Also present but not pictured - Brian Ancell, Bryan Guarente, Bruce Veenhuis, and Steve Willington

2026 Ensemble Clustering and Sensitivity Analysis (ECSA) Testbed Final Report

March 30 - April 3, 2026

Weather Prediction Center (WPC)
Austin Coleman[1,2]

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

1. Overview of the Operations and Science

The Hydrometeorology Testbed (HMT), part of the Weather Prediction Center (WPC), hosted its first Ensemble Clustering and Sensitivity Analysis (ECSA) Testbed Experiment from Monday, March 30 - Friday, April 3, 2026 across a Morning Group and an Afternoon Group. This experiment gathered 32 participants and facilitators from across the National Weather Service (NWS) and academia (Figure 1) to evaluate the utility of WPC ensemble clustering and sensitivity analysis tools for forecasting high-impact weather events at medium-range lead times. All NWS regions were represented in this experiment with the exception of the Pacific Region. Forecast activities focused on two retrospective weather events with varying degrees of data denial to forecast for assigned County Warning Areas (CWAs) at the day 6, day 5, and day 3 lead times. Participants first analyzed provided medium-range guidance on their own for their assigned CWA to write a short Area Forecast Discussion (AFD) to summarize their initial thoughts. From there, participants were stratified into small groups to develop key message graphics for their prescribed area based on the given guidance. Focus group discussions followed the forecast activities each day to assess forecast utility of the tools used, to identify gaps in understanding or training that need to be addressed to use the tools, and to determine optimal strategies for scenario-based forecast messaging. Three seminars from subject matter experts were hosted throughout the week to further supplement the ensemble clustering and sensitivity analysis discussion. Themes and future directions from conversations throughout the week were consolidated into a Jamboard of key takeaways on the final day of the testbed. The testbed schedule may be found below (Table 1). Readers interested in further details on experiment activities are referred to the Operations Plan.

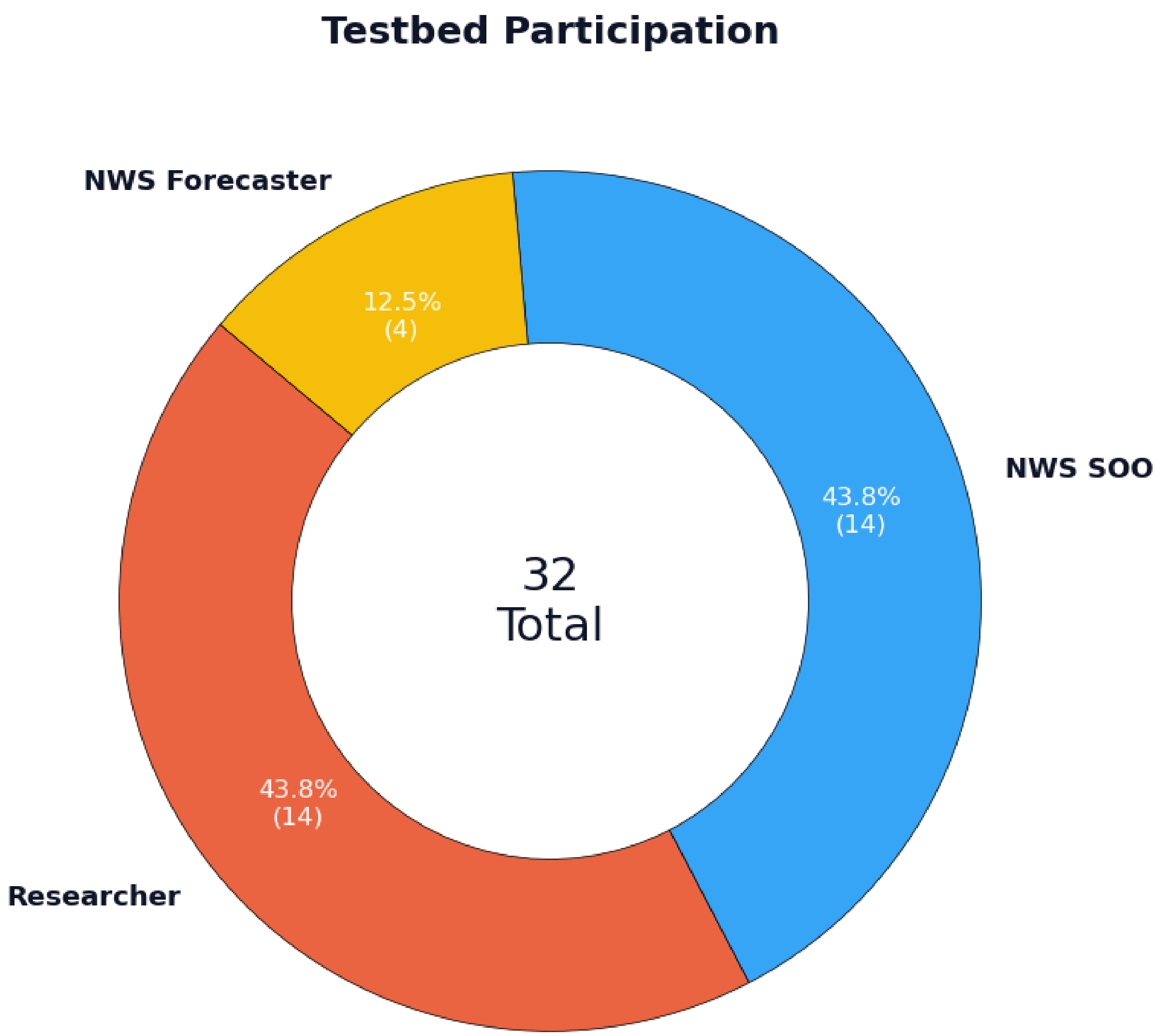

Figure 1. *Pie chart depicting the percentage of testbed participants and facilitators who are researchers, NWS forecasters, or NWS Science and Operations Officers (SOOs) across the Morning and Afternoon Groups.*

