14th Annual WPC-HMT Winter Weather Experiment: *Final Report*

Weather Prediction Center (WPC)
Hydrometeorology Testbed (HMT)

W. Massey Bartolini[1], **James Correia Jr.**[1], and **Sarah Trojniak**[1]

[1]CIRES/CU Boulder, NOAA/NWS/WPC/HMT

February 5, 2025

1 Experiment Overview

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 14th annual Winter Weather Experiment (WWE) during the 2023-2024 winter season. The WWE provides a collaborative research to operations (R2O) experience, bringing together members of the forecasting, research, and academic communities to evaluate and discuss winter weather forecast challenges. Recent WWE successes include improvements to the National Blend of Models, incorporation of snow squalls to the mPING crowd-sourcing observation app, and increased discussion on the creation of winter specific verification metrics.

As the WWE completed its 14th year, HMT staff sought to continue evaluations on the next generation of NWS deterministic and ensemble forecast systems. This year also included evaluation of forecast thermodynamic profiles based around capabilities developed within a new interactive sounding viewer webpage. Building on the success of previous years, the WWE utilized a combination of virtual and hybrid experiment activities and analyzed retrospective case studies to perform the experiment objectives.

2 Science and Operations Objectives

The objectives of the 14th Annual Winter Weather Experiment were:

- Evaluate and compare the utility of operational and experimental deterministic snowfall and precipitation type forecasts from high-resolution convection-allowing models (CAMs)
- Explore modeled and observed thermodynamic profiles via an interactive sounding viewer tool
- Compare CAM snow to liquid ratio (SLR) forecasts to machine learning post-processed SLR forecasts
- Evaluate the accuracy of a gridded freezing rain analysis compared to available observations and CAM forecasts
- Explore experimental CAM ensemble probabilistic forecast products and machine learning post-processing guidance for winter precipitation
- Use both event- and season-long verification to assess the performance of experimental datasets

3 Season Summary

Similar to the previous winter season (2022-2023), the winter of 2023-2024 had well below normal snowfall across much of the eastern Continental U.S. Figure 1 shows the seasonal total snowfall estimated by the National Snowfall Analysis (provided by the National Operational Hydrologic Remote Sensing Center, hereafter, NOHRSC) for the December 2023 through April 2024 period. A few regions had above-normal snowfall, notably interior Oregon and Idaho and eastern Tennessee.

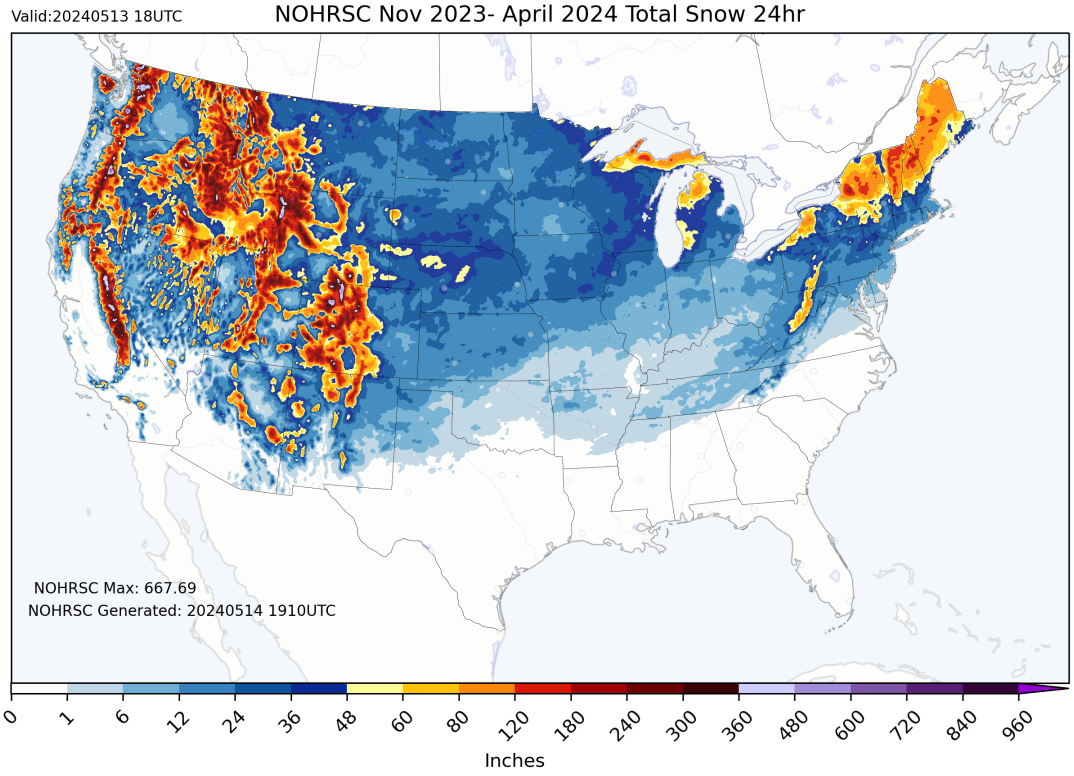


Figure 1: Summary of estimated winter 2023-2024 snowfall using NOHRSC 24h snowfall analysis data.

In terms of freezing rain, a few regions had one or more substantial freezing rain events exceeding a quarter inch of flat ice. Figure 2 provides an estimate of the flat ice accumulations from the Freezing Rain Accumulation National Analysis (FRANA) dataset. The FRANA estimate spans the period from 17 January through 16 April 2024, plus a few selected cases from the winter prior to 17 January.

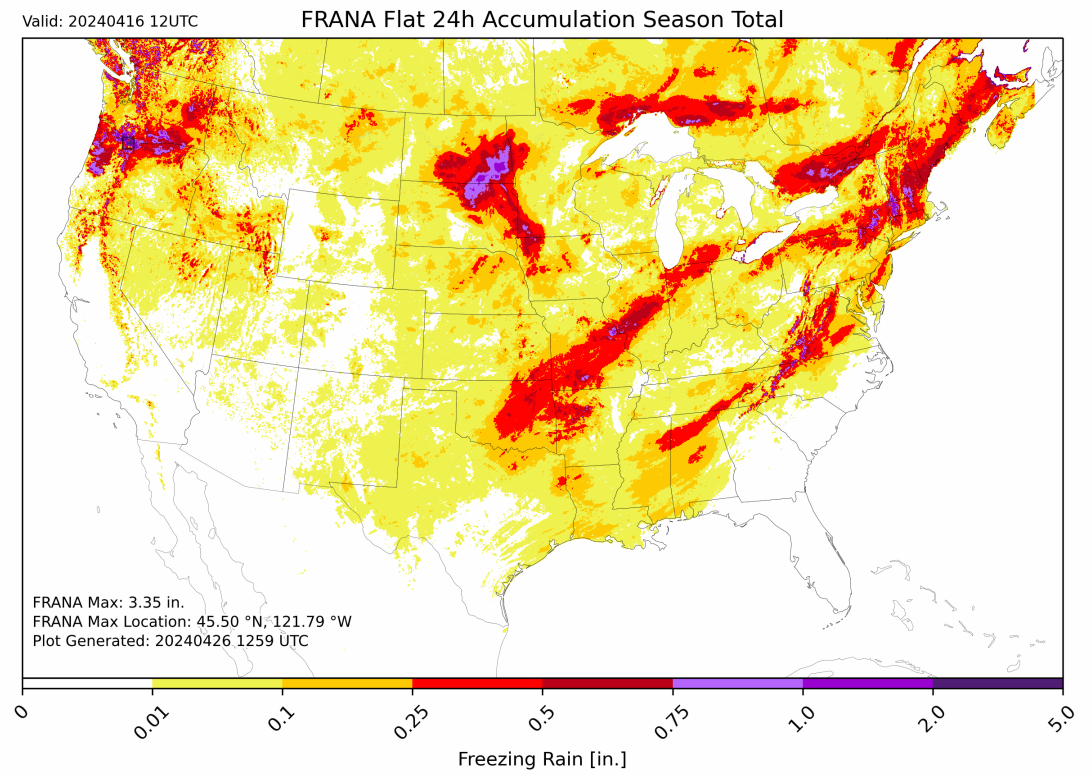


Figure 2: Summary of estimated winter 2023-2024 freezing rain (flat ice) using the Freezing Rain Accumulation National Analysis (FRANA) dataset.

3.1 Overall Case Summary and Intensive Weeks

Overall, this year's WWE season had periods of intermittent winter weather. A slow start to this winter, with widespread above-average temperatures in the Eastern U.S., meant that many of the cases selected occurred in January and February. In particular, mid-January brought an extended series of winter cases across much of the continental U.S. with mixed precipitation and well-below-normal temperatures for a brief time. From 1 December 2023 through 15 April 2024, a total of 25 cases were selected for analysis during WWE, with 9 cases (highlighted in bold in Table 1) studied in-depth during three intensive WWE weeks. Operational and experimental forecast data was collected for all the cases listed in Table 1 to compile seasonal verification statistics.

3.2 Other Notable Cases

Due to a major configuration change by the RRFS development team from a single- to mixed-physics ensemble on 8 December 2023, early-season cases selected by HMT staff prior to this date were not considered for seasonal analysis. These included some moderate lake-effect snow cases impacting

Number	Case Dates	Location	Case Type
1	18-19 Dec 2023	G. Lakes / C. Apps.	Snow
2	7-8 Jan 2024	West Coast	Snow
3	**8-9 Jan**	**G. Plains / G. Lakes**	**Snow**
4	9-10 Jan	G. Plains / G. Lakes	Snow
5	**12-13 Jan**	**G. Plains / G. Lakes**	**Snow**
6	**13-14 Jan**	**Pac. NW**	**Snow, FZRA**
7	**14-15 Jan**	**Northeast**	
8	**15-16 Jan**	**Mid-South / Southeast**	**Snow, FZRA**
9	16-17 Jan	Mid-Atlantic	Snow
10	**17-18 Jan**	**Pac. NW / N. Rockies**	**Snow**
11	19-20 Jan	Eastern U.S.	
12	**22-23 Jan**	**Central U.S.**	**FZRA**
13	**11-12 Feb**	**S. Plains / H. Plains**	**Snow**
14	12-13 Feb	S. Plains / MS Valley	Snow
15	13-14 Feb	Northeast	Snow
16	**16-17 Feb**	**OH Valley / N. Mid-Atl.**	**Snow**
17	26-27 Feb	Rockies	Snow
18	1-3 Mar	Sierras / G. Basin	Snow
19	7-8 Mar	C. Plains / H. Plains	Snow
20	13-16 Mar	C. Rockies	Snow
21	22-24 Mar	Northeast	Snow, FZRA
22	24-25 Mar	U. Midwest	Snow
23	2-3 Apr	U. Midwest	Snow
24	3-5 Apr	Northeast	Snow
25	6-8 Apr	Rockies	Snow

Table 1: All cases span a 24 h valid time period from 12 UTC to 12 UTC, except for case 12 which ended at 00 UTC 23 January. Rows in bold denote cases selected for in-depth study during intensive weeks. Cases 17 through 22 occurred too late in the season for evaluation in the intensive weeks, but were included in the overall seasonal verification discussed in Section 5.3.

portions of the Great Lakes. Furthermore, several high-impact cases around the holidays, such as the late December 2023 Dakotas ice storm, had RRFS model data outages preventing their use as WWE case studies.

4 Experiment Data Summary

WWE participants evaluated a variety of experimental data, which are listed in Figure 3. As in previous years' WWEs, the development of the future Rapid Refresh Forecast System (RRFS) was at the center of evaluation activities. The core of this system is a Limited Area Model (LAM) that has the finite volume cubed-sphere (FV3) dynamic core. The deterministic flagship CAM, referred to as the RRFSp1 or "m1" in Table 2 (also known as "RRFS_A"), was still in active development during

the winter season; thus, the deterministic and ensemble model configuration underwent numerous changes during the season. More information about the experimental guidance that was evaluated can be found in Figure 3. For more detailed descriptions of these products, please refer to Section A of the Appendix.

Model	Provider	Resolution	Forecast Length	Notes
FV3-LAM Ensemble (14 members including TL and HRRR)	EMC	3 km	60 hours	Membership and configurations may change
FV3-LAM Ensemble (11 members)	OU CAPS	3 km	84 hours	Membership and configurations may change
ML Snowfall	OU CAPS	3 km	42 hours	ML input from CAPS RRFS ens. and HREF
ML SLR	Univ. of Utah	3 km	60 hours	ML input from EMC RRFS ens.
FRANA	CIWRO/ NSSL	1 km	-	Analysis product, evaluation only

Figure 3: Summary of ensemble, machine learning post-processing, and analysis datasets planned for evaluation in this year's WWE.

5 Experiment Findings and Results

5.1 Experiment Format

This year's WWE continued with the successful remote interactions of past WWEs and the hybrid modality during two weeks of the 2023 Flash Flood and Intense Rainfall Experiment (FFaIR). The team hosted three week-long WWE sessions during the weeks of **February 12**, **February 26**, and **March 11, 2024**. The final experiment week in March was run in a hybrid format, with participants in-person at NCWCP in College Park and online via Google Meet. As described above, experiment activities were based on retrospective cases captured during the 2023-2024 winter season.

Intensive week attendees across the three weeks represented each of the four CONUS NWS regions, as well as several NCEP centers and NOAA laboratories. Several university/academic partners, including graduate students, also attended the intensive weeks. Figure 4 shows the geographic diversity of the overall attendee list during the 14th WWE intensive weeks.

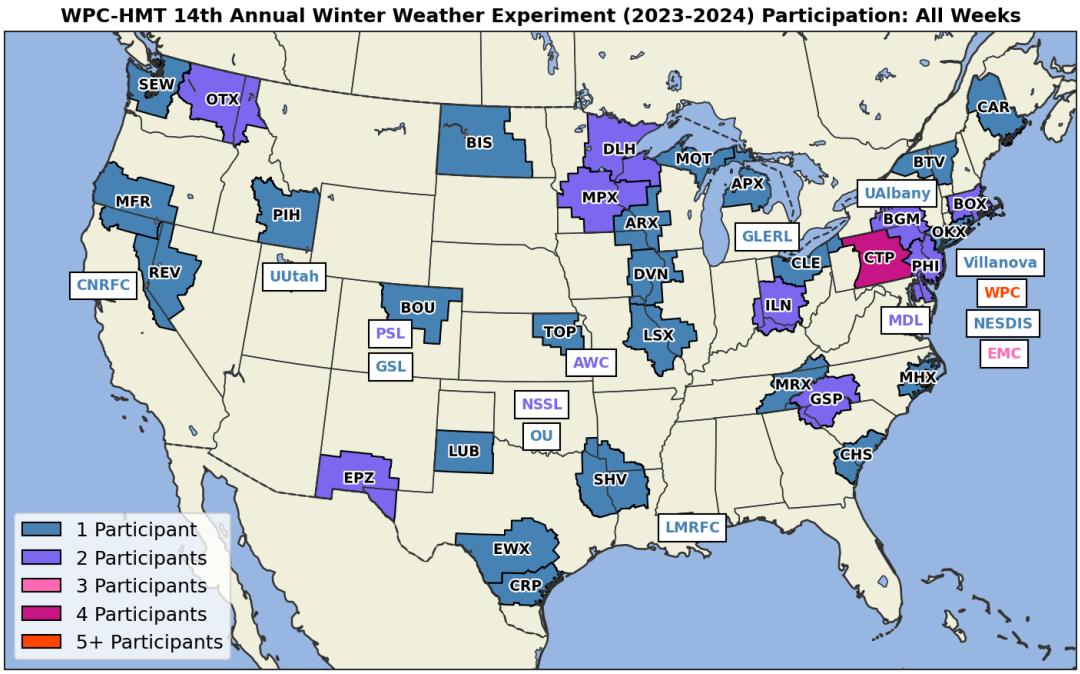


Figure 4: Map of participant affiliations attending the 14th WWE intensive weeks. WFO polygons are color-coded by total attendance counts according to the legend, while research lab and university attendance counts are color-coded by their text box.

Session days were led by the WWE team roughly as follows: each day began with a brief synoptic overview and forecast discussion to orient experiment participants to the forecast setting and predictability challenges of a retrospective case study. The majority of the time was spent analyzing forecast data and discussing the case in terms of footprint, amounts, and timing of winter precipitation. During snow-focused cases, participants completed their own Maximum Snowfall and Timing Product (MSTP) forecast consisting of a snowfall footprint, a highest amount contour based on their confidence in the forecasts, and mark where they forecast the largest snowfall to occur over the region of interest. During mixed precipitation cases, participants did a similar activity forecasting freezing rain footprint, amount, and timing. As part of this activity, participants did a complementary activity forecasting winter precipitation timing for several point locations of interest. Along with plan view maps, participants used thermodynamic profiles from an interactive sounding viewer tool to forecast onset and duration of winter precipitation or the timing of mixed precipitation changeover in certain cases. Each case concluded with a verification activity and time reserved for open-ended discussion of winter weather forecast challenges. Figure 5 shows the general structure of one of the intensive weeks, with a mix of forecasting, evaluation, and discussion activities.

Monday	Tuesday	Wednesday	Thursday	Friday
1415 UTC: Ice Breaker and Orientation	1430: Case 1 Day 2 Verification	1430: Case 2 Day 1 Briefing	1430: Case 3 Day 3 Briefing	1430: Case 3 Day 1 Briefing
1600: Case 1 Day 2 Briefing	1530: Case 1 Day 1 Verification	1500: Case 2 D1 MSTP and Sounding Activity	1500: Case 3 Day 3 MSTP Activity	1500: Case 3 D1 MSTP and Sounding Activity
1630: MSTP Overview				
1700: Lunch	1700: Lunch	1700: Lunch	1700: Lunch	1700: Lunch
1800: Case 1 Day 2 MSTP and Sounding Activities	1800: WWE Seminar	1800: Finish Case 2 Day 1 Activities	1800: WWE Seminar	1800: Case 3 Day 3 Verification
	1900: Case 1 Discussion	1830: Case 2 Day 2 Verification	1900: Case 3 Day 2 Briefing	1900: Case 3 Day 2 Verification
2000: Break	1930: Case 2 Day 2 Briefing	1940: Break	1930: Case 3 Day 2 MSTP and Sounding	2000: Break
2010: Case 1 Day 1 Briefing	2000: Case 2 D2 MSTP and Sounding Activity	1950: Case 2 Day 1 Verification	2030: Break	2010: Case 3 Day 1 Verification
2040: Case 1 Day 1 MSTP and Sounding Activities		2130: Case 2 Discussion	2040: Case 3 Day 2 MSTP and Sounding Activity Continued	2120: Case 3 Discussion and Weekly Wrap-up
2200: End	2200: End	2200: End	2200: End	2200: End

Figure 5: Schedule of activities during one of the 14th WWE intensive weeks.

5.1.1 Maximum Snowfall and Timing Product (MSTP)

In order to facilitate the NWS objective of increasing impact-based decision support services (IDSS) activities, we have employed a number of exercises for our experiment participants. The MSTP forecast activity has two objectives: 1. Give participants a reason to scrutinize various aspects of the model forecasts and challenges, and 2. actually use the model or ensemble data to make specific decisions. This activity allows forecasters to build a little knowledge on the utility of experiment data for specific forecast issues like amount, timing, and location rather than just providing subjective comments about the overall "goodness" of the snowfall. Feedback from developers has indicated that statistics based only on the "goodness" of the forecast, though helpful, does not provide enough information about the deterministic or ensemble model forecast. Developers are interested in things like: if the model had the right idea but had the footprint shifted or if snowfall or freezing rain amounts are too high or low. Along with the product, participants were required to fill out a survey stating which deterministic or ensemble model they based their forecast on, why they forecast what they did, and any other notes they felt might be useful. The MSTPs were evaluated and verified subjectively and objectively using the MODE verification toolkit (Bullock et al., 2016) and NOHRSC data (in freezing rain cases, against a new freezing rain analysis, FRANA. Details of this new analysis will be discussed below).

The MSTP activity had participants draw a minimum snowfall (typically 1") /freezing rain (typically .01") threshold or what we called the footprint. This allows HMT to verify the forecast

and evaluate forecasts objectively on the low end. Participants were also asked to draw a Max Contour threshold of their choice, highlighting the highest threat, or higher confidence or largest areal coverage contour they felt was needed. Each participant was further asked to identify and locate the maximum value of 24h snow or freezing rain they felt would occur for each event. Probabilistic activities included participants recording the probability that 6"/24h or 0.25"/24h would occur for snow/freezing rain, respectively. They also recorded a probability distribution for the maximum snowfall/freezing rain. These activities provide a basis for estimating the utility or relative value of the various models and their specific prediction which should complement objective verification.

5.1.2 Forecast Sounding and Precipitation Timing Activity

New this year, participants also used a mix of plan view forecast maps as well as forecast sounding profiles in an activity designed to assess model thermodynamic profiles and relate them to plan view maps of forecast winter precipitation type. HMT used model forecast information from BUFR profiles at over 300 sounding site locations across the continental U.S. to display individual member forecast profiles from the HREF and REFS ensemble members, including the HRRR, NAMnest, and RRFSp1. An interactive sounding viewer webpage was created by HMT staff, using a code base from collaborations with David Harrison and Israel Jirak at SPC, with sounding displays tailored to winter weather forecasting including specific indices for winter weather.

For the activity, participants were given a list of sounding sites near and within the MSTP regional domain for a specific intensive week case, and instructed to choose 3 sites to forecast for (first-come, first-served). The forecast objectives included the onset and duration of winter precipitation at specific forecast locations near and within the MSTP regional forecast domain. These objectives were defined to capture different types of winter weather events. For example, in the case of an all-snow event, the onset time would be the onset of any precipitation. In the case of rain transitioning to a winter mixed precipitation event, the onset time would be the onset of freezing rain or snow as the rain-snow transition reached a specific site. During the evaluation activity, participants used a METAR precipitation-type viewer created by HMT to assess their site onset and duration forecasts. While designed to get participants to think about the winter weather uncertainty at specific sites, another objective of this activity was to generate conversation about other goals of this year's WWE, and drive discussion about SLR, p-type, freezing rain, snow squalls, and other forecast information that heavily relies on sounding profile analysis.

Following the success of previous years, the WWE also hosted a seminar series with invited presentations throughout the winter season, as shown in Figure 6. These presentations were hosted on Google Meet on **Tuesdays and Thursdays** at 1 pm EST (1800 UTC) and advertised to all NWS personnel. Presentation files were saved from each seminar, and are available on our HMT seminar webpage under "2023-2024 Winter Weather Experiment (WWE) Seminars".

