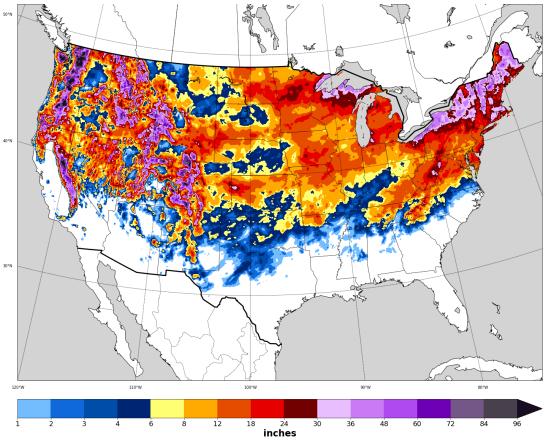
12th Annual Winter Weather Experiment: Findings and Results



15 November 2021 - 15 March 2022
Weather Prediction Center
Hydrometeorological Testbed

Dr. Kirstin Harnos **-** CIRES CU Boulder, NOAA/NWS/WPC/HMT
Dr. James Correia Jr. - CIRES CU Boulder, NOAA/NWS/WPC/HMT
Dr. Benjamin Albright - Systems Research Group, NOAA/NWS/WPC/HMT
Dr. Sarah Trojniak - System Research Group, NOAA/NWS/WPC/HMT
James Nelson - NOAA/NWS/WPC

Table Of Contents

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 12th annual Winter Weather Experiment (WWE) during the 2021-2022 winter season. The WWE provides collaborative research to operations (R2O) experience bringing together members of the forecasting, research, and academic communities to evaluate and discuss winter weather forecast challenges. This year also tasked participants with examining potential updates to the Winter Storm Severity Index (WSSI) and insights into the forecast process. Recent WWE successes include improvements to the National Blend of Models, incorporation of snowsqualls to the mPING crowd-sourcing data app, and increased discussion on the creation of winter specific verification metrics. Building on the success of previous years, the WWE was once again fully remote with two intensive evaluation weeks where retrospective case studies were utilized to examine the experiment objectives. The WWE also hosted invited presentations throughout the entire WWE season.

2. Science and Operations Objectives

While the WWE continues to refine its format and activities, the overall objective of HMT remains the same: to provide a pseudo-operational environment to evaluate operational and experimental guidance and to provide a space for open and unfiltered discussion about winter weather topics. This year's WWE continued that tradition with successful remote interaction during the two intensive evaluation weeks of February 7 and February 28. With the unreliability of real-time cases, activities during the intensive weeks were based on retrospective events captured during the 2021-2022 winter season. This allowed facilitators an opportunity to select cases based on specific storm properties and to better address the experiment's science objectives.

The objectives of the 12th Annual Winter Weather Experiment were to:

- Evaluate the utility of operational and experimental snowfall from high resolution convective-allowing deterministic and ensemble models (CAM).
- Evaluate the redesigned Winter Storm Severity Index (WSSI) in parallel with the operational WSSI.
- Explore precipitation event timing and synoptic parameters in ensemble data to assess predictability of potentially impactful winter weather scenarios
- Explore available precipitation type output from experimental deterministic and ensemble systems to inform forecasts.

- Compare CAM to downscaled snow-to-liquid ratio (SLR) techniques for improvements of forecast snowfall.
- Use both event and season long verification to assess the performance of experimental data sets.

Cases were evenly dispersed east of the Rocky Mountains with a mix of storm tracks from northern clipper systems, Central Plains systems, to Nor'Easters. West of the Rocky Mountains was a different story. While there was an active beginning to the winter before 1 December, there were really only a handful of cases once the experiment began. Aligned with the experiment goals, the team also attempted to capture impactful ice and mixed-precipitation type cases. While there is no official observation dataset to verify against, it is important for WWE to evaluate these types of systems so participants can provide feedback on how best to develop precipitation type guidance in the future. Observations from the National Operational Hydrologic Remote Sensing Center version 2 (NOHRSCv2[1]) snowfall accumulation can be seen in figure 1 for the 29 cases captured for the WWE season.

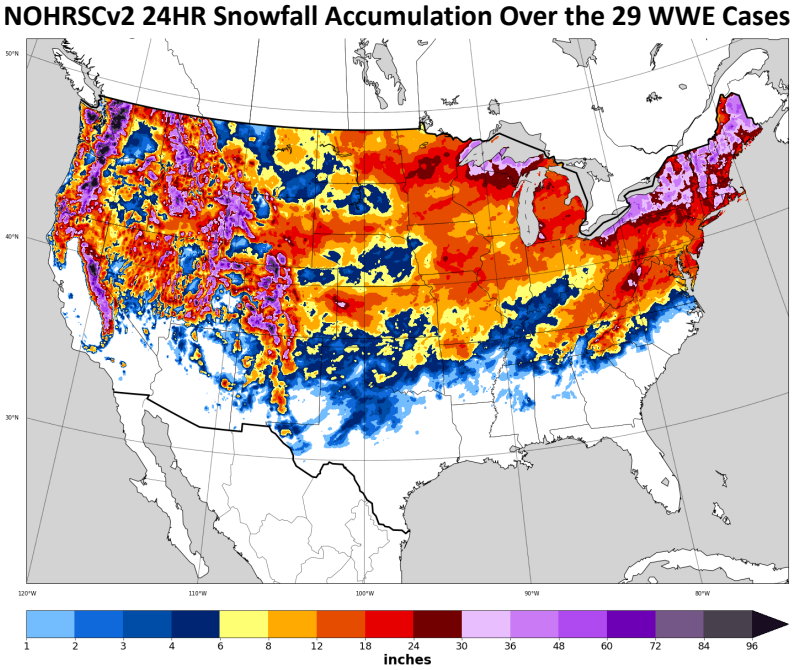


Figure 1: WWE snowfall accumulation over the 29 cases from NOHRSCv2.

While the WWE season started out slowly, by the end of December cases were averaging around 2 per week with many cases considered as multiple day events. From 1 December 2021 through 15 March 2022, WWE collected a total of 29 cases (Table 1). Of the 29 cases, seven were highlighted for evaluation during the intensive week as indicated by the bold and asterisks.

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

Two of those highlighted events will be described in detail later in this report. Additionally, the experimental data for all the cases were used to provide the seasonal statistics and evaluations described in a later section.

Table 1: Summary of the cases captured in the 12th Annual WWE.
Starred and bold indicate intensive week evaluation cases.

Case Number	Case Dates (12z - 12z)	Location
1	6-Dec-2021	Upper MidWest
2	9-Dec-2021	New England
3	10-Dec-2021	Mountain West to Northern Plains
4*	**11-Dec-2021**	**Central Plains to Upper Midwest**
5	15-Dec-2021	Intermountain West
6	16-Dec-2021	Colorado to Upper MidWest
7*	**26-Dec-2021**	**Pacific NW to Sierras**
8	27-Dec-2021	Northern Plains to Upper MidWest
9	2-Jan-2022	Central Plains to Great Lakes
10*	**3-Jan-2022**	**Mid-Atlantic**
11*	**4-Jan-2022**	**Mid-Atlantic to New England**
12	7-Jan-2022	Ohio River Valley to Mid-Atlantic
13	15-Jan-2022	Upper MidWest
14	16-Jan-2022	Central Plains to Ohio River Valley
15	17-Jan-2022	Ohio River Valley to Mid-Atlantic
16	18-Jan-2022	New England
17	21-Jan-2022	Mid-Atlantic to New England
18	26-Jan-2022	Central Plains
19	29-Jan-2022	New England
20*	**30-Jan-2022**	**New England**
21	3-Feb-2022	Central Plains to OH River Valley
22*	**4-Feb-2022**	**Central Plains to OH River Valley**
23*	**18-Feb-2022**	**Central Plans to Great Lakes**
24	22-Feb-2022	Northern Plains to Upper MidWest
25	25-Feb-2022	Great Lakes to New England
26	7-Mar-2022	Central Plains to Great Lakes
27	9-Mar-2022	Intermountain West
28	10-Mar-2022	Central Plains
29	13-Mar-2022	Mid-Atlantic to New England

2.1 Experiment Data

WWE participants evaluated a variety of experimental data, which are outlined in Table 2 including the number of cases run for day 3, day 2, and day 1. As the NWS continues to work towards a Unified Forecast System[2], operational model development is centered around the finite volume cubed-sphere (FV3) dynamic core. Additionally, the development of an FV3 Limited Area Model (LAM) is the basis for the future Rapid Refresh Forecast System[3] (RRFS). This CAM, referred to as the FV3-LAM, is still in development and the model physics have not been officially set. Experiments like WWE allow model development centers like Environmental Modeling Center (EMC), Global Systems Lab (GSL), and University of Oklahoma-Center for Analysis and Prediction of Storms (OU-CAPS) to coordinate and gain feedback on the proposed configurations and are a key component in determining the final configuration for the RRFS. For specific configuration information, please refer to the 12th Annual WWE Science and Operations Plan.

Table 2: Summary of the experimental model guidance evaluated in the 12th Annual WWE

Model	Provider	Resolution	Forecast Hours	Day 1 Case Count	Day 2 Case Count	Day 3 Case Count	Notes
FV3-LAM	EMC	3 km	60 hours	26	22	N/A	Control FV3-LAM
FV3-Cloud	EMC	3 km	60 hours	22	19	N/A	Cloud-based data assimilation
FV3-LAM Ensemble (13 Members)	OU - CAPS	3km	84 hours	12	22	12	SSEF control member evaluated as deterministic model
GFSv16	UUtah	2.5 km	84 hours	23	23	24	Full CONUS
NBMv4.0	MDL	2.5 km	84 hours	29	29	29	Deterministic and Probabilistic
NBMv4.1	MDL	2.5 km	84 hours	22	23	19	Deterministic and Probabilistic

There were two NOAA-OAR-WPO funded projects that were unable to be tested in this WWE. The first project to test physics parameter perturbations in the HRRRe was unable to be included due to data requirement miscommunications that led to delays which could not be accommodated due to lack of HMT computing resources. However, the project PIs did receive data and will continue their analysis as funding ends in June. The second project is for testing the implementation of FVCOM in the RRFS. GSL had not completed the transition for the RRFS

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

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

development in time to complete retrospective case studies, thus it was not possible to examine the FVCOM influences on Lake Effect snows in the Great Lakes region.

In addition, EMC was able to provide a 60 hour deterministic forecast using a recently upgraded version of the RRFS with cycled ensemble data assimilation (table 2: FV3-Cloud), however, the initial data contained a bug, and a rerun was attempted 3 days prior to the start of the experiment. While successful, this version also had a bug, and thus cases 1-12 (table 1) had to be rerun before the data could be included in the seasonal evaluations shown later in the report. HMT staff worked with EMC to make sure the best possible data were available, and communicated with forecasters that EMC was providing cutting edge development software and thus these data were prone to revision. This unavoidable consequence of testing *during* intense development has its own unique challenges and we thank each Team for attempting to do the best they could under the circumstances.

2.2 Intensive Weeks

As stated earlier, this year's WWE was structured around two intensive evaluation weeks. Over the course of the week, participants were asked to evaluate and discuss several cases from incrementally decreasing lead time (day 3 to day 1). The cases would begin with a day 3 briefing from a WPC forecaster, followed by the forecast activity. Then the case would be briefed at day 2, followed by the forecast activity, and again briefed at day 1 followed by the activity. More details on the forecast activities can be found later in this section. The case would then be completed by holding a subjective verification session and discussion. The general outline of the week's activities can be found below in Table 3.

Table 3: Timetable of intensive week activities

Monday	Tuesday	Wednesday	Thursday	Friday
Orientation	Dr. Farrar Drop-in	Case 2 Briefing & Forecast Activity	Case 3 Briefing & Forecast Activity	Case 3 Verification
Case 1 Briefing	Case 1 Forecast Activity	Break	Break	Break
Break	Break	Case 2 Forecast Activity	Seminar	General Discussion
Case 1 Forecast Activity	Seminar	Case 2 Verification	Case 3 Forecast Activity	
	Case 1 Verification			

The first intensive week occurred February 7-11, 2022. Participants ranged from NWS forecasters from both WFO and National Centers, FEMA personnel, University researchers, and

model developers. Figure 2 outlines an approximate geographic representation of the week's participants, as well as a photograph taken on the final day. During this first intensive week, participants evaluated cases 4, 7, 10, and 11 (Table 1). Case 4 was a snow only event over the upper Midwest, mainly focused in Central MN. The discussion and evaluation were focused on messaging and drawing snow gradients through a large metro area (Twin Cities, MN). Case 7 was a snow only event over the Intermountain West. The discussion and evaluation were focused on elevation and coastal snow. With most of our participants from the Eastern CONUS, this was their first exposure to the challenges faced in Western Region. The final cases 10 and 11 focused on a Nor'Easter that tracked from the Mid-Atlantic to New England over two days. For this case, timing and location of the heaviest snow was again the focus of discussion and evaluation.

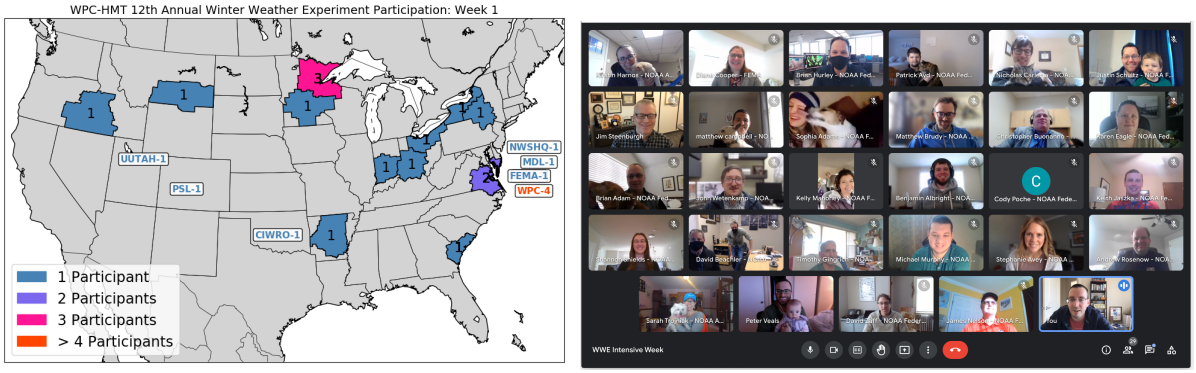


Figure 2: Intensive week 1 participants

The second intensive week occurred February 28 - March 4, 2022. Participants again ranged from NWS forecasters from both WFO and National Centers to University researchers and model developers. Figure 3 outlines an approximate geographic representation of the week's participants, as well as a photograph taken on the final day. During the second intensive week, participants evaluated cases 20, 22, and 23 (Table 1). Case 20 was a blizzard in New England. Discussion and evaluation were focused on timing, location, and snow-to-liquid ratios (SLRs). The other cases for the week, 22 and 23, were focused on precipitation type. These cases were the first attempt at reintegrating precipitation type issues back into WWE evaluations. Without official observations to validate against, it was difficult for the participants to give insights into how the experiment data dealt with ice and freezing rain. These cases also provided the opportunity to craft ideas for future experiment exercises. For example, should participants draw regions of different precipitation types as separate freezing rain, sleet, and snow shapes, should they match model output with snow and sleet grouped together, or as sleet/freezing rain with a separate snow shape? Alternatively should participants only focus on the leading edge of the transition zone?

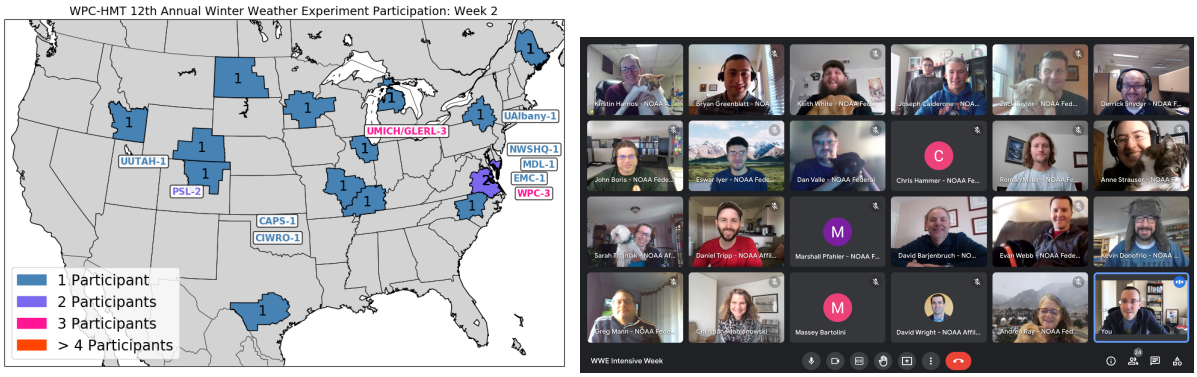


Figure 3: Intensive week 2 participants

As stated earlier, session days were structured roughly as follows: led by a WPC forecaster, each case began with a synoptic overview to orient the experiment participants to the winter weather forecast issue of the day. The majority of the time was then spent on the forecast activity, called the Maximum Snowfall and Timing Product (MSTP), where 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, precipitation type (sleet or freezing rain), and mark where the largest precipitation rates will occur within their snowfall footprint. New to WWE this year, the team split the participants into breakout groups to complete the MSTP exercise. While the MSTP is not new to WWE, the attempt to use breakout groups and collaborate a forecast was an experiment within the experiment this year. The idea behind the breakout groups was to use two different collaboration methods, a "WFO group" where participants would focus on smaller areas within our forecast region and then collaborate to stitch together the forecast. And a "regional group" where participants would look at the region as a whole and work together to create the forecast over the full region. Overall, the breakout groups were more successful once participants got comfortable with the format and understood what the experiment objectives were. Lessons learned and recommendations for these breakout groups will be addressed in the summary section.

3. Experiment Findings and Results

At the completion of each case, participants were asked to provide subjective evaluations on the experiment data. These verification sessions compared the 24 hour change in snow depth from the experimental data to the NOHRSCv2 snowfall analysis and were completed through the use of google surveys. In addition to the NOHRSCv2 snowfall accumulations, participants had access to objective statistics from the Method for Object-Based Diagnostic Evaluation

(MODE) to aid in the evaluation process[4]. MODE is part of the Model Evaluation Tools (MET) package[5] (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. An example slide that was presented during WWE and outlines what information is provided from the MODE maps can be found in figure 4.

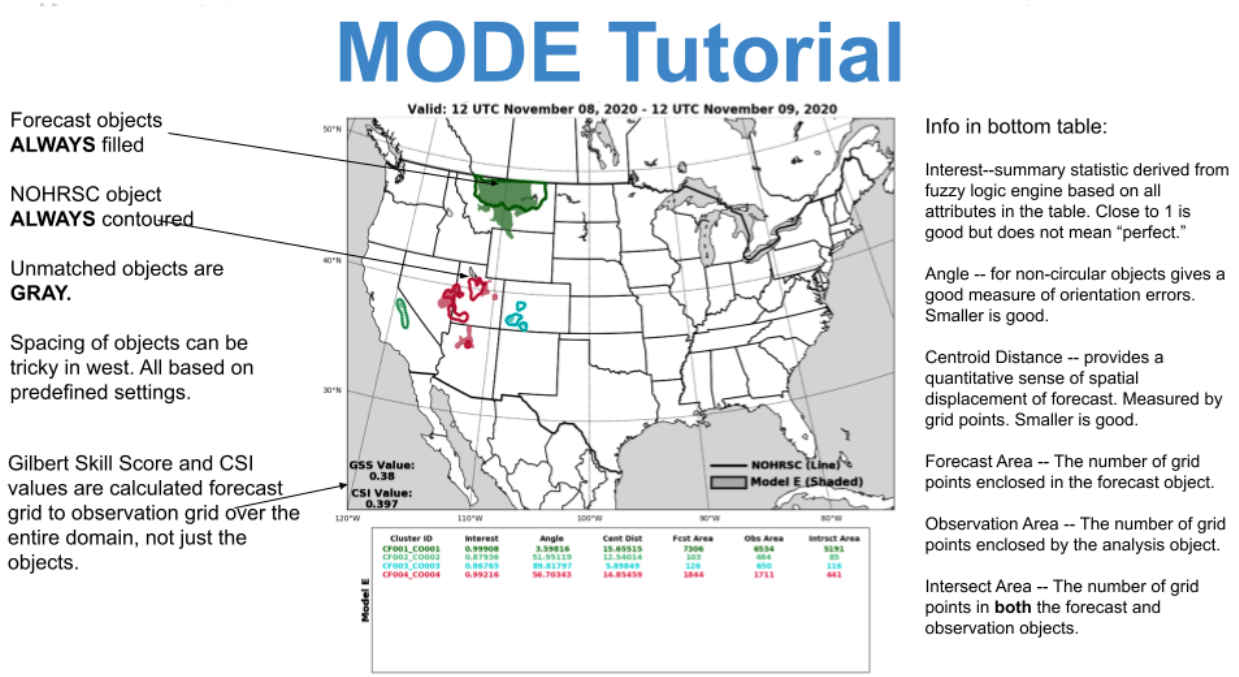


Figure 4: MODE tutorial slide presented during WWE.