Monday | 3.30.26	Tuesday | 3.31.26	Wednesday | 4.1.26	Thursday | 4.2.26	Friday | 4.3.26
09:00am ET — Morning Icebreaker Activity	09:00am ET — Event 1 Day 3 WPC Briefing	09:00am ET — Event 2 Day 6 WPC Briefing	09:00am ET — Event 2 Model Trends and ESA Demonstration - Austin Coleman	09:00am ET — Individually Review Slide Deck of all Forecast Messages from Morning Group
09:30am ET — Orientation Begins	09:15am ET — Individual Forecaster Assessments	09:15am ET — Individual Forecast Assessments	09:45am ET — Short Model Trends / ESA Exercise	09:15am ET — Discuss Overall Themes from Forecast Messaging
10:30am ET — Break	10:00am ET — Break	09:45am ET — Breakout into Three Different CWA / Cluster Variable Groups. Develop D6 Messaging	10:30am ET — Event 2 Post-Event WPC Overview	09:45am ET — Focus Group Discussion
				10:15am ET — Break
10:45am ET — Event 1 Day 5 WPC Briefing	10:15am ET — Breakout into Three Groups for Different CWAs. Adjust Day 5 Messaging for Day 3	10:35am ET — Break	10:45am ET — Break	**10:30am ET — Brian Ancell Presentation**
11:00am ET — Individual Forecast Assessments	11:15am ET — Discuss D3 Forecast Messaging / Collab	10:45am ET — Event 2 Day 3 WPC Briefing	**11:00am ET — Steve Willington Presentation**	11:00am ET — Open up the "Jam Board" Group Document of Best Practices. Participants can reflect and contribute notes
11:40am ET — Breakout into Three Groups for Different CWAs. Generate Day 5 Forecast Messaging	11:30am ET — Event 1 Post-Event WPC Overview	11:00am ET — Individual Forecast Assessments	11:30am ET — Post-ESA Activity Discussion	11:45am ET - Break
12:30pm ET — Break	11:45am ET — Break	11:30am ET — Adjust D6 to D3 within CWA/Variable Groups	12:00pm ET — Break	12:00pm ET — Consolidate Jam Board Themes into a Group Document of Best Practices
		12:20pm ET — Break	12:15pm ET — Discuss D3 Messaging. Evaluate Effectiveness of Morning Group Messaging	
12:40pm ET — Discuss D5 Forecast Messaging	12:00pm ET — Evaluate Effectiveness of Morning Group CWA Forecast Graphics. Compare with Afternoon Group	**12:30pm ET — Ryan Torn Presentation**	12:45pm ET — Compare and Critique Afternoon Group Event 2 D6 / D3 Messaging	12:50pm ET — Closing Remarks
01:00pm ET — Morning Group Adjourns	01:00pm ET — Morning Group Adjourns	01:00pm ET — Morning Group Adjourns	01:00pm ET — Morning Group Adjourns	01:00pm ET — Morning Group Adjourns
02:00pm ET — Afternoon Icebreaker Activity	02:00pm ET — Event 1 Day 3 WPC Briefing	02:00pm ET — Event 2 Day 6 WPC Briefing	02:00pm ET — Adjust Event 2 D6 to D3 within CWA/Variable Groups	02:00pm ET — Compare and Critique Morning Group Event 2 D6 / D3 Messaging. Closing out Event 2 Discussion
02:30pm ET — Orientation Begins	02:15pm ET — Individual Forecaster Assessments	02:15pm ET — Individual Forecast Assessments	03:00pm ET — Break	02:30pm ET — Individually Review Slide Deck of all Forecast Messages from Afternoon Group
03:30pm ET — Break	03:00pm ET — Break	03:00pm ET — Break	03:15pm ET — Event 2 Model Trends and ESA Demonstration - Brian Colle and Austin Coleman	02:45pm ET — Discuss Overall Themes from Forecast Messaging
03:45pm ET — Event 1 Day 5 WPC Briefing	03:15pm ET — Breakout into Three Groups for Different CWAs. Adjust Day 5 Messaging for Day 3	03:15pm ET — Breakout into Three Different CWA / Cluster Variable Groups. Develop D6 Messaging	04:00pm ET — Short Model Trends / ESA Exercise	03:15pm ET —Break
				3:30pm ET — Focus Group Discussion
04:00pm ET — Individual Forecast Assessments	04:15pm ET — Discuss D3 Forecast Messaging / Collab	04:15pm ET — Discuss D6 Forecast Messaging	04:45pm ET — Break	04:00pm ET — Open up the "Jam Board" Group Document of Best Practices. Participants can reflect and contribute notes
04:40pm ET — Breakout into Three Groups for Different CWAs. Generate Day 5 Forecast Messaging	04:30pm ET — Event 1 Post-Event WPC Overview	04:45pm ET — Event 2 Day 3 WPC Briefing	05:00pm ET — Event 2 Post-Event WPC Overview	04:45pm ET —Break
05:30pm ET — Break	04:45pm ET — Break	05:00pm ET — Break	05:15pm ET — Post-ESA Activity Discussion	05:00pm ET — Consolidate Jam Board Themes into a Group Document of Best Practices. Merge with Morning Document

Monday | 3.30.26	Tuesday | 3.31.26	Wednesday | 4.1.26	Thursday | 4.2.26	Friday | 4.3.26
05:40pm ET — Discuss D5 Forecast Messaging	05:00pm ET — Evaluate Effectiveness of Afternoon Group CWA Forecast Graphics. Compare with Morning Group	05:15pm ET — Individual Forecast Assessments	05:35pm ET — Evaluate Effectiveness of Afternoon Group Messaging	05:50pm ET — Closing Remarks
06:00pm ET — Afternoon Group Adjourns	06:00pm ET — Afternoon Group Adjourns	06:00pm ET — Afternoon Group Adjourns	06:00pm ET — Afternoon Group Adjourns	06:00pm ET — Afternoon Group Adjourns

Table 1. *Schedule for the Morning and Afternoon Groups during the ECSA Testbed Experiment. Entries highlighted in green indicate a WPC Forecaster Briefing, purple indicates discussion time, and light orange indicates a seminar time slot.*

2. Weather Event and Forecast Activity Synopses

a. Event 1 | 00 UTC January 24, 2026 - 00 UTC January 25, 2026

The first weather event that participants forecasted consisted of a very significant winter storm that impacted the entire eastern half of the CONUS. Figure 2 gives the reader a sense of geographic scope for the occurrence of high-impact precipitation, even over a single 24-h window and without distinguishing mixed precipitation types. For the Central U.S., a strong, shallow surface Arctic high pressure system surged southward, under-cutting a much warmer subtropical air mass (Figure 3). Small variations in the depth of the cold air had large implications for the type of precipitation that ultimately fell, making this event challenging to forecast, even at shorter lead times when forecasters typically expect greater predictability.