Seminar Date	Name	Affiliation	Topic
Tue, December 5, 2023	Massey Bartolini	CIRES/WPC	14th WWE Overview
Thu, December 7, 2023	Daniel Tripp	CIWRO/NSSL	FRANA
Tue, December 12, 2023	Andrew Rosenow	CIWRO/NSSL	MRMS snow rate
Tue, December 19, 2023	Ben Blake	SAIC/EMC	RRFSv1
Thu, January 4, 2024	Justin Minder	Univ. at Albany	WINTRE-MIX
Tue, January 9, 2024	Peter Veals / Jim Steenburgh	Univ. of Utah	Machine learning for SLR
Thu, January 11, 2024	Greg Carbin	WPC	Recent history of WPC Winter Weather Desk
Thu, January 18, 2024	Laura Tomkins / Sandra Yuter	NCSU	Snow banding observations
Tue, January 23, 2024	Keith Brewster	CAPS/OU	CAPS FV3-LAM ens. and ML snow products
Thu, January 25, 2024	Andrew Lyons	SPC	SPC winter program, snow squall research
Tue, February 6, 2024	Dana Tobin	CIRES/WPC	Winter Storm Severity Index development
Thu, February 8, 2024	Christiane Jablonowski	Univ. of Michigan	RRFS/FVCOM coupling for lake-effect snow
Tue, February 13, 2024	Bruce Veenhuis	WPC	PWPF research
Thu, February 15, 2024	Geoff Manikin	MDL	NBM winter updates
Tue, February 20, 2024	Eric Guillot	NWS HQ	NWS Winter Program
Tue, February 27, 2024	Nick Leonardo / Brian Colle	SBU	Idealized simulations of snow multi-bands
Thu, February 29, 2024	Anna Wilson / Jay Cordeira	CW3E	2022-2023 Western U.S. record snowfall
Tue, March 12, 2024	Tracy Hertneky	NCAR	Ensemble verification tools
Thu, March 14, 2024	Daniel Cobb	NWS	SLR observations and verification

Figure 6: 14th WWE seminar schedule, with presentations during both intensive and non-intensive weeks.

5.2 Representative Cases

In this section, several cases are discussed that were studied during the WWE intensive weeks. The selected cases represent high-impact winter weather events that capture some of the key themes and forecast insights of this past winter's experiment.

5.2.1 Southern Plains Ice Storm: 23 January 2024

During this event, precipitation was overrunning a shallow subfreezing airmass in place across the Southern Plains to Midwest region during late January 2024. Precipitation mainly fell as liquid (freezing rain or rain) and light mixed precipitation in the north, with some event-total flat ice freezing rain accumulations of greater than 0.25 inches observed. Figure 7 shows FRANA estimated flat ice amounts and local storm reports in the region studied as a WWE intensive case, focusing on the swath of heaviest freezing rain from Oklahoma to Missouri.

Surface observations of precipitation type from selected areas were examined for this case in and around the 0.25 inch flat ice areas. In many locations, unknown precipitation typing was

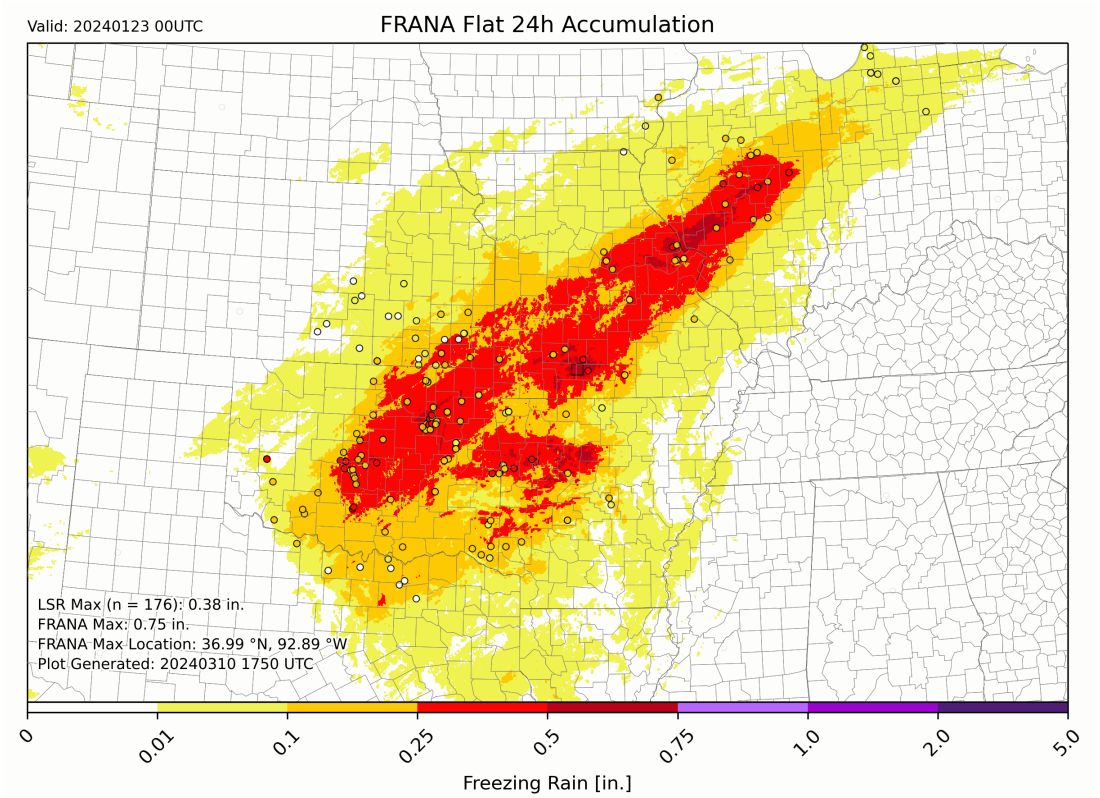


Figure 7: Flat ice freezing rain accumulation estimated by FRANA for the 24h period ending 00 UTC 23 January 2024 (fill, in). Circles denote local storm reports of freezing rain according to the same color scale.

alternating with freezing rain, while other stations had more continuous bouts of freezing rain (Fig. 8). These time series depictions were used by participants to gauge the continuity and duration (not shown) of the events during evaluation sessions. At most of the stations depicted, the rain intensity was consistently light and intermittent and spotty thus emphasizing just how hard it can be to predict these events.

The FRANA estimates of flat ice and freezing rain LSRs differ in their depiction of the freezing rain in this case, especially northeast of the Ozarks (Fig. 7). Note: while LSRs shown are all assumed to be reported as flat ice, it is possible that some are actually reports of radial ice (radial ice amounts are less than flat ice). While the footprint of 0.01 inches of freezing rain is generally well-captured by FRANA compared to the LSRs, there are greater differences in the reports for ice amounts greater than 0.25 inches. Where FRANA estimates show a broad region of greater than 0.25 inches from Oklahoma northeast to Illinois, relatively few LSRs had similar accumulations except in isolated areas. This suggests that FRANA overestimated the flat ice amounts, particularly from central Missouri northeast into Illinois.

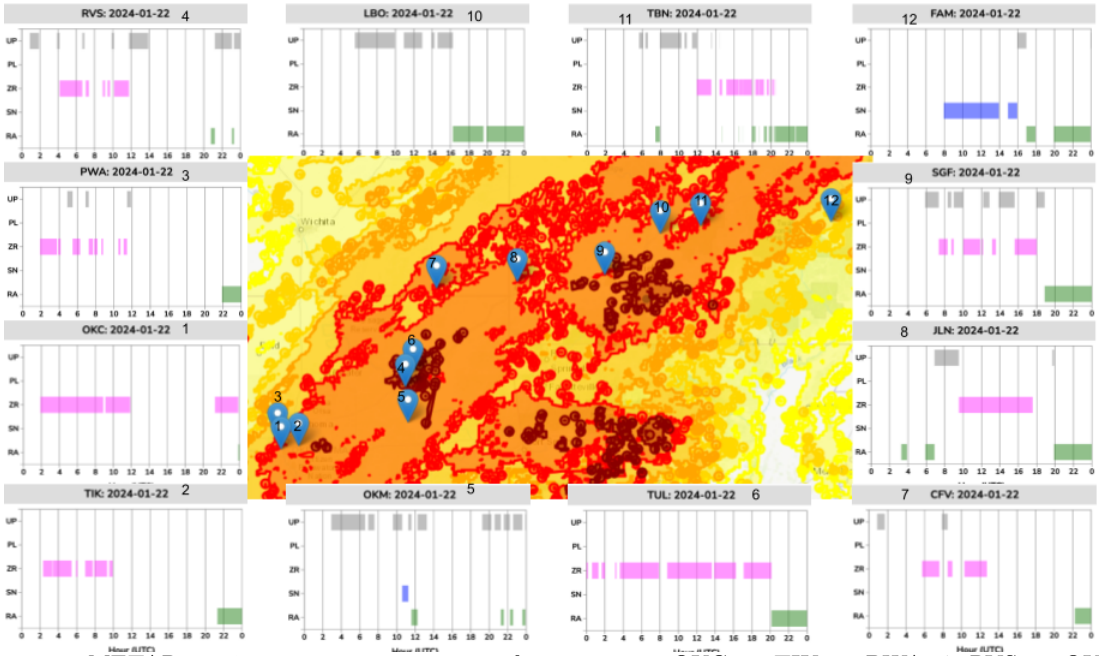


Figure 8: METAR precipitation-type time series for stations 1. OKC, 2. TIK, 3. PWA, 4. RVS, 5. OKM, 6. TUL, 7. CFV, 8. JLN, 9. SGF, 10. LBO, 11. TFN, and 12. FAM along with a map of their respective locations relative to the FRANA 24h analysis (polygons contoured, dots marking local maximum in each respective object). RA is rain, SN is snow, ZR is freezing rain, PL is ice pellets, and UP is unknown precipitation.

For the three freezing rain cases examined in detail during the WWE intensives (January 14, 16, and 23: cases 6, 8, and 12 in Table 1), we found some common differences in the FRANA estimates compared to LSRs. While FRANA often captured the ice footprint reasonably well, it tended to overestimate the higher freezing rain amounts in some areas. Figure 9 summarizes participant feedback on the FRANA performance across all three cases, in terms of footprint of flat ice freezing rain (0.01 inches) and maximum amounts.

Explicit predictions of freezing rain QPF amounts were a new capability of the RRFS and its ensemble members for the past winter season, so WWE examined several freezing rain events in detail including this late January case. WPC-HMT also ran post-processing on the HRRR, RRFSp1, and two other RRFS ensemble members, to generate forecasts of FRAM-adjusted flat ice amounts from each of the CAM guidance, in addition to the raw freezing rain QPF fields available from each model. We were thus able to examine the differences between the freezing rain QPF and FRAM forecasts for multiple cases.

In this case, FRAM post-processing resulted in generally similar swaths of freezing rain compared to the original freezing rain QPF fields as seen in Figure 10. However, there were some increases in the footprint of light freezing rain (0.01-0.1 inches) on the northwest side of the FRAM

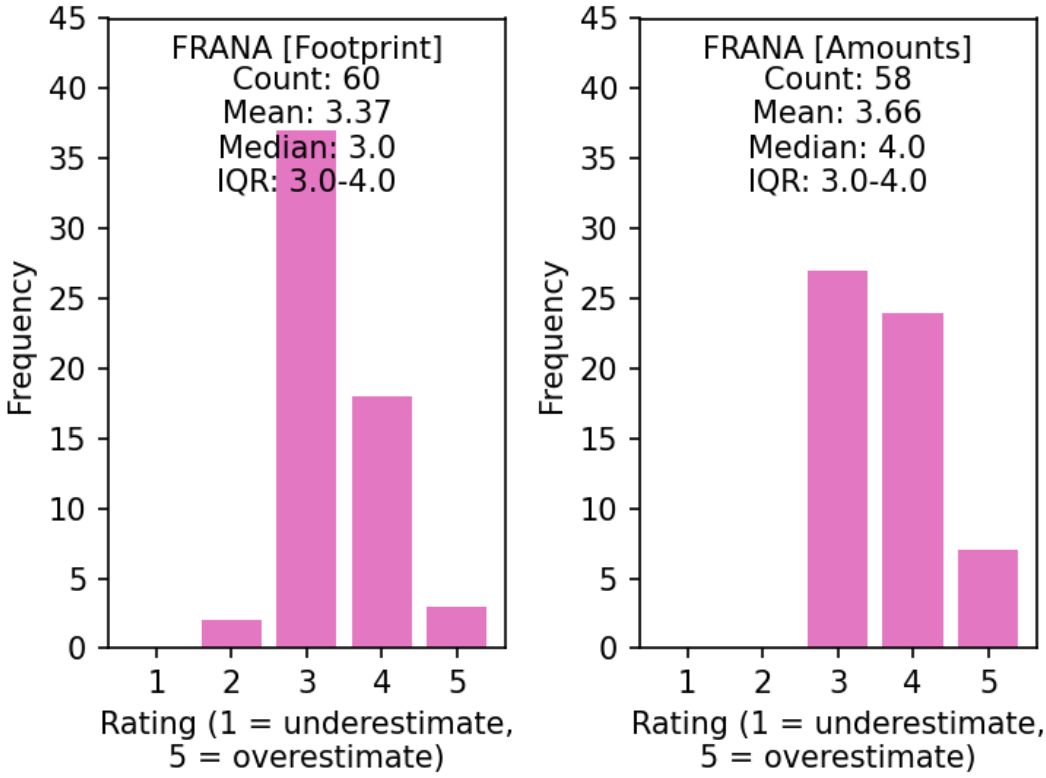


Figure 9: Aggregate participant subjective ratings of 24h FRANA flat ice estimates compared to freezing rain local storm reports (LSRs) across three freezing rain cases studied. Participants scored FRANA both in terms of footprint (0.01 inches) and maximum amounts.

forecast, particularly in the Kansas and northwest Missouri area. The largest influence of using FRAM was in the max amount of freezing rain forecast across the Ozarks, which resulted in about a 10-20 percent reduction in the largest freezing rain amounts relative to the raw freezing rain QPF amounts. Overall, using FRAM resulted in more accurate 24h freezing rain forecast swaths from both HRRR and RRFSp1, both compared to FRANA and the available LSRs across the region.

Participant comments provided some more insight into the challenges the models had with predicting this freezing rain event. While the HRRR and RRFSp1 were best among the guidance in terms of the overall footprint and amounts, they still had challenges resolving some of the marginal freezing rain areas across TX and AR. Participant survey responses indicated they tended to rely on HRRR, RRFSp1, and the REFS ensemble probabilities in the forecast drawing activity. However, the performance of other members like RRFSm2 and RRFSm4 sometimes suffered from greater low or high biases in the maximum freezing rain amounts across the Ozarks (not shown) which may have limited their usefulness.

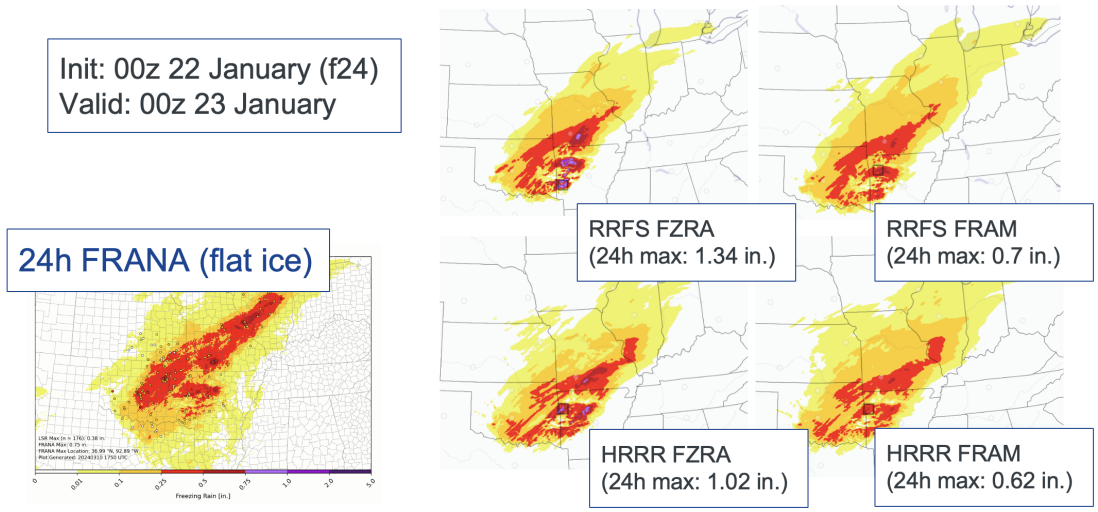


Figure 10: Model forecasts of 24h freezing rain QPF (center) versus FRAM-adjusted flat ice freezing rain (right). Forecasts initialized at 00 UTC 22 Jan 2024, valid for the 24h ending at 00 UTC 23 January 2024.

The differences in the footprint and maximum amount swaths of freezing rain seen in this January case across the Ozarks are representative of the types of adjustments the FRAM post-processing provides compared to the raw freezing rain QPF field. In general, these differences yield positive improvements in CAM forecasts of freezing rain swaths. Since the new capabilities of the RRFS system added freezing rain QPF to the product output stream (in addition to HRRR, which already did so), WWE was able to perform quantitative forecast and evaluation activities for freezing rain for the first time, since other HREF members besides HRRR do not have explicit freezing rain amounts predicted. We recommend that FRAM be added as a post-processing option to RRFS/REFS to provide more comparable output guidance to what NWS WFOs and WPC forecast for freezing rain amounts.

By using the deterministic and ensemble guidance provided to them, participants were generally able to make accurate forecasts of freezing rain footprint and maximum amounts. Objective verification of participant MSTP forecasts for freezing rain was performed by assessing individual forecasts against FRANA for this case, despite the biases with FRANA noted above in this event. Performance diagrams shown in Figure 11 summarize the participant results.

Even considering the predictability challenges posed by freezing rain at CAM guidance lead times, participants were able to forecast accurate freezing rain footprints at Day 2 and Day 1, with CSI values around 0.75 for both activities. Maximum amount forecast contours did not verify as well, with lower CSI values especially at Day 2. However, a number of participants achieved CSI values greater than 0.2 at the Day 1 lead time for a freezing rain max amount contour of 0.25 inches (warning-level ice amounts). In general, participants decided to draw for higher maximum amount

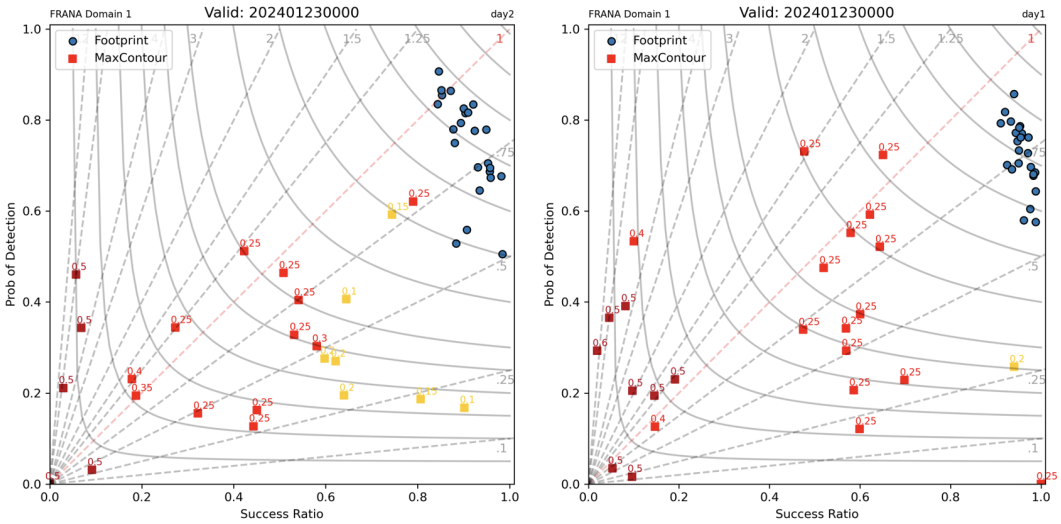


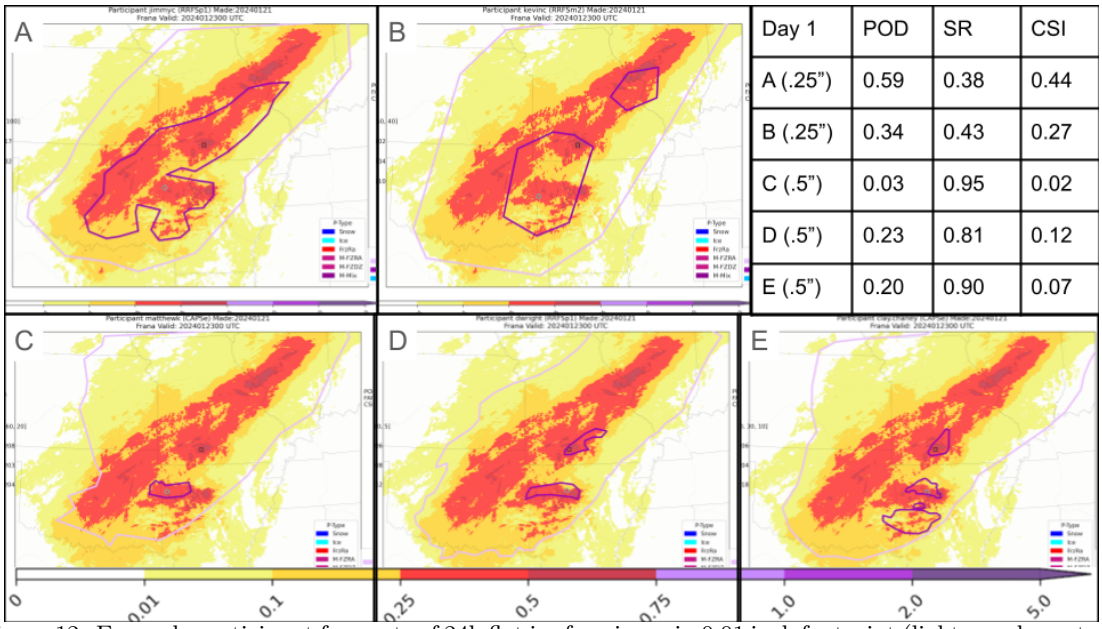
Figure 11: Participant forecasts of 24h flat ice freezing rain 0.01 inch footprint and max amount for Day 2 (left) and Day 1 (right) lead times, verified against FRANA. Blue circles denote 0.01 inch footprint forecasts while squares denote max amount forecasts, annotated by the value individual participants decided to forecast at each lead time. All forecasts valid for the 24h ending at 00 UTC 23 January 2024.

values at Day 1 compared to Day 2, perhaps indicative of higher confidence in the model forecasts at the shorter lead time.

The Day 1 participant example forecasts highlighted the very large area of freezing rain quite well (Figure 12). The higher amounts of 0.25 and 0.5 each had unique challenges with areal coverage in the main southwest to northeast oriented band and the sidelobe across AR. The model forecasts showed areal coverage lower for the higher amounts, and shifted southward across southern IL. The terrain effects in AR were captured well by participants, but the 0.5" max contour forecasts all missed the heavy area in and northeast of St Louis into IL.

5.2.2 Ohio Valley and Mid-Atlantic Mesoscale Snow Banding: 16-17 February 2024

In an ongoing active weather pattern during mid-February, a low pressure system moved from the Tennessee Valley eastward to Virginia on 16-17 February 2024. On the north side of this system, several rounds of narrow mesoscale snowbands developed, impacting areas from St. Louis, Missouri, eastward to north-central New Jersey with heavy snow. While overall snowfall and QPE totals were relatively light across the region, several areas of localized snowfall amounts exceeded 6-10 inches in 24 hours. The most notable part of this event was the extremely high SLRs that were observed in portions of Pennsylvania and New Jersey, exceeding 20:1.



Day 1	POD	SR	CSI
A (.25")	0.59	0.38	0.44
B (.25")	0.34	0.43	0.27
C (.5")	0.03	0.95	0.02
D (.5")	0.23	0.81	0.12
E (.5")	0.20	0.90	0.07

Figure 12: Example participant forecasts of 24h flat ice freezing rain 0.01 inch footprint (light purple contour) and max amount (purple contour) for Day 1 lead times, verified against FRANA for A. and B. 0.25" forecasts, and C-E. 0.5" forecasts. All forecasts valid for the 24h ending at 00 UTC 23 January 2024.