Along with the MODE maps showing the experimental and NOHRSCv2 spatial comparisons, participants were also given access to full CONUS performance diagrams derived from the statistics output from MODE. These diagrams are calculated over the full CONUS domain and provide the success ratio on the x-axis, the probability of detection on the y-axis, bias values on the dashed diagonal lines, and CSI values as the curved lines. In general for a specified threshold, the closer a forecast is to the upper right corner, the better the forecast. Figure 5 provides the tutorial for how to interpret the diagrams which were computed for the same 24 hour snowfall accumulation thresholds as the spatial MODE maps.

[4] WWE MODE verification page can be found here: https://origin.wpc.ncep.noaa.gov/hmt/hmt_webpages/mode/wwemode_int.php

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

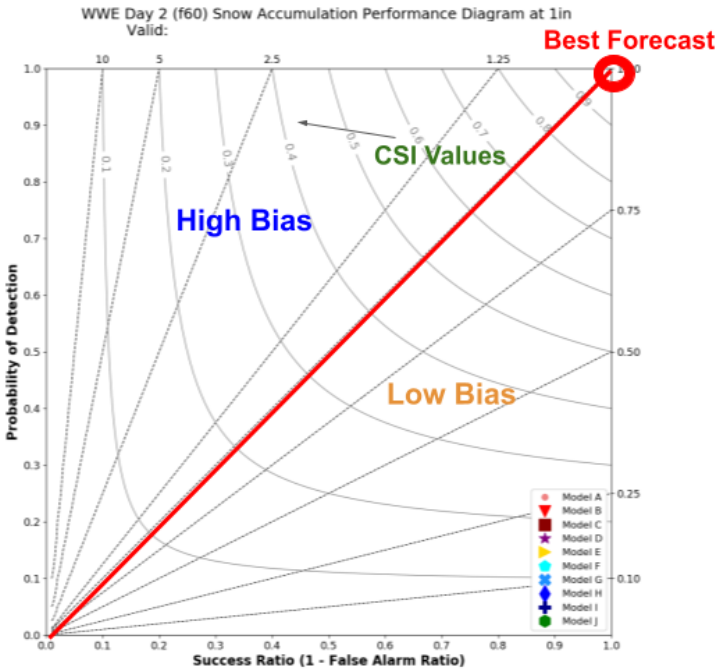


Figure 3: Roebber Performance diagram. X-axis represents the success ratio, y-axis represents the probability of detection. Dashed diagonal lines are bias values. Curved lines are CSI values. In general for a specified threshold, the closer a forecast is to the upper right corner, the better the forecast.

3.1 Subjective Survey Results

In previous experiments, the subjective survey asked participants to either rank the models or score the experiment models on a scale from 1-10. The comments from this method were described as an “eye-test” where people felt they were sometimes selecting arbitrary ranks due to a lack of discernible differences and scores generally hovered around middle values. Therefore it was difficult to provide strong recommendations based on participant feedback for which were the preferred datasets. This year the Google survey used for evaluations was structured differently in that it asked participants to directly compare the related experimental datasets to each other at the differing lead times. The first question asked was to indicate which case they were evaluating. Figure 4 shows the distribution of the evaluations is relatively evenly spread amongst the six cases evaluated in the verification sessions. Note that the 2 January - 3 January case does not have any responses. Due to time constraints and the fact that this was a multi-day event, the WWE team decided to focus solely on the 3 January - 4 January case for the verification session.

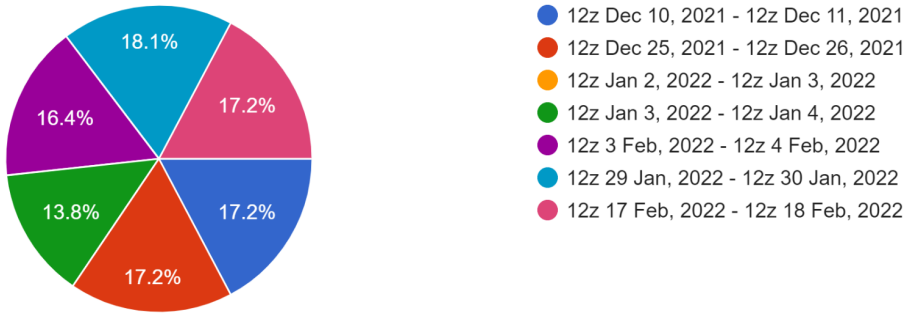


Figure 4 : Distribution of Case Evaluations

Evaluation questions were grouped based on the experimental objectives and configuration of the models. For each question group, participants were asked to choose which model configuration performed better at the day 3, day 2, and day 1 lead time. Based on participant feedback, a selection of "no significant difference" was also added to remove the "eye-test" situations when the model results were very similar. They were also asked "How did you make that decision and why?" The first grouping focused on the NBM (Figure 5) where they were asked to first compare the deterministic NBMv4.0 to NBMv4.1.

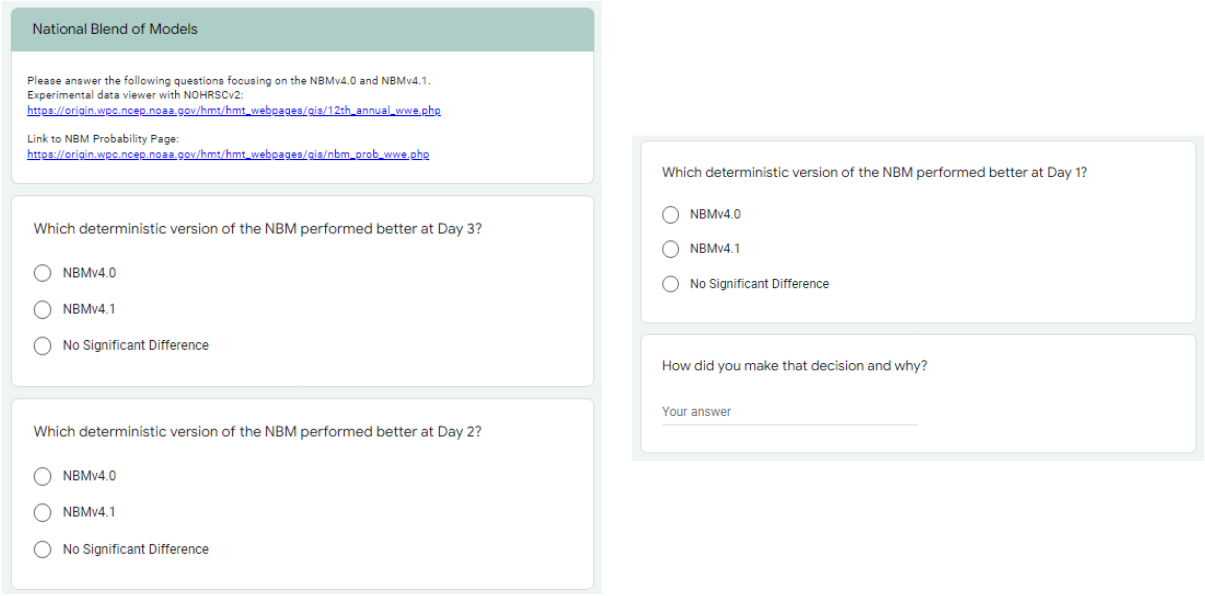


Figure 5: Subjective evaluation questions related to the deterministic NBM .

Results to these questions are shown in figure 6. Interestingly, at the day 3 and day 1 lead time, NBMv4.0 was overwhelmingly preferred to the NBMv4.1, but at day 2 NBMv4.1 was nearly tied with NBMv4.0 in preference. It should also be noted that participants found "no significant differences" in at least 10% of the cases for each lead time. Responses to "how did you make

that decision and why?" were consistent across cases and intensive weeks. Participants found NBMv4.1 to be consistently overdone in snow amounts. Footprints were similar between the two versions but the higher amounts in NBMv4.1 were disappointing to many of the respondents. This led to discussion about how SLRs are handled differently in the versions and if the same SLR reduction technique was applied to NBMv4.1 as is currently done in NBMv4.0, the amounts would probably be similar. This survey comment also highlights this point:

> *"I'm unconvinced that there is much of a difference between the two in terms of usability. NBM 4.1 does better at capturing the high end snow totals, but a good piece of that can probably be explained by the removal of the SLR weighting factor that brings down v4.0's SLR. Remove that from v4, and the forecasts are probably closer in performance (and they're already close)."*

It was also noted that NBMv4.1 amounts changed dramatically from day 2 to day 1. This is probably due to the influx of more CAM information but was concerning from a forecast consistency point of view. Some WFO participants stated this was an eye opening exercise in showing that more ensemble members does not necessarily equate to a better forecast.

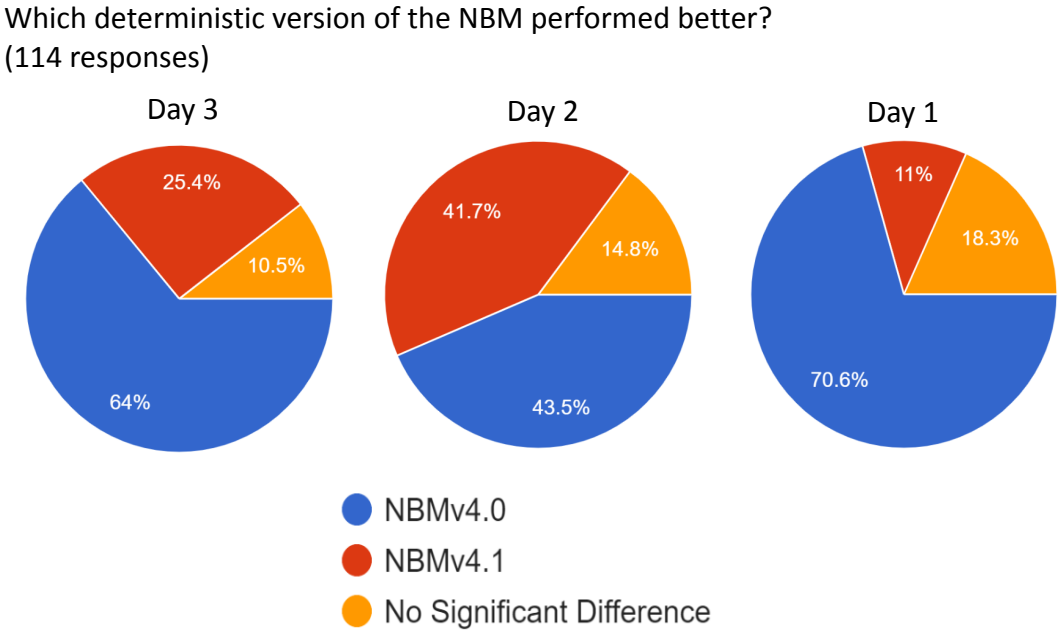


Figure 6: Deterministic NBM subjective survey responses

The second part of NBM subjective evaluation focused on the probabilities. Participants were asked (figure 7) if they used the probabilities in the forecast activity and if so, which version.

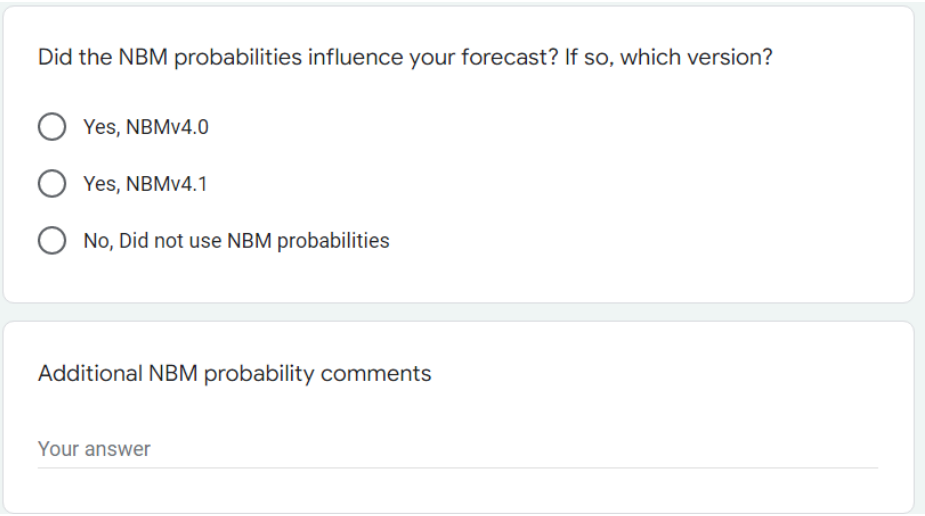


Figure 7: Subjective Evaluation questions related to the probabilistic NBM.

Over 80% of the participants responded that probabilities from either versions of the NBM did influence their forecasts (figure 8). As to which version was preferred, there was a slight preference for the NBMv4.0 (44%) to the NBMv4.1 (39%). However, most of the respondents stated that the footprints were similar so when using the probabilities to draw their snowfall footprints, it did not really matter which was selected to use. This survey comment outlines the general trend in responses to this question:

> *"The NBM probabilities were used as a first guess for drawing the location of both the 1" outline and the highest snowfall outline for days 2 and 3. I also used the NBM probs as a confidence booster for day 1 in drawing the contour of highest snowfall which worked out pretty well."*

As stated above, the NBMv4.1 consistently produced higher amounts, so when using probabilities to inform the maximum contour and amounts, participants tended to lean on the higher threshold probabilities from 4.1.

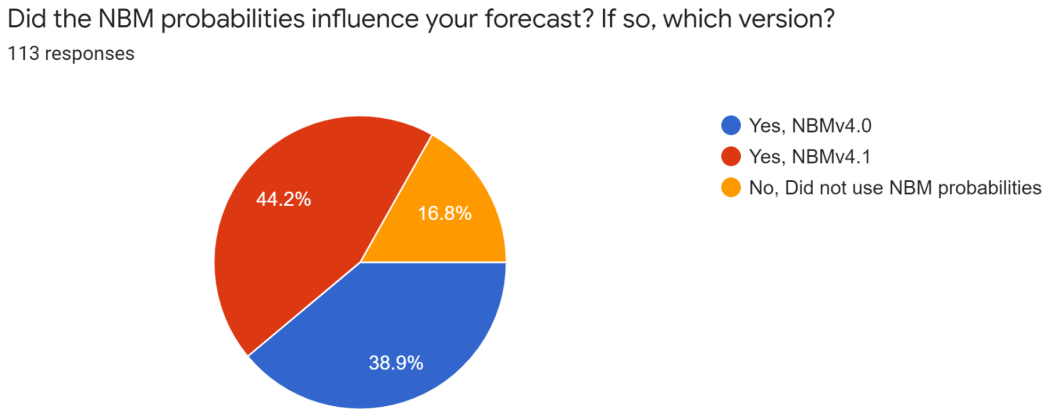


Figure 8: Probabilistic NBM subjective survey responses

The second section of subjective evaluation was focused on the different CAM configurations. Similar to the NBM evaluation, participants were asked which configuration performed better at the day 3, day 2, day 1 lead time as appropriate (figure 9; note that the EMC configurations were not run to the day 3 time), and how they made the decision they did. For the day 1 and day 2 comparisons the EMC FV3-LAM was treated as the control data set for which to compare.

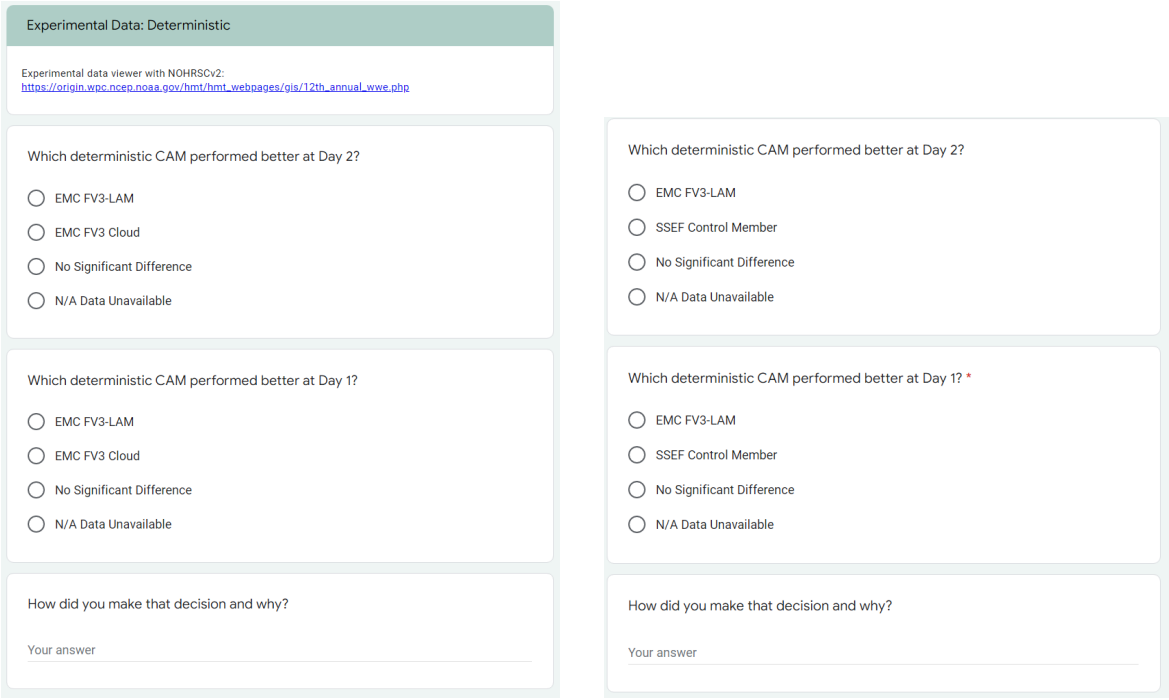


Figure 9: Subjective evaluation questions related to the CAM configurations

The first comparison was between the EMC FV3-LAM and the EMC FV3-Cloud (figure 10). At the day 2 lead time, almost 75% of respondents preferred the performance of the EMC FV3-LAM over the EMC FV3-Cloud. When evaluating these models, participants often noted a jump in the snowfall amounts in FV3-Cloud from day 2 to day 1. This is reflected in the survey responses as at day 1 both configurations are tied for which was considered better versus the strong FV3-LAM preference at day 2. This is a sample comment reflecting the jump in the FV3-Cloud and "flip-flopping" of performance between these two models:

> *"The FV3-LAM more closely resembled the observed snowfall totals for day 2 whereas the FV3 Cloud was too high for totals. Interestingly, this issue was reversed for day 1 with the FV3-LAM too aggressive with snowfall totals compared to the FV3 Cloud which more closely resembled observed snowfall totals."*

A lack of similarities in the simulations at day 2 is also reflected in only around 8% finding no significant difference with that value increasing to around 16% at day 1. It should be noted that

during the first intensive week an error was found in the FV3-Cloud simulations. EMC re-ran the simulations for the seasonal objective evaluations below, but those erroneous simulations were included in a few of the first week cases. That may explain some of the differences especially at day 2. Participants noted in their comments that they found the FV3-Cloud simulations overdone on the snowfall amounts and the large jump from day 2 to day 1 made the configuration hard to trust.