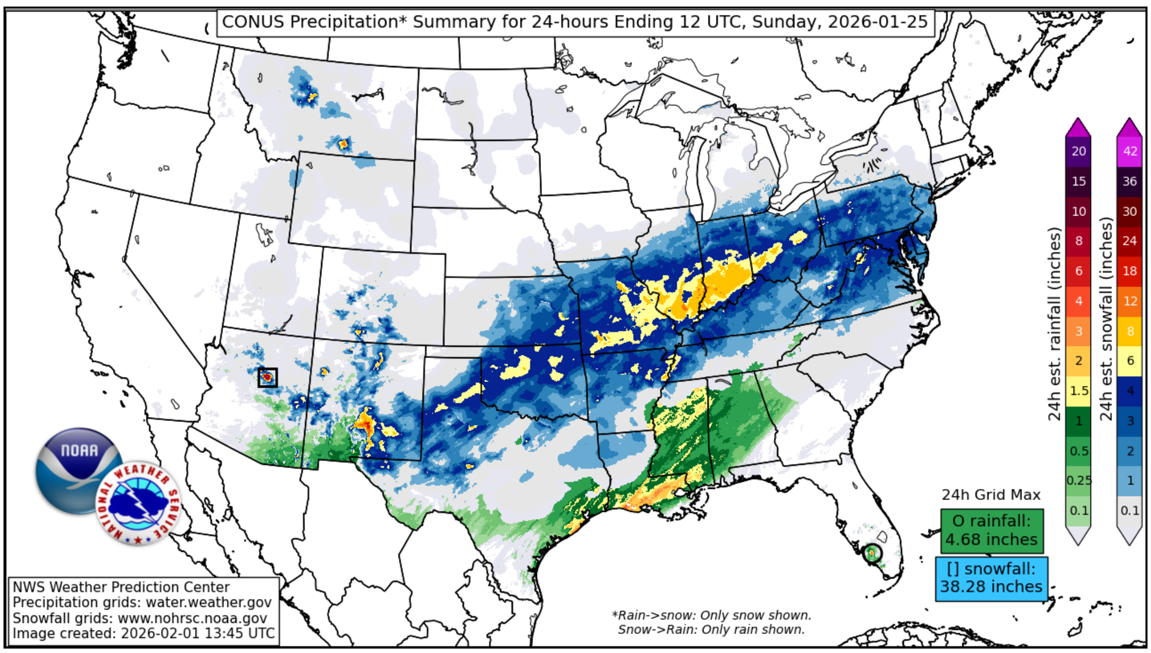

Figure 2. *Observed precipitation valid between 1200 UTC January 24 and 1200 UTC January 25, 2026 from the WPC Winter Storm Archive.*

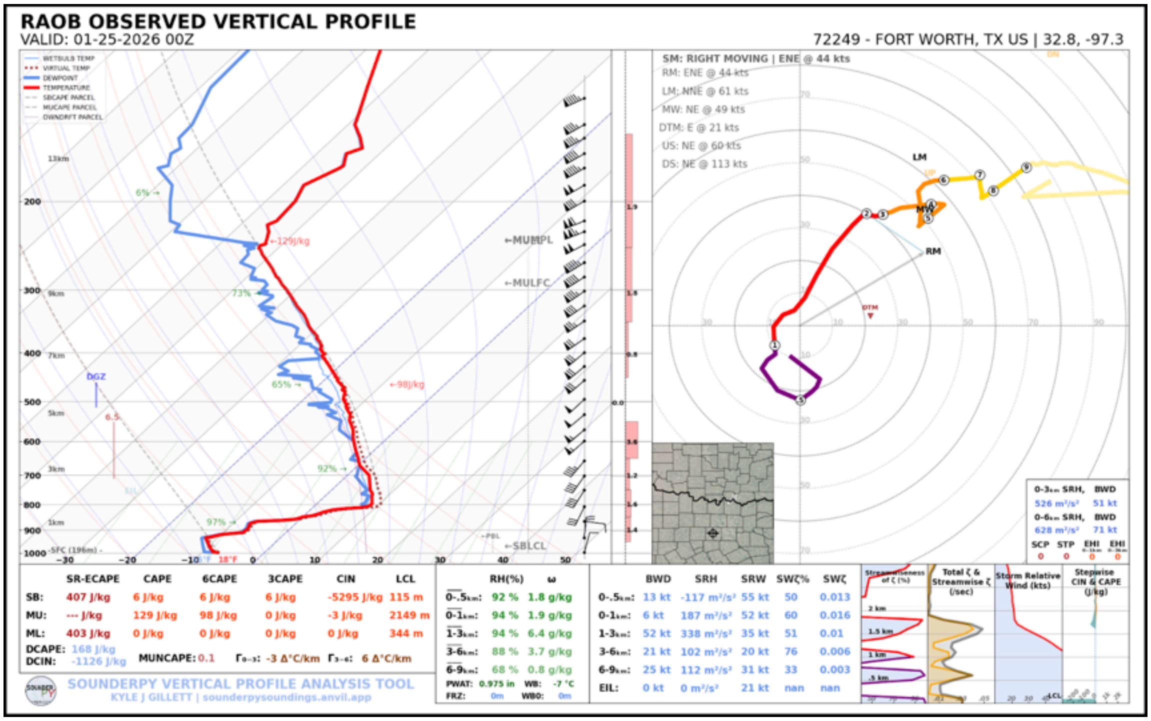

Figure 3. *Observed radiosonde sounding profile valid at 0000 UTC on January 25, 2026. The very cold Arctic air mass brings 18 degree Fahrenheit surface temperatures to Fort Worth, Texas. An incredibly sharp temperature inversion begins at approximately 925 hPa, with significantly warmer temperatures aloft. This sounding may be found in the [NWS Fort Worth review of the late January winter storm](#).*

The forecast activities focused on developing key message graphics for one of three assigned County Warning Areas (CWAs; shown below by Figure 4) at the Day 5 lead time, using the 0000 and 1200 UTC January 19, 2026 forecast initializations, and then at the Day 3 lead time, using the 0000 and 1200 UTC January 21, 2026 initializations. A WPC forecaster — Cody Snell for the Morning Group and Mussie Kebede for the Afternoon Group — provided 15-minute forecast briefings at each lead time based on available deterministic and ensemble guidance to set the stage for the forecast activities. After the briefing, participants had between 40 and 60 minutes to analyze available deterministic and ensemble forecast data within their assigned CWAs and share their initial thoughts on the event through a mini AFD and survey. For each set of lead times, participants had access to the [ECSA Testbed Website](#), which hosts WPC 500-hPa Height Clusters, National Blend of Models (NBM) Quantile Mapping and Dressing (QMD) members Quantitative Precipitation Forecast (QPF) Clusters, Mean Sea Level Pressure (MSLP) Clusters, and Ensemble Sensitivity Analysis (ESA) for 500-hPa Height, QPF, and MSLP Principal Components (PCs). They additionally had access to archived cluster data via the [Dynamic Ensemble-based Scenarios for Impact-Based Decision Support Services (IDSS) (DESI) platform](#). The facilitators would like to acknowledge Travis Wilson for his efforts in adding realtime and archived ensemble clusters onto the public-facing DESI platform as well as the NOAA-only DESI platform so that non-NOAA academic participants could still incorporate DESI clusters into their forecast process.