CoCoRaHS 24h observations across eastern Pennsylvania and New Jersey showed a small band of enhanced precipitation exceeding 0.3-0.4 inches, with a broader region of 0.1-0.2 inches of precipitation across the northern Mid-Atlantic. Combined with the enhanced SLRs and precipitation falling as all snow, a narrow swath of 30-40 km in north-south extent (at the scale of a county or less across eastern PA and NJ) was impacted by high snowfall totals with a few locations reporting 11-13 inches of snow. The small-scale nature of snow banding and snowfall amounts highlights the predictability challenges posed for even CAM guidance in this event.

During this case examined in WWE intensive week 3 (March 11-15, 2024), we evaluated both deterministic and ensemble model snowfall guidance at 2- and 1-day lead times to understand the predictability of this event. Given the small-scale nature of the snow banding, individual models struggled to accurately forecast both the magnitude and location of the heaviest snowfall amounts. Figure 14 shows snowfall forecasts from two of the REFS members and the HRRR for several lead times up until event onset.

While each of the models captured a swath of light to moderate snowfall exceeding 4 inches, several key differences separated the HRRR and REFS members. RRFSp1 and RRFSm4 had the heaviest snowfall forecasts at the Day 2 lead time (forecasts initialized at 00 UTC 15 February), but decreased in magnitude while also shifting southward in the position of the west-east oriented snowfall swath with decreasing lead time. These patterns also occurred in the QPF fields (not

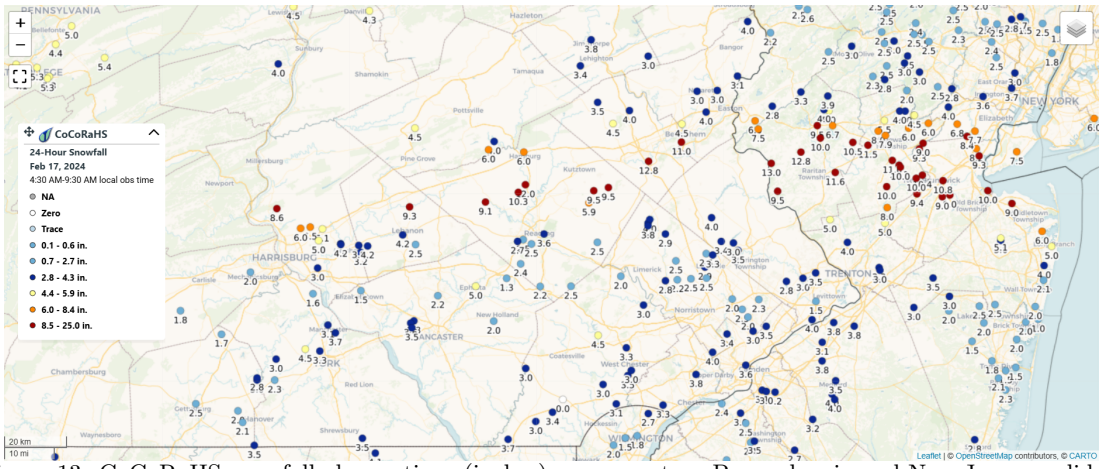


Figure 13: CoCoRaHS snowfall observations (inches) across eastern Pennsylvania and New Jersey, valid for the 24 h period ending at 12 UTC 17 February 2024.

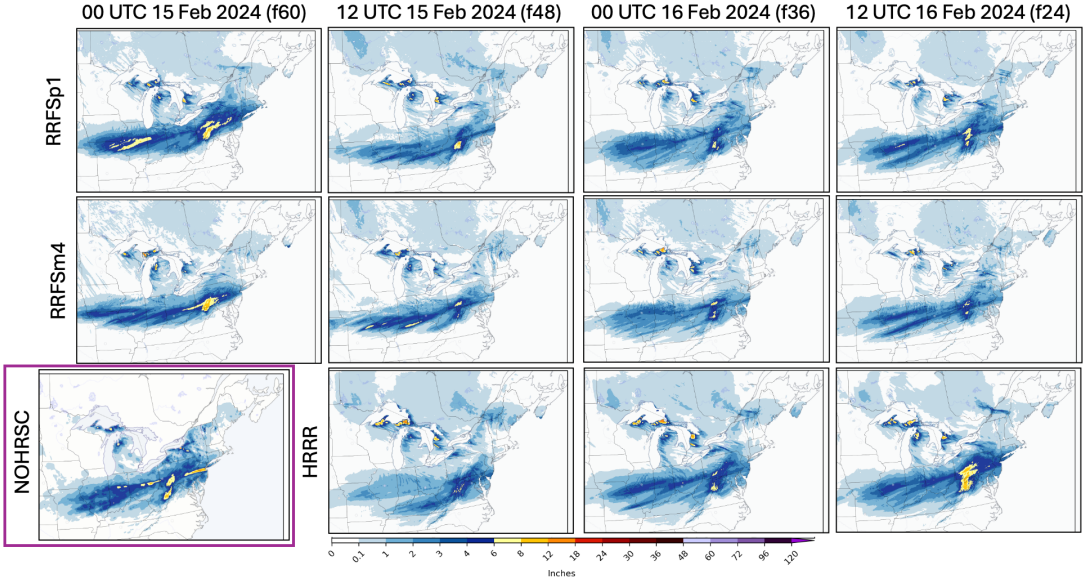


Figure 14: Variable density snowfall forecasts for the RRFSp1 (top row), RRFSm4 (middle row), and HRRR (bottom row), for 24h snowfall valid at 12 UTC 17 February 2024. Each column depicts forecasts at different lead times, for the same valid end time. NOHRSC snowfall estimates are shown for the same 24h period.

shown), indicating this was not just a SLR error. This generally resulted in larger forecast differences relative to NOHRSC at short lead times. In contrast, HRRR overall had less snowfall than RRFSp1 and RRFSm4 at the 12 UTC 15 February lead time, but trended higher with overall snowfall amounts as well as a northward shift in the west-east snowfall, especially towards the eastern portion of the region in central Pennsylvania and New Jersey. By the "nowcast" lead time of 12 UTC 16 February,

HRRR had the most snowfall of the models shown, even overpredicting snowfall in some areas like the West Virginia mountains relative to NOHRSC.

To explore overall run-to-run trends in the individual REFS ensemble members for this case, a JTTI funded project ran MODE object-based forecast verification metrics for each lead time as shown in Figure 15, relative to the 24h NOHRSC analysis. This figure was provided by Tracy Hertneky (National Center for Atmospheric Research). In each panel, the time series values indicate attribute changes for each member relative to the previous forecast cycle. For the 90th percentile of the snowfall object intensities (top row), the REFS members show little run-to-run consistency but the variations decrease somewhat with decreasing lead time. For the displacement errors in the forecast snowfall centroids, run-to-run changes also were large at longer lead times at both 1 inch (lower left) and 4 inch (lower right) thresholds, but decreased with decreasing lead time. Other differences are evident between members for this event, with some of the perturbed REFS members exhibiting larger trends in their snowfall footprints than the control member, RRFSp1, which had slightly greater run-to-run consistency for this event. For example, RRFSm4 (green lines) had an eastward shift in centroid location at two consecutive forecast lead times of 42 and 36 h, whereas RRFSp1 (black lines) had smaller displacement changes, or more run-to-run consistency overall.

Despite the magnitude and displacement errors in snowfall shown by the CAMs at Day 2 and Day 1 lead times, analysis of other model fields and thermal profiles still highlighted the potential for heavy snow and mesoscale snow banding in advance of the event. Figure 16 shows sounding profiles at 06 UTC on 17 February in a north-south cross-section across the Mid-Atlantic from model forecasts initialized at 00 UTC 15 February 2024, 30 h in advance of the event.

While the magnitude and north-south placement of forcing features for mesoscale snow banding (e.g., a band of warm air advection at 700 hPa) varied across forecast cycles at Day 2 and Day 1 lead times between the HRRR and RRFSp1, they existed nearby a deep, saturated isothermal layer around the 700-800 hPa level in the model sounding profiles, varying from -19 C at BGM to -11 C at ABE in Figure 16. Many model forecasts, such as the 00 UTC 15 Feb RRFSp1, showed a brief but robust signal for warm air advection and lift, but displaced too far south into Maryland instead of where thermal profiles were more ideal for efficient snow growth in Pennsylvania. The best overlap between thermal profiles and forcing wasn't apparent until some of the last "nowcast" cycles of the HRRR as the snow was beginning, where the warm air advection signal trended north (not shown). This could help explain the underprediction of snow across most of the CAM guidance during this event (Fig. 14) in Pennsylvania and New Jersey. The forecast uncertainty and displacement errors in one or more ingredients for mesoscale bands of snow remains an open research question for NWP, and/or post-processing, to better express high-SLR potential in cases such as this one.

In the last forecast cycle prior to event onset, the forecast environment from the 12 UTC 16 Feb 2024 HRRR initialization depicted a deep dendritic growth zone over the OH to NJ corridor

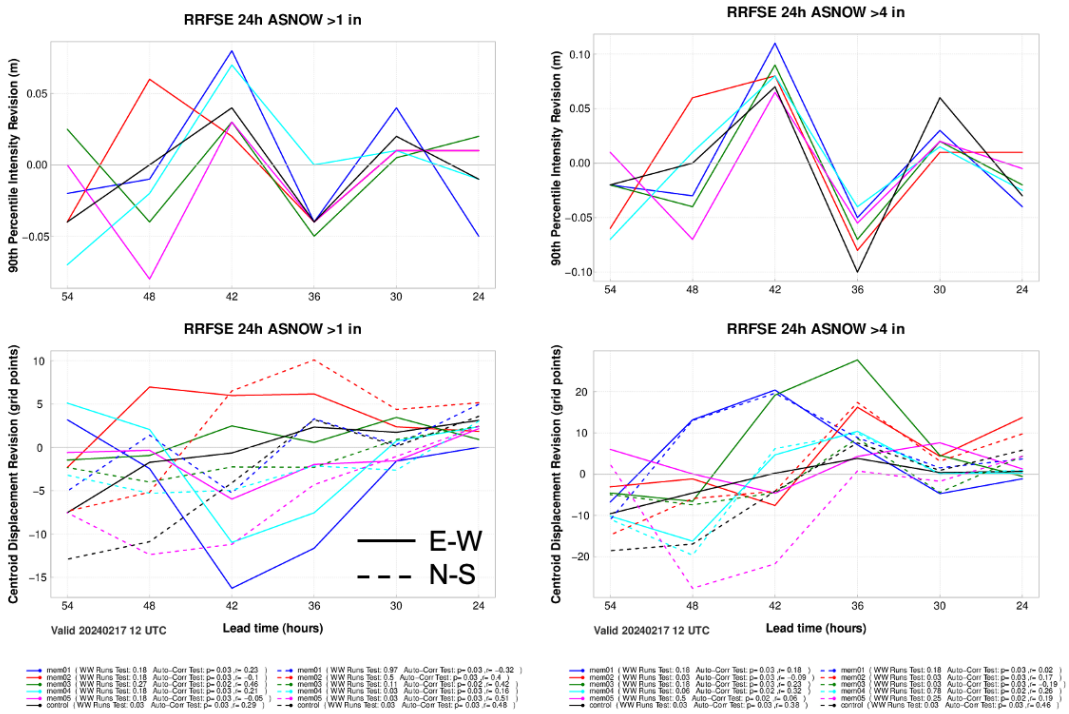


Figure 15: Time series for REFS member consistency attributes at each forecast lead time prior to the 17 February 2024 snowfall event. Object-oriented attributes are evaluated relative to 24h NOHRSC snowfall estimates. In all panels, line colors vary by ensemble member, while in the lower panels (bottom row) solid lines indicate east-west centroid displacements while dashed lines indicate north-south centroid displacements relative to NOHRSC. Figure provided by Tracy Hertneky.

(not shown) with forecast precipitation moving from west to east between 00-12 UTC on the 17th (not shown). By 10 UTC on the 17th, the dendritic growth zone had filled with reflectivity and shrunk from a depth of 2 km to just under 1.5 km (Fig. 17) but it was still present resulting in (variable density) snowfall above 4" in a narrow band across PA and into NJ. The snow water equivalent was uniformly above a quarter inch, with peak values just above half an inch in 6 hours. The 6h snow to liquid ratio was generally in the range of 11-13 to 1. As mentioned earlier, the actual observed snow ratios were above 20-25 to 1 with between 0.25 and .5 inches of precipitation within the snow band. The HRRR values were only inconsistent with respect to SLR. The HRRR variable density snow algorithm is only based on 2m temperature and is limited to between 8 and 17:1 (not shown). For this case, the SLR is roughly a factor of two in error and is the most likely culprit for the snow under-forecast. For the same initialization cycle and valid time, the corresponding RRFSp1 has much reduced SLR (7-10:1) and almost half of the variable density snow (3-4") and a narrower swath of 0.25 inches of SWE, though not identical spatially, relative to the HRRR forecast. The most obvious conclusion is that the mechanisms which govern SLR are poorly understood and represented in the models discussed here.

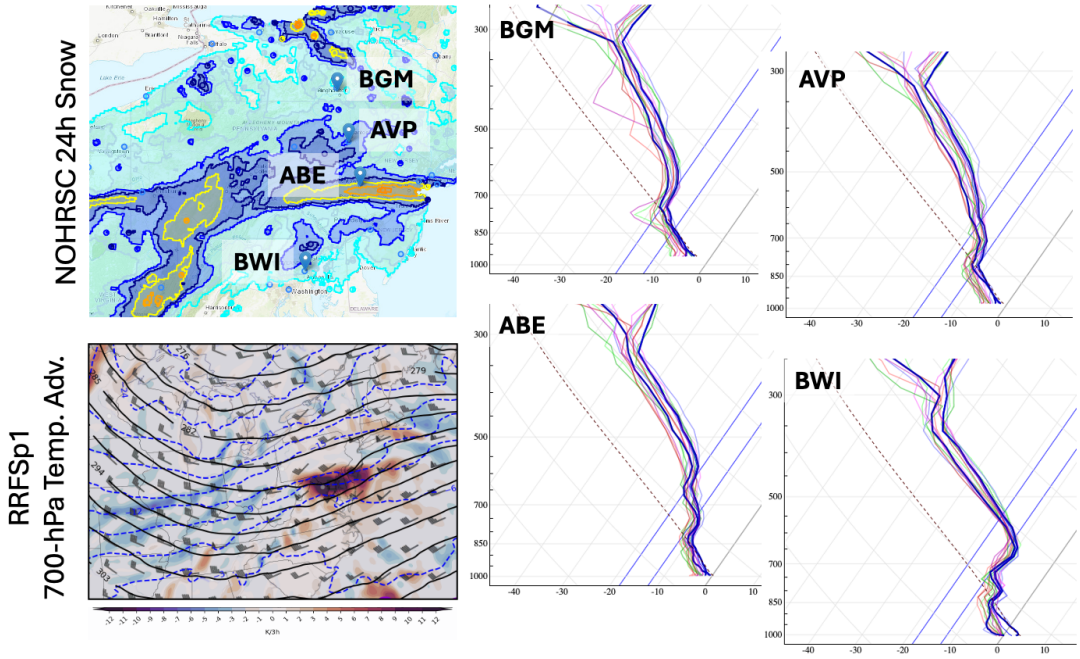


Figure 16: Forecast sounding profiles at Binghamton, NY (BGM), Wilkes-Barre, PA (AVP), Allentown, PA (ABE), and Baltimore, MD (BWI), valid at 06 UTC 17 February 2024. Colors indicate different model forecasts: RRFSp1 (blue, highlighted), HRRR (red), NAMnest (green), and RRFSm2 (pink), all from forecasts initialized at 00 UTC 15 February 2024. Thinner lines indicate soundings from the same models but the corresponding 6-h time lagged forecasts. Also shown are (top left) NOHRSC 24h snowfall estimate valid at 12 UTC 17 February 2024 and (bottom left) RRFSp1 700-hPa temperature advection (K/3h, shaded), temperature (degC, blue dashed lines), height (dam, black lines), and wind (kts, barbs) from the same model forecast and valid time as the soundings.

Due to the University of Utah research project studying ML post-processing for improving SLR forecasts, HMT examined both raw SLR forecast fields from the RRFSp1 and ensemble and the Utah ML SLR forecasts for each of the intensive week cases. The raw SLR technique used in the RRFS members (ported from the HRRR) uses a SLR relationship varying between 8:1 and 17:1 depending on the lowest model level temperature (Benjamin and Collaborators, 2021). The Utah SLR technique uses forecast fields from the RRFS model to predict an hourly SLR of its own, which is combined with raw RRFS QPF forecasts to determine a new snowfall output for each member.

SLR and snowfall forecasts are shown in Figure 18 for this case study. In this example, Utah ML SLR did not deviate much from the raw RRFSp1 SLR forecasts along the main snowfall axis from Illinois into New Jersey, remaining around 12-15:1. However, Utah ML SLR values exceeded the narrow bounds of the RRFSp1 SLR (capped between 8:1 and 17:1) especially in the Great Lakes region where some modest lake-effect snow was forecast. In contrast to the deterministic forecast, the Utah ML ensemble minimum and maximum SLR fields showed large variations in SLR between REFS members (minimum and maximum of the 6 members from a single initialization

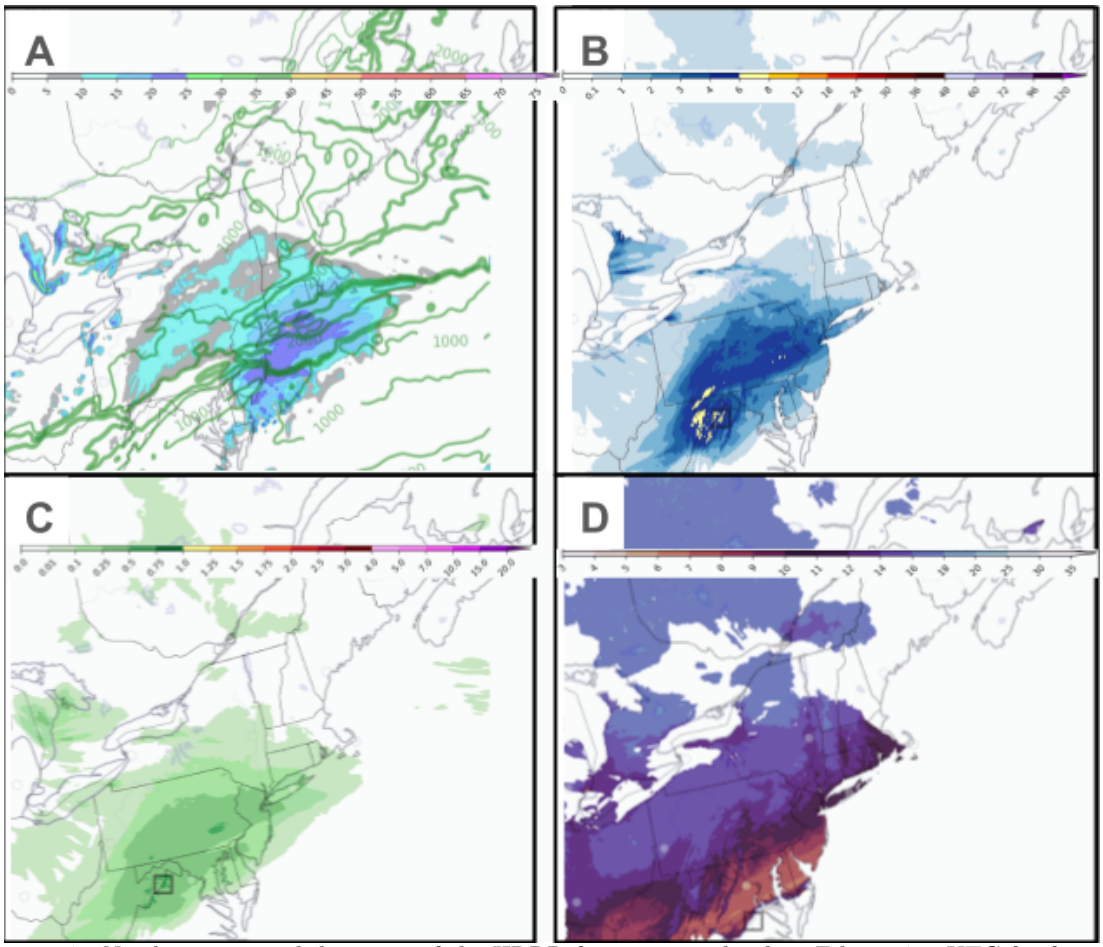


Figure 17: Northeast regional depiction of the HRRR forecast initialized 16 Feb 2024 12UTC for forecast hour 22 valid 10UTC 17 Feb 2024 for A. 263K reflectivity (shaded, dBZ) and dendritic growth zone depth (contours, m), B. variable density snowfall (in), C. snow water equivalent (in) and D. snow to liquid ratio, shading referenced in each panel's colorbar.

cycle, no time lagged membership). When combined with the individual members' QPF differences (not shown), the spread in Utah SLR forecasts led to large differences in ensemble minimum and maximum snowfall forecasts in this case.

In addition to the deterministic model forecasts and forecast trends, we also examined ensemble consensus products such as arithmetic means and probability matched means (PMM) for this case study. An example of the differences between the two for REFS forecasts at a Day 2 lead time is shown in Figure 19. While the arithmetic mean fields generally show the synoptic signal for a band of snowfall across the Ohio Valley and Mid-Atlantic, the details are smoothed relative to the verifying observational analyses. This is expected given member-to-member differences at a Day 2 forecast lead time. In contrast, the PMM technique better captures some of the higher-end snowfall

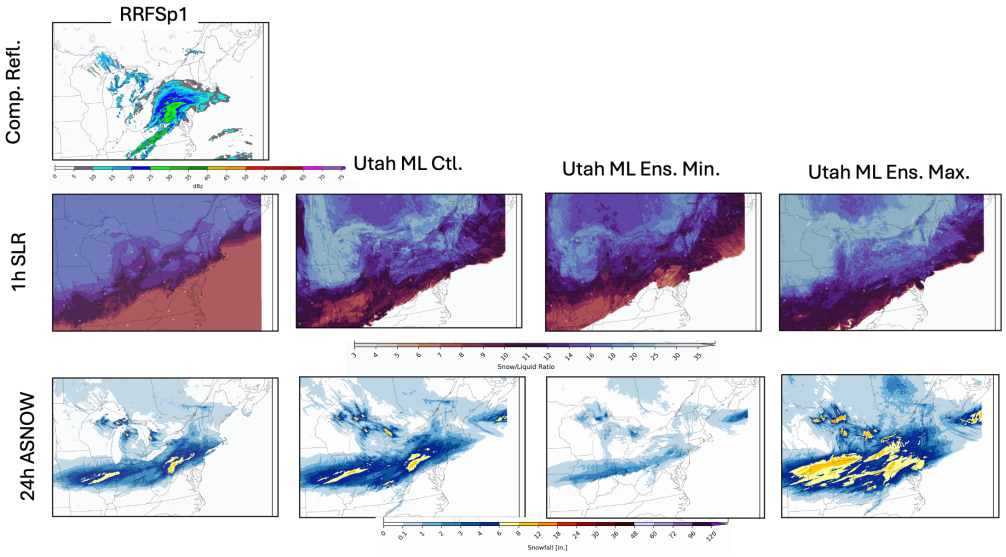


Figure 18: SLR and snowfall output from the RRFSp1, Utah ML control forecast based on RRFSp1 inputs, and the Utah ML ensemble min and max, all from forecasts initialized at 00 UTC 15 Feb 2024. Reflectivity and SLR panel forecasts are valid at 06 UTC 17 Feb 2024 (f54) while 24h snowfall panels are valid at 12 UTC 14 Feb 2024 (f60).

and QPF signal of the west-east translating swath of heavy snow greater than 4 inches, although the forecast maxima by REFS are still less than what was estimated by NOHRSC (except in the upslope snowfall region in West Virginia as well as the Great Lakes lake-effect snow belts).