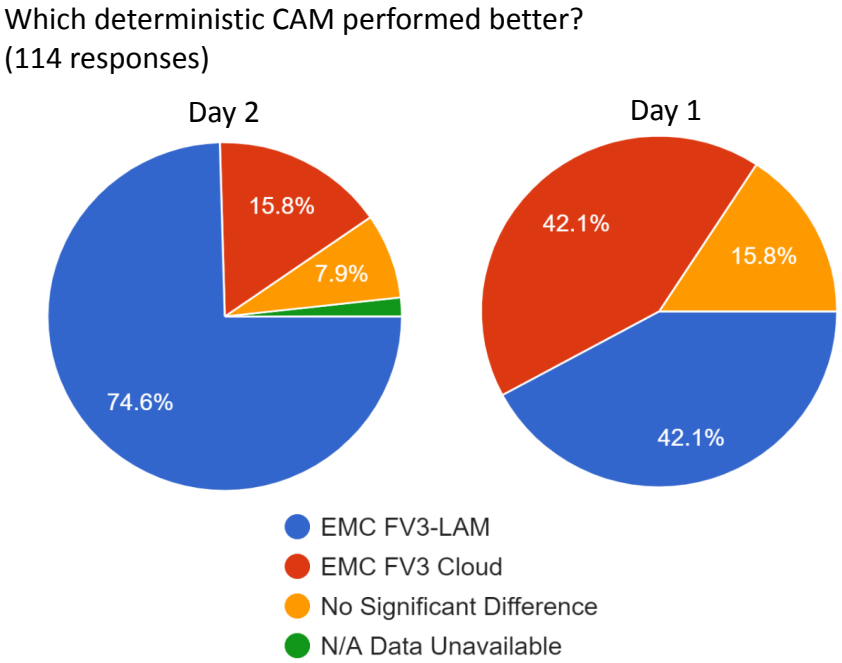


Figure 10: EMC FV3-LAM vs. EMC FV3 Cloud subjective survey responses

The second comparison was between the EMC FV3-LAM and the SSEF control member for day 2 and day 1 lead times (figure 11). Day 2 showed almost 60% of responses favoring the EMC FV3-LAM over the SSEF control member, with almost 16% finding no significant difference. While there was a large amount of missing data on day 1, the majority of the responses still favored the EMC FV3-LAM with more people (~15%) finding no significant difference between the models than favoring the SSEF control member (~7%). Comments for these selections were focused a lot on the higher snowfall amounts. For example a participant commented:

> *"In general, both the FV3-LAM and SSEF CNTL were too far south with the snowfall band. However, the most notable pitfall for the SSEF CNTL was that it was very hot with the snowfall totals. Each piece of guidance was fairly similar but the lower snowfall totals from the FV3-LAM did verify better as well as the more northern track of the band of snowfall as you headed northeast deeper into New England."*

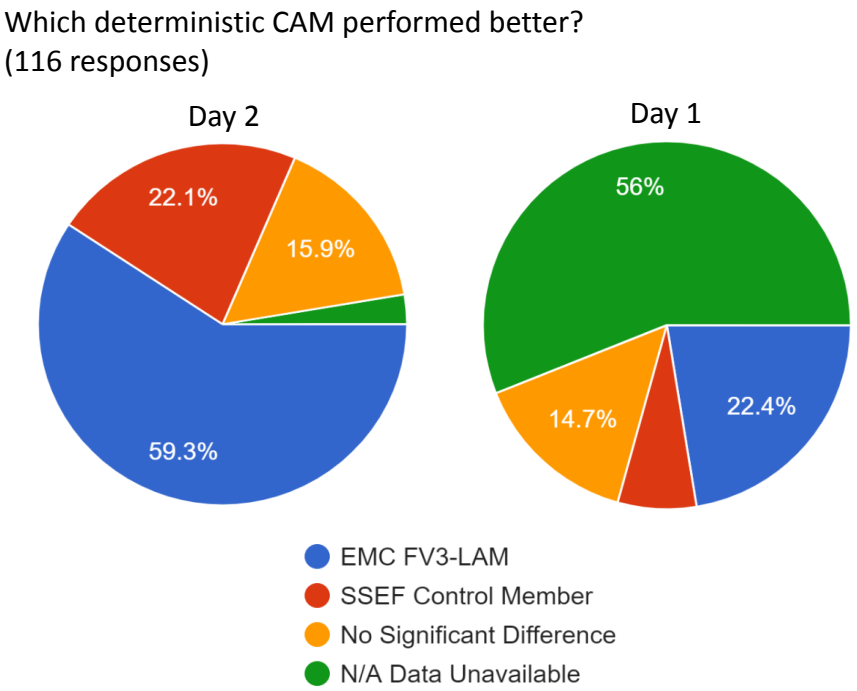


Figure 11: EMC FV3-LAM vs. SSEF Control Member subjective survey responses

One of the science objectives of this year's WWE was to compare the downscaled GFSv16 over the full CONUS to the CAMs. The past two winters have shown that this downscaling technique was preferred over the Western US, but has yet to be tested east of the Rocky Mountains. Again treating the EMC FV3-LAM as the control simulation at the day 2 and day 1 lead times, participants were asked (figure 12) to compare the downscaled GFSv16. For day 3, they were asked to compare the downscaled GFSv16 to the SSEF control member.

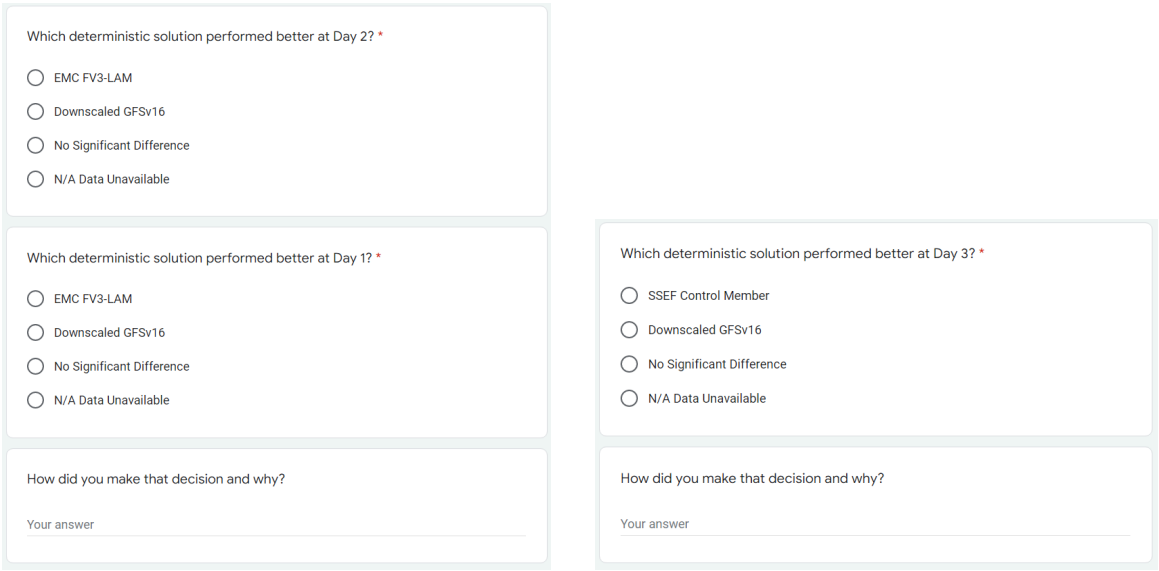


Figure 12: Subjective evaluation questions related to the downscaled GFSv16

As expected, at day 2 and day 1 lead times, participants overwhelmingly preferred the EMC FV3-LAM to the downscaled GFSv16 (figure 13). A few caveats to this result should be noted. First, precipitation types are not accounted for in the downscaled GFSv16 method. This is where the majority of the comments and issues occurred for participants in providing evaluations. Since this method is calibrated for SLRs over the Western CONUS, it was not expected to perform well over the east. However, the WWE team felt it was important to apply the method over the full CONUS this year as a starting point for future development. As one participant stated:

> *"At low-end amounts, the GFS did have higher Success scores relative to the FV3-LAM. However, once you got above 6 inches, the GFS really tailed off, largely due to its underprediction of the snowfall. Even though the LAM had a somewhat lower Success rate, its POD was higher than the GFS at all thresholds. On the map comparison, the GFS definitely underpredicted snow totals, but it did show its highest amounts in central ME, consistent with the NOHRSC. So while it had a good handle on the highest amounts, it was vastly low on totals."*

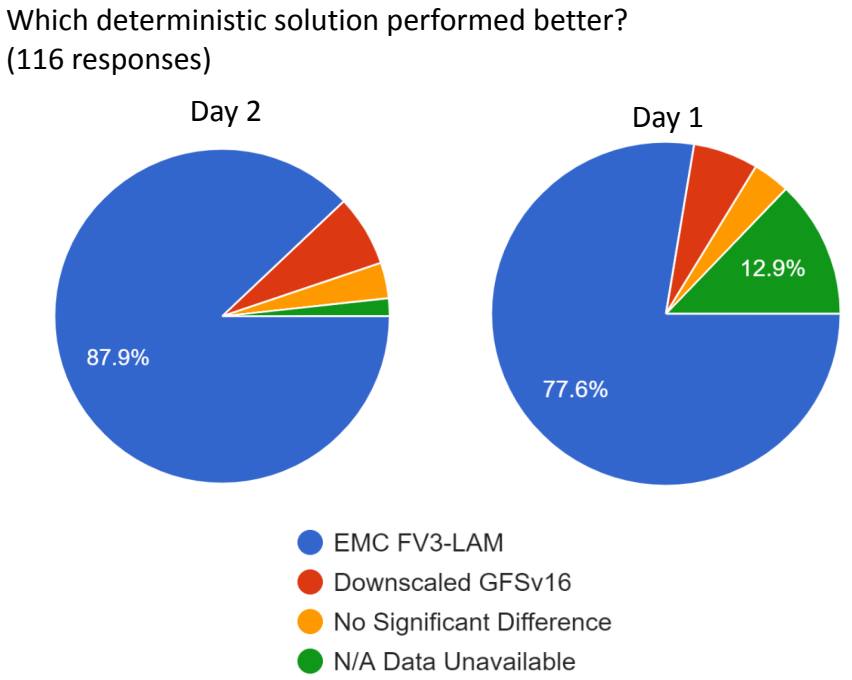


Figure 13: EMC FV3-LAM vs. Downscaled GFSv16 for Days 1 and 2 subjective survey responses

There were similar results for day 3 (figure 14) when the downscaled GFSv16 was compared with the SSEF control member. While more responses favored the SSEF control member (~38%) to the downscaled GFSv16 (~26%), many people commented that the downscaled GFSv16 had a better handle on the footprint. The issue on day 3 was that the under forecasting of amounts by the downscaled GFSv16 was a larger error than any over forecasting errors from the SSEF control member. Here is a sample comment submitted by one participant that highlights this issue:

> *"Both were not good but for different reasons. The SSEF member was an overforecast while the GFSDS was an under forecast. Picked the SSEF member as the better of the two because at least it signaled that this would be a high snowfall total event from Lake Erie to Maine."*

This again alludes to issues in the SLR methods from the West applied to the Eastern CONUS.

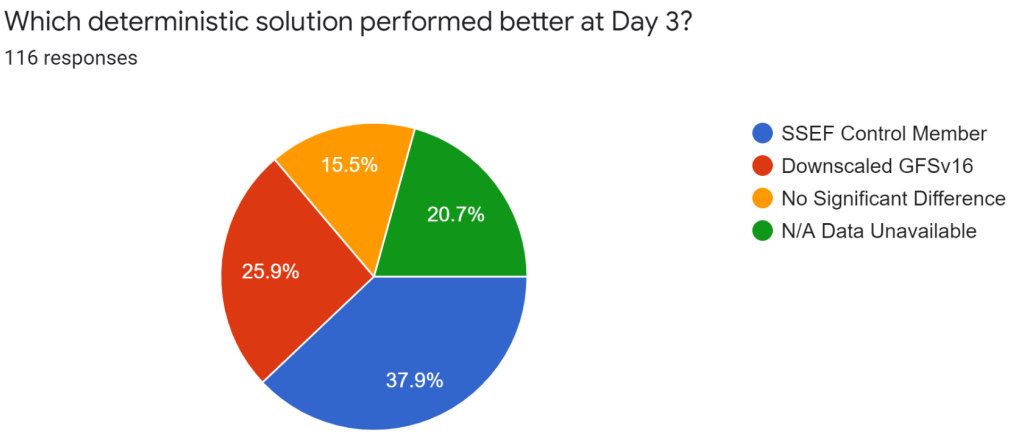


Figure 14 : SSEF Control Member vs. Downscaled GFSv16 for Day 3 subjective survey responses

The survey closed with asking which of the experimental models provided the most utility to the forecast activity (figure 15). This question highlighted the comfort and familiarity of the forecasters with the NBMv4.0. Many of the responses focused on how they used one of the NBM versions as the basis of their forecast especially when looking at the footprint. It was difficult for them to buy-in and trust the experimental CAMs for the forecast activity, but after these verification sessions, a few commented that perhaps they should have relied more on the EMC FV3-LAM. This comment stands out as a good summary:

> *"On the whole, the FV3-LAM and Cloud are useful guidance but they do some flip-flopping from time to time and it's hard to have much confidence in their deterministic solutions unless you use them in tandem with/as a nudger for NBM probabilistic information."*

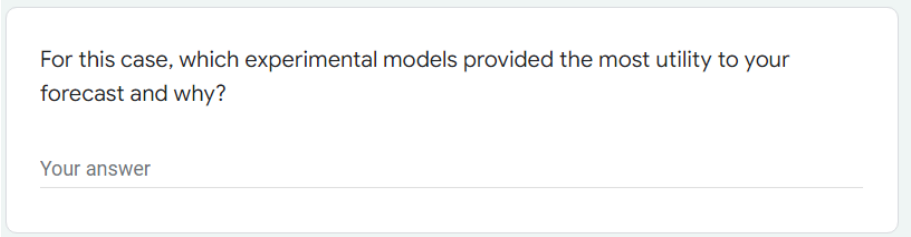


Figure 15: Final subjective evaluation survey question.

3.1.1 WSSI Results

Another component of this year's WWE was an evaluation on some potential redesign features of the WSSI. For the past year, the Nurture Nature Center (NNC) has been working with WPC and regional WFO staff to host a series of scenario-based focus groups with professional users from six regional WFO areas: Boston, MA; Hanford, CA; Grand Rapids, MI: Boulder, CO; Jackson, MS; and Omaha, NE. These focus groups were completed in two rounds: first in January 2021, second in October 2021. Additionally, NNC hosted focus groups with WFO representatives, professionals in the logistics/transportation industry and national NWS office representatives. Results from these sessions collectively informed modifications to the WSSI. These modifications were examined within the WWE in parallel with the operational WSSI. An example of what was shown can be seen in figure 16. For a single case during the intensive week (case 11 for week 1 and case 22 for week 2), participants were provided with images of the current NDFD WSSI to compare with the proposed NDFD WSSI for the overall winter storm impacts as well as any appropriate components like snow amount or ice accumulation.

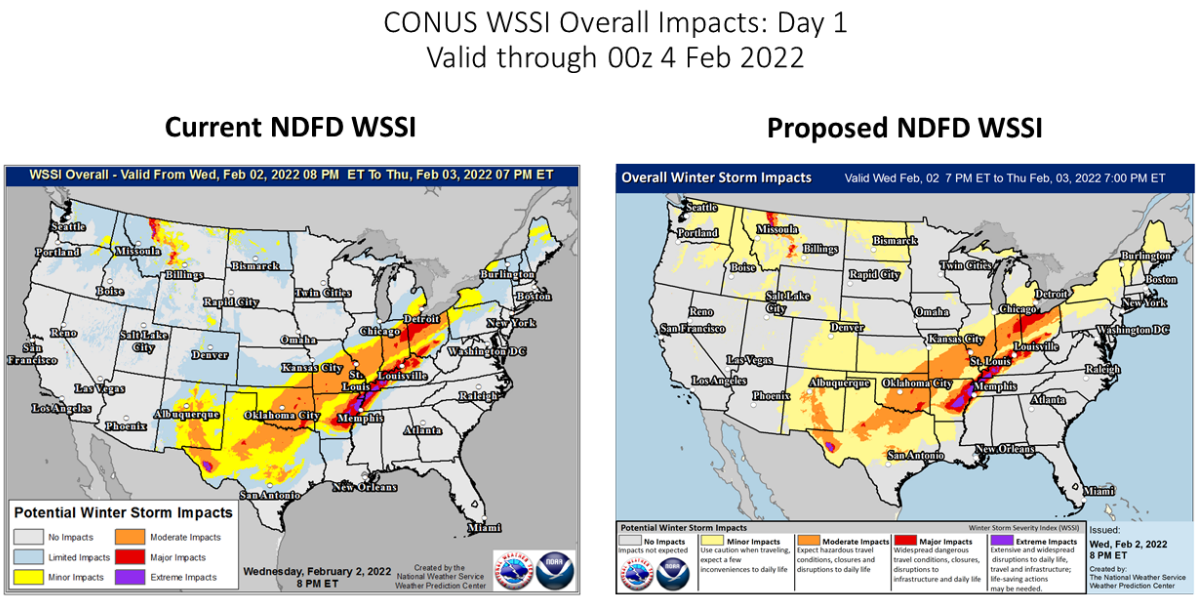


Figure 17: Sample event that compared the current operational NDFD Overall WSSI (left) to the proposed NDFD Overall WSSI (right) as shown to WWE participants.

Participants were then asked to compare the current and proposed impact definitions (figure 17) from which nearly 85% preferred the proposed legend, colorscale, and definitions. During both intensive weeks, participants noted the clarity in wording with the proposed scale for example, one participant commented:

> *"The proposed impact wording is so much more relatable."*

and the removal of the "limited" category as the biggest improvements:

> *"The public doesn't understand the difference between limited and minor, in my experience. Combining them into minor is a great idea."*

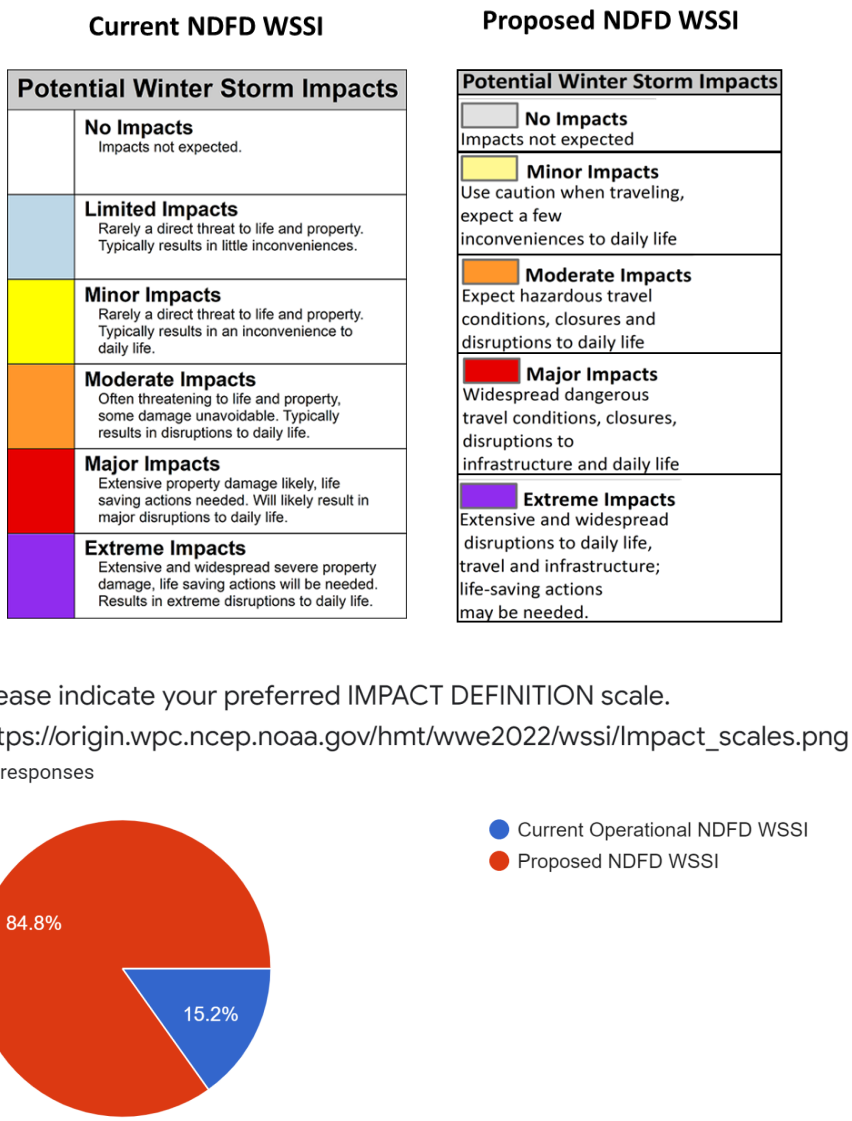


Figure 17: Comparison of current operational NDFD WSSI color scale and legend (top left) to the proposed NDFD WSSI color scale and legend (top right) as shown to WWE participants. Survey result for the comparison of the WSSI legends and color scales (bottom).

This participant comment articulated the purpose of this study and reflected the thoughts of the general WWE participants well:

> *"Thanks for your hard work on improving this tool so that 1) we as forecasters can feel more comfortable using it in our messaging and 2) our partners can feel more comfortable with the impact statements in informing their decision making."*

3.2 Seasonal Performance

Performance diagrams and seasonal snowfall difference maps were computed over the 29 case WWE season, data availability dependent, to provide a more objective evaluation into how well the experiment datasets did this year. It should be noted that these diagrams are calculated over the full CONUS domain at the day 3 (FH84), day 2 (FH60), and day 1 (FH36) lead times. In addition to the experimental models listed in Table 2, the diagrams also include three additional values from the CAPS SSEF ensemble: the ensemble mean, the probability-match mean (PMM), and the local probability-match mean (LPMM). While these ensemble means were not subjectively evaluated, they were available to participants during the forecast activity with many participants choosing one of these means as an influence to their forecast. Also aligned with the science and operations objectives to evaluate both deterministic and ensemble products, the team felt it was important to incorporate the ensemble values in this evaluation.