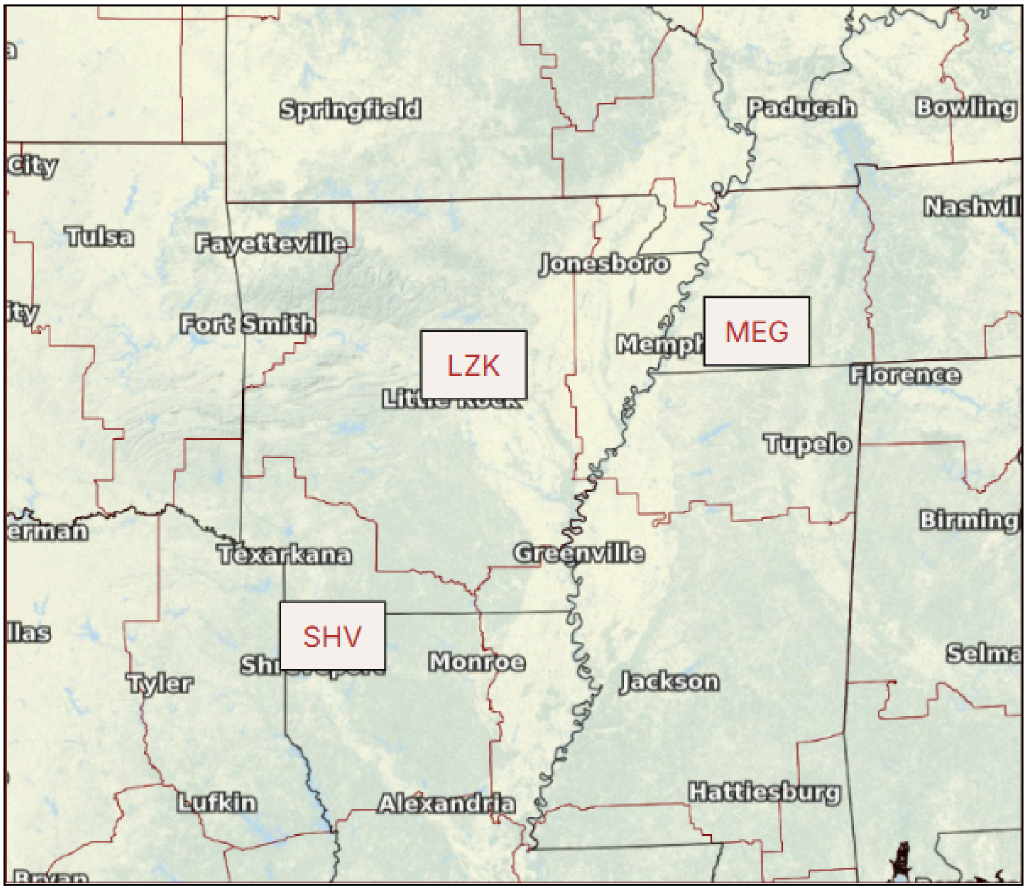

Figure 4. *Map describing the three CWAs used for Event 1 forecast activities. CWA boundaries are outlined in maroon, and abbreviated CWA titles are annotated in text boxes. The underlying geographic figure was generated with the Whole Story Uncertainty & Probabilities (WSUP) Viewer.*

I. Day 5 Forecast Activities

For the Day 5 lead time, participants' AFDs generally reflected a consensus that there was high confidence in a general synoptic pattern that would be conducive to a potentially impactful winter system affecting their respective CWAs. However, significant uncertainty remained regarding precipitation amounts, placement, and type. Most participants identified the phasing of a cutoff low near Baja, California with a northern trough advancing from Canada as the primary source of precipitation uncertainty. While 6 / 21 participants specifically mentioned using ESA to come to or confirm this conclusion, many of the participants inferred this from the cluster scenarios. A representative example of an AFD for this event at Day 5 can be found below:

"The Day 5 period features a complex and potentially high-impact winter weather setup for the ArkLaTex region. Ensemble Cluster Analysis reveals significant spread regarding

the timing and amplitude of an advancing trough, as well as the southward progression of an arctic surface high. A critical factor for our CWA will be the thermal profile, specifically the interplay between a shallow sub-freezing airmass at the surface and a warm nose near 850 mb.

If the system trends toward a stronger northern stream (Cluster 2), we could see our wettest and snowiest solution, with 2-3 inches of snow possible and high-end spread reaching 3-7 inches across northern zones. Conversely, a slower trough with a strongly phased southern stream (Cluster 4) would limit snowfall but introduce a historic, high-impact freezing rain threat across our north due to a tight surface temperature gradient and warmer air aloft. MSLP clustering shows uncertainty in how fast and far the arctic high builds south, further complicating exactly where that transition zone will set up. Regardless of precipitation type, QPF placement remains somewhat uncertain, with variance leaning heavily on the magnitude and NW-to-SE placement of the heaviest precipitation band."

Figure 5 describes how heavily participants used each forecast tool when creating their mini AFDs ranked from highest to lowest average use score of each product at the Day 5 lead time for this event. Non-cluster DESI visualizations and WPC ESA fields had the lowest average use score, with only 25% of participants reporting a use score of 4 or more. Cluster scenarios in all forms, particularly the WPC 500-hPa Height Clusters where 66% of participants reported a use score of 4 or more, were used the most extensively. It should also be noted that WPC QPF Clusters were erroneously excluded from the options in this survey until mid-week. For the sake of comparison, the results from QPF Cluster usage for Event 1 are excluded from this report.

