LSM configurations were consistently ranked among the highest within the subjective evaluation. CAPS_MP2 was found to be among the lowest ranked models especially in the post-event evaluations when the objective information was available. Due to these results the team **recommends further testing and development** as well as continued coordination with EMC to refine the FV3-LAM configurations.

- **University of Utah** provided a downscaled methodology on the GFSv15 snowfall over the Western CONUS. Participants were split on ranking this dataset either as first or last due to the high bias and POD. However, the majority of comments were favorable to this methodology especially when compared to the CAM data. These results are consistent with last year's WWE. Due to the comments and results this methodology is **recommended for transition into operations**. In fact as of March 2021, this methodology has been successfully implemented for GFSv16 and 12km NAM within WPC's Winter Weather Desk AWIPS platform. However, the team also is **recommending**

further development and testing as their proposed machine learned methodology is of great interest and could be expanded beyond the Western CONUS.

- As with the 2020 FFaIR experiment, the interactive MSTP drawing activity was a great success. It engaged the participants on a level not seen previously in the WWE. The discussion generated during the activities helped the WWE team gain more insights into what forecasters thought of each dataset as well as the thought processes behind capturing extreme snowfall amounts. The HMT team will continue to refine and improve this activity and will again be featured in the upcoming 2021 FFaIR experiment.

- Since the WWE has been held virtually for several years, the COVID19 pandemic did not change how the underlying experiment functioned. However, the team was unable to invite participants to NCWCP for intensive in person sessions. These were adapted virtually and may lay plans for future WWEs where specific science questions can be targeted with retrospective cases and selected experts to provide evaluations.

Table 5: Research to operations transition recommendations for the 11th Annual WWE.

Evaluated Dataset	Recommended for transition to operations	Recommended for further development and testing	Rejected for further testing	Provider/Funding Source
EMC FV3-LAM EMC FV3-LAMDA EMC FV3-LAMX		X X X		EMC
CAPS_CNTL CAPS_LSM1 CAPS_LSM2 CAPS_MP1 CAPS_MP2		X X X X X		OU-CAPS
University of Utah	X	X		University of Utah

5. Acknowledgements

The WWE team would like to sincerely thank the WPC forecasters (Bryan Jackson, Peter Mullinax, Laura Pagano, Josh Weiss, Frank Pereira, Greg Carbin, and Josh Kastman) for helping with the forecast briefings before each session. Your expertise and insights into each case was invaluable to making our experiment a success! Also thank you to Alex Lamers for coordinating the forecasting schedule. Additional thanks to Josh Weiss and Frank Pereira for participating and helping with the briefings for the intensive sessions. Josh Kastman for his immense help in getting the WWE website up and running. Mark Klein for his contributions to the science plan and vision for the experiment. Joe Nettesheim for helping us troubleshoot the MSTP drawing activity when everything broke. Ben Albright for his dedication to keeping the MODE websites current and providing all of the objective verification information. We would also like to recognize Geoff Manikin for continuing to foster collaboration with EMC, not only for data expertise but also with detailed scheduling of EMC participants. Lastly, the team would like to thank all of HMT and WPC staff that helped us prepare with data troubleshooting and support throughout the experiment. See you all next WWE!

6. References

Aligo, E., B. Ferrier, J. Carley, E. Rodgers, M. Pyle, S. J. Weiss, and I. L. Jirak, 2014: Modified microphysics for use in high-resolution NAM forecasts. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 16A.1, https://ams.confex.com/ams/27SLS/webprogram/Paper255732.html.

Bullock, R. G., Brown, B.G., & Fowler, T. L. (2016). Method for Object-Based Diagnostic Evaluation (No. NCAR/TN-532+STR). doi:10.5065/D61V5CBS.

Harris, L. M., and S.-J. Lin, 2013: A two-way nested global-regional dynamical core on the cubed-sphere grid. *Mon. Wea. Rev.*, **141**, 283–306, https://doi.org/10.1175/MWR-D-11-00201.1.

Mansell, E. R., C. L. Ziegler, and E. C. Bruning, 2010: Simulated electrification of a small thunderstorm with two-moment bulk microphysics. *J. Atmos. Sci.*, **67**, 171–194, https://doi.org/10.1175/2009JAS2965.1.

Olson, J. B, and Coauthors, 2019: A Description of the MYNN-EDMF Scheme and the Coupling to Other Components in WRF-ARW. NOAA Technical Memorandum. https://doi.org/10.25923/n9wm-be49

Roebber, P.J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601-608. DOI: https://doi.org/10.1175/2008WAF2222159.1

Tewari, M., and Coauthors, 2004: Implementation and verification of the unified Noah land surface model in the WRF model. *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction*, Amer. Meteor. Soc., Seattle, WA, 14.2a, https://ams.confex.com/ams/84Annual/techprogram/paper_69061.htm.

Thompson, G., P. R. Field, R. M.Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

WPC-HMT, 2020: The 10th Annual Winter Weather Experiment: Final Report. Accessed 5 April 2021, https://www.wpc.ncep.noaa.gov/hmt/10th_Annual_WPC_HMT_WWE_Final_Report.pdf

WPC-HMT, 2020: The 11th Annual Winter Weather Experiment: Program Overview & Operations Plan. Accessed 5 April 2021, https://origin.wpc.ncep.noaa.gov/hmt/wwe2021/11th_WWE_Plan_Draft.pdf

Appendix A: MODE Configuration

MODE was used to objectively analyze Day 2, 00Z cycle (f60), 24 hour snowfall forecast objects to 24 hr observed snowfall from NOHRSC. MODE was run each day from November 1, 2020 through March 15, 2021 for all WWE guidance. Because the CAPS FV3 members were not run every day of the experiment, there were only approximately 29 days analyzed. The GFSv15 downscaled from the University of Utah domain covered only the western portion of the CONUS and was analyzed against the other model guidance over the same western domain. All snowfall accumulation forecasts and NOHRSC observations were regridded to a 5 km grid regardless if it was only the western CONUS or the full CONUS. Table 6 contains select settings that were used to identify the objects.

Table 6. *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

Grid statistics were harvested from daily MODE CTS. The daily MODE CTS were aggregated over the whole season to compute the monthly and seasonal statistics shown in the performance diagrams.