For day 3 (figure 18), the seasonal difference between the experiment data and NOHRSCv2 highlights the regional biases present in the experimental models. High biases as indicated by the blue shading are very prevalent in both versions of the NBM over the majority of the CONUS. The magnitude of the high bias is slightly larger in NBMv4.1. This difference could be attributed to the different treatment of the SLR between the versions and was a topic participants wanted to discuss frequently during the intensive weeks. Low biases indicated by red shading are mostly seen from North Dakota to Wisconsin and some of the terrain over the West. The NBMv4.0 also has a low bias along the east coast from Virginia to Maine. The other deterministic datasets are more regional in their biases with the CAPS SSEF control member displaying a low bias over most of the CONUS with the exception of the Northern Plains and 2 storm tracks where WWE sampled mixed precipitation type cases. The downscaled GFSv16 is divided by the Rocky mountains, with a high bias in the Western CONUS, consistent with the results of past WWEs, and a low bias in the Eastern CONUS. Throughout the intensive weeks, participants routinely commented on the low bias and discussed how the SLR techniques did not translate east of the Rockies. The bottom row of the panel displays the three ensemble means from the CAPS SSEF and shows the smallest biases among the experimental models throughout the CONUS, with the mixed precipitation cases showing a high bias similar to the

control member. Performance diagrams for day 3 give additional insights into whether these biases exist at various accumulation thresholds.

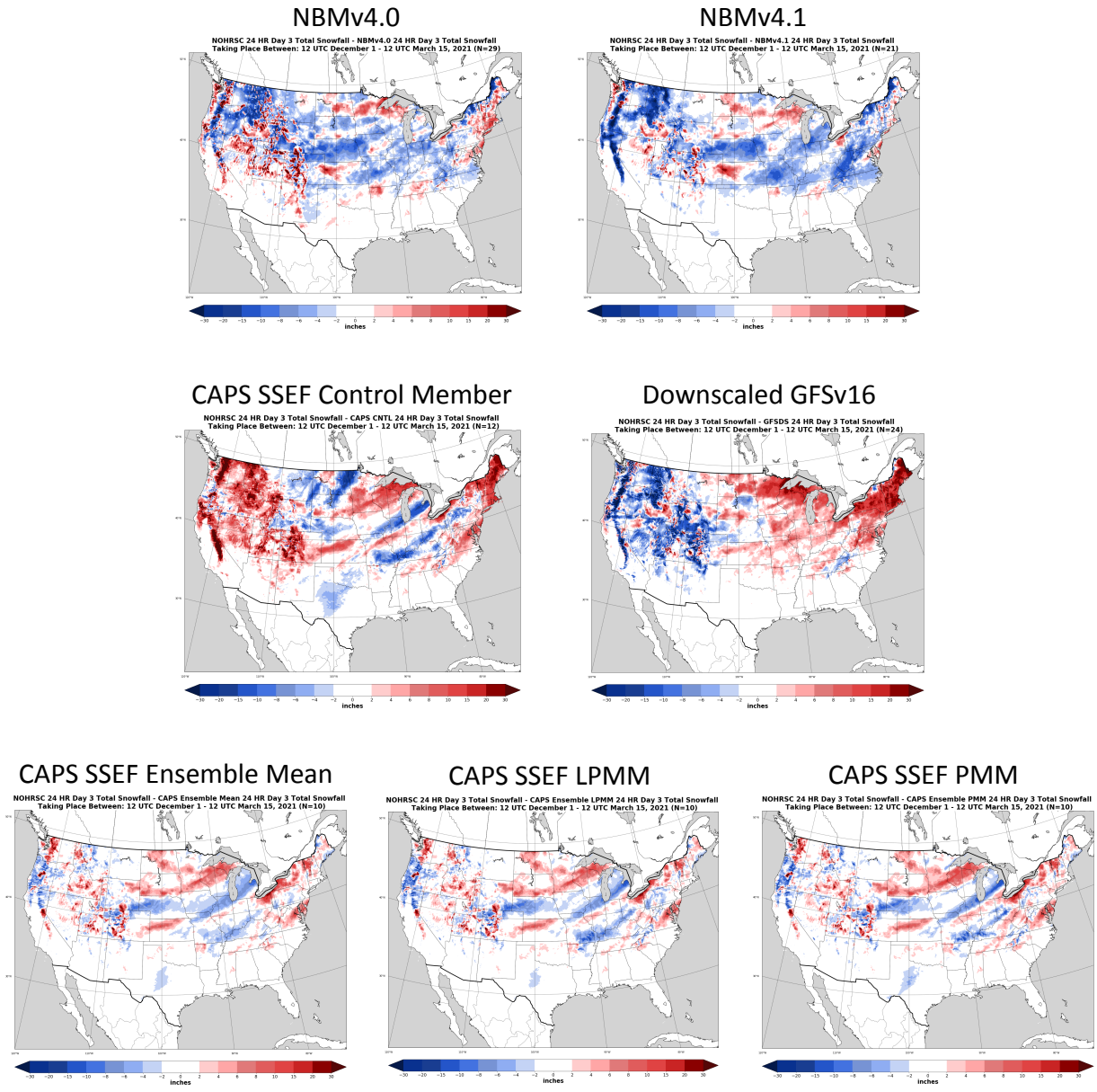


Figure 18: Differences between experimental models and NOHRSCv2 over the available WWE cases at day 3 (FH84). Blue shading indicates experimental model high bias, red shading indicates low bias.

At day 3 (figure 19), all models show a high bias for the 1 inch threshold (upper left). The CAPS SSEF ensemble mean, PMM, and LPMM are clustered together (purple markers) and closest to the upper right corner of the diagram, suggesting that these were the best performers for capturing the snowfall footprint. Interestingly, the CAPS SSEF control simulation (purple square)

has the lowest bias, but also the lowest CSI and diagram position although it should be noted there are fewer cases included in the CAPS SSEF evaluation (Table 2). As was noted throughout the subjective evaluation, the NBMv4.1 (red triangle) exhibits a strong high bias which is also present in NBMv4.0 (pink circle). In the middle of the pack is the downscaled GFSv16 (blue triangle) which is in line with the NBMv4.1 for CSI values with a slightly smaller high bias.

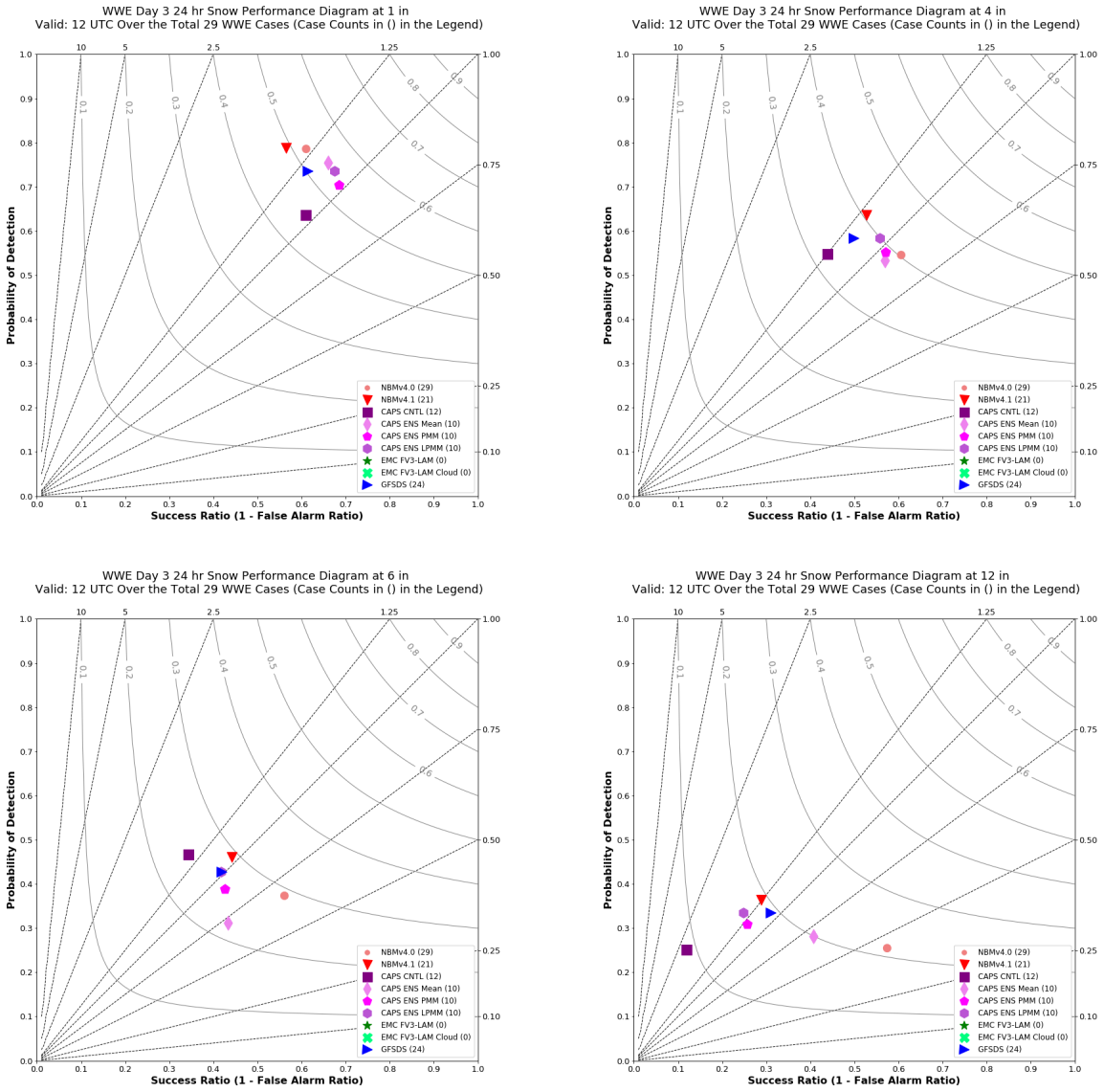


Figure 19: Seasonal performance diagram at the day 3 lead time at the 1 inch (upper left), 4 inch (upper right), 6 inch (lower left), and 12 inch (lower right) thresholds. The number of cases included for each model is listed in the legend.

As the threshold value increases to 4 inches (figure 19, upper right), the CAPS SSEF control member remains the lowest for CSI and position on the diagram. The biggest improvement is

the NBMv4.0 which decreased the bias and is now in line with the CAPS SSEF ensemble mean, PMM, and LPMM. The downscaled GFSv16 performance remains consistent in relation to the other models. At the highest thresholds of 6 inches (figure 19, lower left) and 12 inches (figure 19, lower right) the models start to diverge from one another, probably due to sample size. Notably the NBMv4.0 has a distinct low bias where the NBMv4.1 retains the slight high bias. This is again consistent with the subjective evaluations that found the NBMv4.1 to have higher values at the higher thresholds and the SLR reductions in NBMv4.0 decreasing the amounts. The downscaled GFSv16 shows minimal bias, one of the highest CSI scores, and one of the best positions on the performance diagram at these larger thresholds as well. The high bias in the CAPS SSEF control member continued to increase with the increasing threshold. It also continued to be among the worst for positioning on the performance diagram. The CAPS SSEF ensemble mean remained relatively steady in its position, bias, and CSI value for the 6 and 12 inch thresholds. The others display a noticeable high bias jump from 6 inches to 12 along with a decrease in positioning on the diagram.

As the lead time moves to day 2 (FH60), the spatial maps (figure 20) show similar bias patterns as were seen in day 3. Both versions of the NBM still have the CONUS-wide high bias, the magnitude of which has increased from day 3 to day 2 possibly due to the configuration and number of models included in the blend as the lead time decreases. The downscaled GFSv16 sees a slight decrease in bias values with the decreasing lead time with the continued split along the Rocky Mountains. For day 2, the FV3-LAM and FV3-Cloud data are now available and show similar bias patterns to the CAPS SSEF control member. All three of these FV3 based CAM models have a pronounced low bias over the terrain in the West, Great Lakes region, and New England. This is countered by the high bias again seen over the mixed precipitation type case storm tracks in the central US to the Ohio River Valley. One possible explanation for these bias patterns could be due to displacement of the heaviest snows, something that was noted in several cases during the intensive weeks. The smallest bias values again belong to the various ensemble mean techniques from the CAPS SSEF with the mixed precipitation storm tracks standing out as a swath of high bias among all 3 of the ensemble means.

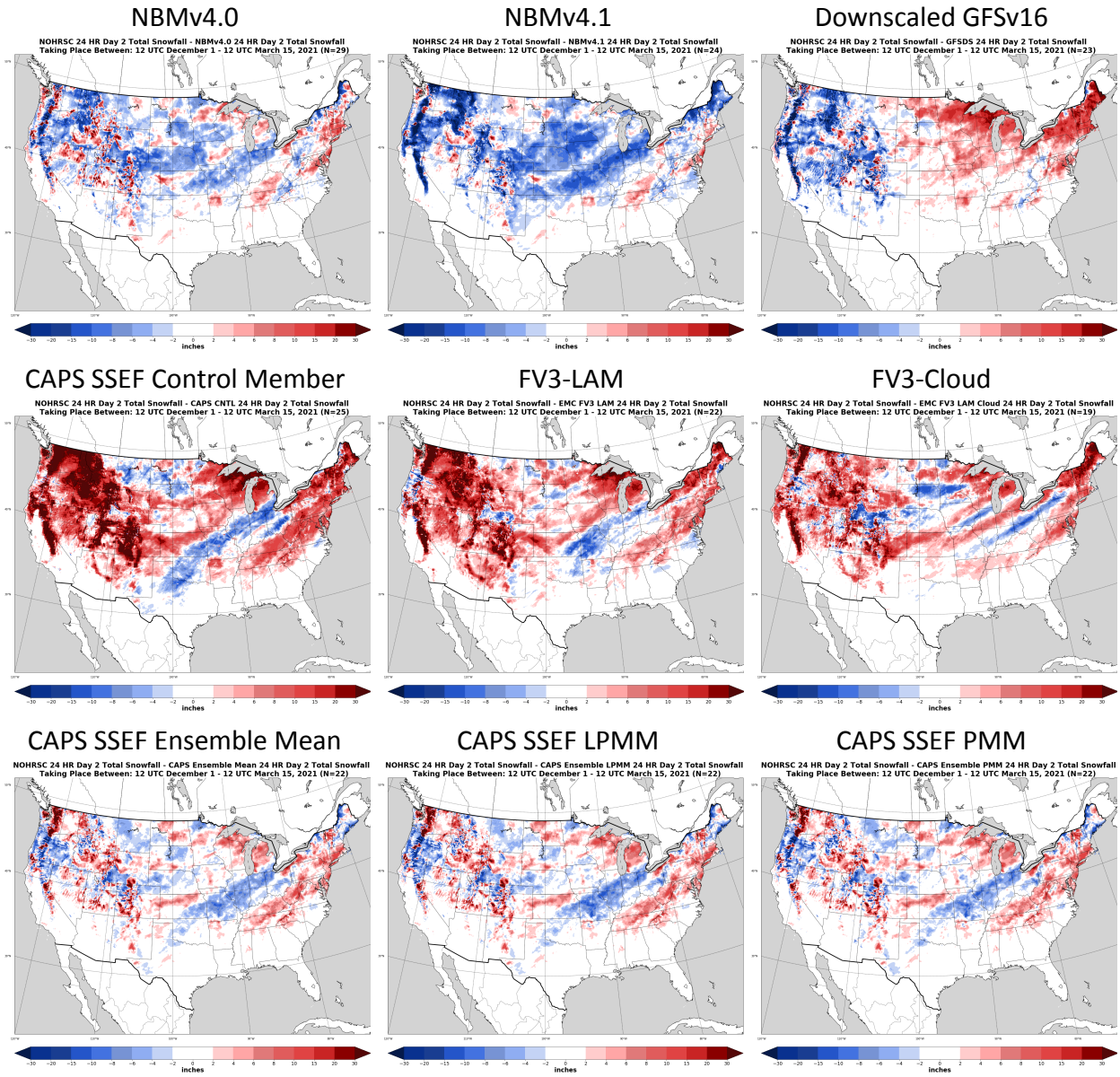


Figure 20: Differences between experimental models and NOHRSCv2 over the available WWE cases at day 2 (FH60). Blue shading indicates experimental model high bias, red shading indicates low bias.

At day 2 (figure 21) NBMv4.1 has the largest high bias among the experiment data at all thresholds. That bias also increases slightly with increasing snowfall thresholds, while CSI decreases. The NBMv4.0 has the second highest bias at the 1 inch threshold; however that bias disappears at the higher amounts and even becomes a low bias at 12 inches. The other consistent model that stands out is the CAPS SSEF control member. While it has a relatively minimal bias value, it has the lowest position on the performance diagram with the lowest CSI value for all thresholds. The remaining CAPS SSEF ensemble means are among the best for highest CSI value and diagram placement. They also remain clustered together at all the amount

thresholds. The downscaled GFSv16 bias remains relatively consistent with performance diagram placement in the middle of the pack. Finally, day 2 includes the EMC FV3-LAM (green star) and FV3-Cloud (green square). Both of these models are relatively similar, with bias and CSI values closest to the downscaled GFSv16.

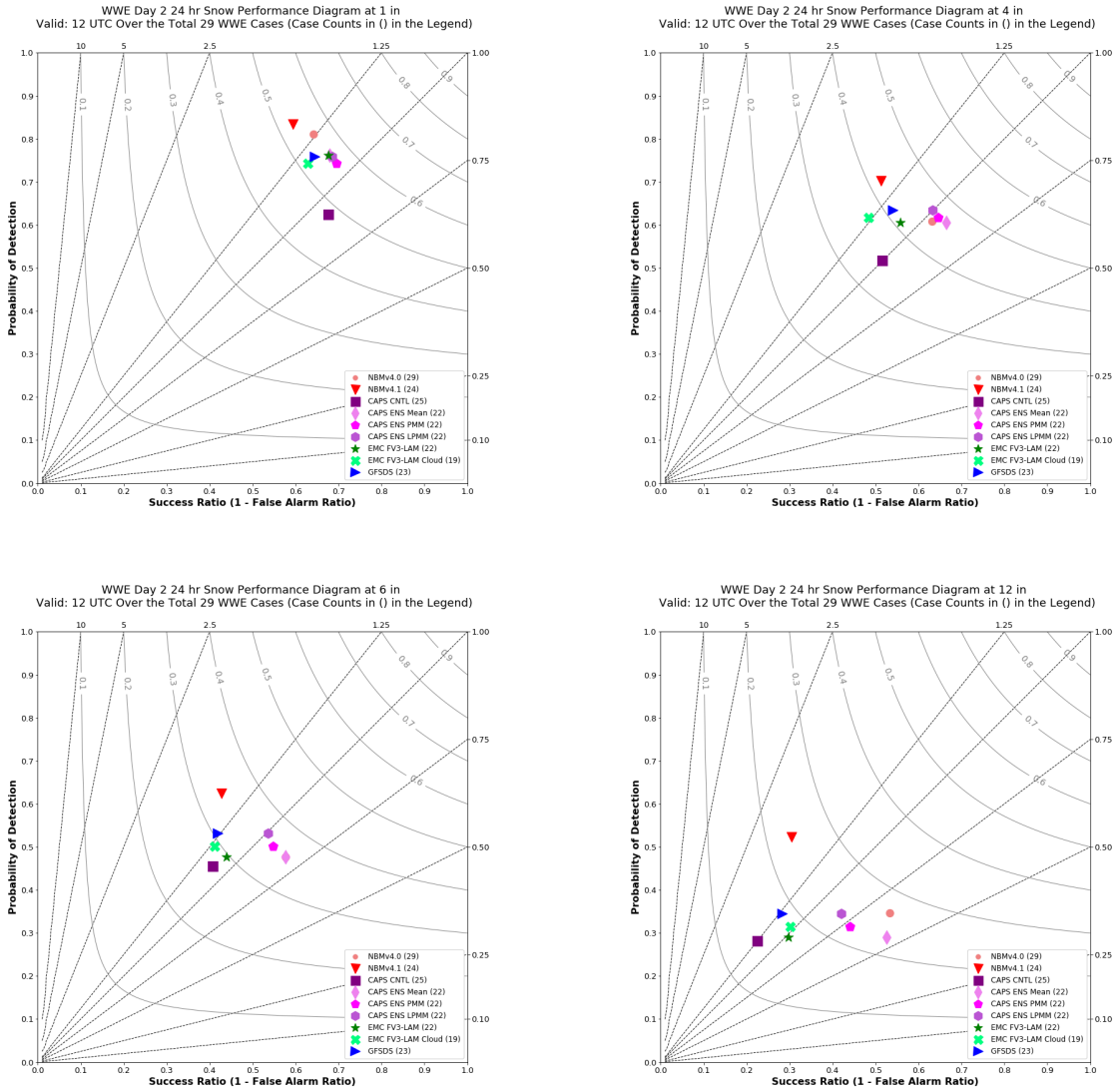


Figure 21: Seasonal performance diagram at the day 2 lead time at the 1 inch (upper left), 4 inch (upper right), 6 inch (lower left), and 12 inch (lower right) thresholds. The number of cases included for each model is listed in the legend.

Finally, as the lead time decreases to day 1 (FH36), the spatial maps show the same patterns of biases as seen in day 2 and day 3 with a decrease in the magnitude. The main standout feature among the CAMs continues to be the high bias over the mixed precipitation storm track from

the central plains to the Ohio River Valley. This alludes to an issue in how precipitation types are handled within the CAMs and will need to be addressed in the future. The FV3-Cloud appears to have the smallest bias over this storm track. These spatial maps also show how poorly the deterministic CAMs handle change in snow depth over the Western CONUS which again, will need to be addressed.