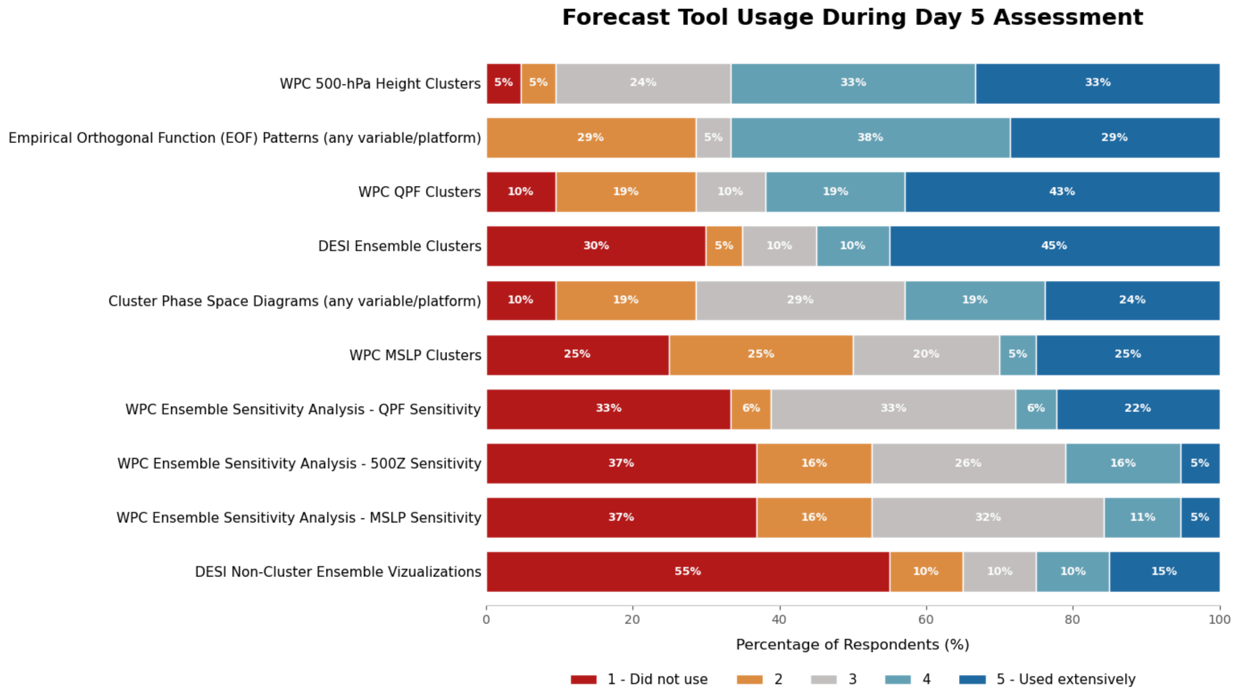

Figure 5. *Participant-reported ensemble product usage for the Day 5 forecast of Event 1 ranked by average use score in descending order from top to bottom.*

The last survey question for Day 5 pertained to choosing a preferred forecast scenario that the participant felt had the best chance of coming to fruition. The facilitators hypothesize that it is against best practice to identify a single forecast scenario ahead of the event and focus messaging on that singular outcome. This hypothesis is due to the fact that the ensemble cluster approach is based on an uncalibrated ensemble of opportunity with known dispersion issues rather than an ensemble of equal likelihood, which means that the cluster scenario consisting of the highest ensemble membership is not necessarily the most likely solution. However, we also sought to understand if participants exhibited any forecast skill at selecting the best-performing scenario in advance. If so, one could consider honing forecast messaging towards a particular scenario at an earlier lead time than previously thought.

To identify whether participants are skillful at correctly identifying the best-performing forecast scenario ahead of the event, we asked participants to answer the following question in the survey: “Even though we hypothesize that it's against best practice to do this, if you *had* to choose a single 500Z Cluster Scenario for Day 5 to message for the overall forecast of your chosen region, which scenario would you choose?” A breakdown of participants’ responses reveal that Clusters 1 and 2 were chosen most often, with 7 responses each, followed by Cluster 3 (Figure 6). While verification was not the primary goal of the testbed experiment, Mark Klein assembled post-event overviews for both events and summarized the Mean Absolute Error (MAE) metrics and spatial comparisons of the forecast scenarios relative to Global Forecast System (GFS) and Unrestricted Mesoscale Analysis (URMA) analyses. Cluster 3 from the Day 5 500-hPa height ensemble clusters possessed the lowest MAE for 500-hPa height and derived 2-meter surface temperature. However, it performed relatively poorly compared to other scenarios in its handling of QPF. To understand which 500-hPa height cluster scenario performed the best overall for the event, a Google Meet poll was conducted with the mode of the poll used as best-performing cluster scenario. In both Morning and Afternoon Groups, Cluster 3 was selected as the best Day 5 forecast scenario overall, although it was not chosen unanimously. For Event 1 at the Day 5 lead time, participants were unable to correctly identify the optimal forecast scenario in advance of the event, constituting a data point that using ensemble clusters as a means to message forecast uncertainty rather than to pick a single winning scenario should remain the best practice for using this tool.

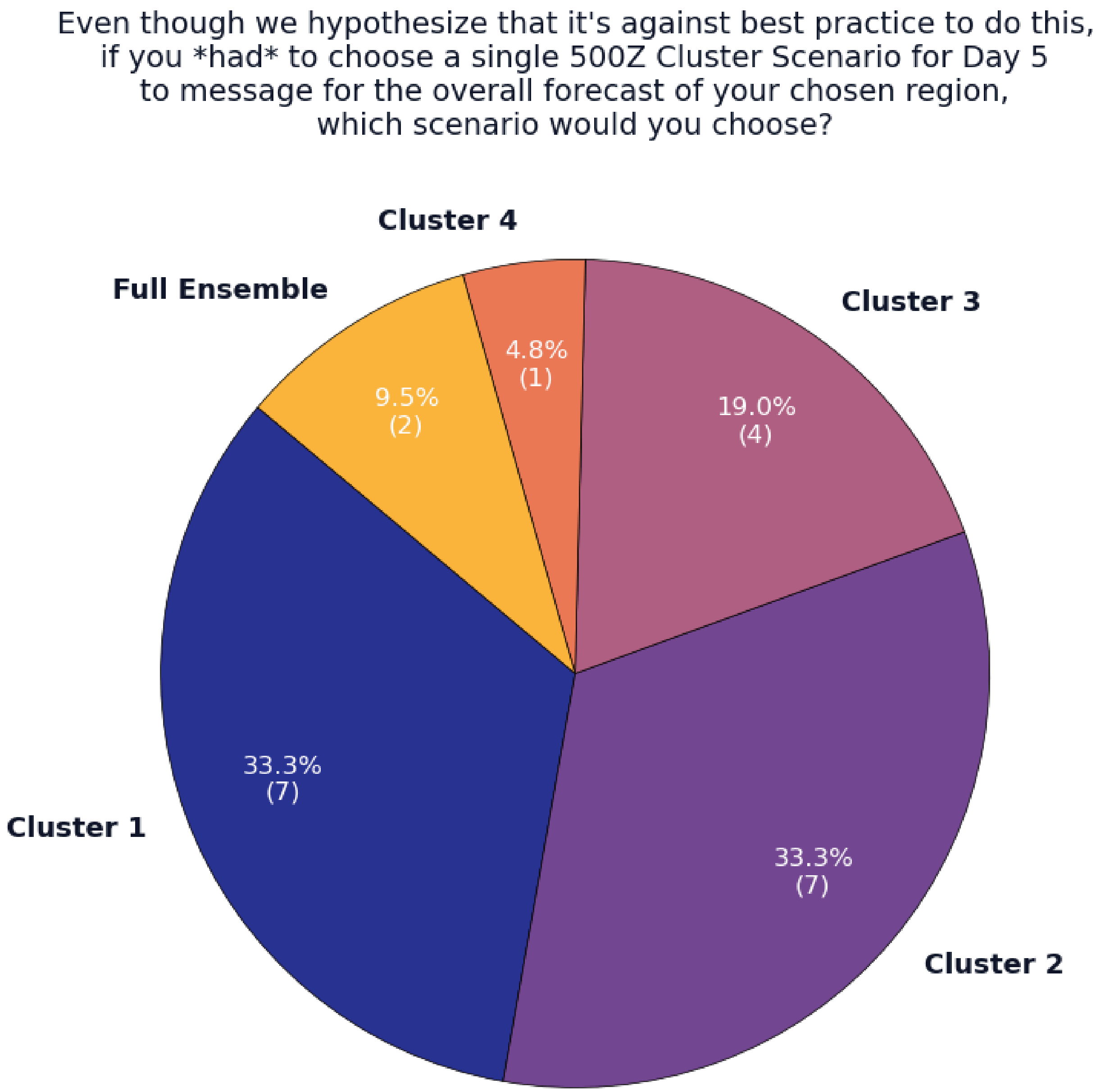

Figure 6. *Participant-reported preferred forecast scenarios based on their understanding of the event at the Day 5 lead time.*