For REFS forecasts initialized 24 h later (Day 1 lead time), some of the mesoscale details were in greater focus as shown in Figure 20. In comparison to Fig. 19, both the arithmetic mean and PMM QPF and snowfall amounts trended higher in the later forecast initialization on the 16th. While still low-biased and shifted about 50 km too far south compared to the observed snowfall axis, the REFS ensemble consensus products trended in the right direction from the Day 2 to Day 1 lead time. Also, the differences between the arithmetic mean and PMM are less at the Day 1 lead time (Fig. 20), indicating less spread between the individual REFS member solutions.

In addition, the CAPS group provided their FV3 ensemble (Fig. 42) and ensemble consensus forecasts initialized at 00 UTC for each lead time during this case, with forecasts extending out to Day 3. New this year, we evaluated the spatial aligned mean (SAM) and spatial aligned mean combined with local probability-matched mean (SAM-LPM) techniques (Lee et al., 2024). An example of CAPS's different ensemble products for snowfall at 6h intervals is shown in Figure 21.

Overall, the Day 3 CAPS ensemble consensus products were generally able to capture the extent of the snowfall footprint, with some spatial errors, despite the predictability challenges in this event. However, the consensus products all underrepresented the maximum snowfall amounts,

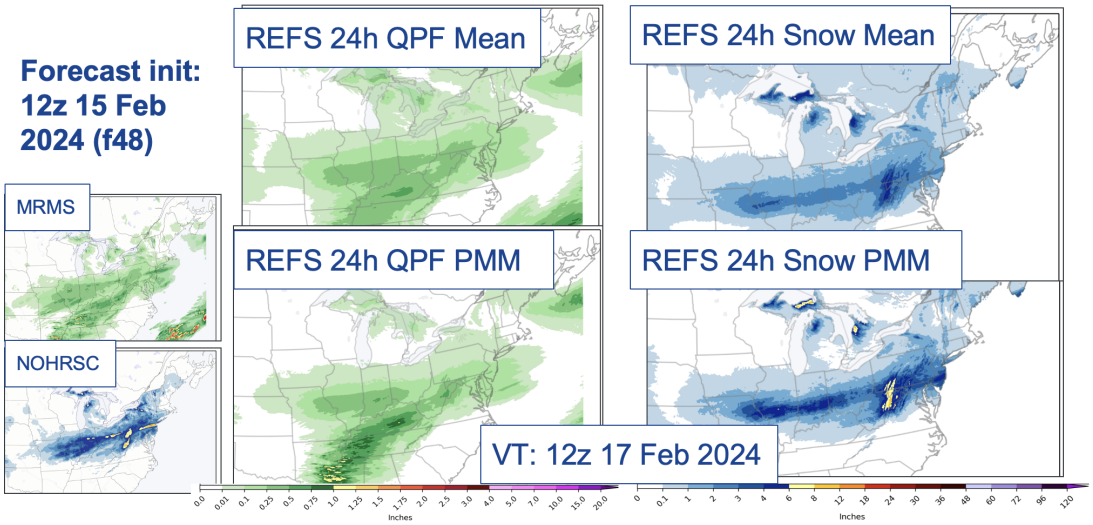


Figure 19: REFS 24 h QPF and variable density snowfall means, forecast initialized at 12 UTC 15 February 2024 valid at 12 UTC 17 February (forecast hour 48). Arithmetic means are shown in the left column while probability matched means are shown in the right column, for QPF and snowfall. MRMS QPE and NOHRSC estimates are shown for comparison, valid for the same 24 h period.

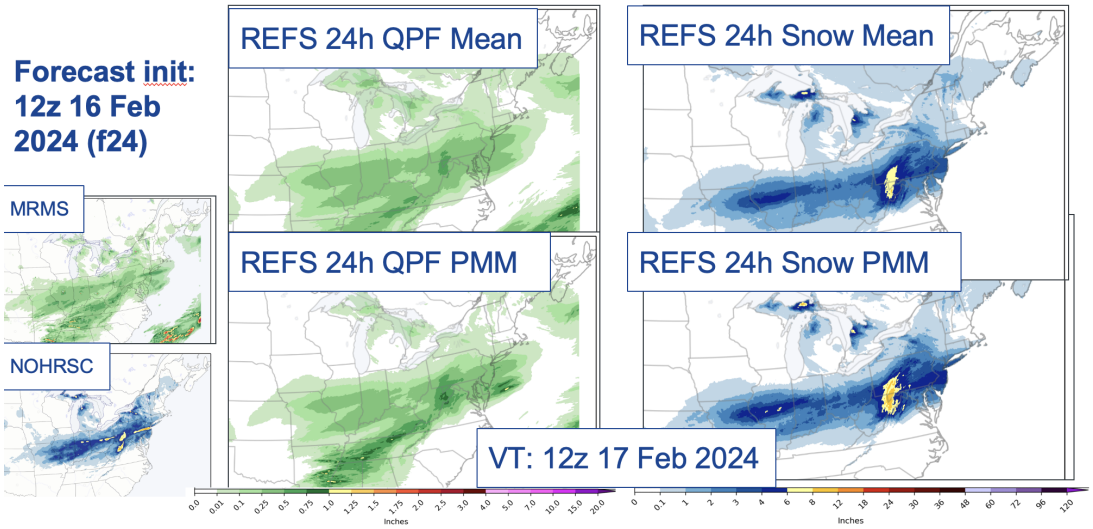


Figure 20: As in Fig. 19, but for REFS forecasts initialized at 12 UTC 16 February 2024.

with the arithmetic mean smoothing out the snow signal the most. Both the methods applying the local probability-matching technique (LPM and SAM-LPM) fared the best, retaining more of the higher-end forecast snowfall amounts in excess of 6 inches. While the SAM and SAM-LPM had more of a signal for a mesoscale snowband across eastern Pennsylvania, they also struggled to capture the southern extent of the heavy snow at this forecast hour across the central West Virginia mountains, relative to the other means.

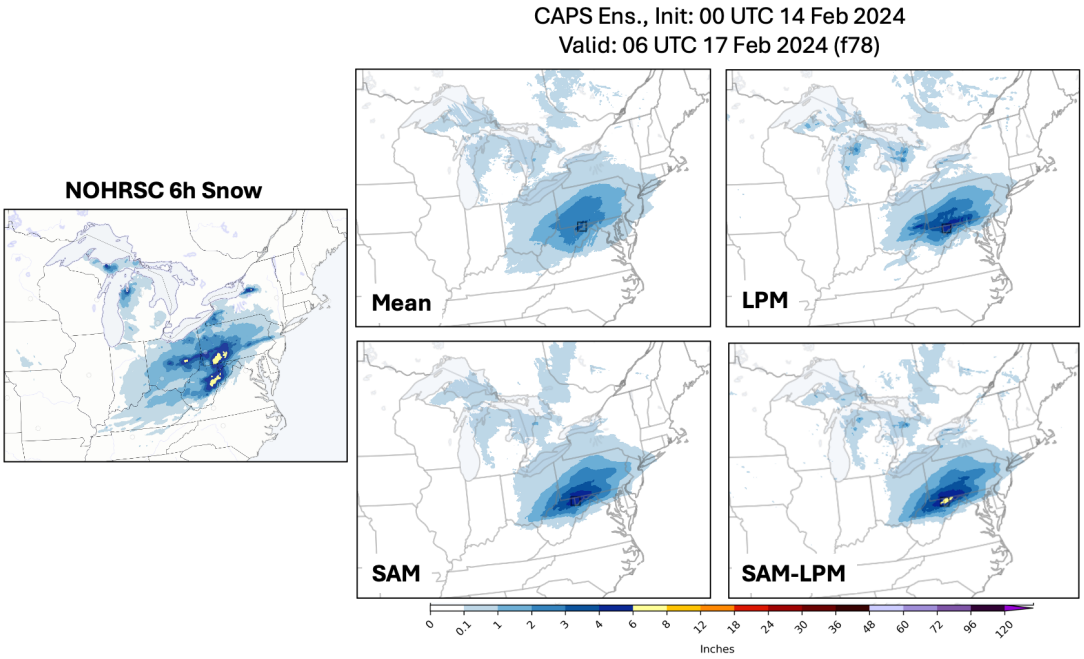
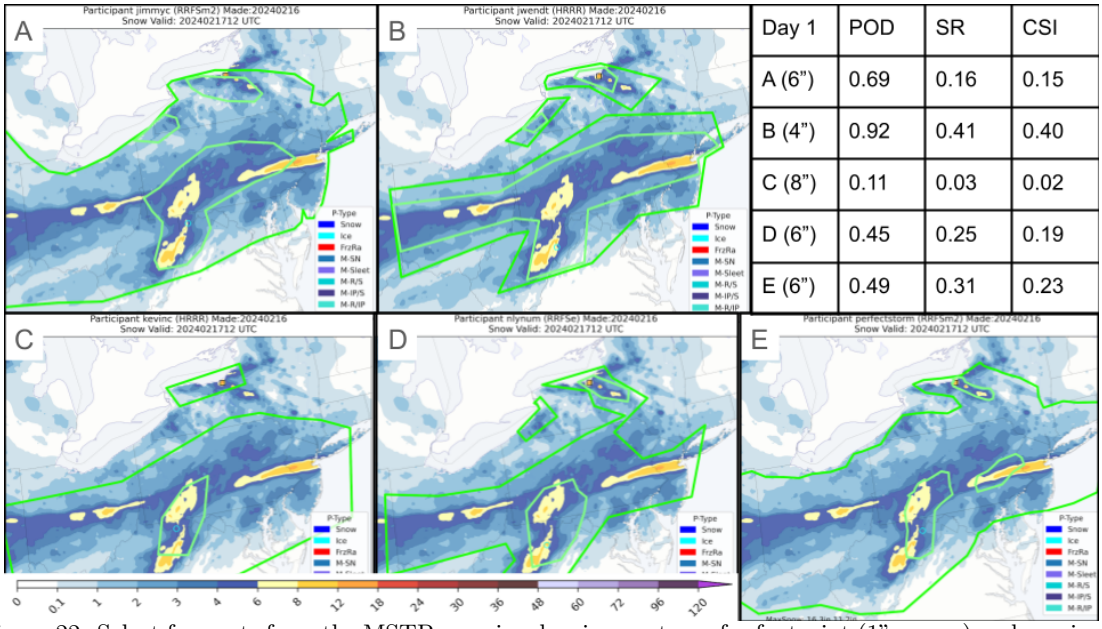


Figure 21: CAPS ensemble consensus products for 6h snowfall, from forecasts initialized at 00 UTC 14 February 2024, valid at 06 UTC 17 February. NOHRSC 6h snowfall estimates are also shown at left.

During this event, the SAM technique didn't seem to substantially improve on other methods like the LPM. This may be due to more member-to-member consistency in banded snow placement across the CAPS ensemble, limiting the realignment of the individual member position differences for a winter weather case. Forecaster comments varied as to what aspects (snowfall footprint or maximum amounts) they valued most in the consensus fields, and which mesoscale aspects they focused on, from the synoptic snowfall to the lake-effect bands also present in this case. From the Day 3 to Day 1 lead times, participants generally favored the SAM-LPM for 6h snowfall since it improved on some attributes of the forecast over the other techniques. However, many participants noted that the consensus products underestimated the maximum amounts in the synoptic snow region in the Mid-Atlantic and also struggled to capture the mesoscale details of the lake-effect snow bands. This is unsurprising as much of the deterministic guidance (e.g., Fig. 14) also underforecast the lake-effect potential, especially downwind of Lakes Erie and Ontario.

Considering the range of deterministic and ensemble guidance available to participants during the snow forecasting activities, select MSTP forecasts for this event showed (Fig 22) forecasters accurately depicting the footprint and choosing maximum contour amounts above 4". All forecasters chose to depict the Appalachian Mountain maximum generally and 3 of the 5 further emphasizing a band of heavier snow across PA and NJ, and 2 of the 5 adding a band across OH, and 3 of the 5 adding a lake-effect band off of Lake Ontario and/or Lake Erie. These forecasts were in part driven

by multiple models, since participants had access to a range of experimental forecast guidance as discussed throughout this section.



Day 1	POD	SR	CSI
A (6")	0.69	0.16	0.15
B (4")	0.92	0.41	0.40
C (8")	0.11	0.03	0.02
D (6")	0.45	0.25	0.19
E (6")	0.49	0.31	0.23

Figure 22: Select forecasts from the MSTP exercise showing contours for footprint (1", green) and maximum amount (lighter green) for forecasts with A. 6 inches, B: 4 inches, C. 6 inches, D. 8 inches and E. 6 inches of snow with performance metric table emphasizing the "maximum amount" contour. The underlay in A-E is the NOHRSC 24h snow analysis contoured according to the color bar.

5.3 Objective Verification

Since a total of nine cases were evaluated in detail across the three intensive weeks, we collected case data for an extended period during the 2023-2024 winter season. The larger set of snowfall cases used for objective verification in this section (cases 1-22 in Table 1) covered a total of thirty-two unique 24 h verification days, since some cases spanned multiple days or overlapped in different regions of the CONUS on the same day. Cases were chosen to span a variety of geographic regions and synoptic and mesoscale phenomena, to the best of our ability given the limited amount of snowfall that fell in the CONUS outside the mountains during the 2023-2024 winter season (Fig. 1). The period of time chosen for objective verification was also constrained by the extensive ongoing RRFS development efforts, so we limited our analysis to the winter period with the fewest changes to the RRFS ensemble.

Object-based verification at Day 2 and Day 1 lead times was performed using MODE (Bullock et al., 2016), comparing model forecasts of 24 h variable density snow to NOHRSC snowfall estimates. Since HREF members other than HRRR do not output variable density snow, we limited our analysis to the HRRR, RRFSp1, and two perturbed REFS members all using the same technique (rather

than introducing SLR assumptions varying by model). Objective verification was performed over the full CONUS domain covered by NOHRSC.

Day 2 verification results in performance diagram format are shown in Figure 23. The highest critical success index (CSI) values around 0.5 occur at a snow threshold of 1 inch, with lower CSI values and increasing bias at the higher snow thresholds. Compared to the REFS members at 00 and 12 UTC, the 12 UTC HRRR had the highest probability of detection (POD) but also slightly higher bias at each threshold. Differences in forecast performance between the 00 UTC and 12 UTC REFS members are most evident at the 12 inch threshold, with the 12 UTC forecasts performing better in terms of CSI.

Similar to the Day 2 verification, Day 1 verification performance diagrams are shown in Figure 24. Relative to the results of the Day 2 verification, snowfall performance in terms of CSI and bias for each of the models is improved at the 1, 4, and 6 inch thresholds. However, the cluster of model forecasts remains around a CSI of 0.2 at both Day 2 and Day 1 lead times. While HRRR and RRFSp1 performance remained relatively consistent across each threshold and forecast cycle, RRFSm2 and RRFSm4 varied in performance across the Day 1 forecasts as they were occasionally the worst or best member of the four models verified in terms of CSI.

Between the two deterministic "flagship" models, HRRR had slightly higher CSI than RRFSp1 for Day 1 snowfall across the thresholds, but at the expense of higher bias. The gap in performance between RRFSp1 and HRRR increased for higher snowfall thresholds, with RRFSp1 worse at thresholds greater than 6 inches. However, these results should be noted in context of this winter not having many heavy snow events outside of the mountainous western U.S. **Finally, there is a systematic high bias in snowfall forecast for both HRRR and REFS members at higher thresholds (> 6 in.) which may indicate a mismatch between the model variable snow density snowfall field we are verifying against NOHRSC snow depth. That is, NOHRSC estimates may account for snowfall settling or other observed factors that the model-predicted field does not.**

Due to substantial issues with missing REFS member data for individual forecast cycles and cases, we were unable to run an objective evaluation of ensemble probabilistic data for the WWE cases this winter. Instead, we highlight ensemble probabilities and consensus products from individual cases that were generally representative of notable positive or negative qualities found during the experiment.

5.4 Subjective Evaluation

During each intensive week, participants were tasked with evaluating model snowfall predictions after each set of case study forecast activities. These evaluations focused on rating the 24 hour

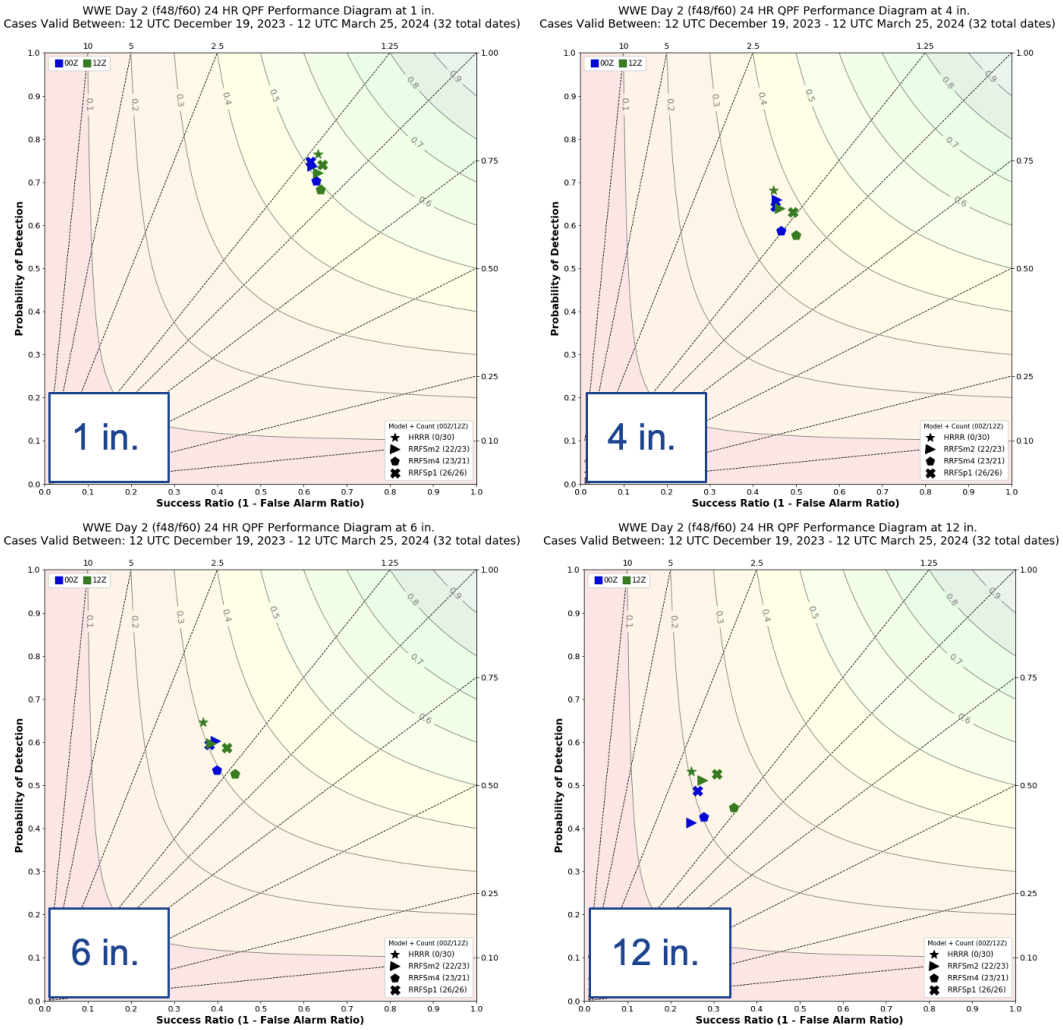


Figure 23: Seasonal MODE object-based verification for several 24h snowfall thresholds for Day 2 deterministic model forecasts: HRRR (star), RRFSp1 (cross), RRFSm2 (triangle), and RRFSm4 (pentagon). Results are color-coded by forecast initialization cycle with dark blue denoting 00 UTC and green denoting 12 UTC forecasts.

variable snow density snowfall (variable "asnow" in the GRIB2 output fields), from each of the CAMs that provided explicit variable density snowfall forecast output as in the HRRR (Benjamin and Collaborators, 2021). Given this choice, we focused evaluations on the HRRR, RRFSp1, and two other REFS members (RRFSm2, RRFSm4), as well as the CAPS control member and Utah ML snowfall forecast, since these CAMs all had the desired snowfall output. Participants were also able to evaluate ensemble consensus products, including the REFS mean snowfall and CAPS ensemble probability matching fields (PMM/LPMM). By focusing on variable density snow, this meant that we omitted evaluation of deterministic models which did not have variable density snowfall available

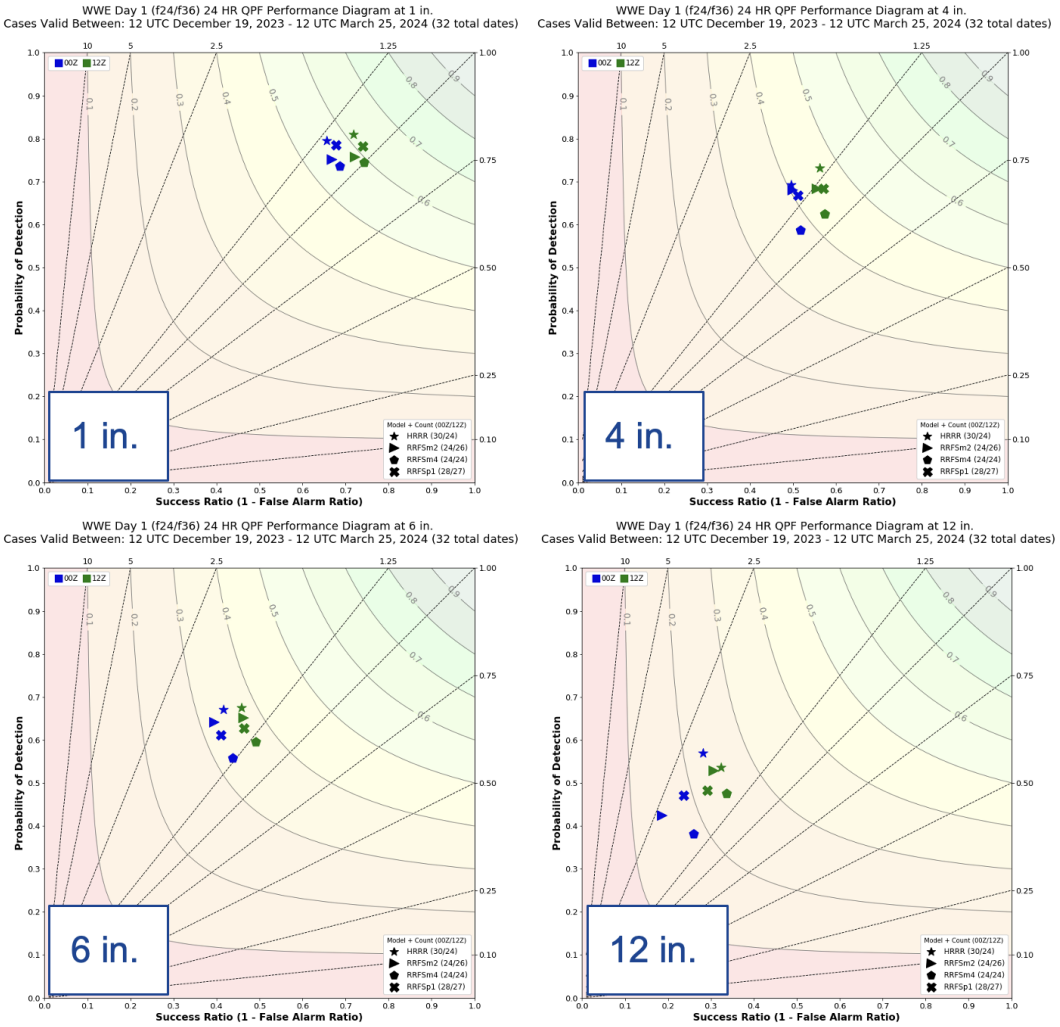


Figure 24: As in Fig. 23 but for Day 1 forecasts.

as part of their output (e.g., NAMnest, GFS), which would have required introducing other SLR assumptions (e.g., liquid equivalent snowfall multiplied by a fixed 10:1 SLR to obtain snow depth, or positive accumulated snow depth change by the land surface model). Hereafter, discussion of model "snowfall" refers to variable density snowfall.