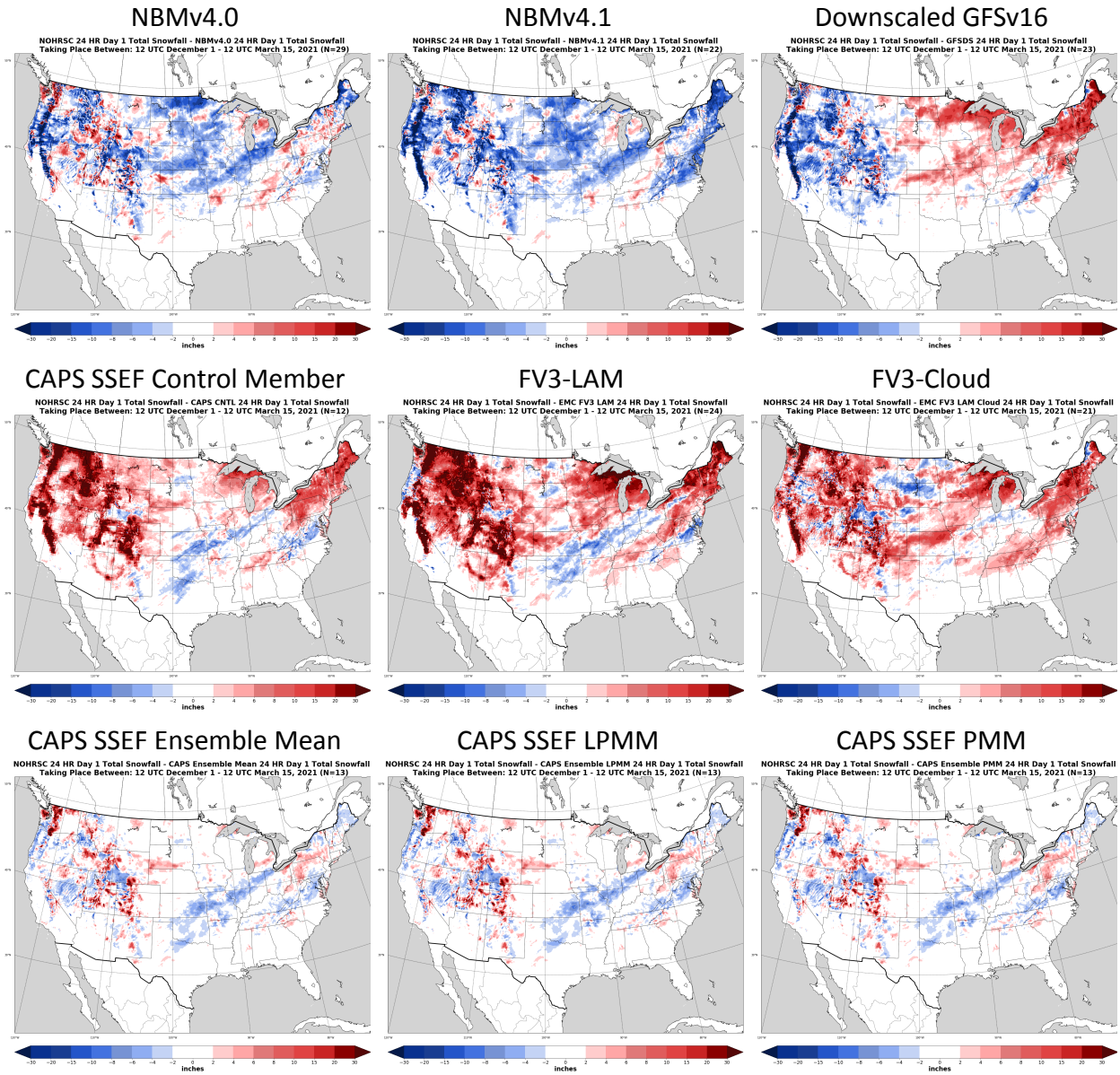


Figure 22: Differences between experimental models and NOHRSCv2 over the available WWE cases at day 1 (FH36). Blue shading indicates experimental model high bias, red shading indicates low bias.

The 1 inch threshold for day 1 (figure 23) shows all the models clustered together with the exception of the CAPS SSEF control member which has the lowest bias and CSI value for all

amounts. Similar to day 2, the NBMv4.1 retains the high bias with increasing threshold amounts. The NBMv4.0 also retains the high bias at all thresholds, but again it is not as large as the NBMv4.1. The remaining three CAPS SSEF ensemble means are again clustered together. They have the lowest bias values for the 4 and 6 inch thresholds with a low bias more pronounced at 12 inches. This may be due to sample size, as there were only 13 out of the 29 cases for the CAPS contributions.

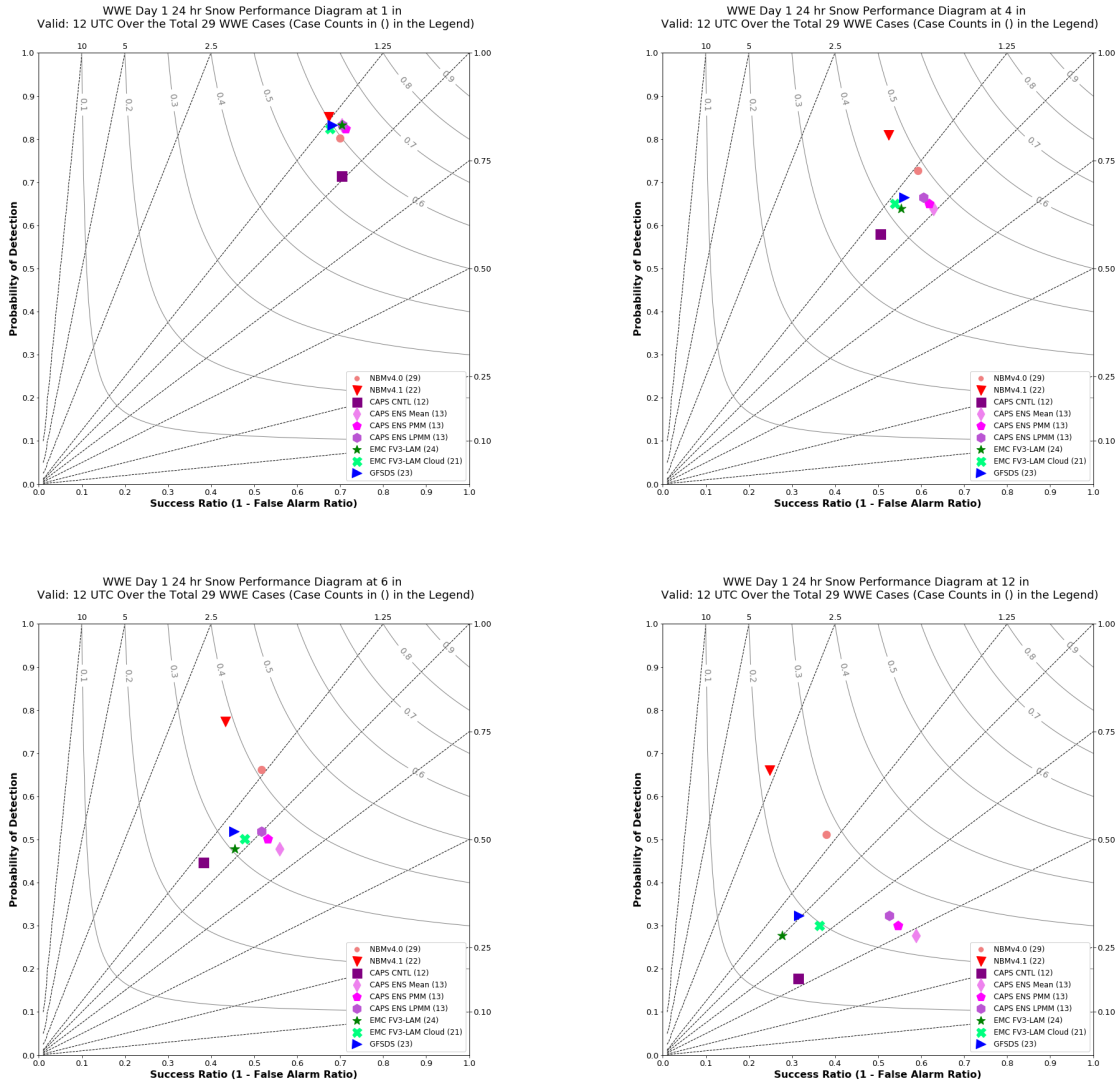


Figure 23: Seasonal performance diagram at the day 1 lead time at the 1 inch (upper left), 4 inch (upper right), 6 inch (lower left), and 12 inch (lower right) thresholds. The number of cases included for each model is listed in the legend.

Generally speaking, the NBMv4.1 had a noticeable high bias at all lead times and thresholds. This was something that was echoed in the subjective evaluations as well. The CAPS SSEF

control member had one of the smallest biases but also had one of the lowest CSI values at all lead times and thresholds. The downscaled GFSv16 was relatively in the middle of the pack for CSI values and seemed to have a slight high bias for all lead times and thresholds. Past WWEs have highlighted how well this method works over the west, so while the subjective evaluations found issues in applying this downscaling method over the eastern CONUS, the performance diagram metrics do not reflect poor performance due to the calculations occurring over the full CONUS. Finally there was not a huge difference between the FV3-LAM and FV3-Cloud configurations. The FV3-LAM may have a slightly better CSI value at day 2 with the FV3-Cloud slightly better at day1, but both models seemed to be clustered together at both lead times and all threshold values.

3.3 Highlighted Events

While the WWE managed to capture and evaluate several impactful events, there are two events the team would like to highlight. These cases were representative of the intensive week evaluation process and provide an opportunity to describe details of the performance of some of the experimental models.

Case 11: 12z 3 January 2022 - 12z 4 January 2022

From the first intensive week, case 11 (table 1) was an impactful Nor'Easter that proved challenging for the models in terms of amounts and position. Figure 24 shows the NOHRSCv2 24 hour snowfall analysis for this event. For the WWE evaluations, we focused on the snowfall event in the Mid-Atlantic where a 1 inch snowfall footprint runs from North Carolina to New Jersey with an 8 inch maximum contour from Northern Virginia to southern New Jersey.

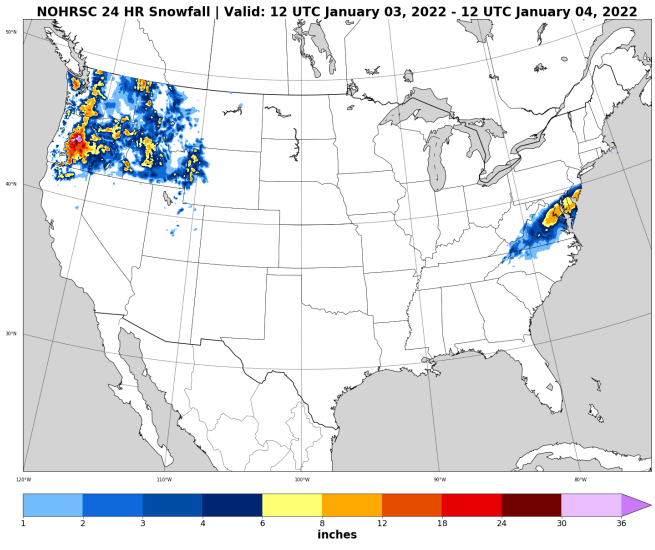


Figure 24: NOHRSCv2 24 hour snowfall analysis for case 11

Overall, the experimental data did not have a great handle on this event. For day 3 (not shown), all of the experiment models have the low track too far offshore resulting in little to no accumulation along the coast. By day 2, a few of the models begin to show the possibility of snow, the CAPS control member (figure 25) has a swath of precipitation that extends farther north and east than the NOHRSCv2 analysis for both the 1 inch footprint and the 4 inch threshold.

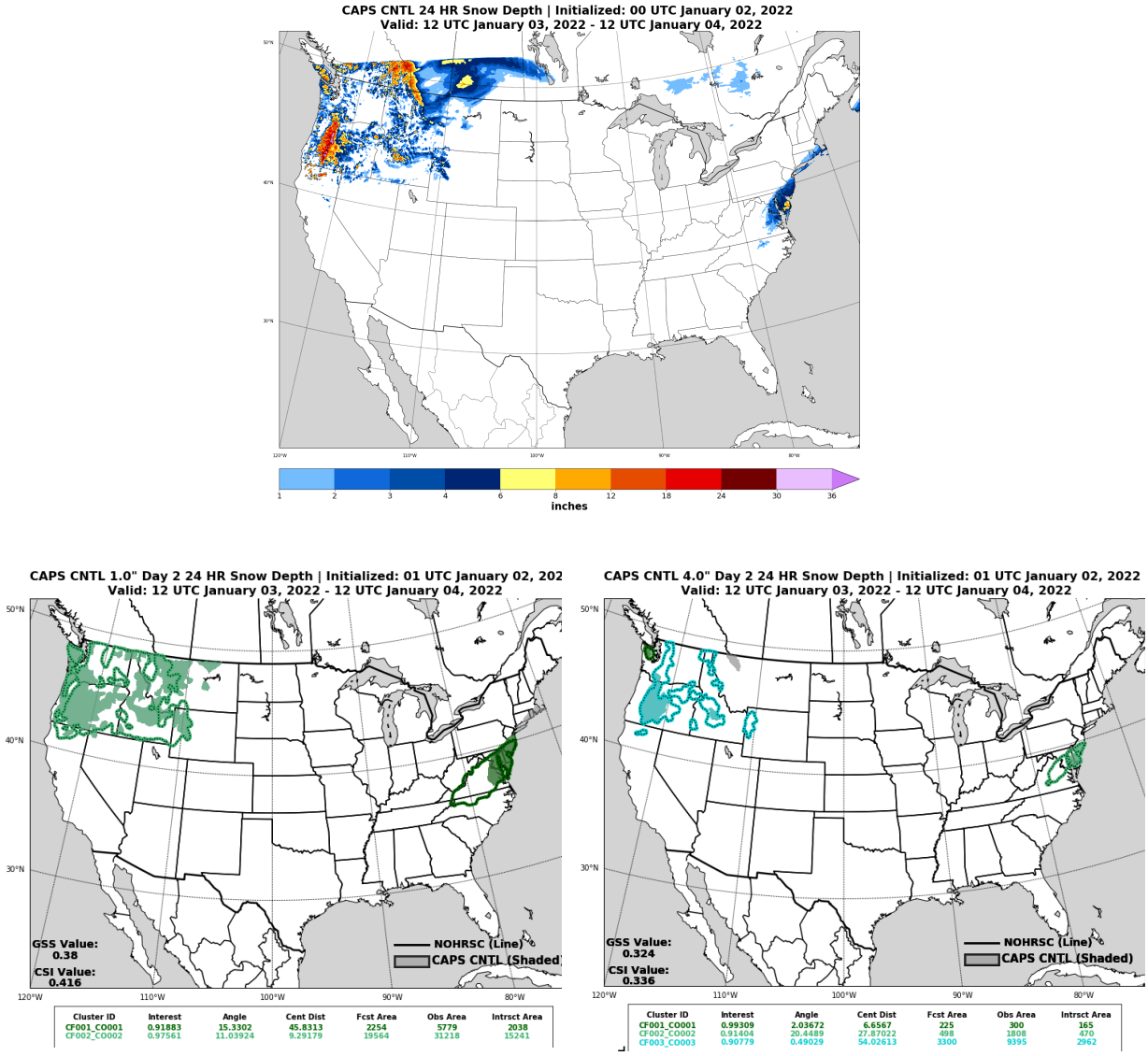


Figure 25: Case 11 24 hour snowfall accumulation at the day 2 lead time for the CAPS CNTL (upper) with the MODE verification information at the 1 inch (lower left) and 4 inch (lower right) thresholds. For MODE maps, filled objects indicate CAPS CNTL shapes while contours indicate NOHRSCv2 shapes.

The EMC FV3-LAM (figure 26) has slightly larger amounts but is also notably too far east and not far enough south with the 1 inch footprint. The 4 inch threshold extent is pretty well matched to

the NOHRSCv2 just displaced too far eastward. Note that the EMC FV3-Cloud presented in this case was the version before errors in the coding were identified, so it will not be shown here.

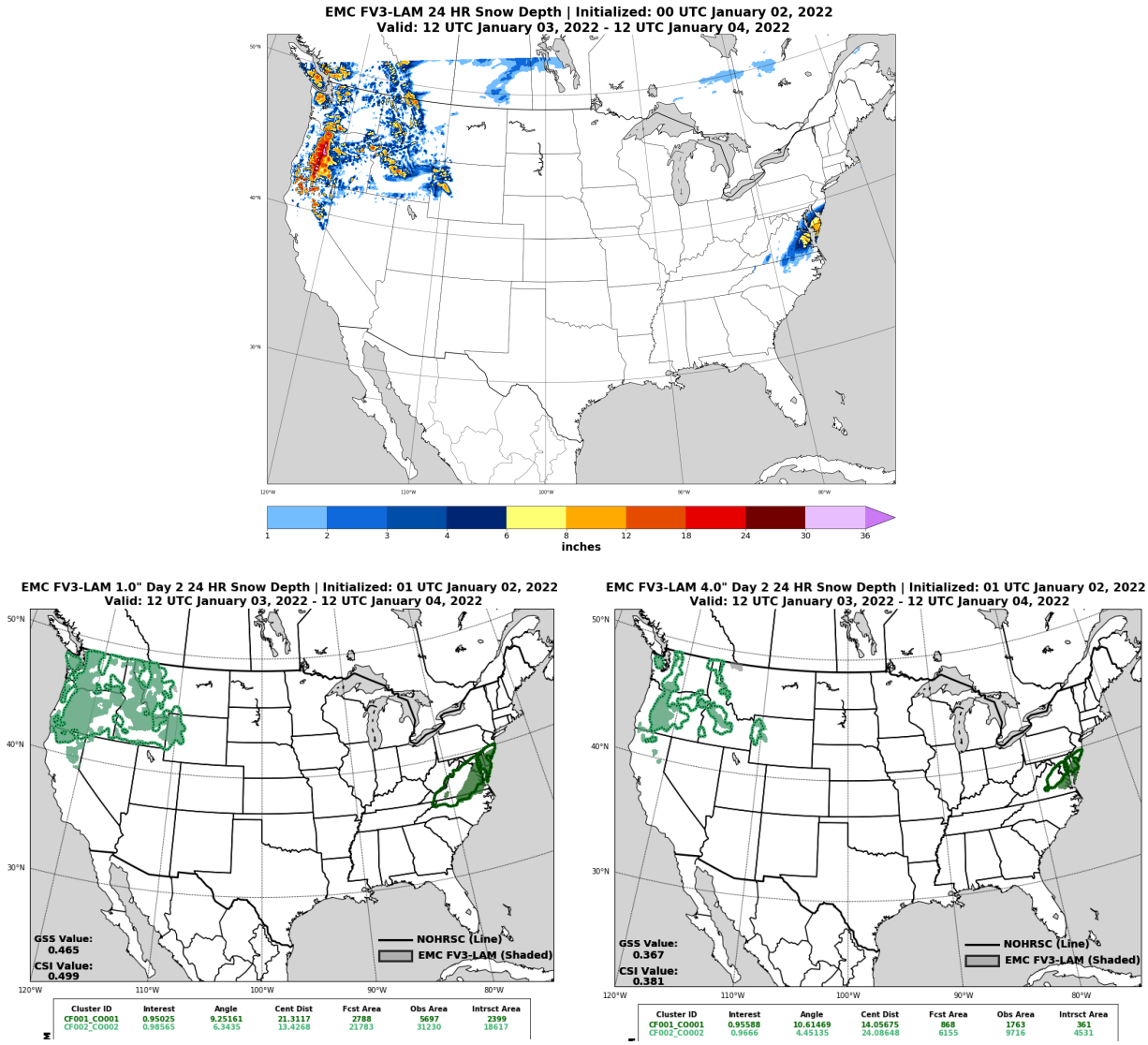


Figure 26: Case 11 24 hour snowfall accumulation at the day 2 lead time for the EMC FV3-LAM (upper) with the MODE verification information at the 1 inch (lower left) and 4 inch (lower right) thresholds. For MODE maps, filled objects indicate EMC FV3-LAM shapes while contours indicate NOHRSCv2 shapes.

This case was an example showing the SLR and precipitation type issues present in the downscaled GFSv16 over the eastern CONUS. Figure 27 shows that while the geographic extent of the 24 hour accumulation is similar to NOHRSCv2, it does not have a continuous 1 inch footprint over the area of focus in the Mid-Atlantic. Meanwhile, over the western CONUS, the footprint is well captured. Participants quickly disregarded this model due to its outlier status amongst the experimental models.