After completing the individual forecaster assessments, participants joined their full CWA groups to build Day 5 Key Message graphics for their area and later briefly present their Key Message slides. All CWA groups across the Morning and Afternoon sessions displayed DESI data in their Key Message graphics (example shown by Figure 7). Half of the CWA groups across sessions chose to display probability of exceedance graphics while the other half chose to display scenario-based graphics. Discussion revealed that groups who did not incorporate scenario-based graphics either felt that there were too many potential outcomes at this lead time to whittle down the messaging to two or three forecast clusters, or they felt that the precipitation type uncertainty was too complex to convey in a public-friendly manner with scenario-based graphics. Finally, both sessions drew comparisons between their Key Message slides and actual Weather Forecast Office (WFO) messaging from similar lead times for that event. Social scientist, Leticia Williams, later noted that she really appreciated the actual NWS Memphis WFO's simple, actionable messaging with colorblind-friendly color palettes and public-friendly icons for this lead time (Figure 8).

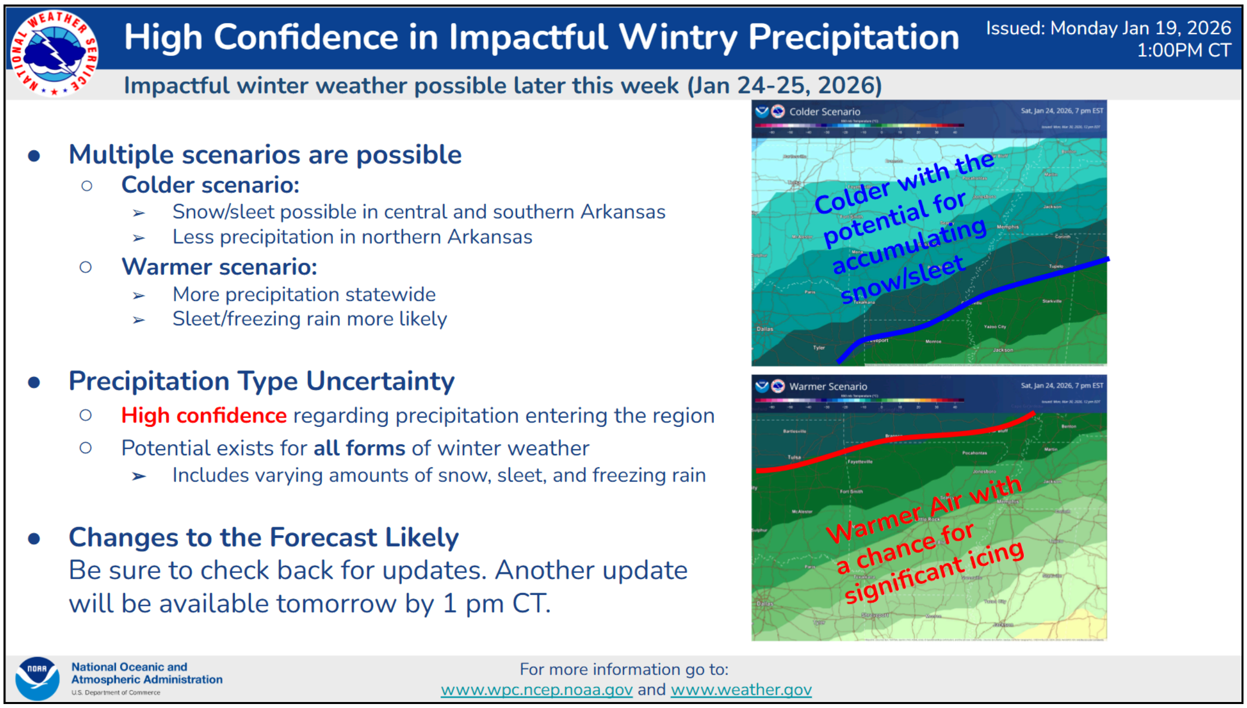

Figure 7. *Example Day 5 Key Message slide created for the Little Rock, AR (LZK) CWA during the Morning Group session.*

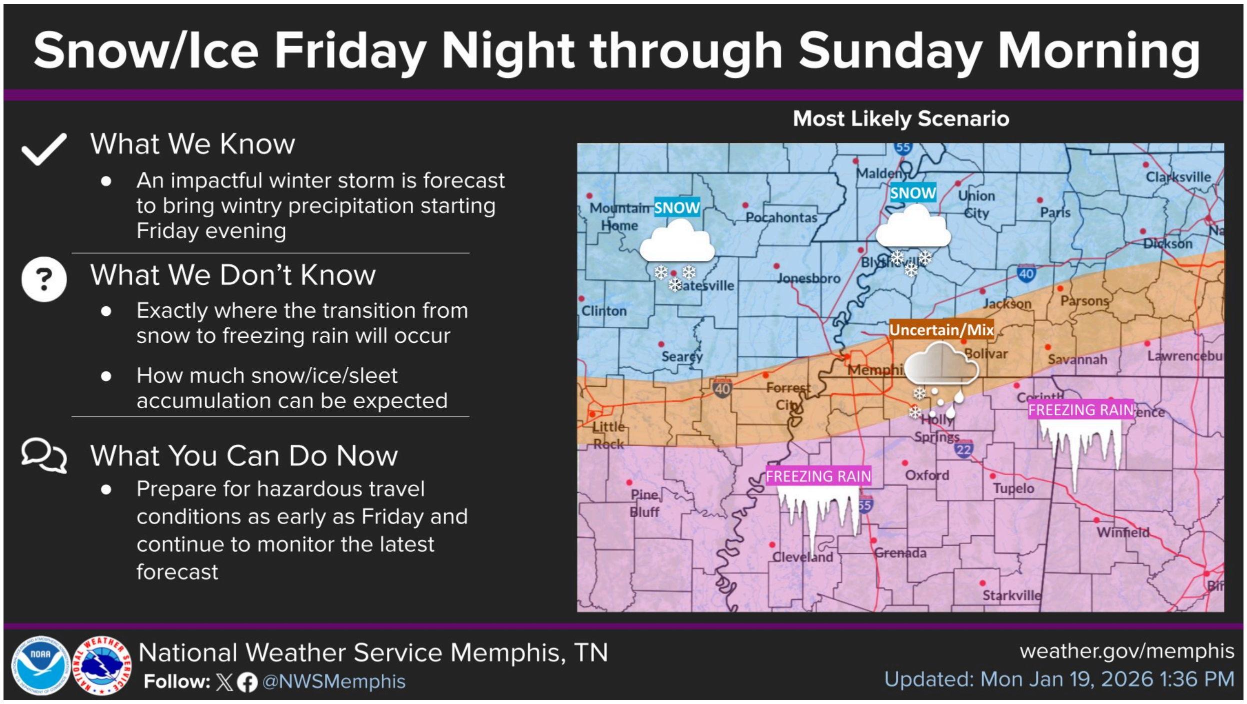

Figure 8. *Exemplary actual WFO messaging from NWS Memphis at the Day 5 lead time.*