During our verification sessions, participants compared the CAM snowfall forecasts at several different lead times to the corresponding NOHRSC 24 hour snowfall analysis. MODE was used on the variable density snowfall field at many snow thresholds. Participants were asked to consider the full-CONUS objective statistics as part of their subjective evaluations of each model, while also focusing on the mesoscale forecast domain of interest for each case study.

To evaluate model snowfall forecasts against NOHRSC gridded snowfall estimates, participants used the interface shown in Figure 25. MET/MODE was used to provide object-oriented verification statistics as described in section "Objective Verification". Participants were asked to provide subjective evaluations of each of the primary model 24 h snowfall forecasts, focusing on 00 UTC initialization times for the Day 2 and Day 1 periods. A screenshot of the MODE snowfall evaluation webpage is shown in Figure 25.

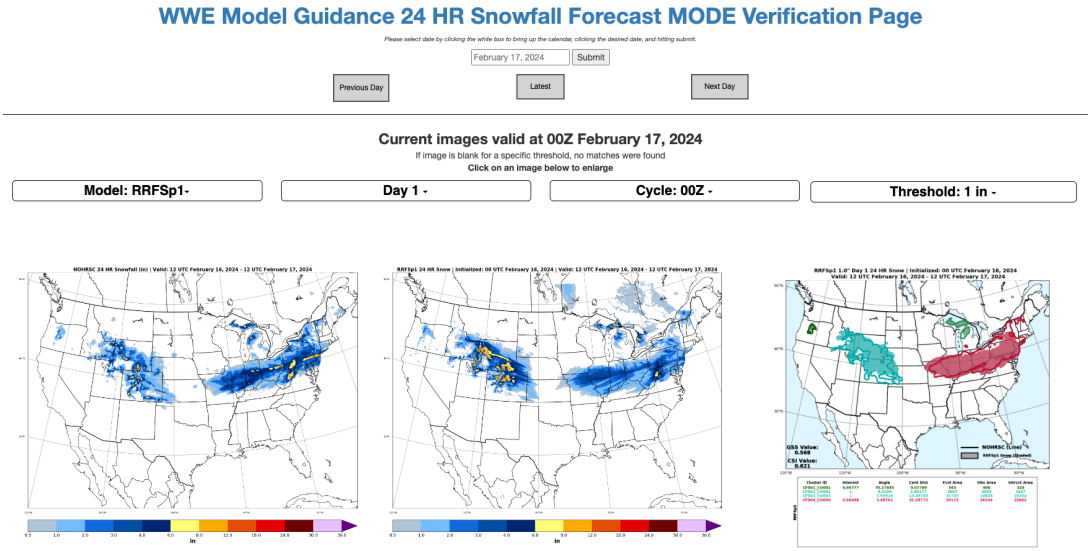


Figure 25: Example of web interface that participants used to evaluate deterministic model 24h snowfall (left panel) against NOHRSC gridded estimates (center panel), with MODE-based object-matching performance statistics (right panel). Participants were able to select different models, lead times, forecast cycles, and verification thresholds using the dropdown menus.

Figure 26 shows overall participant ratings for the 24 h snowfall for several deterministic models, aggregated across the nine snowfall cases examined across the intensive weeks. At Day 2 lead times, rating results were more similar across all of the forecast models. RRFSp1 scored highest (mean score of 2.57), followed by CAPSdet (3.01) and HRRR (3.33). Participants compared models across the 00 and 12 UTC cycles, since CAPSdet and Utah were only available at 00 UTC while HRRR was only available at 12 UTC (48 h lead time).

At Day 1 lead times, HRRR (2.64) and RRFSp1 (2.30) emerged as the consistently highest rated models according to the participant surveys. The other REFS members we examined, RRFSm2 and RRFSm4, performed worse than the RRFSp1 control member. The CAPS control member (CAPSdet) and Utah ML SLR (based on the RRFSp1 model output) snowfall forecasts were rated lowest. Interestingly, the CAPSdet forecast was rated lowest at Day 1 even though it was second-highest at Day 2 lead times. *Since CAPS forecasts were cold-started from GEFS initial and boundary*

conditions, this might have put them at a disadvantage for shorter Day 1 lead times compared to the other forecasts which use radar data assimilation.

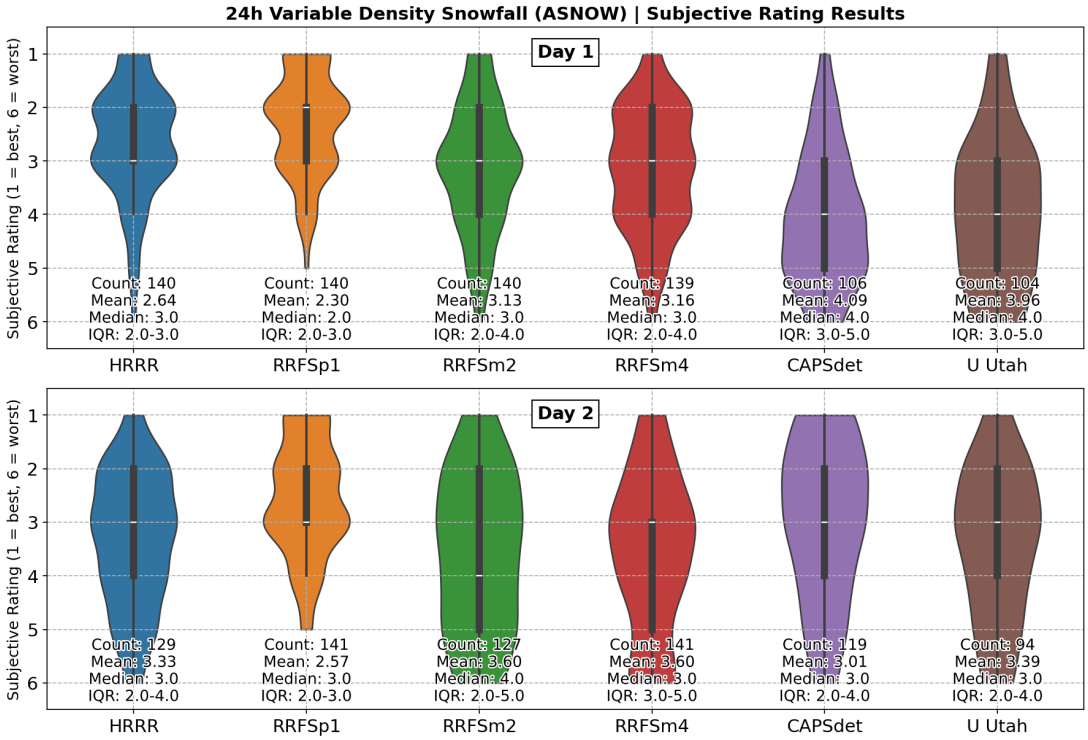


Figure 26: Aggregate statistics for participant rating of 24h snowfall forecasts.

In addition to the participant ratings of each model snowfall forecast, participants were also asked to rank each model compared to the others. Overall results from this ranking are shown in Figure 27. Similar to the rating results, HRRR and RRFSp1 were generally comparable at Day 2 lead times but with RRFSp1 having a slight edge (RRFSp1 mean ranking of 2.31 versus HRRR mean of 3.25). RRFSm2 had more of a bimodal ranking distribution than the others but was the lowest-ranked member at Day 2.

At Day 1 lead times, HRRR and RRFSp1 were ranked as the best models among the group. RRFSp1 again had a slight edge over HRRR, since it was rarely ranked as the worst or second-worst model.

Since overall rating and ranking results are only able to capture broad impressions of model "goodness", we asked participants to elaborate on other positive qualities of the model forecasts in a binary checkbox matrix format. These attributes included overall coverage of snowfall at various thresholds (1, 4, and 6 inches), location accuracy of the snowfall areas (1, 4, and 6 inches), and the forecast maximum amount, all compared to NOHRSC 24h snowfall estimates. Figure 28 shows

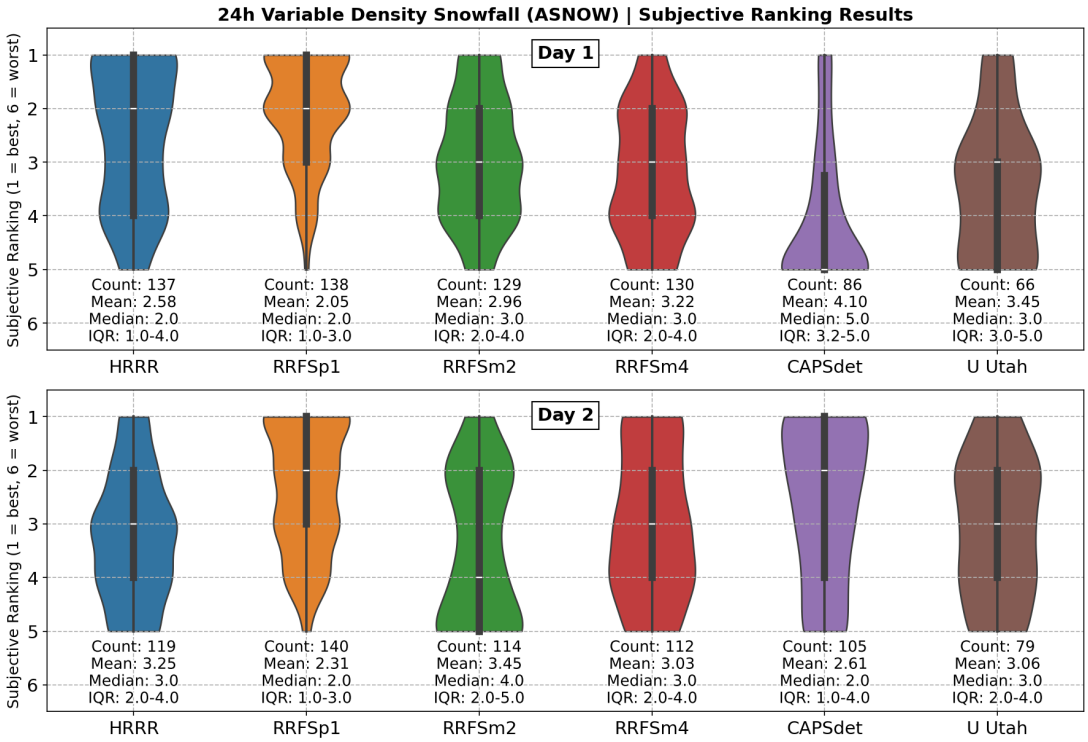


Figure 27: Aggregate statistics for participant ranking of 24h snowfall forecasts.

these results, normalized by count. A value of one indicates that a model was evaluated to always have a given quality, while a value of zero indicates that a model never had that quality.

Overall, at both Day 2 and Day 1 lead times, models generally had an easier time capturing lighter snowfall amounts, with higher frequencies of positive qualities (coverage/location) at 1 inch of snowfall than 6 inches. At Day 2, RRFSp1 slightly outperformed HRRR at most of these attributes in terms of snowfall coverage and location. At Day 1, RRFSp1 and HRRR were more equally matched in terms of the attributes we evaluated.

5.5 Discussion Topics

In this section, we discuss major themes that emerged from the variety of snow,freezing rain, and winter mixed precipitation cases examined during the WWE intensive weeks. Some subsections summarize the range of conversations and discussion we had about certain topics, while others offer representative case examples to illustrate specific attributes or biases we noted in the model guidance.

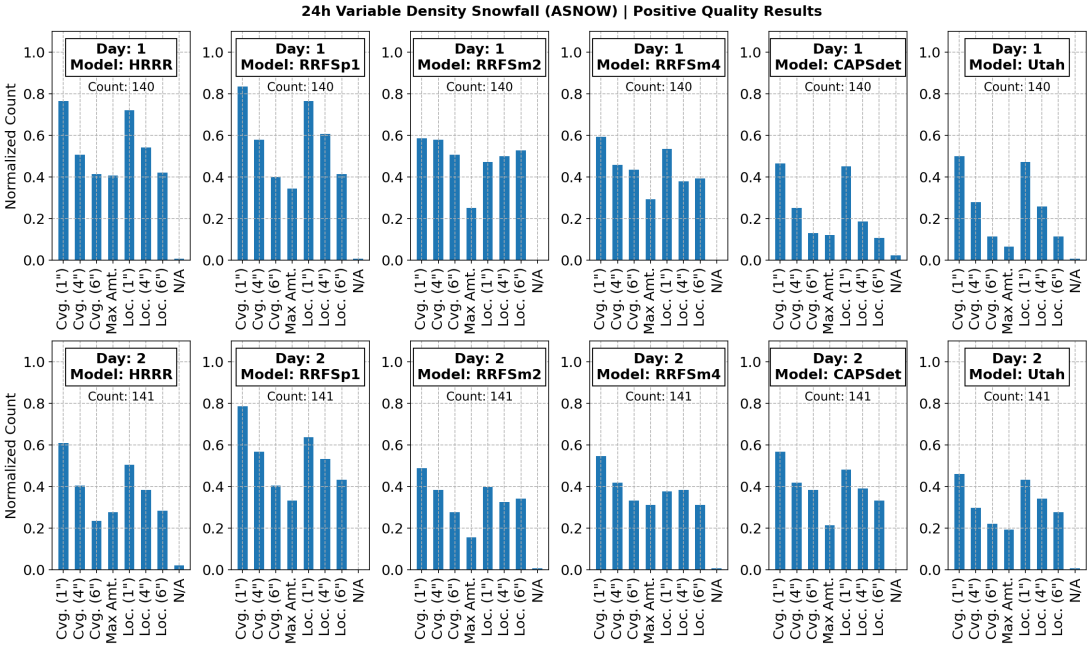


Figure 28: Aggregate statistics for participant assessments of positive qualities in the deterministic 24h snowfall forecasts.

5.5.1 Snow-to-Liquid Ratio Techniques

During the experiment, forecasters were interested to learn about the simple SLR techniques used natively in the HRRR and RRFSp1 variable snow density snowfall accumulations as these are not well known. Intensive week cases examined the SLR differences between the Utah ML methods and the raw HRRR and RRFSp1 SLR fields in a variety of geographical and thermodynamic environments, from warmer, mixed-precipitation cases in the Central Plains to the higher-SLR cases found in the Mountain West and Great Lakes (lake-effect snowfall). Overall, the Utah ML SLR methods were able to forecast a wider range of SLR amounts, producing a larger range of snowfall forecasts as in the February 17 event in the Northeast (Fig. 18). Despite this spread, sometimes the most anomalously high SLR values were not well forecast.

Another case from 12 February 2024, where heavy wet snow occurred across northern Texas and Oklahoma, highlights an extreme example of some issues we noticed after evaluating the Utah SLR products across several similar events. Figure 29 shows a comparison of 1 h SLR and 24 h snowfall amounts between the RRFSp1 and Utah ML control forecast (based off the RRFSp1 fields).

This case featured a compact upper low which produced heavy precipitation rates leading to snowfall, despite a marginal thermal environment with near-surface temperatures around 0 ° C (not shown). The SLR comparison in Figure 29 shows RRFSp1 SLRs across Oklahoma static at 8:1

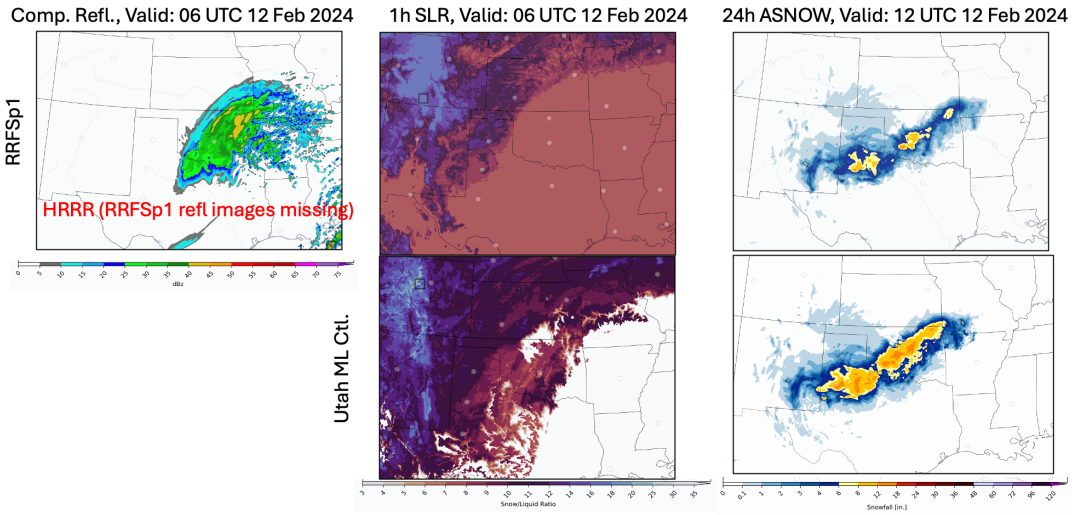


Figure 29: SLR and snowfall output from the RRFSp1, Utah ML control forecast based on RRFSp1 inputs, from forecasts initialized at 00 UTC 11 Feb 2024.

(the lower limit of the native model SLR) while Utah SLRs vary between <5:1 and 10-12:1 across Oklahoma and the Texas Panhandle. In addition, there are substantial jagged artifacts in the Utah SLR, which somewhat resemble topographical contours from west Texas to the Ozark Mountains but appear erroneous as they varied from hour to hour (not shown). The summed differences in SLR between RRFSp1 and the Utah ML forecast lead to substantial differences in the maximum snowfall forecasts across the area (Utah forecast a much larger area of 8 inches), despite relatively similar 1 inch snowfall footprints between the two deterministic solutions. NOHRSC snow estimates (not shown) indicated that both forecasts were overdone in terms of max snowfall amount and footprint, with reality closer to the RRFSp1 solution.

Given the predictability challenges of the heavy snow rates and snowband locations in a marginal thermal environment, forecasters were generally understanding of the model to model discrepancies at the Day 2 and Day 1 lead times during this case, preferring the models with the most accurate 1 inch footprint area. Some representative comments sum up their subjective impressions of the model guidance during evaluation. One participant said, "Overall, most of the models had a decent handle on the location and coverage of the >1" snowfall. However, all models poorly handled both the location and coverage of the maximum snowfall amounts.". Another offered perspective on the RRFSp1 and Utah ML overpredictions for messaging, writing "the RRFSp1 did a bit better perhaps with the general location of 1"+ snowfall. Combined, they were useful and probably all that was needed for the event. I also felt the [Utah] deterministic was useful for worst case scenario forecasting and estimating a high end amount as even though it was overdone for the max amounts, the footprint of 1"+ was still useful."

Across the intensive week WWE cases, forecasters generally found relatively small differences between the Utah ML and raw RRFS snowfall forecasts, despite occasionally large differences in instantaneous SLR in certain cases. The Utah ML was often competitive with or slightly worse than RRFS overall. When the Utah ML snowfall performed well, participants found it was more accurate and useful for forecasting snowfall footprint amounts rather than the snowfall maximum contour, since the Utah ML method tended to be somewhat biased high on max snowfall amounts. Knowing that the model was trained on data from the Western U.S., participants expressed some hesitation to trust the Utah ML forecasts for intensive week cases in the central and eastern regions where more mixed precipitation types were present, noting some biases. One representative quote mentioned "I would really like to know more about the application of the Utah MLP SLR with more cases. SLRs can be critical to getting a forecast right, and the central US case ... really showed the potential value it could have for mixed precipitation events. There were certainly times where there were odd signals over terrain and within mixed precipitation zones. I'd like to see more cases before increasing my confidence."

5.5.2 Freezing Rain and FZRA Post-Processing

During the intensive weeks, we were able to discuss and compare quantitative freezing rain forecasts from HRRR as well as the RRFS members, which all had the HRRR freezing rain algorithm implemented in their output (Benjamin and Collaborators, 2021). In addition, the REFS had new freezing rain QPF probability forecasts which we examined during each of the intensive week freezing rain cases (the HREF has no corresponding probability field).

In a Western U.S. case from January 14, 2024, we discussed excessive freezing rain probabilities in the REFS over the higher elevations of the mountain west, as shown in Figure 30. With light reflectivity in the RRFSp1 forecast, forecast soundings from the RRFSp1 and other REFS members in northern Nevada (WMC) did not seem to indicate classic freezing rain profiles, unless there was freezing drizzle over the higher mountain peaks. P-type algorithms categorized these profiles as snow, not freezing rain. Given this setup, REFS ensemble probabilities of categorical freezing rain seemed excessively high. These freezing rain probabilities also were suspect because participants noted that the mountain west rarely receives freezing rain per climatology (e.g., McCray et al., 2019).

While not an intensive week case, another highly impactful mixed precipitation event from 23-24 March 2024 illustrates the overall utility of the REFS freezing rain QPF probability fields. Figures 31-32 show two different REFS forecasts 24 h apart, to depict how the probabilities evolved with decreasing lead time.

Overall, the REFS forecast probabilities generally capture the spatial pattern of the FRANA observed ice accumulations. REFS forecasts from 21 March (Fig. 31) have high probabilities ex-

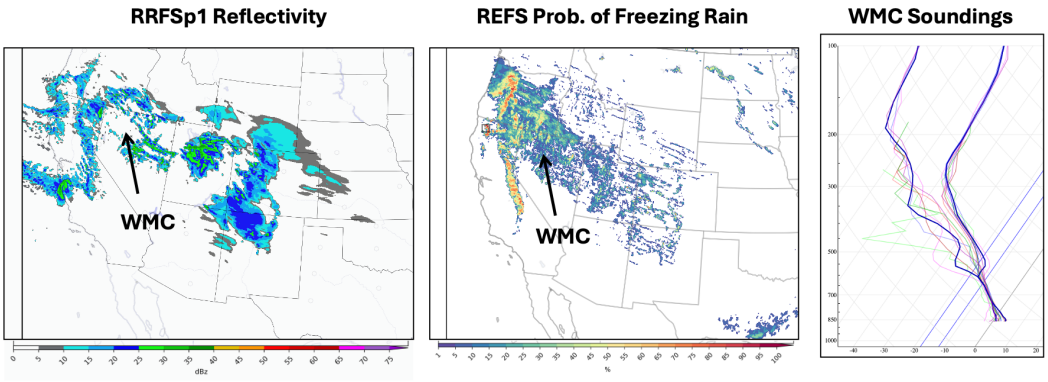


Figure 30: Example of light freezing rain accumulations, from REFS forecasts initialized 00 UTC 13 January 2024, valid at 12 UTC 14 January (f36). (left) RRFSp1 composite reflectivity. (center) REFS categorical probabilities of freezing rain. (right) Forecast soundings at Winnemucca, Nevada (WMC), with RRFSp1 highlighted in blue.