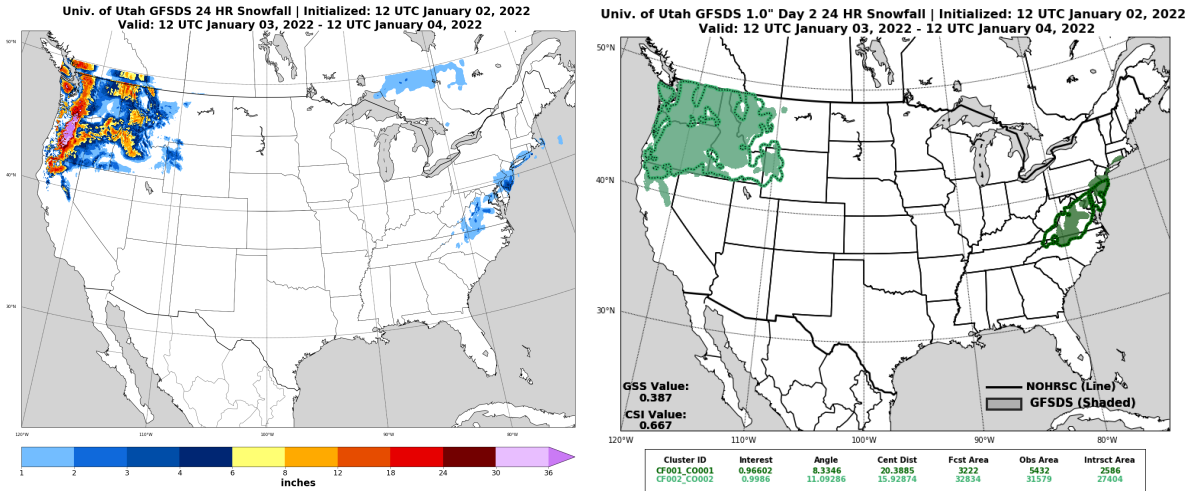


Figure 27: Case 11 24 hour snowfall accumulation at the day 2 lead time for the Downscaled GFSv16 (left) with the MODE verification information at the 1 inch (right). For MODE map, filled objects indicate Downscaled GFSv16 shapes while contours indicate NOHRSCv2 shapes.

Participants leaned heavily on both NBM versions for the MSTP activity, and as seen with the other experiment models, the footprint was too far east (figure 28).

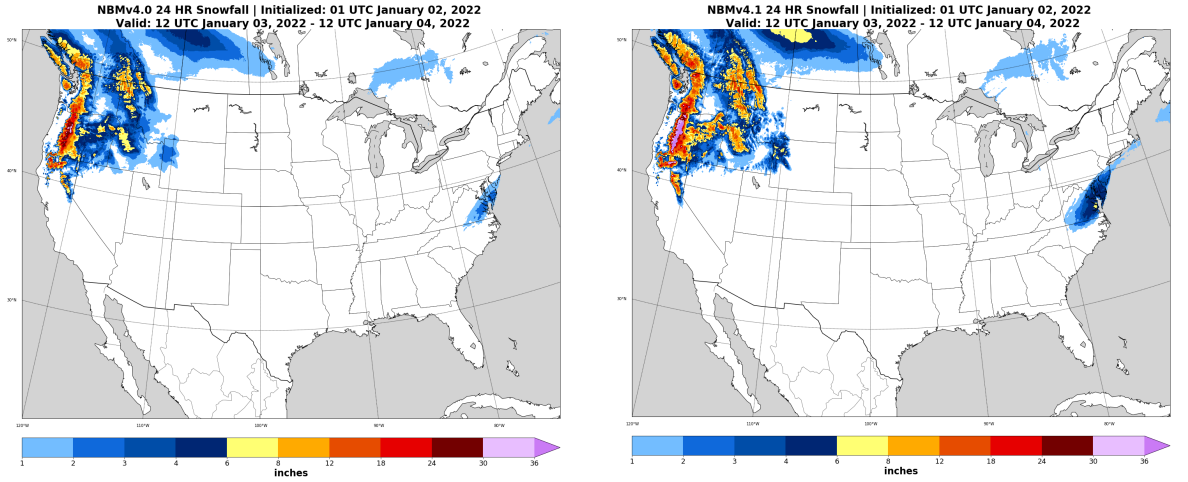


Figure 28: Case 11 24 hour snowfall accumulation at the day 2 lead time for the NBMv4.0(left) and NBMv4.1(right).

By day 1 (figure 29), both NBM versions and the EMC FV3-LAM come into line with both the footprint and amounts. The NBMv4.1 has a higher accumulation area than the NBMv4.0 which was seen in the seasonal performance diagrams as well as discussions throughout the WWE intensive sessions. The downscaled GFSv16 remains relatively consistent from day 2 to day 1 with a lack of snowfall accumulation over the area of interest.

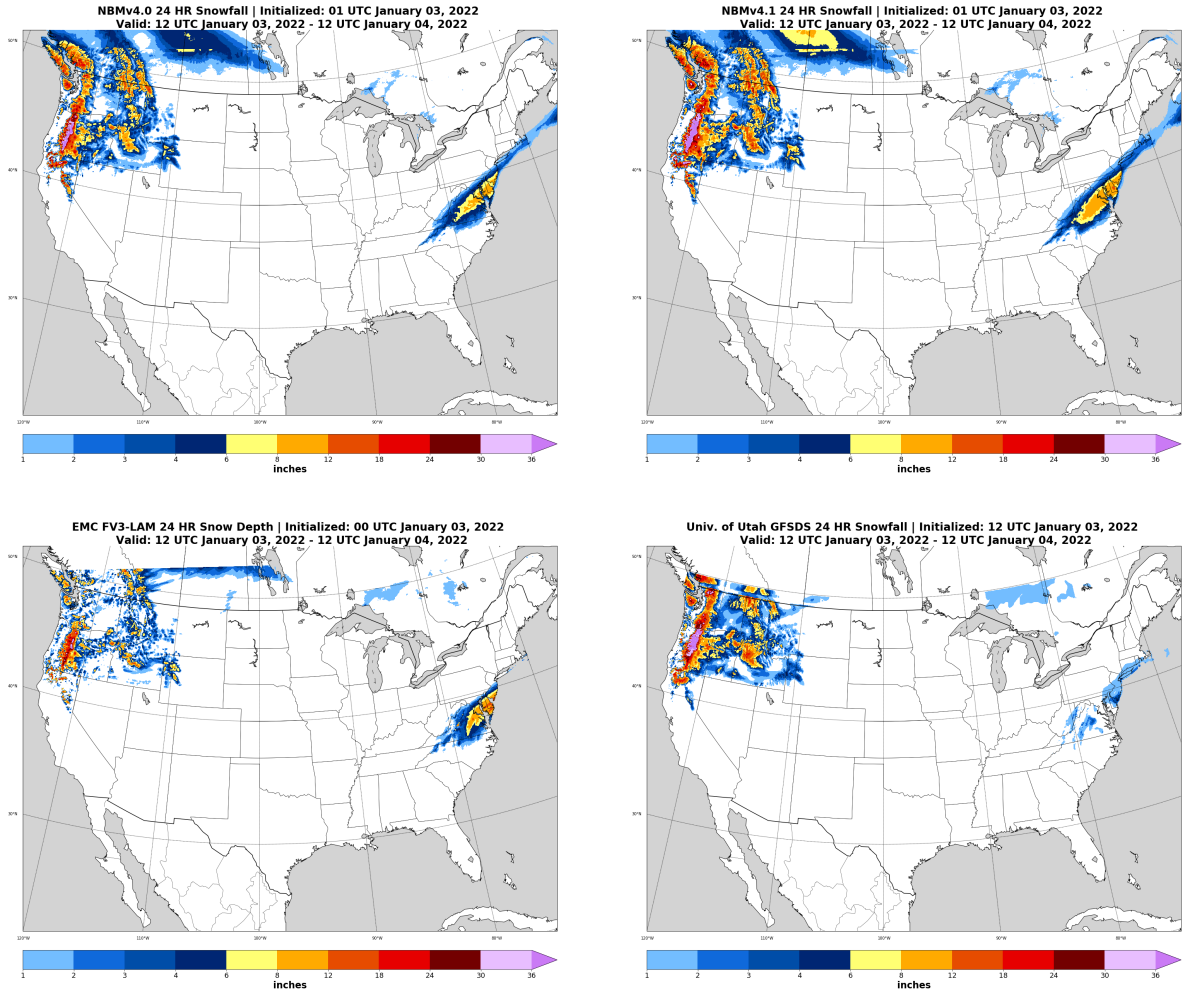


Figure 29: Case 11 24 hour snowfall accumulation at the day 1 lead time for the NBMv4.0 (upper left), NBMv4.1(upper right), EMC FV3-LAM (lower right), and Downscaled GFSv16 (lower right). CAPS SSEF data was not available for day 1.

The performance diagrams for this case represent a recurring issue within the WWE intensive weeks. Due to the full CONUS calculation of these performance diagrams, they do not accurately represent the performance of the experiment models over the focused event. For example (figure 30) shows the downscaled GFSv16 as one of the best performing models for this case, but from the maps above in figures 24 and 26 show a lack of snowfall accumulation in the eastern CONUS. The other experiment models' performance diagrams show similar behaviors to the seasonal statistics.

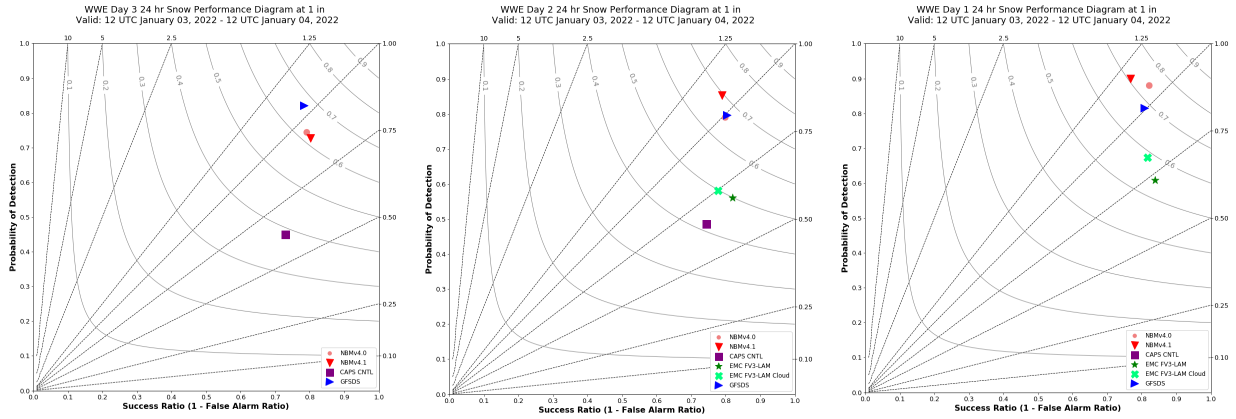


Figure 30: Performance diagram for case 11 at the 1 inch threshold day 3 lead time at the 1 inch (upper left), day 2 (upper right), and day 1 lead times (lower).

As outlined earlier, participants were broken up into 2 breakout groups with the goal of collaborating a forecast at the day 3, 2, and 1 lead times. The results of these collaborations are shown in figure 31. The top row is the day 3 forecast and shows both groups drawing their footprints larger than the NOHRSCv2 analysis. For the regional group (left column) participants were heavily influenced by the NBMv4.1 for the footprint at day 2 and thus the footprint was moved farther north to reflect what was seen in the experimental models, then on day 1 it was extended back to the south following the updated guidance. For the maximum contour, the regional group followed the CAPS SSEF control member. Even though it did not verify well, the group was convinced the CAPS control member looked reasonable due to its positioning farther to the east. The EMC FV3-LAM was farther west, which turned out to be correct, and was not in line with what the other models were outputting. This accounts for the position of 4" maximum contour at day 3 and day 2, and 8" at day 1. For the WFO group, uncertainty at Day 3 was the biggest driver for producing expansive areas for the 1" and 3" amounts, especially to the west in the Appalachian mountains. On day 2, the WFO group participants rapidly increased maximum snow amounts and trimmed the corridor for the heaviest snow. This continued into the Day 1 period, where the location of the heaviest snow was narrowed and refined to be at and just southeast of the DC metro. Numerous CAM solutions were used to justify both the location and orientation of the heaviest snow.

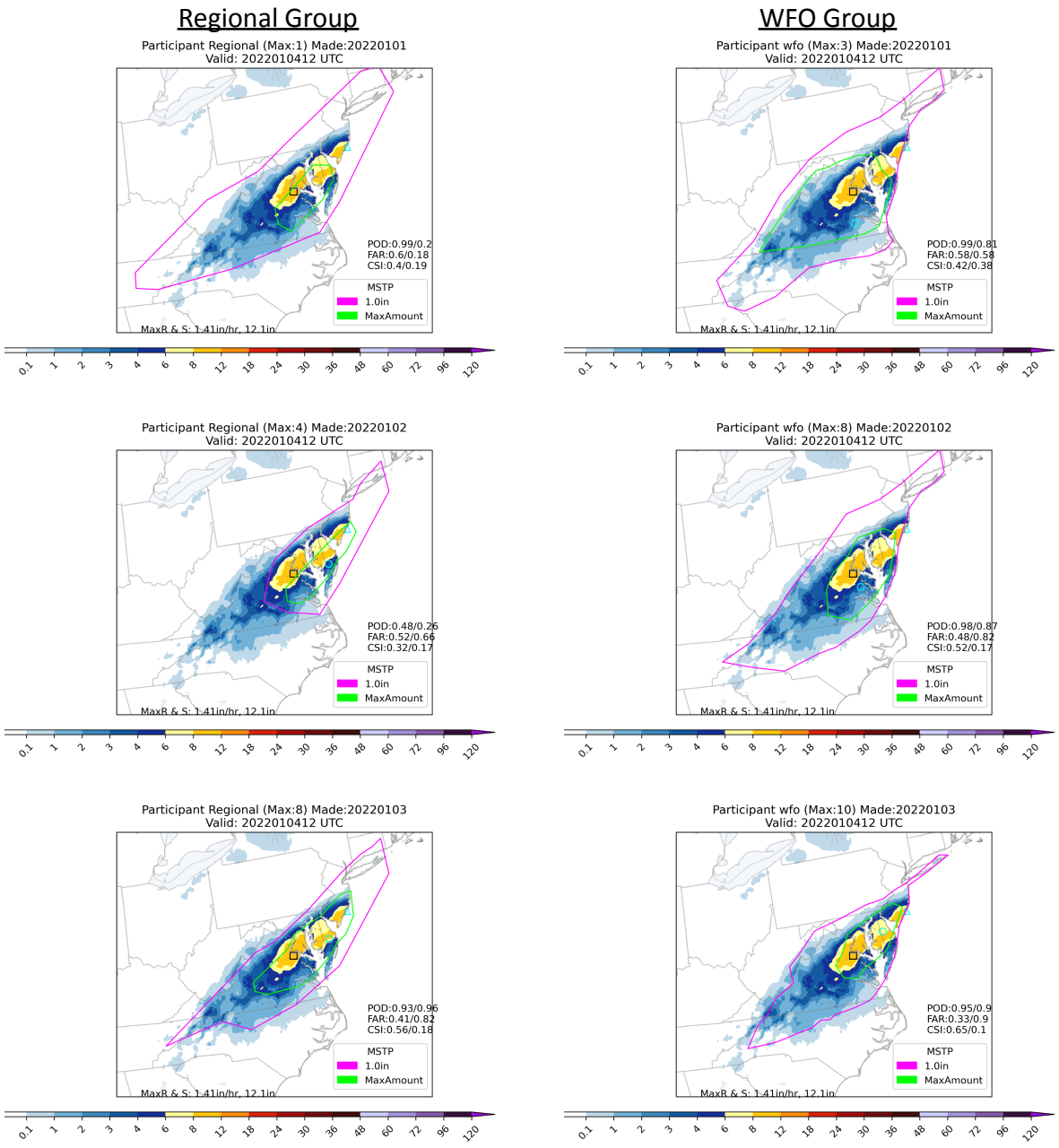


Figure 31: MSTP collaborated drawings for case 11, the left column is from the Regional breakout. Right column is the WFO breakout. Top column is at the day 3 lead time, middle column is at the day 2 lead time, and the bottom column is the day 1 lead time. Pink outline is the collaborated 1 inch footprint, green outline is the maximum contour.

Case 22: 12z 17 February 2022 - 12z 18 February 2022

From the second intensive week, case 22 (table 1) was one of several mixed precipitation cases sampled by the WWE. This case proved challenging not only for the models in terms of timing, amounts, and precipitation type, but also for the facilitators who were figuring out how to

incorporate the precipitation type information into the MSTP activity. Figure 32 shows the NOHRSCv2 24 hour snowfall analysis for this event. The forecast activity focused on the precipitation from Missouri through Michigan.

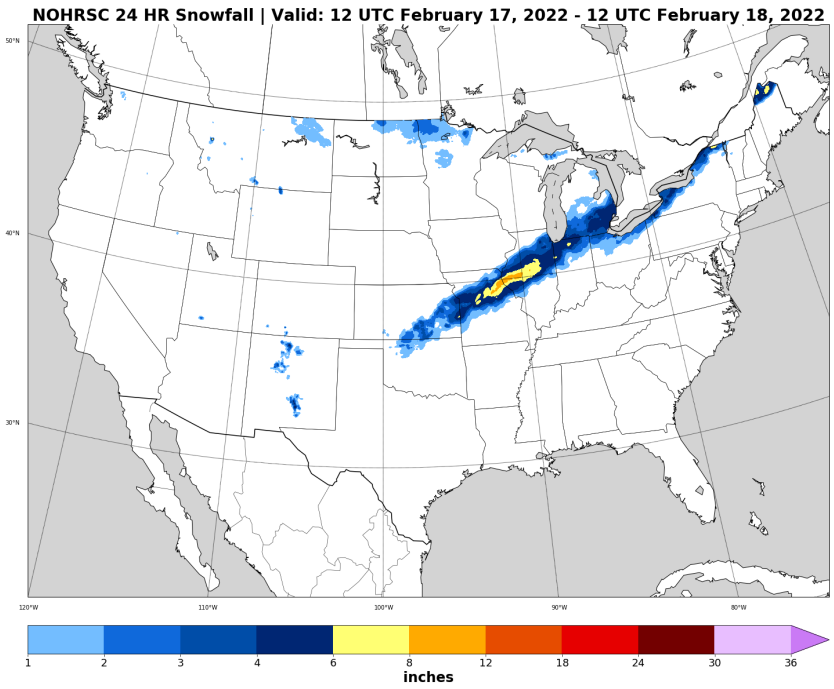


Figure 32: NOHRSCv2 24 hour snowfall analysis for case 22

The biggest issue with this case was the position of the precipitation axis. Many of the experiment datasets were shifted to the northwest at the longer lead times. This is reflected in the increasing probability of detection on the performance diagrams with decreasing lead time (figure 33). It should be noted that while the performance diagrams are calculated over the full CONUS, there was little to no other snowfall outside of the case region. For the 1 inch footprint (figure 33), both NBM versions continue to have the highest bias and probability of detection of all the experiment models at all lead times, with NBMv4.1[6] higher than NBMv4.0. This is due to the wide swath of snow found in the NBM at day 3 that narrows until day 1 (figure 34). The CAPS SSEF control member also has a slight high bias throughout with a peak in CSI value at day 2, in line with the NBMv4.0 and the EMC FV3-LAM at day 1. On day 2, the EMC FV3-LAM has the smallest bias and one of the best positions on the performance diagram without having the shifted precipitation axis seen in other experiment models. Meanwhile, the FV3-Cloud has the lowest CSI value and worst position at day 2 and then dramatically improves at day 1 to the best position and smallest bias. This is due to the main axis of snowfall shifted to the northwest (figure 35) at day 2 that is corrected for the day 1 lead time. While the downscaled GFSv16 has a

[6] NBMv4.1 was unavailable at day 1 for this case.

minimal bias at day 3 and day 2, it does have a low bias on day 1. The day 3 to day 2 increase in probability of detection is explained by the shift of the precipitation axis. Also with the exception of the FV3-Cloud at day 2, it also has the lowest CSI values and worst position on the performance diagram. This again highlights probable SLR and precipitation type issues within the algorithm over the eastern CONUS.