II. Day 3 Forecast Activities

For the Day 3 activity for Event 1, confidence of a significant, if not historic, winter storm increased across the board as reflected in participants' mini AFDs. Enough uncertainty in the degree of phasing between the northern and southern stream troughs existed at this lead time to result in different precipitation type scenarios, as exemplified by this comprehensive AFD:

"Anomalously strong Arctic High poised to dig southward across the Midwest while cutoff low phases back into parent northwesterly flow storm track across Rocky Mountains Friday into Saturday.

Precipitation to rapidly fill in as disturbance rides along the stalled frontal boundary between these two airmasses Saturday into Saturday night.

Mixed precipitation zone right across the heart of the CWA, with greatest confidence for crippling freezing rain across far northern LA and into southern AR.

Uncertainty continues in how quickly the cutoff upper low phases into the incoming amplifying shortwave trough across the Rocky Mountains.

Roughly half of medium range ensemble cluster members (clusters 2 and 3) result in more significant ice accumulations through the CWA.

The other half of the members in clusters 1 and 4 delay the phasing of the cutoff low across Baja California, which results in deeper penetration of the cold front and Arctic air southward across the CWA, bringing higher snowfall totals and more limited ice accumulations.

There is a 50% chance for nearly a half an inch of ice in the Shreveport area with the half of the members with a faster phasing.

Around 20% of medium range ensemble members push heavier liquid precipitation totals southward across the entire CWA with 1.5" or greater, whereas the other 75% maintain a stronger gradient between heavier precipitation across southern AR and northern LA and lighter precipitation into central LA."

A few mini AFDs brought up the incorporation of sleet in snow forecast totals as a challenging aspect of the Day 3 forecast. The Global Ensemble Forecast System (GEFS) and Canadian Meteorological Centre Ensemble (CMCE) systems both include sleet in their snow total output, whereas the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble system treats sleet distinctly from snow. These differences amongst ensemble guidance made it particularly difficult for participants to diagnose precipitation type uncertainty, as it became all the more challenging to understand the extent to which sleet made up the snow totals amongst model guidance.

Forecast tool usage evolved from Day 5 to Day 3, in part due to participants getting more comfortable with some of the newer tools and partly due to the increased forecast confidence with the shorter lead time. In the Day 3 survey, participants were asked to explain how their forecast tool usage changed between the two lead times, and most mentioned that they prioritized DESI more to tease out sensible impacts from the relatively more subtle differences between clusters at Day 3. They also reported feeling more comfortable with their assigned CWA and thus had more time to devote to exploring ESA or other guidance with which they were less familiar. This feedback is reflected by the Day 3 forecast tool usage (Figure 9), with DESI Ensemble Clusters taking the top ranking for average product use score and WPC 500-hPa Height Sensitivity fields moving up from the 8th Day 5 use score ranking to the 4th Day 3 use score ranking. EOF Patterns were extensively used at both lead times for this event.

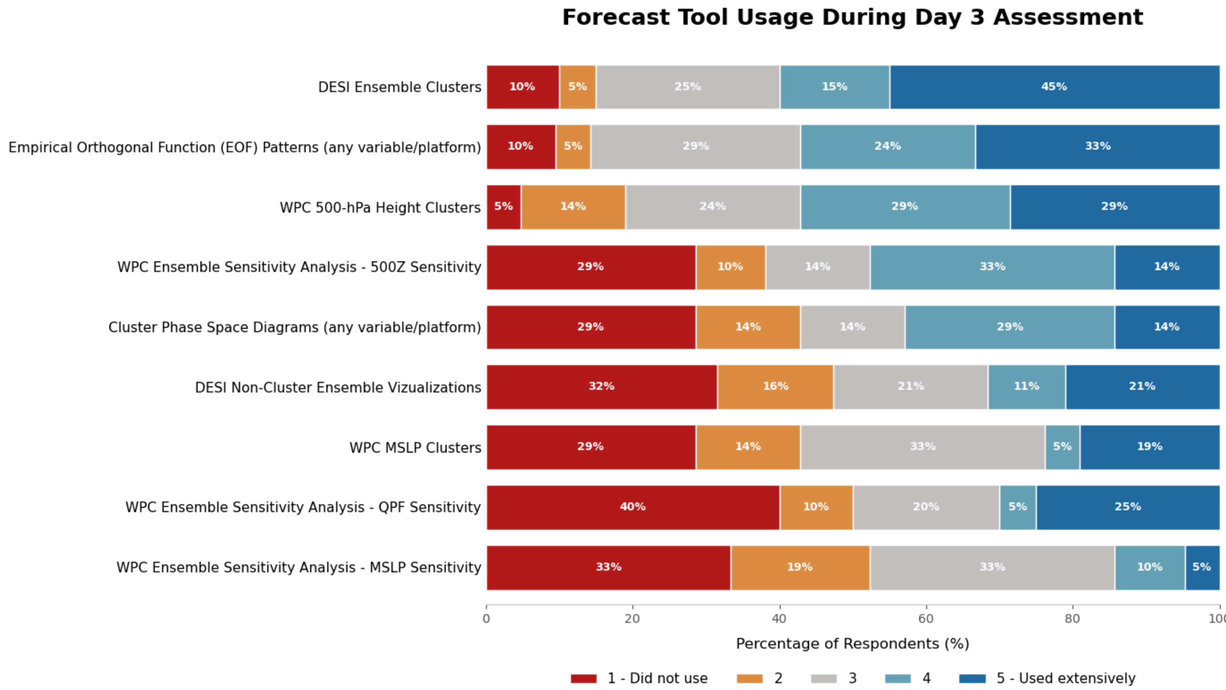

Figure 9. *Participant-reported ensemble product usage for the Day 3 forecast of Event 1 ranked by average use score in descending order from top to bottom.*