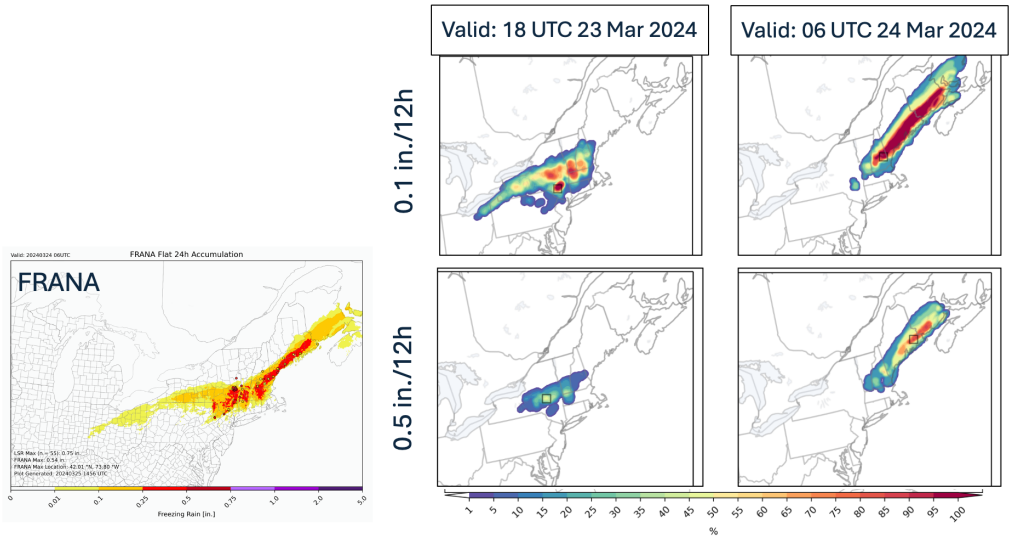


Figure 31: REFS forecasts of 12 h freezing rain QPF probabilities, model initialized at 18 UTC 21 March 2024. Note that FRANA observation for 24 h flat ice accumulations are valid ending at 06 UTC 24 March 2024, spanning the two 12 h forecast periods from the REFS guidance.

ceeding 60-70% of 0.1 inch in 12 h over the region impacted, from New York through Maine, with moderate probabilities of 0.5 inch (especially in Maine). A day later, both thresholds of probabilities have increased in magnitude as forecast lead time and ensemble spread decrease. REFS probability forecasts of freezing rain exhibited a small northward bias at both lead times, since the highest observed freezing rain totals estimates and LSR observations were across the region from Albany, NY, to Portland, ME. Magnitudes of higher-end freezing rain probabilities (0.5 inch in 12 h) were also relatively high, but considering that freezing rain QPF is typically an overforecast compared

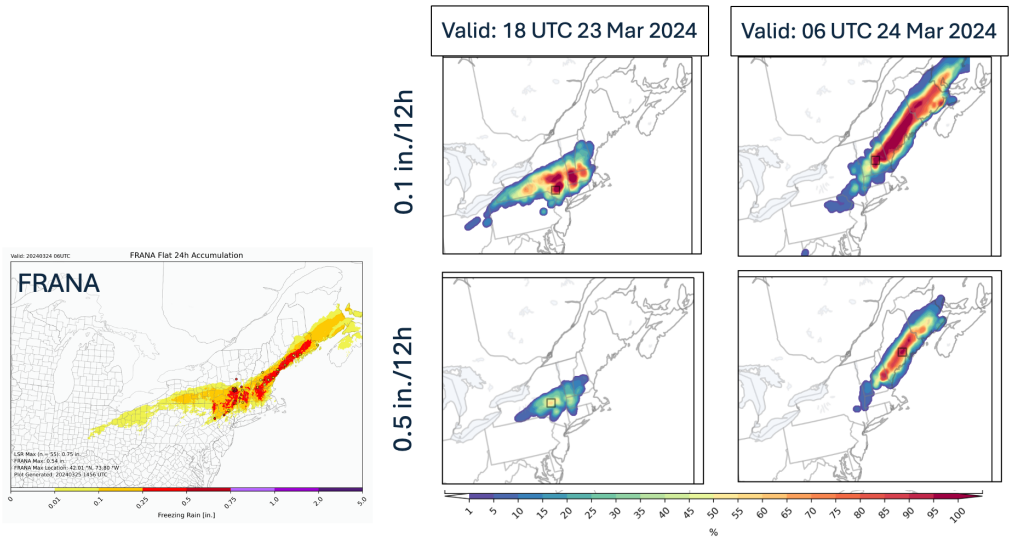


Figure 32: As in Figure 31, but for REFS forecasts initialized at 18 UTC 22 March 2024.

to what accretes on surfaces (Sanders and Barjenbruch, 2016), forecast probabilities may not have been as high-biased as they appear. Given that NWS WFOs and WPC create freezing rain forecasts based on flat ice accumulations, HMT recommends that FRAM post-processed probabilities from REFS should match that of operational outputs and would be of even greater value to operations, perhaps in place of the raw freezing rain QPF probabilities that were examined during last winter's WWE.

Discussions during the WWE intensives also focused on freezing rain observations and the FRANA dataset. After evaluating FRANA for one freezing rain case during each intensive week, participants were asked to summarize their impressions of FRANA in an open-ended question in a survey sent to the participants after the intensive week ended. While the intensive week was the first time many participants had examined FRANA, several people mentioned that their assessment was based on more than one case since they had been able to assess FRANA in an event impacting their local area. Figure 33 summarizes subjective assessments of FRANA performance and operational utility from the participants, where the survey asked people to rate FRANA based on its potential use as an observational dataset and nowcasting tool (the latter being a thought experiment, as HMT did not have a nowcasting activity using FRANA during WWE). Overall, participant feedback was primarily positive regarding FRANA performance as an observational dataset during the WWE cases, with about a third of the participants assessing that FRANA was ready for operational use alongside other data sources, albeit with some need for additional calibration and modest development (focusing on addressing the high bias found in FRANA for higher freezing rain amounts, compared to freezing rain LSRs).

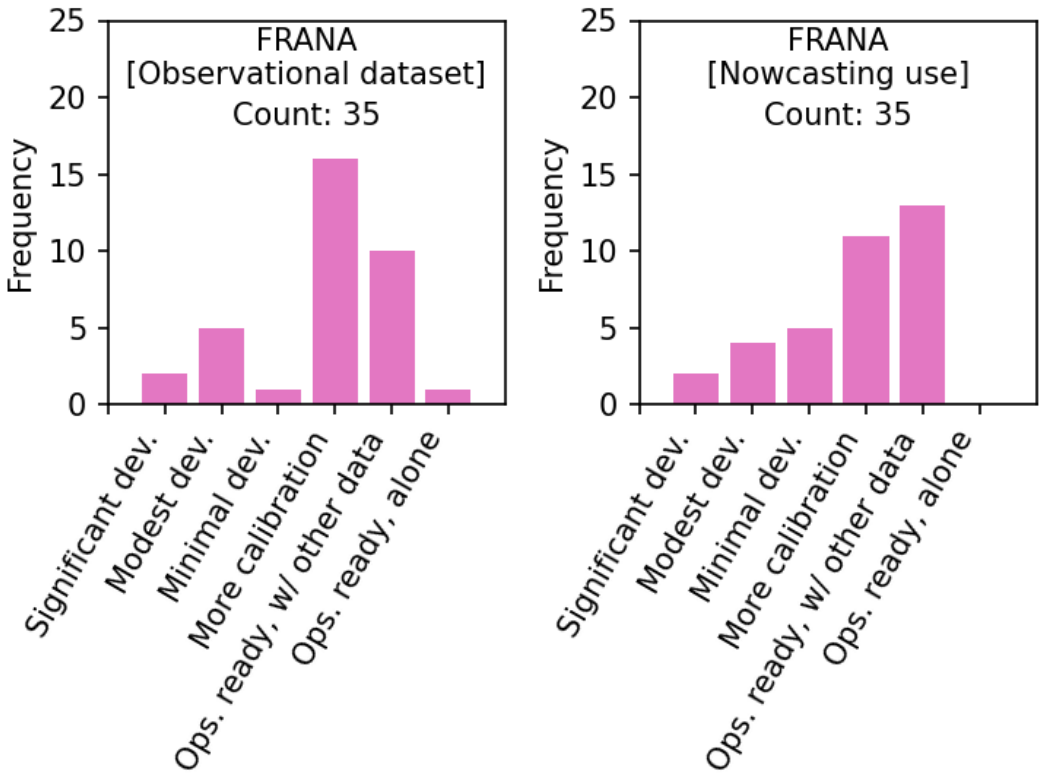


Figure 33: Subjective assessments of overall FRANA quality from end-of-week surveys sent to participants.

Participants were able to expand on their subjective ratings in the survey, describing some of the positive and negative qualities of FRANA in short-answer responses. Overall, participants thought FRANA would be useful alongside other data, such as ASOS and LSR freezing rain amounts, with one person stating, "FRANA did very well with the depiction of ice... on our case. It nailed the area, and generally got amounts correct. Personally, I would put this into operations along with other data. Even if it needs more calibration/development, it's a tool that we've never had, but wanted for some time. Admittedly, I'm providing input based on one case, but I think getting eyes on it on a regular basis will hasten any improvements it may need.". Other remarks focused on the need for additional calibration to improve noted biases seen in individual cases, for example, "Across many locations, I remember there being a large difference between the FRANA compared to observations, with a large high bias or a large low bias so it is hard to conclude which one was more prevalent. There would definitely need to be more case studies done to find why these biases occur.". Finally, participants were more hesitant to trust FRANA in certain regions with sparse radar coverage or beam blockage such as the mountainous western U.S. These factors included both the lack of freezing rain observations in mountainous areas as well as uncertainty in the MRMS-derived precipitation,

with one participant stating, "FRANA performed well in locations with adequate radar coverage and ground truth. I don't see this working well in areas where one or both of those are missing because that has been true for other MRMS products. I do think as an observational dataset it has enough utility to begin use outside of a lab/development setting, but it will need improvements before it works over a wider domain...".

5.5.3 Model Soundings

The development efforts that HMT staff worked on for the sounding viewer tool paid off in several ways, including the site-specific precipitation onset timing and duration activity discussed previously. Figure 34 shows a composite summary of the subjective participant evaluation results of their own performance at diagnosing wintry precipitation onset time from their Day 2 and Day 1 forecasts at three individual sites, where they used MRMS and time series of ASOS observations to verify. Even considering the subjectivity of these results, participants indicated that they were able to determine the winter precipitation onset time within an hour over 40% of the time, with a roughly symmetric distribution around the "less than 1 hour offset" bin. The second-most common result was an assessment that "no winter precipitation" fell at a site. This result was expected because the HMT team intentionally picked some sites outside of the winter precipitation area in the MSTP exercise, to make participants think critically about whether the profiles matched expectations of winter precipitation (or any precipitation) in the model guidance.

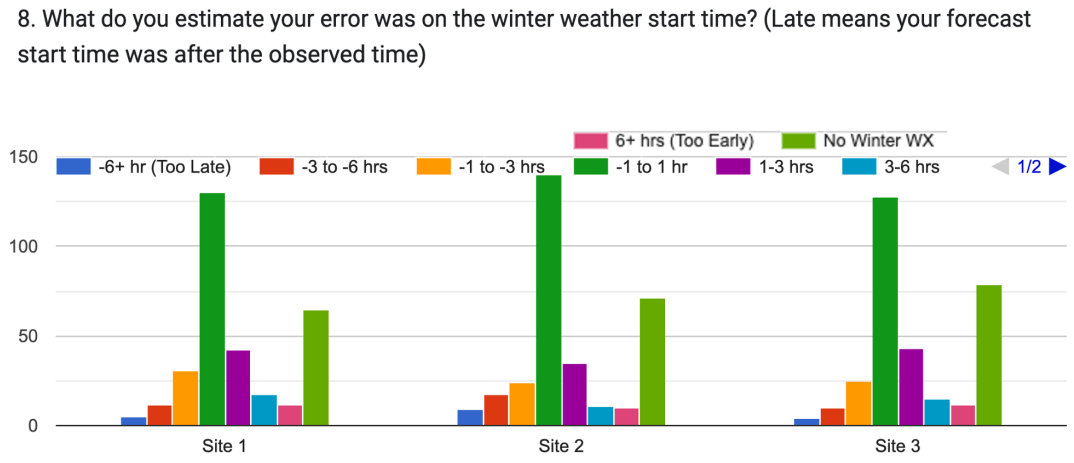


Figure 34: Combined results from the Day 1 and Day 2 winter precipitation onset timing verification activity, where participants subjectively rated their forecast onset times at each of their three sites. Negative values indicate a forecast onset time later than the observed onset time, while positive values indicate a forecast onset time earlier than the observed onset time.

In addition, we diagnosed several model physics oddities among the REFS members since EMC provided sounding profile data from each member starting in January 2024. One consistent example

was a surface-based nocturnal cold bias in two of the REFS members (RRFSm2 and RRFSm5), which we noted happened only where snow was already on the ground or after snow had fallen during the model forecast. We hypothesize that RRFSm2 and RRFSm5 (see Table 2), the only two members that used the GFS planetary boundary layer and surface layer (PBL/SL) land surface model scheme during the winter of 2023-2024, suffered from the cold bias over snowpack as a result of using that particular PBL/SL scheme. An example of the cold bias is shown in Figure 35, where forecast soundings at Caribou, Maine, over existing snowpack were about 10 °C too cold in the lowest 50 hPa relative to the observed sounding on 1 February. HMT communicated this feedback to EMC by April 2024 and they replied that they were working on a fix.

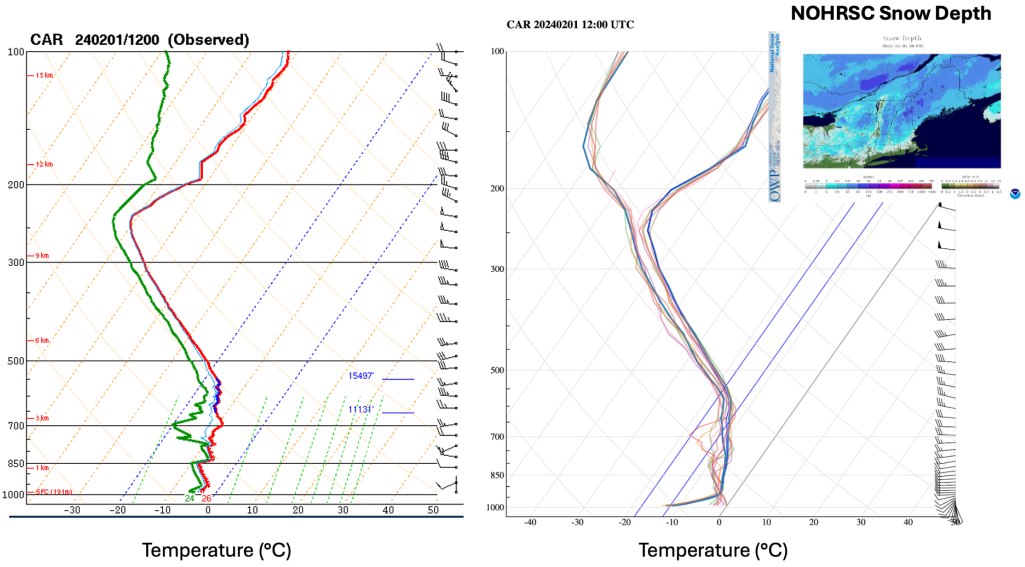


Figure 35: Observed (left) and REFS forecast sounding profiles (right) at Caribou, Maine (CAR), valid at 1200 UTC 1 February 2024. RRFSm2 (blue, highlighted) as well as RRFSm5 (orange) forecast profiles are the ones with a near-surface cold layer. REFS forecasts were from the cycle initialized 00 UTC 31 January 2024. NOHRSC snow depth valid at the same time is shown as an inset panel.

Other uses of the sounding viewer were to integrate past testbed research products such as the Spectral Bin Classifier into the sounding processing routines, for display of detailed p-type information during the WWE intensive weeks. We had valuable discussion about different methods of diagnosing p-type, their utility, and some new ideas for visualizing p-type uncertainty, all in the context of evaluating ongoing products under development like FRANA. In the future, HMT plans to continue developing the sounding viewer displays to integrate additional quantities such as SLR to facilitate discussion of ongoing research products. Figure 36 shows an example of the sounding viewer during the mid-February 2024 event discussed in the case study section, that we used to better understand SLR forecasting challenges.

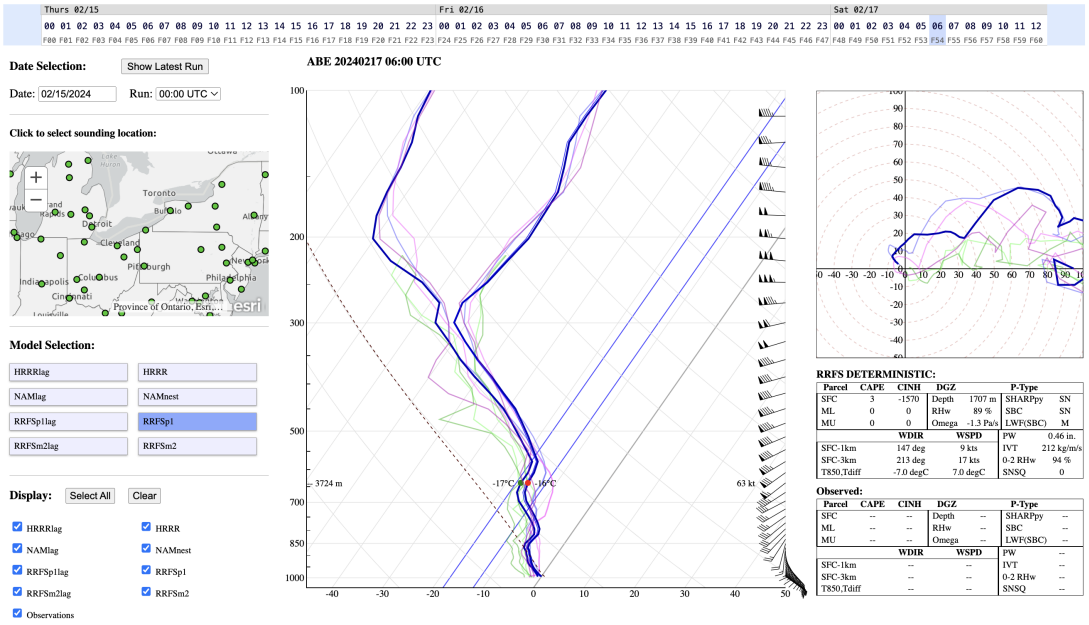


Figure 36: Forecast soundings at Allentown, PA, from HRRR (pink), NAMnest (green), and RRFSp1 (blue) forecasts initialized at 00 UTC 15 Feb 2024, all valid at 06 UTC 17 Feb 2024 (f54).

Plan view maps of dendritic growth zone (DGZ) depth were called into question for a 15 February 2024 case since the RRFSp1 had multiple layers where temperatures crossed the -12C mark (Figure 37). The DGZ depth algorithm appears to assume only 1 possible layer, identifying a first occurrence at or below -12°C and a final occurrence where the temperature profile falls below -16°C. The warmer layer, centered near 700 hPa in the example, is thus counted as part of the depth. A more useful algorithm would potentially need to be designed to capture multiple DGZs and either sum the layers, or only count those layers which proceed to be colder than -16°C. This could make the DGZ depth a more reliable variable in understanding the DGZ and its role in potentially enhancing SLR and thus snowfall.

5.5.4 Ensemble Consensus Fields - Neighborhood Probabilities and Probability-Matched Mean (PMM/LPMM)

During the experiment, we often discussed the utility and value of ensemble probabilities. This often included gathering feedback on the radius of influence used in the EMC-produced REFS ensemble probabilities. Figure 38 shows an example of the neighborhood probabilities of 1 inch and 6 inches of variable density snow in 6 hours, from REFS forecasts initialized at 12 UTC 16 January 2024. The EMC-produced probabilities for the winter used an 18 km radius of influence to perform the smoothing. In this case, the neighborhood probabilities highlighted the potential for heavy snow

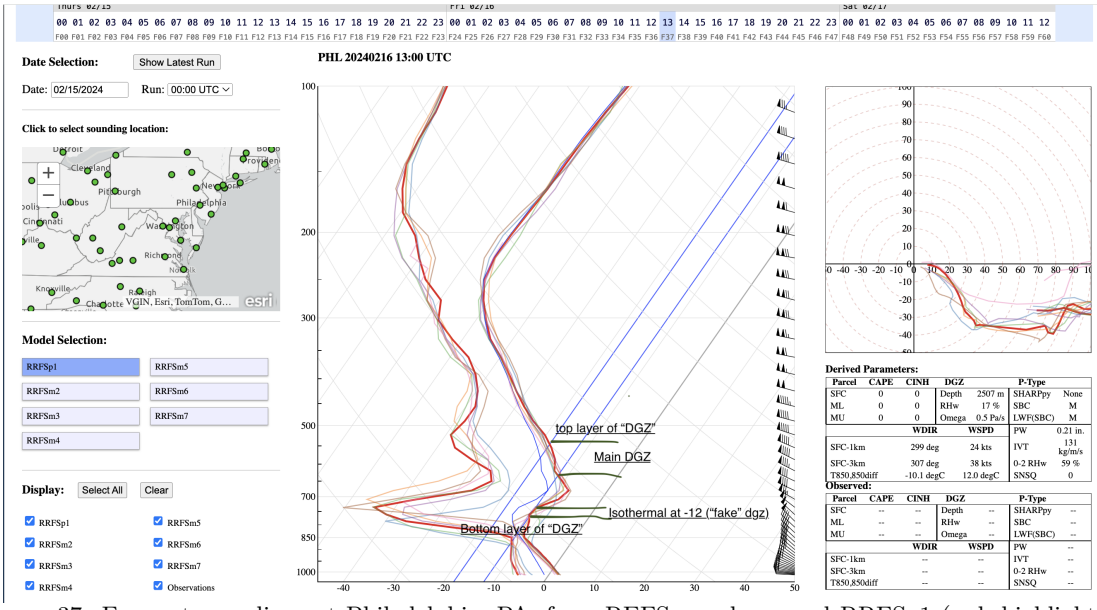


Figure 37: Forecast soundings at Philadelphia, PA, from REFS members, and RRFSp1 (red, highlighted) forecasts initialized at 00 UTC 15 Feb 2024, all valid at 13 UTC 16 Feb 2024 (f37).

across the higher terrain of the Pacific Northwest, but showed high probabilities of 1 inch of snow for most of the region including the lower elevations of interior Oregon and Washington.