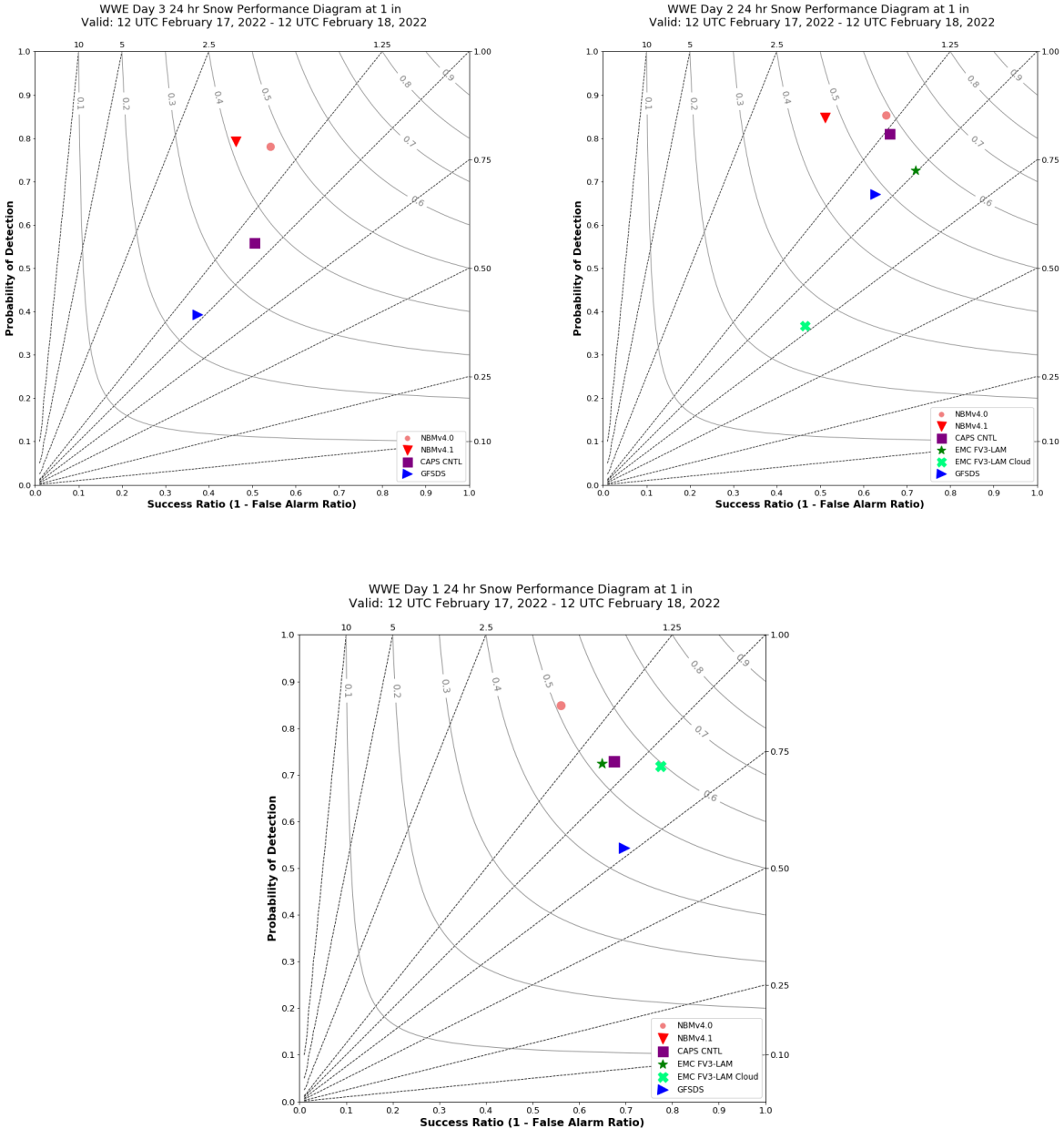


Figure 33: Performance diagram for case 11 at the 1 inch threshold day 3 lead time at the 1 inch (upper left), day 2 (upper right), and day 1 lead times (lower).

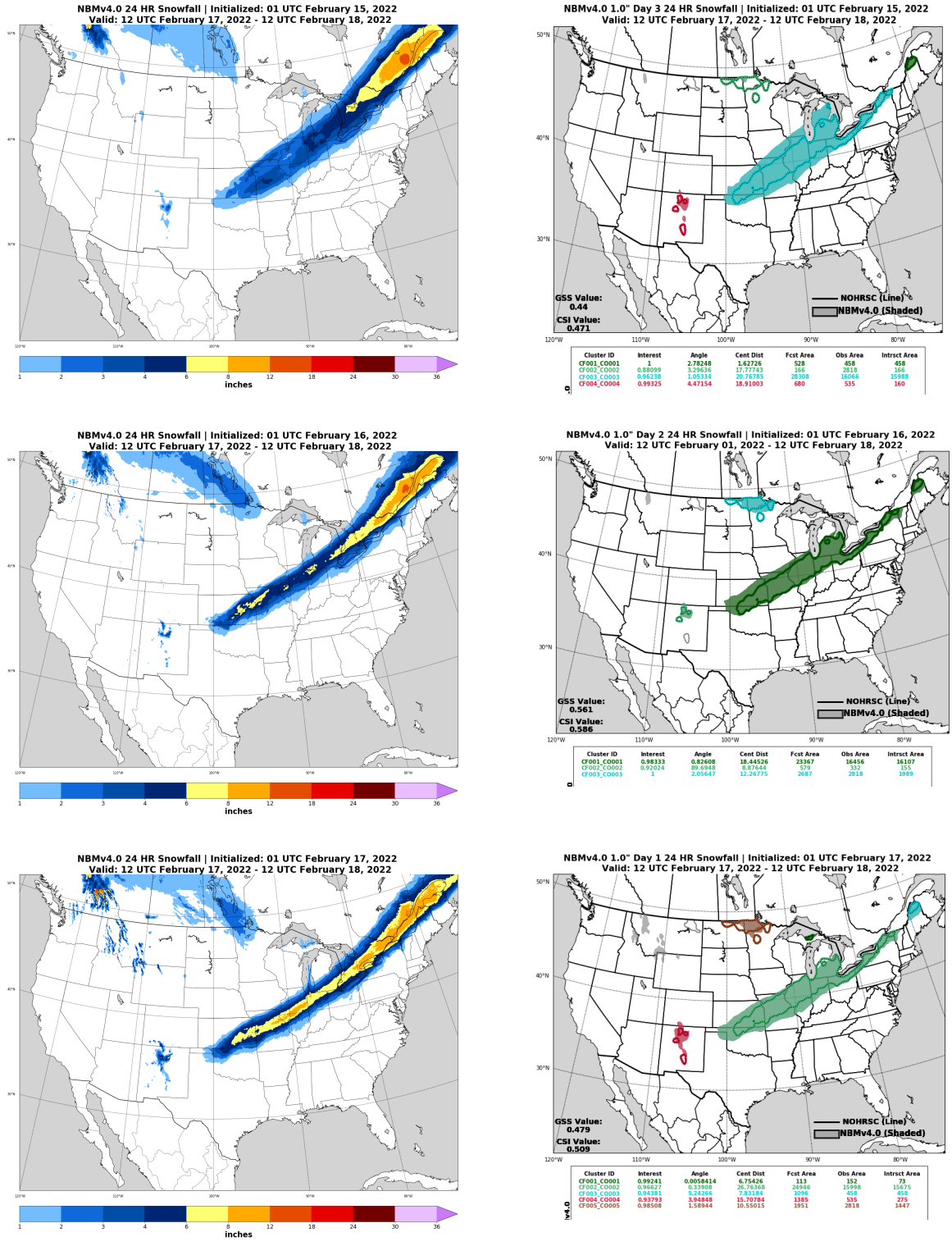


Figure 34: NBMv4.0 24 hour snowfall accumulation (left) and MODE analysis at the 1 inch threshold (right) for case 22 at the day 3 (top), day 2 (middle), and day 1 (bottom) lead times. For MODE maps, filled objects indicate NBMv4.0 shapes while contours indicate NOHRSCv2 shapes.

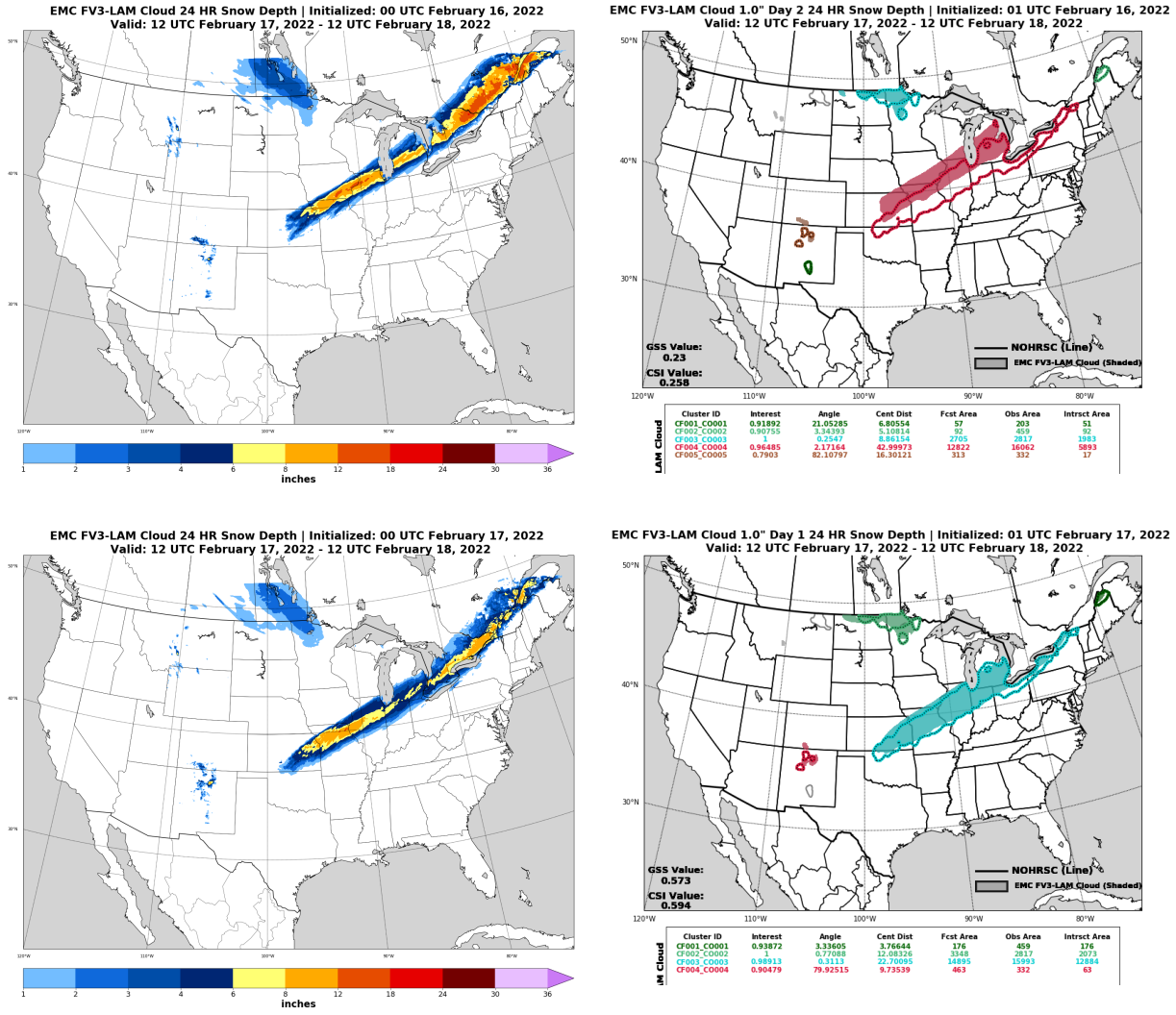


Figure 35: EMC FV3-Cloud 24 hour change in snow depth (left) and MODE analysis at the 1 inch threshold (right) for case 22 at the day 2 (top) and day 1 (bottom) lead times.

In the breakout forecast activity, the two groups were again asked to collaborate on a forecast. However, this forecast would also include a contour for sleet/ice pellets and freezing rain in addition to the snowfall footprint and maximum snowfall contour. The results of these activities are shown for day 3 (figure 36), day 2 (figure 37), and day 1 (figure 38). As stated earlier, most of the experiment models had the precipitation shifted to the northwest, along the track of the decaying upper level low or comma head, therefore the MSTP drawings were also shifted relative to the observations at day 3 and 2 for the regional groups and all 3 days for the WFO group. On day 3, it was also difficult to distinguish different areas between the precipitation type areas, hence the large overlapping areas of snow, ice, and freezing rain. Additionally, while the southeastern edge of the precipitation shield would start as rain, in the transition zones it

would switch to a frozen variety as the cold front drifted southeastward by the end of the period.

Day 3

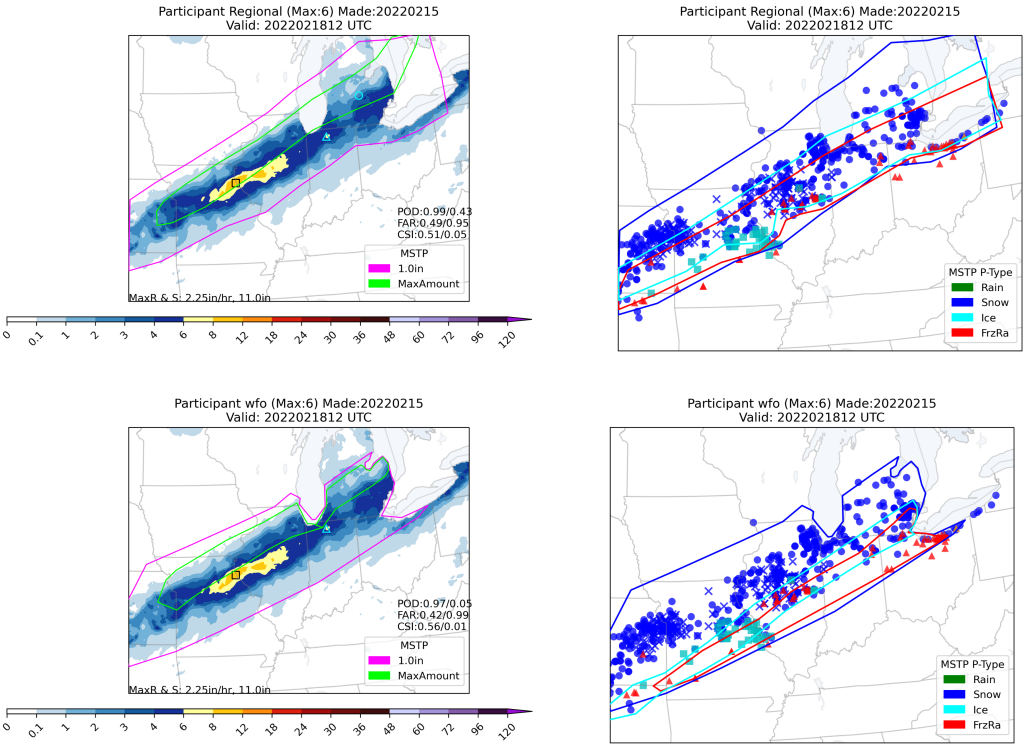


Figure 36: MSTP collaborated drawings for case 22 at the day 3 lead time, the top row is from the Regional breakout. Bottom row is the WFO breakout. Left column is the snowfall forecast: pink outline is the collaborated 1 inch footprint, green outline is the maximum contour. Right column is the precipitation type outlined with the local storm reports.

On day 2 (figure 37), when the guidance begins to shift southeast, the regional group also shifted both snowfall contours. The precipitation type collaboration also becomes more distinct in an area of freezing rain transitioning to an area of sleet and finally into snow. The WFO group retained their original footprint and shifted the maximum contour to indicate the shift in the models. The same can be observed for the precipitation type shapes as well. The southern extent remains consistent while the northern extent of both the ice and freezing rain shifted southeast as did most of the CAMs in the track of the surface low and upper level decaying trough. Most CAMs had consistent high rates of snowfall for a period of a few hours in a narrow corridor, however the location of that corridor was not consistent. The FV3-Cloud, with an older GFS forecast as the initial condition, continued to stay farther northwest in location. This event was subject to small displacements in synoptic features that led to consistent forecast displacements in the heaviest snowfall regions.

Day 2

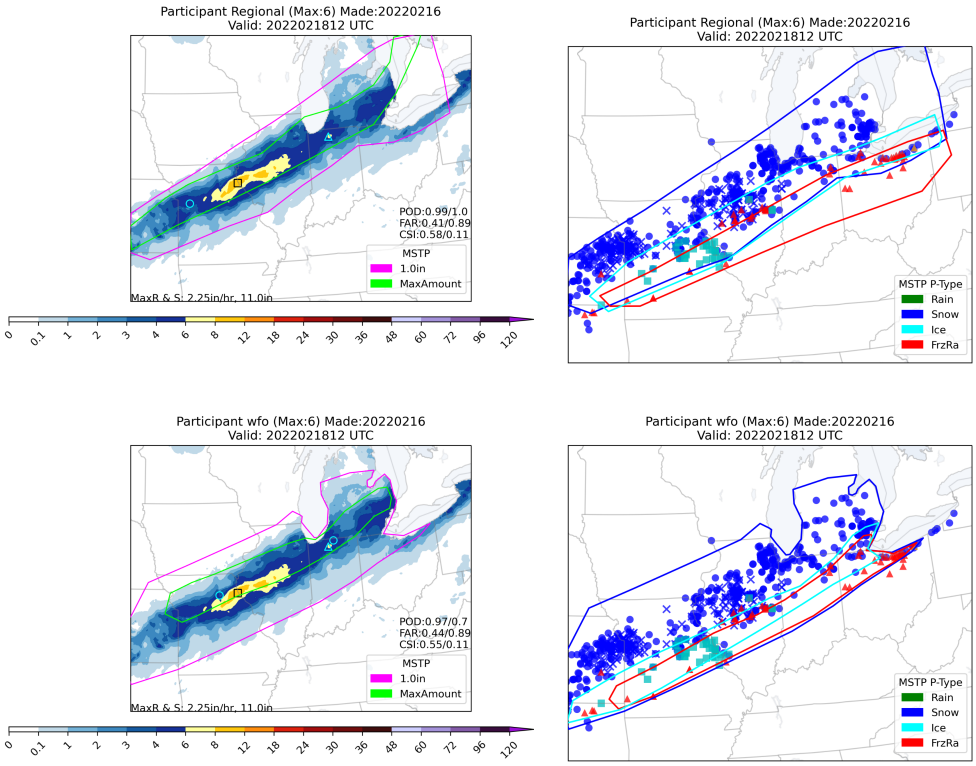


Figure 37: MSTP collaborated drawings for case 22 at the day 2 lead time, the top row is from the Regional breakout. Bottom row is the WFO breakout. Left column is the snowfall forecast: pink outline is the collaborated 1 inch footprint, green outline is the maximum contour. Right column is the precipitation type outlined with the local storm reports.

The day 1 discussion was much of the same with the rest of the experimental guidance converging on accumulating snowfall from Missouri to Michigan. As such, both groups further refined their northern boundary and shifted the maximum contour. With the surface low and decaying upper trough track much better on Day 1, so were the forecasts. Forecasters in the WFO group highlighted the HREF as having 1” per hour snow rates for a period of 4-6 hours where the heaviest snow eventually fell, but not much farther northeast into Chicago as the collaborated forecasts indicated.

In the regional group, the precipitation type discussion focused on the southern edge of the QPF. Some of the data indicated the possibility of freezing rain into the southern Ohio River Valley and as such, that boundary was expanded into Southern Ohio and Indiana. The WFO group did not change their southern extent but instead refined their freezing rain and sleet shapes into southeastern Michigan. The group as a whole discussed the tricky evolution of precipitation types for this event, since in some locations in Michigan sleet was observed but

was reported as snow, per snowfall measurement practices. So it was difficult to trust the LSRs given some of the first hand experiences mentioned in the verification. It is difficult to tell how well the precipitation type forecasts from both participants and models performed. A rigorous accounting of precipitation type evolution and accumulation would be required to make any quantitative statements about forecast quality. As such there are few, if any methods, to display the evolution of precipitation type over time across wide swaths of the transition zone shown here. Methods to evaluate models are sorely needed, however, observations are also sorely lacking.

Day 1

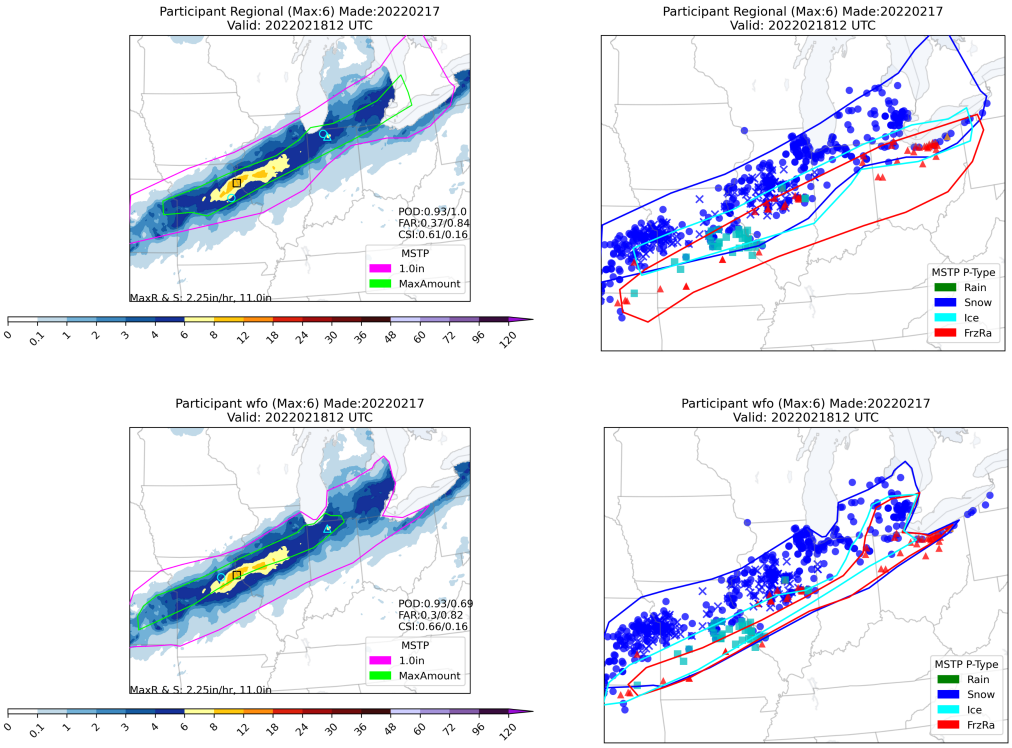


Figure 38: MSTP collaborated drawings for case 22 at the day 1 lead time, the top row is from the Regional breakout. Bottom row is the WFO breakout. Left column is the snowfall forecast: pink outline is the collaborated 1 inch footprint, green outline is the maximum contour. Right column is the precipitation type outlined with the local storm reports.