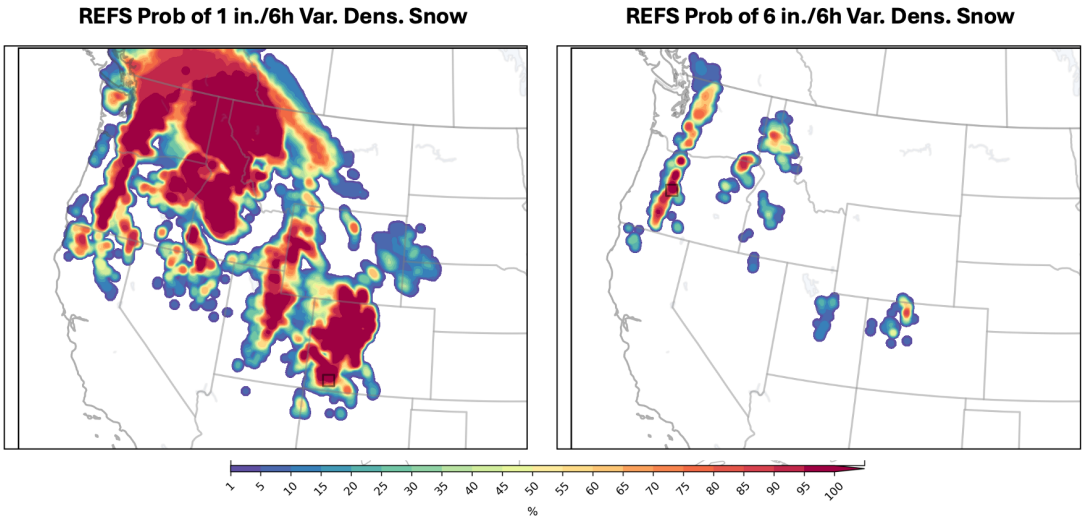


Figure 38: Neighborhood probabilities of (left) 1 inch and (right) 6 inches of variable density snow in 6 hours, from REFS forecasts initialized at 12 UTC 16 January 2024 valid at 00 UTC 18 January (f36).

Overall, some participants noted that it was more difficult to use the probability fields over the Western U.S. since any smoothing makes it difficult to account for impacts in the complex terrain relative to valley locations. We discussed whether other approaches would be useful for

visualization, such as unsmoothed gridpoint probabilities or having more than one set of probability fields (i.e., having at least two different radius of influence smoothing options) for mountainous and non-mountainous areas. Other comments discussed the utility of viewing ensemble information in complementary ways for forecasting and IDSS messaging, such as time series diagrams and probability distributions for individual locations.

Overall, the majority of forecasters enjoyed viewing the REFS probabilities for variable density snowfall and freezing rain, alongside the guidance from the deterministic forecasts. One forecaster's summary describes the tone of the positive feedback: "During the experiment, I primarily looked at the variable density snowfall probabilities and was satisfied with them as a whole. I would use these in an operational setting. I also found the freezing rain probabilities to be extremely valuable when evaluating the freezing rain impacts in the forecast." The participants found more utility in using the ensemble probabilities as a guide for drawing footprints of snowfall (1 inch) and freezing rain (0.01 inch) during the MSTP activity, while relying on other guidance to key in on higher precipitation amounts for each intensive case. Deterministic forecasts and consensus products, such as REFS variable density snow PMM, were more generally used by participants to investigate maximum amounts, as discussed previously in the February 17 case (Section 5.2.2).

Participants also had the opportunity to investigate ensemble probability and consensus products from the CAPS FV3-LAM ensemble, using them for the MSTP activities and evaluating them during the verification sessions. Feedback on the CAPS consensus products was mixed, as many of the techniques ended up producing broadly similar 6h forecasts to each other in the cases we focused on during the intensive weeks (e.g., Fig. 21). Of the techniques examined (arithmetic mean, LPM, and SAM), participants generally favored consensus products using the LPM since that method tended to best preserve snowfall maxima as seen in the individual deterministic forecasts.

5.5.5 Visibility Fields with Blowing Snow Parameterization

During the winter, we learned that EMC implemented an experimental blowing snow parameterization into the RRFS member visibility field. In discussion with forecasters, they were interested to learn more about the visibility field between RRFS (with blowing snow) versus HRRR (no blowing snow reduction). Generally, forecasters were interested in examining the RRFS visibility field but suggested that visibility could be separated into two outputs, one with and one without the blowing snow visibility reduction component.

In a case from 14 January 2024, the overall differences between the HRRR and RRFSp1 visibility fields were evident, as strong winds occurred behind a cold front associated with a deepening surface low pressure system across the Great Lakes. Figure 39 summarizes the differences with the inclusion of the blowing snow parameterization in RRFS, where larger visibility reductions in

RRFSp1 occur due to both blowing snow across the Northern Plains and interior Northeast and in lake-effect snow bands across Michigan.

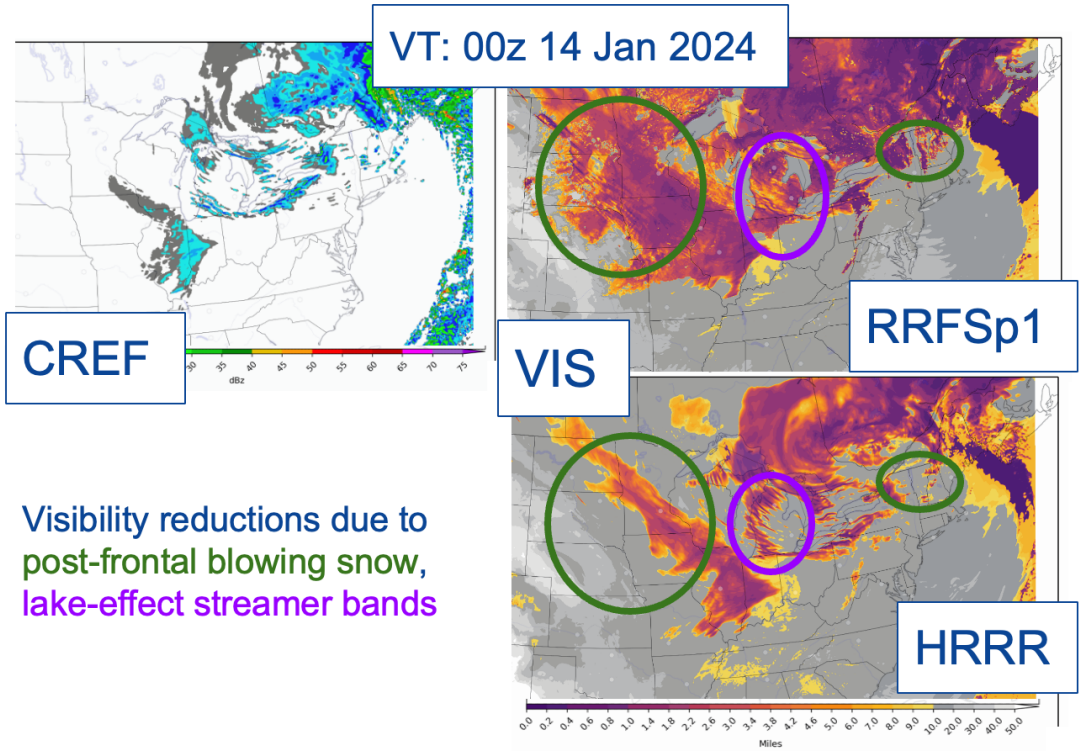


Figure 39: Composite reflectivity and visibility comparison from RRFSp1 (top) and HRRR (bottom), from forecasts initialized at 12 UTC 13 January 2024, valid at 00 UTC 14 January (f12).

5.5.6 Accuracy and Representativeness of Winter Weather Observations

As in past WWEs, each week we had good discussions about the accuracy and potential biases of gridded observational "best guess" datasets such as NOHRSC. For certain cases in the sparsely populated Western U.S., it was acknowledged that fewer observations were likely assimilated into the analysis so that NOHRSC relied more heavily on model nowcast fields from the HRRR to generate snowfall analyses. Thus, there is always some concern among participants in snowfall verification that we are verifying a model (e.g., HRRR, RRFSp1) with a model-based analysis (NOHRSC). For example, Figure 40 shows a case examined during WWE from 12 UTC 18 January 2024 where participants were concerned about NOHRSC estimates relying on relatively sparse observations across western Montana, such as those shown by the CoCoRaHS reports map. Despite these concerns, observational analysis datasets are an extremely useful resource for event review and verification at a regional or national scale, instead of having to parse through individual point reports of snowfall.

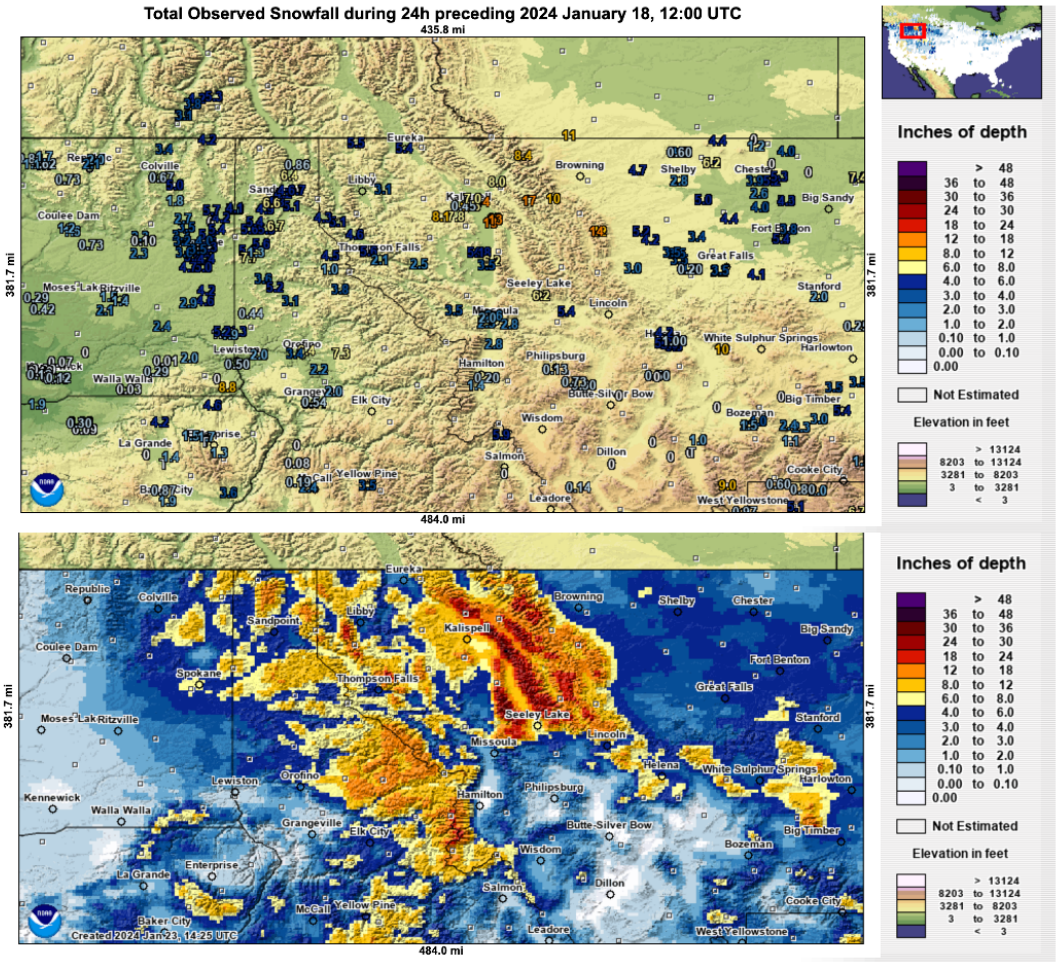


Figure 40: Point 24h snowfall reports (top) versus NOHRSC analysis, for the 24h ending at 12 UTC 18 January 2024.

Other topical discussions about winter weather observations centered around local storm reports (LSRs), particularly those for freezing rain. Currently, there are differing practices among NWS offices in reporting freezing rain and snow amounts. While many reports focus on "flat ice" amounts, sometimes reporting practices are ambiguous between reporting "flat" or "radial" ice amounts unless specifically described in the LSR comments. For snow LSRs, NWS offices have different philosophies. In one WFO they do not report if the snow is less than 2 inches (week 2). There were additional questions about whether NWS Cooperative Observer Program (COOP) sites or CoCoRaHS data make it into NOHRSC. These data questions continue to be asked by forecasters, and we would recommend increasing the transparency of which reports make it into various NOHRSC analyses. Some conference presentations exist on this topic (e.g., Carbin et al., 2020) but the NOHRSC product workflow does not appear to be well-known in the field. There were

many indications that participants had doubts about the quality of the NOHRSC in the western U.S. for a few tested case studies. The use of point observations with varying time and durations may mean that valid reports need to be standardized to maximize their use outside of PNS reports. One of the additional difficulties with LSRs that HMT has is that LSRs do not have standard categories for duration. This makes use of the LSRs difficult in verification. While NWS uses event total, for the purposes of verification, some of these observations can not be used in sub-event contexts, lowering their overall use and utility.

6 Summary and Recommendations

In this section, we provide an overall summary of the primary datasets evaluated during this year's WWE and our recommendations for future research and development efforts.

6.1 RRFS/REFS

Overall, we were encouraged by the new winter capabilities demonstrated in the REFS, such as probability-matching ensemble consensus fields for snow, individual member freezing rain QPF, and ensemble probabilities for freezing rain. However, more ensemble fields are needed to realize substantial benefits for WPC operational products and HMT activities, most importantly, Freezing Rain Accumulation Model (FRAM, Sanders and Barjenbruch, 2016) post-processed output and FRAM probabilities. From the limited sample of winter weather cases we evaluated, RRFSp1 (RRFS deterministic) model results seem comparable to HRRR for variable density snowfall. Given ongoing changes to the RRFS configuration during and after the 2023-2024 WWE, light precipitation biases noted in the 2024 FFaIR experiment (following this experiment's end) could potentially be an issue for winter as well, which needs further investigation. Our recommendations (as of the March 2024 version of the RRFS/REFS that we evaluated) are that more research and development are needed.

6.2 CAPS Ensemble and Consensus Products

Overall, HMT had some challenges in evaluating the CAPS ensemble dataset and consensus products, due to delays in when the data was provided relative to the intensive weeks. However, we were able to generate discussion of the CAPS products during the latter two intensives. The CAPS deterministic member we chose to focus on in our snowfall evaluations performed fairly well at Day 2 lead times relative to the HRRR and REFS members, but struggled at Day 1. We hypothesize this may be due in part to the lack of data assimilation in the GEFS-initialized CAPS ensemble members, unlike the HRRR and REFS forecasts, which was more detrimental at the short Day 1 lead time for CAPS. For the ensemble consensus products, participants had mixed feedback about the new

spatially aligned mean (SAM) technique compared to other methods such as the local probability matched mean (LPM), where the LPM tended to retain higher snowfall maxima that were needed in the MSTP forecast process. We speculate that SAM had a relatively small impact on the forecast output, since winter precipitation features are in greater alignment in winter (relative to summertime deep convection), so there were not as many adjustments made by the SAM method to the CAPS ensemble during the WWE cases. Participants generally indicated they preferred the consensus products that included the LPM technique. Our recommendation is for the CAPS ensemble and consensus products to continue research and development.

6.3 Utah SLR Post-Processing

Overall, the University of Utah machine learning snow-to-liquid ratio (SLR) and snowfall products showed promise, and we appreciated the efforts of the research team to successfully provide a full-CONUS dataset this past winter season. However, several persistent biases were noted across the cases examined during this year's WWE. One such issue was streakiness in SLR values, also affecting snowfall, in marginal p-type cases along the boundaries of the snowfall footprint (the Utah research team identified this as a bug in their methods in the post-season). We also noted a consistent high bias in maximum snow amounts across multiple cases, suggesting that the ML methods struggled to keep SLRs low enough in marginal thermal cases with partial melting of accumulating snow. Finally, the ML methods had trouble increasing forecast SLRs enough in extreme high-SLR outlier cases (e.g., 17 February 2024 in PA/NJ), so we encourage further research to better leverage model fields suggesting potential for extreme SLRs. Our recommendation is for this project to continue additional research and development.

6.4 FRANA

Overall, we thought the first CONUS-wide demonstration of FRANA for the 2023-2024 winter season was successful, with the development team demonstrating it can run in a near-real time, quasi-operational environment within the MRMS product suite. Building on the efforts of the Spectral Bin Classifier (SBC), a product previously tested in past WWEs, FRANA and SBC showed great utility, accurately describing p-type across many cases and regions. While we noted some issues with FRANA's calibration, namely an overestimate of maximum freezing rain amounts and errors in the freezing rain footprint in certain cases, the CIWRO/NSSL team has actively pursued improvements to the FRANA algorithms during the offseason. The HMT team believes FRANA development would be further accelerated by greater standardization of freezing rain local storm reports within the NWS (i.e., all- "flat" or all-"radial" reports). FRANA was well-reviewed by participants in the WWE intensive weeks and focus groups, and the analysis fills a big gap in winter observational capabilities for freezing rain. If FRANA continues towards operations, the HMT team anticipates

FRANA will have a large positive impact for both research and operations, as a highly-valuable dataset for NWP evaluation/verification of freezing rain as well as in operational settings including nowcasting during mixed precipitation events. One participant comment summarizes our views well: "Even if [FRANA] needs more calibration/development, it's a tool that we've never had, but wanted for some time... I think getting eyes on it on a regular basis will hasten any improvements it may need." Our recommendation for FRANA is that it is nearly ready for transition to operations, following further testing and evaluation during the upcoming winter season and 15th WWE.

Evaluated Dataset	Nearly ready for transition to operations	Needs further development and testing	Rejected for further testing	Provider / Funding Source
RRFS, REFS		X		EMC
FV3-LAM Ensemble		X		CAPS/OU
ML SLR, Snowfall		X		Univ. of Utah
FRANA	X			CIWRO/NSSL

Figure 41: Graphical summary of research to operations transition recommendations for the 14th WWE.

7 Other Experiment Activities

7.1 Focus Group Summaries

As part of this year's WWE, two sets of focus groups were convened to discuss winter weather hazards. The first set focused on evaluation and discussion of the Freezing Rain Analysis (FRANA) product being developed within the Multi-Radar Multi-Sensor (MRMS) system by NSSL/CIWRO. The second set of focus groups discussed winter weather forecast messaging issues related to transportation and road hazards, led by a team from CIRES/CU Boulder and CIWRO/NSSL. Focus group summaries are provided by the PIs of each respective funded project.

7.1.1 FRANA

Author: Daniel Tripp (CIWRO/NSSL)

The FRANA focus groups were each comprised of NWS forecasters from various geographic regions and were advertised to forecasters who had experience with FRANA during the 2023-2024 winter season. The discussion in the focus groups was centered around 4 main themes: (1) FRANAs' subjective performance for NWS forecasters, (2) flat versus radial ice thickness, (3) stakeholders needs of ice accumulation information and (4) FRANA footprint/accumulation errors.

For the first theme, forecasters shared examples where FRANA was highly skillful and aided NWS decision-support. Other examples were also provided of poor FRANA performance that were typically tied to precipitation-type or NWP temperature error. Overall, forecasters expressed that they would like to see verification of FRANA for events in their CWA to become more familiar with its performance. After forecasters become familiar with the product, they indicated that they are more likely to use it for nowcasting – but in the meantime are primarily using it for post-event verification.

The discussion for the second theme revealed that forecasters understand the different needs of stakeholders/public. They strive to provide decision-support using flat and radial ice to meet the needs of each group. For FRANA, forecasters indicated that they prefer to see both flat and radial grids, though flat ice seemed to be the preferred choice by the majority since that is the standard for ice verification.

The conversation about stakeholders revealed that forecasters see FRANA as having a strong utility for messaging. Forecasters shared that maps of accumulation are an easy way to communicate areas of high impact and also to solicit reports in data sparse areas. It was reported that FRANA has been used by stakeholders in crafting a federal disaster declaration for an ice storm during the 2023-2024 winter season.

In the last part of the focus groups, forecasters discussed FRANA errors and what level of tolerance they have for them. It was shared that they are familiar with footprint errors in precipitation-type forecasts and are accustomed to adjusting things over by a few counties if needed. When asked about the accumulations, forecasters shared that errors of 0.05 inches are typically trivial, errors of 0.1 inches are acceptable for some events, but that errors of 0.2 inches are unacceptable. Several made the point that events with lower storm-total accumulations come with a forecaster expectation of higher precision.

Other comments indicated that forecasters value FRANA being updated hourly in real-time so they can provide this information quickly. In addition to the rapid updates, one forecaster shared that if FRANA proves to be skillful, the product would likely be used in real-time to decide whether to upgrade to an ice storm warning or not.

7.1.2 Road Hazards

Author: Dana Tobin (CIRES/WPC)

Road hazards focus group findings are summarized in Section B of the Appendix.

8 Acknowledgments

We thank numerous people who helped make this year's WWE successful, starting with the HMT team who provided facilitator and logistical support throughout the winter and during the WWE intensive weeks. WPC staff, especially winter weather desk forecasters Josh Weiss, Bryan Jackson, and Tony Fracasso who contributed to forecast briefings and discussion during each intensive week, with additional DTB helpers assisting with note-taking and participant questions during the hybrid and virtual intensives. Ben Albright (Lynker/WPC) provided MODE verification results and graphics on an interactive webpage. Kirstin Harnos (WPC) provided an interactive REFS probability webpage. Tomer Burg (CIRES/WPC) provided an interactive ASOS p-type visualization webpage. Project PIs Daniel Tripp (CIWRO/NSSL) and Dana Tobin (CIRES/WPC) provided summaries of the focus group findings in this report. David Harrison and Israel Jirak (SPC) shared a code base and offered vital collaborations for HMT to construct its own sounding viewer for use in accomplishing key WWE objectives. Thanks to all attendees of the WWE intensive weeks and focus groups who provided valuable insight into the winter weather forecast process and challenges experienced in their individual regions across the U.S., and for their help in evaluating the products examined during this year's WWE.

A Featured Numerical Guidance and Data

WWE participants will be evaluating a variety of experimental data, which are listed in Figure 3. As in previous years' WWEs, the development of the future Rapid Refresh Forecast System (RRFS) is at the center of evaluation activities. The core of this system is a Limited Area Model (LAM) that has the finite volume cubed-sphere (FV3) dynamic core. The deterministic flagship CAM, referred to as the RRFSp1 (also "RRFSa" or "m1" in Table 2), is still in development and the model configuration is subject to change. More information about the experimental guidance being evaluated can be found in Figures 3-42, Table 2, and in the following subsections.

Name	Core	MP	PBL/SL	LSM	Conv.	Radiation	GWD	IC/LBC
m1 (CTRL)	FV3	Thompson	MYNN	RUC	G-F deep	RRTMG	GSL	RRFS hybrid/GFS
m2	FV3	Thompson*	GFS	RUC*	G-F dp*+sh	RRTMG*	GSL*	RRFS EnKFm1/GEFSm1
m3	FV3	Thompson*	MYNN*	RUC*	saSAS deep	RRTMG*	GSL*	RRFS EnKFm2/GEFSm2
m4	FV3	NSSL#	MYNN*	RUC*	G-F deep*	RRTMG*	GSL*	RRFS EnKFm3/GEFSm3
m5	FV3	NSSL#	GFS	RUC*	G-F dp*+sh	RRTMG*	GSL*	RRFS EnKFm4/GEFSm4
m6	FV3	NSSL#	MYNN*	RUC*	saSAS deep	RRTMG*	GSL*	RRFS EnKFm5/GEFSm5
m7 (HRRR)	ARW	Thompson	MYNN	RUC	None	RRTMG	GSL	HRRRDAS/RAP

Table 2: EMC RRFS ensemble design. Schemes denoted by an asterisk are those running with stochastic parameter perturbations (SPP). Schemes denoted by a pound sign are those running with fixed parameter perturbations. Members 8-14 (not shown) are the same as members 1-7 except they are from 6-h old cycles (i.e., time lagged).