3.4 Discussion topics

At the end of each week, HMT facilitated discussions regarding the state of forecasting, operations, and the future of the WWE activities. Generally the HMT is a pseudo-operational environment in that we cannot replicate a full AWIPS2 environment. During the pandemic, HMT has an increased web presence to facilitate forecasting activities using experimental models, techniques, and tools that are new to most participants. As such, forecasters in both weeks

mostly discussed their AWIPS2 procedures for winter weather forecasting using the NBM and WPC as primary sources of information. Discussion topics are summarized into eight themes below.

A. Procedures and Data Inputs

NWS forecasters have different procedures, across the regions, to accomplish their tasks in creating forecast grids, generating information, and perspective for IDSS. While each procedure has a different workflow, each procedure has certain data inputs regardless of source: QPF, snow to liquid ratio (SLR), probability of weather type (POWT).

SLR was by far the most discussed, and contained a large amount of uncertainty both in forecasting and verification. From an observational perspective, snow is already difficult to measure and SLR measurements are even more problematic. SLR varies on timescales of minutes. From a NWP perspective, SLR is not a direct model output but is instead inferred from both precipitation type and QPF, water equivalent amount of snow depth, and snow depth (determined from the land surface model). Most procedures have set time periods or averaging intervals (typically 6 hour bins for NDFD), however when faced with putting together IDSS briefings, snow rates computed below the averaging interval can only be based on QPF, and depending on the SLR, add some error and uncertainty into an already uncertain field. Forecasters have SLR climatology (local or CONUS) that they use to infer if the forecast SLR looks acceptable. They discussed potentially relying more heavily upon WPC SLR grids during the Winter Weather collaboration process so they could focus on the CFP principle of Targets of Opportunity (TO). This is the principle that assumes the first guess fields are in the ballpark of what is expected over most of the forecast area and only those areas with perceived flaws should be further examined and subsequently modified.

> *"I think focusing on those targets of opportunity is definitely the future, but to go along with some of the prior comments we could all benefit from a robust NBM training to understand the methodology behind its QPF and snowfall to better diagnose a target of opportunity. I think these go hand in hand."*

B. Training

Many forecasters expressed interest in more direct training about the NBM, its elements, methods, and strengths and weaknesses. Most importantly, forecasters expressed some technology fatigue. With the yearly cadence of model updates and changes, it was noted as difficult to keep up with changes in general. These comments were directed at the entire model suite since multiple models can have changes in any given year. When coupled with an increasing variety of experimental models in development or parallel, the information on model changes is difficult to keep in mind. As such, forecasters had suggested reference material cheat sheets for core model information or changes.

C. Role of the forecaster

As TO is becoming the norm only in the last 2 years, the role of the forecaster was a frequent topic of discussion. With the addition of more and frequent collaboration (WFO <-> WFO, WFO <-> WPC) and IDSS, there is more time pressure to produce quality products. Developing key messages requires more information and knowledge about the forecast, yet there is less time for both IDSS and forecasting. Creating a coherent story going beyond existing forecast graphics was seen as more important than the grids (minus targets of opportunity).

Forecasters were split in their general role of either being a forecast producer (grid production and editing) or forecast interpreter (NBM plus targets of opportunity as directed in CFP training). Many still felt that production was still the key ingredient in being an interpreter of the forecast in the content generation for IDSS. There are still events, especially a few of the events demonstrated in this experiment, where guidance could not be taken literally all the way through the Day 1 period.

D. Consistency

Regarding NWP and predictability, consistency across forecast days was mentioned frequently. Individual models tend to wax and wane, and so did the NBM in terms of location, snow amounts, precipitation types, etc. Forecasters also noted inconsistency from day 2 to day 1 in the NBM as more CAMs contributed to an increase in snow amount. While this was not unexpected, forecasters had expected less of a jump.

E. Messaging

Issues of consistency sparked a vigorous discussion about the temporality of winter weather messaging. As more long range guidance becomes publicly available and transmitted through various social media channels, there is more pressure to message early for “snow storms”, in a rumor control posture. Forecasters expressed their discomfort of this additional responsibility as some of the time ranges were beyond predictable. However there were also times when model inconsistency developed closer to day 2 and forecasters found themselves walking back their previous certainty of no snow, only to find their area was once again at risk for snow. Thus, consistency was even more prominent in their increased IDSS role.

F. On being more “probabilistic”

Forecasters were mildly frustrated with drawing deterministic snowfall, wanting to add a probabilistic layer to the forecast and forecast process. This was largely discussed as ways of using the probability fields to adjust the deterministic forecast, or ways in which one could visualize the range of possible outcomes. The ModelCertainty tool, which is designed to quickly

provide statistical information from datasets provided in AWIPS Graphical Forecast Editor (GFE), fell into the latter category as a way of seeing if the forecast fell within the inter decile range (10-90th%) as a sanity check. In the former, forecasters were invoking TO as they adjusted areas of snowfall up or down based on the respective snowfall probabilities of exceedance. In either discussion there was no real consensus on what it meant to be "more probabilistic". The one aspect that was not met with frustration was using probabilities to convey the key messages. The forecasters felt more comfortable using probabilities for messaging. It was mentioned that grids and graphics can only do so much to convey the story, but that probabilities of amounts, rates, and timing were seen as being key elements in IDSS briefing packages and communications.

G. Verification

For snowfall, we had many discussions involving the NOHRSC analysis and its credibility with respect to measured snow observations. Forecasters in both weeks noted some discrepancies between the analysis and the LSRs. They had specific software discussions about how to generate Standard Hydrometeorological Exchange Format (SHEF) reports and the tricks of making sure LSRs were encoded properly so that it could be incorporated into the NOHRSC analysis system:

> *"Also worth noting, NOHRSC is only able to assimilate LSRs if offices remember to SHEF encode them in IRIS." ; "...there's a checkbox on the last step that says something like "SHEF encode as RRM message." IMO it should be the default but it's not. "*

For winter precipitation type, there exists relatively little direct observations that are easily available for model verification. The sources of precipitation type range from citizen science reports (MPING, LSRs) to direct ASOS observations. In comments from forecasters, snow and sleet are lumped into one snow category. It was discussed that maybe the NWS offices could disaggregate the "snow" and list each precipitation type and accumulation separately in the metadata, if possible, to continue abiding by NWS snowfall measurement standards[7].

H. Adapting HMT for the future

New capabilities for the Precipitation Type forecast activities will be required, including the viewing of model soundings. Forecasters were, understandably, frustrated by lacking this key functionality with respect to interrogating model data.

New derived data were also requested to help with winter weather forecasting, especially time of arrival, departure and duration, peak snowfall rates, and SLR. Forecasters liked the

[7] Snow Measurement Guidelines for National Weather Service Surface Observing Program, NWS, 2013.

precipitation type activity and felt that improving it could be beneficial as precipitation type and its uncertainty play a large role in key message development and forecasting.

Improvements to the MODE evaluations were also discussed as forecasters wanted the performance diagrams to be focused over the forecast area, not just full CONUS, since it was difficult to use this objective information when there was snowfall throughout the CONUS.

New ideas to improve the HMT activities included:

- Asking for partners to be included so that key messages could become a possible activity within the Testbed,
- Adding a capability to view model soundings specifically for precipitation type forecasting,
- Incorporating more NBM graphical displays typical of operations, and
- Developing and testing strategies to incorporate probabilistic thinking into the forecast.

4. Summary and Recommendations

The 12th annual WWE was conducted over the 2021-2022 winter season through two intensive evaluation weeks and the use of retrospective cases. This year participants were asked to not only evaluate the experiment data but also to use the information to collaborate on a forecast. Experiment data was again 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 along with exploring precipitation event timing and how to explore precipitation type output. There was also some focus on evaluating potential redesigns for the WSSI and expanding the downscaled SLR techniques over the full CONUS. 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.

- **FV3-LAM** was the control CAM configuration from which the others were compared at day 2. The seasonal performance diagrams place this model in the middle of the pack with relatively low bias values, but also lower CSI scores especially toward the larger snowfall amounts. Due to the continued refinement of the FV3-LAM for RRFS development, the team **recommends further development and testing**.
- **FV3-Cloud** was a brand new configuration with the first runs completed only a few days before the WWE intensive sessions. As sometimes happens with new science, there were errors found and reruns had to occur after the first intensive week. Participants

were very excited to look at this data as they felt directly involved in the RRFS development. This configuration had some issues in event timing, which was especially evident in case 22. Seasonally, it overall has a slightly higher bias than the FV3-LAM but also higher CSI values. Similar to the FV3-LAM, the team recommends **further development and testing** as NOAA works toward the RRFS configurations.

- **SSEF Ensemble** was run as a 13 member ensemble from OU-CAPS. The control member was evaluated among the other deterministic CAMs. This member did not perform well when compared to the other experimental models in terms of CSI values and performance diagram position. However, it also had one of the lowest model biases as was evident when participants favored using it in the MSTP activity due to the more realistic snowfall amounts. Seasonally the ensemble mean, LPMM, and PMM were clustered together with the best positioning on the performance diagrams especially at the lower snowfall amounts. As with the CAMs from EMC, the team **recommends further development and testing** in coordination with EMC working toward the development of RRFS.
- **Downscaling methodology** provided by University of Utah was expanded over the full CONUS for this year's WWE. As expected, the algorithm performed well over the Western CONUS, but there were apparent issues east of the Rocky Mountains. The snowfall amounts were too low especially in cases where mixed precipitation types were involved. This was an expected result, since the algorithm does not currently include any training over the Eastern CONUS. As such, the team **recommends further development and testing** by including Eastern CONUS information for the SLR training and more direct handling of precipitation types.
- **NBMv4.1** was provided by MDL in parallel to the NBMv4.0. Participants were excited to see what this version looked like for winter. Many ended up surprised at the extremely high bias found for most of the cases and seen on the seasonal performance diagrams. Similar to the NBMv4.0, the 4.1 relied heavily on the probability of detection for position on the performance diagram. Discussion was focused on SLR calculations within NBMv4.1 with concern from participants and confusion for how the calculations are performed. Due to this, the team **recommends further development and testing** especially to reduce the high bias found at all lead times and amounts. NOTE: Development is underway between MDL and WPC to address the high bias
- **WSSI redesign** project evaluation in the WWE was the first opportunity for NWS forecasters to look at new design elements including a new color scale and impact definitions. Participants heavily favored the proposed colors and definitions, as such, the team recommends these updates to WSSI be **transitioned to operations**.

Table 5: Research to operations transition recommendations for the 12th Annual WWE.

Evaluated Dataset	Recommended for transition to operations	Recommended for further development and testing	Rejected for further testing	Provider/Funding Source
FV3-LAM FV3-Cloud		X X		EMC
SSEF Ensemble		X		OU-CAPS
Downscaled GFSv16		X		University of Utah
NBMv4.1		X		MDL
WSSI Redesign	X			WPC/Nurture Nature Center
ProbSR/Travel WSSI		X		WPC/NSSL

Another aspect of the experiment was the continued effort to find the best format for the WWE. This year focused solely on retrospective cases and evaluated selected events over the two intensive weeks. During these intensive weeks, we tested the use of breakout groups to encourage participants to collaborate on the forecast activity. Comments from participants were that they enjoyed this approach to the experiment with suggestions for future attempts. One main suggestion was to be more precise in assigning people to the breakout rooms. This year, it was left to the Google meet randomization which caused people to repeat the same group the entire time. Another suggestion was to make sure the differing approaches in the breakout groups are made more clear. The original idea was to have the WFO group focus on smaller areas with the facilitator stitching together their forecast from those while the regional group focused on the entire area, similar to how a national center operates.

Other comments for future WWEs include being more focused on an evaluation goal each day. For example, one day to examine only probabilistic information where participants craft a “most-likely” or “90th percentile” forecast, and another day for precipitation type information. Since the precipitation type was a new venture for the MSTP, participants were happy to provide comments on potential improvements. The main suggestion is for the inclusion of

vertical profile information. It is difficult to evaluate how well the models depict freezing rain or ice pellets without access to the vertical profiles. Participants also requested the performance diagrams to be calculated over the forecast region, not just full CONUS to make them more relevant to the evaluations. Future WWEs hope to be able to provide this information.

One issue facilitators found was the focus of participants on the NBM as the basis for their forecasts due to familiarity. Future WWEs should directly assign models to individuals to evaluate or make breakout groups focused on inclusion and exclusion of the NBM.

5. Other HMT Winter Weather Experiment Activities

5.1 ProbSR Focus Group

In addition to the intensive weeks, this year's WWE was responsible for organizing several other winter related activities. The first relates to hosting a series of focus groups where NWS forecasters were invited to provide feedback on the development of a decision support tool that provides a 0 - 100% probability that roads are subfreezing. This tool was developed by OU/ Cooperative Institute for Severe and High-Impact Weather Research and Operations (CIWRO) as part of a funded Joint-Technology Transfer Initiative (JTTI) project and is called Probability of Subfreezing Roads (ProbSR). ProbSR was originally developed and tested in an experimental real time version of the Multi-Radar Multi-Sensor (MRMS) system supported by NOAA/NSSL. As a part of the initial evaluation of ProbSR, a data feed was made available to NWS WFOs and the tool was socialized through regional headquarters. Preliminary feedback was positive.

The developers for ProbSR have been working collaboratively with developers of the WSSI at WPC to integrate a prognostic version of ProbSR into a special travel component of the WSSI as a part of a 2020 JTTI project. An element of the JTTI was to evaluate ProbSR in the WPC's Winter Weather Experiment. Because ProbSR is a novel tool that most forecasters have not been previously exposed to, both CIWRO and WPC felt the optimal approach was to stage a series of focus groups wherein the product was introduced, a case study highlighting the utility of the tool was presented, and targeted questions presented to the participants. Three focus groups were conducted from 2021 December - 2022 February. Themes and recommendations from these focus groups are presented below. The complete report and findings can be found linked here.

ProSR Focus Group Themes and Recommendations

Some recurring themes emerged in all discussion groups and even in many of the questions.

1. Lack of clear guidance on what can and cannot be messaged externally. All of the forecasters are aware of NWS directives[8] limiting the dissemination of road weather forecasts, but were unclear what they could message. This made it difficult for them to fully embrace a product like ProbSR. Most forecasters who raised this concern said they would like to use it in their messaging, they just are not sure if they can. A key recommendation stemming from these discussions is for a clear agency-wide statement on what can and cannot be messaged.

2. While the forecasters expressed optimism for the potential benefits of ProbSR, they really need to "play" with it in their CWA before they can fully buy into it. They appreciate the emergent WDTD training, but want more. They specifically requested a 1-page summary for regular reference and WES simulations. Given that the majority of participants noted that a tool like ProbSR would only be used occasionally, additional training resources are recommended to ensure the successful adoption of ProbSR in operations.

3. Many participants admitted to using anecdotal experiences in diagnosing road weather threats (e.g. when there's FZRA, we always have trouble in this area) and additionally asked several questions along the lines of, "How does the algorithm perform when…" suggesting a need to understand specific scenarios. There was also the concern raised about ProbSR values being too low and a discomfort with mid-range probabilities in general. Many of these questions raised a higher concern in our minds – that being a general lack of education on road weather and how road temperatures vary or respond to various forcings. This is not well researched in the literature and is an area of inquiry so dominated by the private sector that information is difficult to come by. We recommend additional research identifying key relationships between road weather and sensible weather as well as how these factor into how ProbSR performs.

4. One loud and clear recommendation was the need for trends. Forecasters felt very strongly that trends are more illuminating than actual probabilities and wanted a graphical interface that allows them to view trends. We recommend additional capabilities within the WSSI that will allow forecasters to more easily assess trends.

5. Last, the need for variable thresholding based on time of day and geographic location in any downstream WSSI impacts index was clearly conveyed. It is unknown to what extent these impacts should be heightened. Additional research to support this capability is recommended.

5. 2 Seminars

Following previous years success, the WWE again hosted a weekly seminar series. Seminars were held weekly on Tuesdays except during the intensive weeks when an additional Thursday session was held. Topics ranged over a variety of winter weather issues with 50 - 100 attendees

[8] https://www.nws.noaa.gov/directives/sym/pd01024005curr.pdf

each week. Due to storage constraints, these seminars were not recorded; however, the slides have been archived on the experiment website linked here and in the table below.

Table 6: *List of seminar presenters and titles for the 12th Annual WWE. Presentations generally occurred weekly on Tuesdays throughout the winter season. During the WWE Intensive weeks, there was an additional seminar presentation on Thursday. Slides for the presentations are linked in the table below.*

Presenter	Presentation Title
Brian Filipiak (U. Albany)	**Data Fusion: A Machine Learning Tool for Forecasting Winter Mixed Precipitation Events**
Huan Meng (NESDIS)	**NESDIS Satellite and Radar-Satellite Merged Liquid Equivalent Snowfall Rate Products**
David Zaff (NWS Winter Program Office)	**The NWS Winter Program: Enabling Innovation to Achieve Consistent, Collaborated Products and Messaging**
Jeff Craven (MDL)	**NBMv4.1 Winter Overview: WWE 2022**
David Foster Hill (OSU)	**Community Snow Observations: From Concept to Operations**
Jim Steenburgh (UUtah)	**Snowfall Prediction over the Western CONUS**
Rachel Hogan Carr (Nurture Nature Center)	**Winter Storm Severity Index Improving Winter Storm Readiness through Severity and Social Impact Forecasting: A Social Science Study**
Andrew Rosenow (NSSL)	**Snow Rate Research at the National Severe Storms Laboratory**
Christiane Jablonowski/ David Wright (UMich)	**Improving Lake-Effect Snow Forecasting Capabilities via Advanced Coupling techniques in NOAAs' Unified Forecast System (UFS)**
Ben Blake (EMC)	**The Rapid Refresh Forecast System (RRFS): Current Status and the 2022 WWE**
Keith Brewster (OU CAPS)	**FV3-LAM CAM Ensemble Forecasts and Ensemble Consensus Products for the HMT Winter Weather Experiments**
Massey Bartolini (U. Albany)	**Evaluating Stochastic Parameter Perturbations in High-Resolution Rapid Refresh Ensemble (HRRRE) Forecasts of Mixed-Precipitation Events.**

Jason Cordeira (Plymouth State University)	**East Coast Atmospheric Rivers: Examples of their hydrometeorological impacts.**

6. Acknowledgements

The WWE team would like to sincerely thank WPC forecasters Josh Weiss, Brain Hurley, and Zack Taylor for helping with the forecast briefings during the intensive weeks. Your expertise and insights into each case was invaluable to making our experiment a success! Ben Albright for his dedication to running MODE and providing all of the objective verification information, and Sarah Trojniak for updating the MSTP drawing tools and building the HMT Retro image website. 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. Teams from EMC, GSL, and CAPS are gratefully acknowledged for developing, implementing and testing the RRFS and giving participants a first look at potential RRFS datasets. 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!

7. References

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.

WPC-HMT, 2021: The 11th Annual Winter Weather Experiment: Final Report. Accessed 29 April 2022, https://www.wpc.ncep.noaa.gov/hmt/11th_Annual_WPC_HMT_WWE_Final_Report.pdf

WPC-HMT, 2021: The 11th Annual Winter Weather Experiment: Program Overview & Operations Plan. Accessed 29 April 2022, https://origin.wpc.ncep.noaa.gov/hmt/wwe2022/12th_WWE_Plan.docx

Appendix A: MODE Configuration

MODE was used to objectively analyze the 00Z forecast cycle (12z cycle for GFSv16) at Day 1 (f36), Day 2 (f60), and Day 3 (f84) for the 24 hour snowfall forecast objects to 24 hr observed snowfall from NOHRSC. MODE was run for each case identified by the WWE facilitators, 29 cases in total. Some datasets were missing cases as noted within the sections above. All snowfall accumulation forecasts and NOHRSC observations were regridded to a 5 km grid over 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.