A.1 EMC - RRFS Ensemble

The RRFS is a rapidly-updated, high-resolution (3 km) ensemble forecast system that runs over a domain covering North America. Its ensemble includes multiple physics schemes, stochastic parameter perturbations, time lagging, initial/lateral boundary condition diversity, and two time-lagged (TL) members from the operational HRRR system (see Table 2). Deterministic and ensemble forecasts are provided to 60 hours four times per day at 00, 06, 12, and 18 UTC. Deterministic forecasts are provided to 18 hours at all other times.

Ensembles are initialized using 3-km ensemble perturbations drawn directly from the RRFS Data Assimilation System's (RDAS) ensemble Kalman filter analysis members. The control member forecast is initialized from the hybrid 3DEnVar analysis. The RDAS uses a wide variety of conventional observations along with radar reflectivity. It also includes a nonvariational cloud analysis. Note that owing to different forecast horizons among membership the ensemble size will decrease as a function of forecast lead time. This means that the full 14 members will be available up to forecast hour 42 when the HRRR-TL will drop out, 13 members until hour 48 when the on-time HRRR drops out, 12 members to hour 54 when the RRFS-TL members drop out, and 6 members to hour 60.

Disclaimer: The RRFS will be under active development through most of the WWE period and may undergo changes to its underlying scientific package that impact results.

A.2 OU CAPS - FV3-LAM Ensemble and ML Post-processing

A.2.1 Ensemble Forecasts

For the 14$^{\text{th}}$ WWE, CAPS will run an 11-member FV3-LAM ensemble on Frontera at the Texas Advanced Computing Center at 00 UTC for selected winter storm days. The members will be 3-km grid-spacing (CAM resolution) FV3-LAM models over the Contiguous United States (CONUS) with various physics configurations paralleling planned members of the future Rapid Refresh Forecast System (RRFS), including those similar to the present High Resolution Rapid Refresh (HRRR), the

Global Forecast System (GFS), FV3 version of Warn on Forecast System (WoFS) and Hurricane Analysis and Forecast System (HAFS) as shown in Figure 42. Forecasts will be run out to 84 h (3.5 days) and will use initial and boundary conditions from the operational GFS and GEFS. A recent code release of the FV3 Short Range Weather (SRW) App will be used.

Experiment	Microphysics	PBL	Surface	LSM	IC/LBC (like system)	AI member
GFS IC for Baseline Configuration						
M0B0L0_P	Thompson	MYNN	MYNN	NOAH	GFS /GFS	AI-1
M1B0L0_P	NSSL	MYNN	MYNN	NOAH	GFS/GFS (WoF)	AI-2
M0B0L2_P	Thompson	MYNN	MYNN	RUC	GFS/GFS (RRFSm1)	
M1B2L2_P	NSSL	TKE-EDMF	GFS	RUC	GFS/GFS (RRFSmphys8)	
M0B2L1_P	Thompson	TKE-EDMF	GFS	NOAHMP	GFS/GFS (GFSv16)	AI-3
Physics + IC Perturbation Ensemble						
M0B1L0_PI	Thompson	Shin-Hong	GFS	NOAH	GEFS_m1	
M0B2L1_PI	Thompson	TKE-EDMF	GFS	NOAHMP	GEFS_m2	
M0B2L2_PI	Thompson	TKE-EDMF	GFS	RUC	GEFS_m3	AI-4
M1B1L0_PI	NSSL	Shin-Hong	GFS	NOAH	GEFS_m4	
M1B2L1_PI	NSSL	TKE-EDMF	GFS	NOAHMP	GEFS_m5	
M1B2L2_PI	NSSL	TKE-EDMF	GFS	RUC	GEFS_m6	

Figure 42: Planned CAPS FV3-LAM ensemble member configurations for this year's WWE.

Days to run will be decided in consultation with the WPC WWE Team and other collaborators. CAPS plans to be prepared to run cases by mid-December 2023 and can run cases until mid-March 2024, depending on the cases selected for WWE. Cases prior to December 1 may still be run as needed as initial and boundary condition data on AWS can be accessed after-the-fact. Up to 30 forecast days will be run in order to provide sufficient case days for objective scoring and for training of machine learning algorithms.

A.2.2 Ensemble Consensus Post-Processing and Combination with HREF

Ensemble Mean, Local Probability Matched-Mean (LPM mean), and Spatial Aligned LPM (SAM-LPM) precipitation and snowfall (10 times snow-liquid equivalent precipitation) at 6h and 24h (12z-12z) intervals will be provided for the CAPS CAM ensemble. The Spatial Aligned Mean (SAM) algorithm has recently been upgraded to a 2-stage process to account for forecast displacements at the synoptic and meso-alpha scales. The SAM will be run to align all members to a common position as determined by an initial application of PM mean. This configuration has produced better

structure in 3- and 6-h rainfall fields than other approaches and that we expect will apply to the winter period as well.

A.2.3 Machine Learning Calibrated Ensemble Forecast

A machine learning (ML) algorithm (a convolutional neural network deep learning system known as U-Net) will be trained using four selected members of the CAPS FV3-LAM ensemble plus HREF ensemble forecasts from the past three winters to predict snowfall at 6h intervals as verified by the NOHRSC snowfall analyses. Algorithms will be trained independently for each member and the ML forecasts for each member will be combined into probabilistic forecasts (Neighborhood Ensemble Probability and Neighborhood Maximum Ensemble Probability) of 6-h snowfall accumulation exceeding 1, 2, and 3 inches. Forecast lead times for these are 6-36 h due to the use of time-lagged HREF members.

Since last year, the ML algorithm has been updated and re-trained to include derived fields among its inputs; these include moisture convergence, upslope/downslope wind components, high-resolution terrain height, and sub-grid-scale terrain variation. These derived fields are used in addition to the model state variables that had previously been employed.

CAPS forecast results will be posted to the web on their real-time weather page and shared with the WPC for use in the WWE. Some experimental products that are not yet ready for forecaster evaluation will be included on the website, scored objectively and possibly retuned before releasing to the WWE next year.

A.3 Univ. of Utah - ML SLR

The UU group has developed an SLR algorithm trained on high-quality observations from 12 sites across the Western U.S. where snow safety professionals manually measure snowfall and liquid-equivalent on a snow board 1 or 2 times daily. Statistics show that, over the Western U.S., the UU algorithm outperforms the SLR techniques currently used, including MaxTaloft, Kuchera, and Cobb. The next version of the algorithm will be trained on carefully-quality-controlled SLR observations from the entire CONUS, and will therefore have increased skill for CONUS-wide forecasts. This algorithm is currently being applied to output from the GFS and HRRR. It is also being applied to the RRFS ensemble, with probabilistic SLR and snowfall products generated. Another area of focus for development on the UU algorithm is precipitation-type recognition. For the RRFS and HRRR, UU has implemented a precipitation-type algorithm from Birk et al. (2021), an improvement on the Bourgouin method, so that snowfall forecasts avoid conflating freezing rain QPF with snowfall.

A.4 NSSL/CIWRO - FRANA

The FRANA is a gridded analysis-of-record for ice accumulations across the CONUS. FRANA is derived from the Freezing Rain Accumulation Model (FRAM; Sanders and Barjenbruch, 2016), Multi-Radar Multi-Sensor (MRMS) Pass 1 QPE (MSQPE; Martinaitis et al., 2020), and the surface hydrometeor phase from the Spectral Bin Classifier (SBC; Reeves et al., 2022). FRAM requires wet bulb temperature, wind, and a precipitation rate to calculate an ice accumulation. The HRRR wind/wetbulb is used along with the MSQPE to calculate measurable ice accumulations. MRMS base reflectivity and the SBC precipitation type (ptype) are also used to determine where freezing rain is falling and to identify trace ice. FRANA ice accumulations update hourly with new MSQPE, SBC, and HRRR data.

B Road Hazards Focus Groups

Three focus groups were conducted in partial fulfillment of deliverable 2. These were done as an embedded activity in the WPC's Winter Weather Experiment. The purpose of the focus groups was to allow the CIWRO and CIRES researchers to better understand how road hazards are anticipated and communicated by NWS forecasters. Because this hasn't been specifically studied before, we didn't have a jumping off point for research and had limited appreciation of the different ways road weather may be forecast and messaged. Therefore, we did not use a standard set of questions for each focus group, but rather adapted questions in the wake of each focus group to cover topics that we felt were missing in the previous discussion. But all three focus groups followed a similar structure. All had about 45 minutes of group discussion followed by a 30-minute jam board activity and a 15-minute wrap up.

The responses to questions sometimes drifted away from the original question as forecasters centered about certain themes. Therefore, a thematic analysis is performed on the responses to all of the questions to highlight important findings from these focus groups rather than present a question-by-question analysis.

B.1 Theme 1: Kinds of weather and specific challenges presented by each

Forecasters often reverted to talking about specific winter-weather phenomena and the particular challenges presented by these as described below.

- *Freezing rain/freezing drizzle (FZRA/FZDZ)*: This was cited by several participants as the most impactful form of weather in their CWA. Sometimes FZRA and FZDZ were treated as different phenomena, perhaps because one is more likely to result in black ice and, therefore,

has different impacts. The reason FZRA/FZDZ is impactful is self-evident, but participants elaborated noting that "all you need is one icy spot to cause accidents." Some stated that "gotcha" events with ice are a challenge defining these as a little ice for just a couple of hours. Others noted the start time for FZRA is especially important stating that once FZRA starts, DOTs can't easily get on top of it as treatment is generally needed before the precipitation begins. The order of precipitation phase also plays an important role. Some forecasters noted that if FZRA falls on bare pavement, that that is more challenging from an impacts perspective than if it falls onto snow (SN). Forecasters also noted that FZRA is an especial challenge because it may freeze onto elevated objects, but not roads. Being able to anticipate this is very difficult with current decision-support guidance.

- *Snow (SN)*: This is another obvious issue for road hazards, but forecasters noted that with SN, there are several degrees of freedom in how SN impacts roads. Some noted that minor to moderate snowfall can be highly impactful because people are still active on the roads whereas with bigger events, people stay home. Snow rates, as opposed to storm-total accumulations, were also mentioned several times. Thresholds for snow rates varied with several noting that 1 to 2 inches per hour is a threshold where plows struggle to keep up. Very low amounts were also noted as problematic as SN can be "too shallow to even plow." Many forecasters stated that the character of snow (i.e., density) is as important as the rate. High density snow can be more difficult to plow, whereas low density snow may blow back onto the road immediately after plowing, if the wind is sufficient. Locality issues with SN were also highlighted. Forecasters with CWAs that include complex terrain noted that in high-elevation passes, the snow rates can be very high and lead to a singular location where SN is impactful while the remainder of the CWA is fine. Lake-effect snow was another local form of heavy SN that can impact just portions of a CWA. Last, heavy snow from mesoscale banding was mentioned several times. Both lake-effect and mesoscale banding were noted as a unique forecasting challenge because the exact locations where heavy snow will occur are difficult to pinpoint in advance.

- *Freezing fog (FZFG)*: Several forecasters noted this is one of their more impactful hazards for driving, which makes sense as it is a hazard that includes two threats (icy roads and reduced visibility). However, unlike with other hazards, they had very little to say about it except they lack necessary decision support to anticipate when and where it will form.

- *Blizzards/blowing snow*: These were mentioned as a difficult forecast challenge for a number of reasons. First, some forecasters noted that current guidance is not reliable, particularly for ground blizzards, they also highlighted that blowing snow is dependent on non-meteorological forcing (namely land use) so can be geographically variable. Last, they noted that the public and some stakeholders are challenged by the fact that blizzards have a duration requirement. This makes storms that don't meet the 3-h minimum difficult to message. These have similar impacts to blizzards, but that term is not applied. This causes confusion for the public.

- *Snow squalls*: While mentioned several times as a threat, issues with snow squall prediction and detection were rarely discussed.

B.2 Theme 2: Complicating factors that impact roads

Several non-meteorological contributions to road hazards during winter weather were highlighted as discussed below.

- *Mountainous terrain*: Forecasters with complex terrain in their CWA noted that snow rates at which mountain roads are impacted can be different than in flat terrain because the snow removal equipment is different. They noted that mountain passes are "travel choke points" that can sometimes lead to profound economic impacts. The position of the rain/snow line in elevated terrain was highlighted as another very difficult forecasting challenge.

- *Urban versus Rural land use*: Land use was noted as a control on blowing snow, but forecasters also noted other important impacts of land use. For example, one forecaster noted that people are more prone to drive too fast for the weather in rural areas. The impacts of winter weather when combined with rush-hour traffic in urban areas were also noted as it's more challenging for the DOT to operate in heavy traffic and the higher vehicle count introduces more opportunity for accidents.

- *Population changes*: Some forecasters noted that in recent years, population shifts have occurred that have resulted in new residents moving to their area having little familiarity with winter driving conditions. This, they argue, has had a noticeable impact on safety on roads during winter weather.

B.3 Theme 3: Communication of winter weather road hazards

Forecasters treat communication with the public differently than with key stakeholders such as EMs or DOTs.

Communication with DOTs has some variability. In some instances, the relationship with the state DOTs is very collaborative, marked by routine telecons, strategies to develop joint messaging, and shared information. In other instances, forecasters said they communicate weather threats to their DOT, but get very little information in return. Many forecasters noted that the kinds of information they share with DOTs is broader than with the public. This was particularly evident in discussions of probabilistic guidance. Forecasters noted that DOT partners are appreciative of probabilistic guidance such as best and worst-case-scenarios. Some DOTs have private-sector weather support, which was mentioned by several forecasters as a key part of the collaboration. In particular, they noted that when private-sector entities are involved, the NWS role is limited to what

the weather will do as opposed to how the weather might impact roads. When private-sector entities are not involved, forecasters note that they do provide information on how weather might impact roads, but shy away from making specific recommendations about whether/how to treat roads.

Communication with the public appears to be more limited and refined. Forecasters note that the public is more interested in knowing how the weather will impact them personally, such as the best time to travel through certain areas, rather than on the regional impacts. Forecasters were also quite reserved about the use of probabilistic data for the public. They noted that their constituencies are able to digest probability of exceedance graphics, but thresholds such as the 10th and 90th percentiles are often misused. Last, forecasters were varied in how much detail they share about roads with the public. In some cases, they described their media as "blurry" so as not to cross lines in the NWS directives. In other cases, they noted they have strong collaborations with their DOT and private sector partners and have coordinated messages prepared for the public. Last, some said they simply refer the public to their state's 511 pages if they want more information about the road conditions.

B.4 Theme 4: Timelines, content of communication, and sources of weather information

What and when hazards are communicated as well as how forecasters derive this information was another theme that emerged. The specific time ranges mentioned and how forecasters communicate at these ranges are as follows.

- *Greater than one week*: Forecasters noted that they are sometimes pressured to discuss the potential for major winter weather 7+ days in advance. They note that this makes them uncomfortable because there's still a high degree of uncertainty at these lead times, but social media often alerts stakeholders and the public to these storms and there is an expectation that the NWS would address it. There are specific decision points that some stakeholders make at these longer lead times that forecasters note are legitimate reasons to want advance guidance, even if there is uncertainty. In particular, they noted that staffing can be adjusted and routing schedules for ground shipping can be modified to work around the weather.

- *Days 4-7*: Forecasters noted an increased comfort with discussing the potential for winter storms on these time horizons. At these lead times, they focus primarily on headlines such as the potential for a winter storm and highlight what is uncertain, such as the track, precipitation type/amount. Several of the forecasters noted that if the storm is expected to exceed certain thresholds, this triggers specific collaborative sessions with stakeholders. Collaborative sessions may include webinars, briefings, emails, and chat rooms. For some forecasters, collaborative sessions with stakeholders are more of a one-way interaction – they present what they know

and the stakeholders listen and ask questions. For others, collaborative sessions are truly collaborative – the parties involved share what they know, their mitigation strategies for dealing with the weather, and work together to develop shared messaging for the public. The kinds of decision support guidance they seek at this time range is varied. Some tools/forecast systems that were specifically mentioned include the Winter Storm Outlook, WSSI, NBM forecasts, and DESI post-processed fields such as its cluster analysis. The WSUP viewer was also mentioned by a couple of forecasters. Some forecasters noted that days 4-7 present a special challenge in providing graphical messages. They said that when a graphic is created that shows other CWAs, there can be conflict over whether that graphic is appropriate to share. One forecaster said that often what happens is no one can reach a compromise and no graphics are shared, just a text message produced describing what may happen.

- *Days 1-3*: At these timelines, forecasters become more specific about the spatial/temporal details of the event. They try to pin down how the weather may impact certain activities such as travel or major community events. Several forecasters noted that in this window of time, their messaging becomes increasingly less probabilistic and more deterministic in content. Forecasters listed the same tools/forecast systems as they reference in the Days 4-7 time frame, but they begin to fold in CAM guidance from the HREF. They also mentioned some WPC forecasts such as Prob Snow.

- *As the event is unfolding or is imminent*: Forecasters noted that at these time horizons, they shift to a nowcasting mode and rely most heavily on current observations of the storm as it is moving in their direction or is over their area. Mostly, they message what is happening, the impacts the storm is having (where appropriate), and when to expect the storm to end. Forecasters noted a clear transition toward using observational trends to help guide their messaging. Some model output was still indicated, such as the HRRR and HREF, but forecasters seemed to rely much more strongly on upstream evolution of the weather and their own pattern recognition as the basis for their messaging.

B.5 Theme 5: Current and desired IDSS capabilities

The final theme to emerge from these discussions was on what kinds of decision-support capabilities forecasters most use and need. Far-and-away the biggest demand was for time of arrival and time of cessation graphics. However, forecasters didn't have specific thresholds for what constitutes arrival or cessation. Some forecasters were aware of and using experimental products that provide this kind of diagnostic while others were not. Several forecasters said they used simulated radar loops to provide time of arrival/time of cessation information.

Information about road temperature was also desired. This was noted to be controversial as forecasters are keenly aware of the need to follow NWS directives, but they felt they lacked key

information that could help them be more effective collaborators with DOTs by not having road temperature guidance. One forecaster mentioned that they use MetRo output, but noted it only provides point forecasts, so doesn't quite fulfill their needs. Others mentioned they used ProbSR or had heard of it, but since they can only access an analysis of ProbSR at present, they were uncertain whether it can truly fill their needs.

Quality observations during an event were also desired. Two forecasters mentioned that solicitation of LSRs is a challenging task, but much needed as an event is unfolding, especially in areas where there was heightened uncertainty. One forecaster even mentioned they have a dedicated person for soliciting LSRs during events.

While many noted that tools for probabilistic weather type exist, some complained that these perform poorly for freezing rain and asked for improved methods to predict the surface precipitation-type probabilities.

Improved decision support for blowing snow and ground blizzards was also highlighted. Forecasters noted that while tools for traditional blizzards (i.e., both falling and blowing show) exist, saltation after the snow has stopped is a major hazard that they lack adequate guidance on.

Last, decision support during the clean-up phase of storms was mentioned by several forecasters. They noted that DOTs may be in an accelerated mode for several days after a major winter storm to clean up side streets and these can be hampered by flash freezes.

C Key recommendations for future research

1. The kind of weather that is happening is pivotal. For example, the playbook for how to message and respond to FZRA is different than for a blizzard. Therefore, *as the HMT evolves to consider stakeholder engagement (or even to address advances for NWS forecasters), a balance in phenomena is recommended.* We were surprised at how important FZFG was in the eyes of forecasters as this is not typically mentioned when winter weather is discussed.

2. The forecasters were very cognizant of non-meteorological considerations and how those may impact whether a weather hazard generates impact. This creates a challenge for synthetic decision-making environments as forecasters may find it difficult, if not impossible, to speculate on how a certain tool or advance may benefit them in operations. In this regard, synthetic decision-making environments may give a misleading or incomplete conclusion about the viability of new products. Such a finding suggests products should be evaluated not just in a testbed setting, but in real time in the WFOs. However, the cost of such an endeavor can be prohibitive for individual research teams. *Hence, we recommend that products evaluated in the WWE that demonstrate success also be subject to a real-time CONUS-wide testbedding so that*

forecasters can try new products in their home offices. Whether this is supported through the WWE or the Operations Proving Ground is debatable.

3. Communication is very nuanced and depends on the stakeholder. *A simple jumping off point that builds on recent advances in the WWE is to have participants message content after doing an activity.* This can be done through a private twitter account and/or private slack chat. This content could then be evaluated by stakeholders to ask whether the messaging is clear and actionable.

4. Timelines for communication showed two areas where future work in the WWE could provide benefit. In days 4-7, forecasters were less concerned about guidance and more about messaging with graphics. They noted that creating graphical content is a challenge due to differences of opinion among neighboring offices on what is appropriate to share. *A future WWE activity could be to have different forecasters create content for days 4-7 and use social science experts to have these graphics vetted by stakeholders and the public.* Perhaps science-driven guidance on best practices may help diminish this issue. The other timeline of concern is within 24 of and during the event. Forecasters expressed a desire for more targeted decision support (e.g., road temperatures) during this time frame. There are some emergent tools that could be evaluated as a part of the WWE and even partnered with the messaging activity in point 3 above. *These include the hourly WSSI and Warn-on-Forecast. We recommend that future research seek to develop these capabilities and evaluate them as a part of the WWE.*

References

Benjamin, S., and Collaborators, 2021: Diagnostic fields developed for hourly updated noaa weather models. *NOAA Tech Memo*, 55, https://doi.org/https://doi.org/10.25923/f7b4-rx42.

Birk, K., E. Lenning, K. Donofrio, and M. T. Friedlein, 2021: A revised bourgouin precipitation-type algorithm. *Weather and Forecasting*, **36 (2)**, 425 – 438, https://doi.org/10.1175/WAF-D-20-0118.1.

Bullock, R., B. Brown, and T. Fowler, 2016: Method for object-based diagnostic evaluation. Tech. rep., National Center for Atmospheric Research. https://doi.org/10.5065/D61V5CBS.

Carbin, G., D. Petersen, and G. Fall, 2020: Use of gridded snowfall from noaa's office of water prediction at the weather prediction center. *Use of Gridded Snowfall from NOAA's Office of Water Prediction at the Weather Prediction Center*, URL https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/369382.

Lee, C., K. A. Brewster, N. Snook, P. Spencer, and J. Park, 2024: Spatial aligned mean: A method to improve consensus forecasts of precipitation from convection-allowing model ensembles. *Weather and Forecasting*, **39 (11)**, 1545 – 1558, https://doi.org/10.1175/WAF-D-23-0229.1, URL https://journals.ametsoc.org/view/journals/wefo/39/11/WAF-D-23-0229.1.xml.

Martinaitis, S. M., and Coauthors, 2020: A physically based multisensor quantitative precipitation estimation approach for gap-filling radar coverage. *Journal of Hydrometeorology*, **21 (7)**, 1485 – 1511, https://doi.org/10.1175/JHM-D-19-0264.1.

McCray, C. D., E. H. Attalah, and J. R. Gyakum, 2019: Long-duration freezing rain events over north america: Regional climatology and thermodynamic evolution. *Weather and Forecasting*, **34 (3)**, 665 – 681, https://doi.org/10.1175/WAF-D-18-0154.1.

Reeves, H. D., N. Lis, G. Zhang, and A. A. Rosenow, 2022: Development and testing of an advanced hydrometeor-phase algorithm to meet emerging needs in the aviation sector. *Journal of Applied Meteorology and Climatology*, **61 (5)**, 521 – 536, https://doi.org/10.1175/JAMC-D-21-0151.1.

Sanders, K. J., and B. L. Barjenbruch, 2016: Analysis of ice-to-liquid ratios during freezing rain and the development of an ice accumulation model. *Weather and Forecasting*, **31 (4)**, 1041 – 1060, https://doi.org/10.1175/WAF-D-15-0118.1.