To identify whether participants are skillful at correctly identifying the best-performing forecast scenario ahead of the event, we again asked participants to answer the following question in the survey: “Even though we hypothesize that it's against best practice to do this, if you *had* to choose a single 500Z Cluster Scenario for Day 3 to message for the overall forecast of your chosen region, which scenario would you choose?” At this lead time, the breakdown of participant responses reflects less variability, with most responses indicating Cluster 2 (8 responses) was their preferred scenario, followed by Cluster 1 (6 responses) and the Full Ensemble (4 responses; Figure 10). This result is particularly interesting since post-event verification revealed that Cluster 2 was the worst-performing 500-hPa height cluster with incorrect phasing of the northern and southern streams, which resulted in the highest MAE for 500-hPa heights as well as QPF. Cluster 1 and the Full Ensemble visualizations tied for lowest 500-hPa height MAE, although Cluster 3 possessed the lowest QPF MAE which matched participants’ subjective evaluations of the cluster derived QPF performance. Conflicting verification characteristics of different fields from the 500-hPa Height Clusters led to different best-performing forecast scenarios being chosen by the Morning and Afternoon Groups. The Morning Group most often voted for Day 3 Cluster 1 as the best forecast visualization for the event overall, while the Afternoon Group most often voted for Cluster 3 as the best forecast visualization for the event overall. Neither of these scenarios were chosen most often prior to event verification, further supporting the notion that “picking a winner” forecast scenario in advance of the event should not become a best practice.

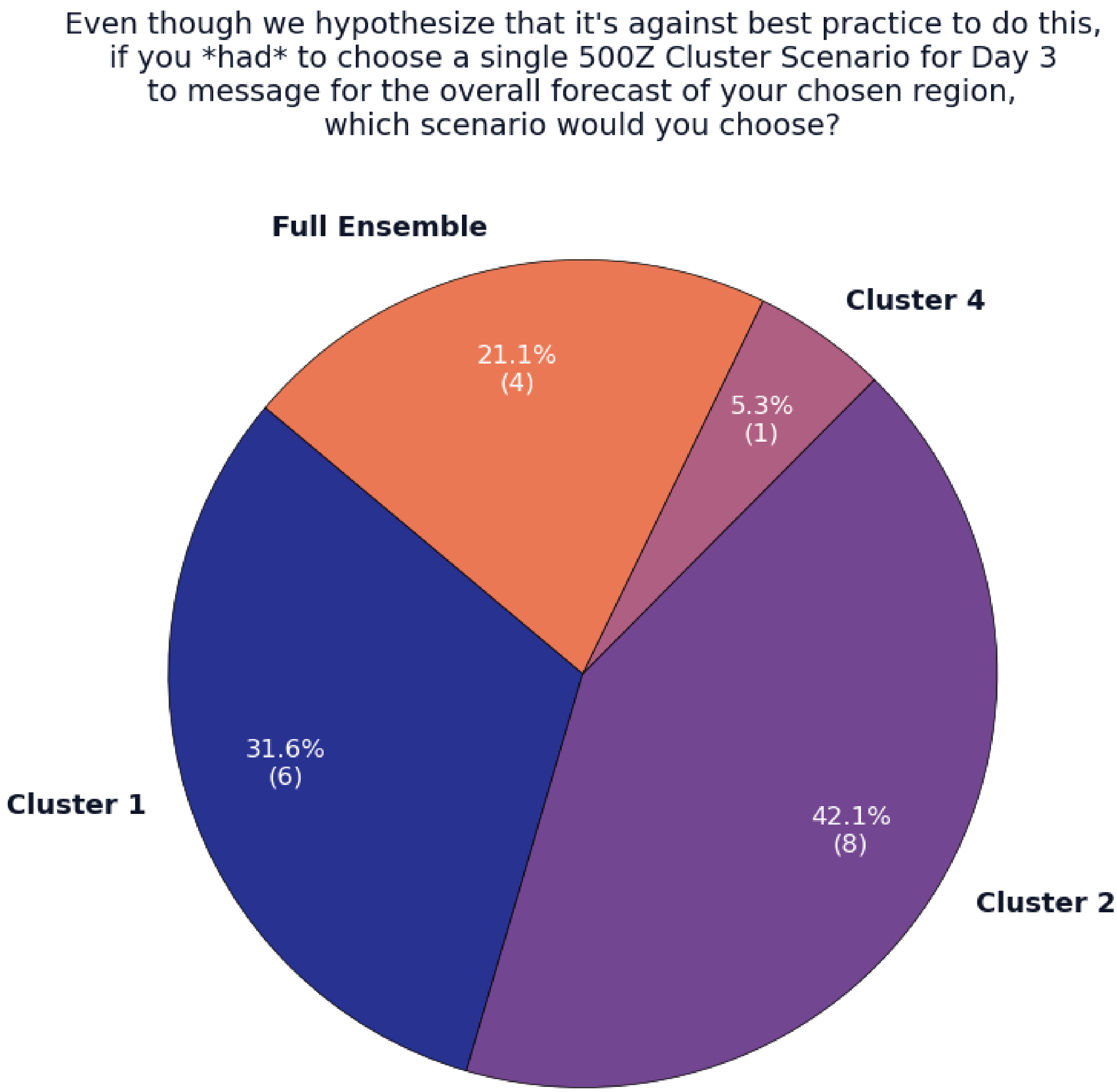

Figure 10. *Participant-reported preferred forecast scenarios based on their understanding of the event at the Day 3 lead time.*

After the individual assessments, participants joined the same CWA Breakout Groups to adapt their Day 5 messaging for Day 3. Most groups were able to dig further into DESI data, especially with point soundings. Many spent time pointing out the high intracluster spread at many locations for this case. A few participants who used DESI ensemble cluster soundings noted 8-10 degree Fahrenheit spread in some forecast scenarios, and even some bi-modal distributions within single clusters for some locations. These all suggested that some of the ensemble clusters were averaging together physically-distinct atmospheric scenarios. Certain CWA groups combined multiple cluster scenarios together to message common impacts, even if the meteorological mechanisms to reach a given impact were slightly different. Other themes of discussion included the need for clusters that properly reflect the uncertainty of lower-level dynamics. Multiple participants from the Morning Group suggested incorporating Potential Vorticity (PV) viewpoints into the cluster product suite. Similarly, participants from the Afternoon Group noted that when the clusters are not based on the impactful feature of interest, it tends to be easier to use the grand ensemble to tell the story.

For this lead time, CWA Groups opted to use fewer scenario-based graphics in favor of probability of exceedance graphics for different precipitation types (example shown by Figure 11). Only 2 / 6 CWA Groups displayed cluster graphics in their Key Messages across the Morning and Afternoon sessions. Lastly, participants from the Afternoon Group pointed out that using 90th percentile precipitation graphics to highlight a worst-case scenario would not have worked in messaging either, since the scenarios were linked. Had the 90th percentile snowfall verified further south, the 10th percentile snowfall would have verified further north. Given this relationship between scenarios, messaging the 90th percentile snowfall across any particular CWA would have resulted in a large area of incorrect precipitation type or false alarm.

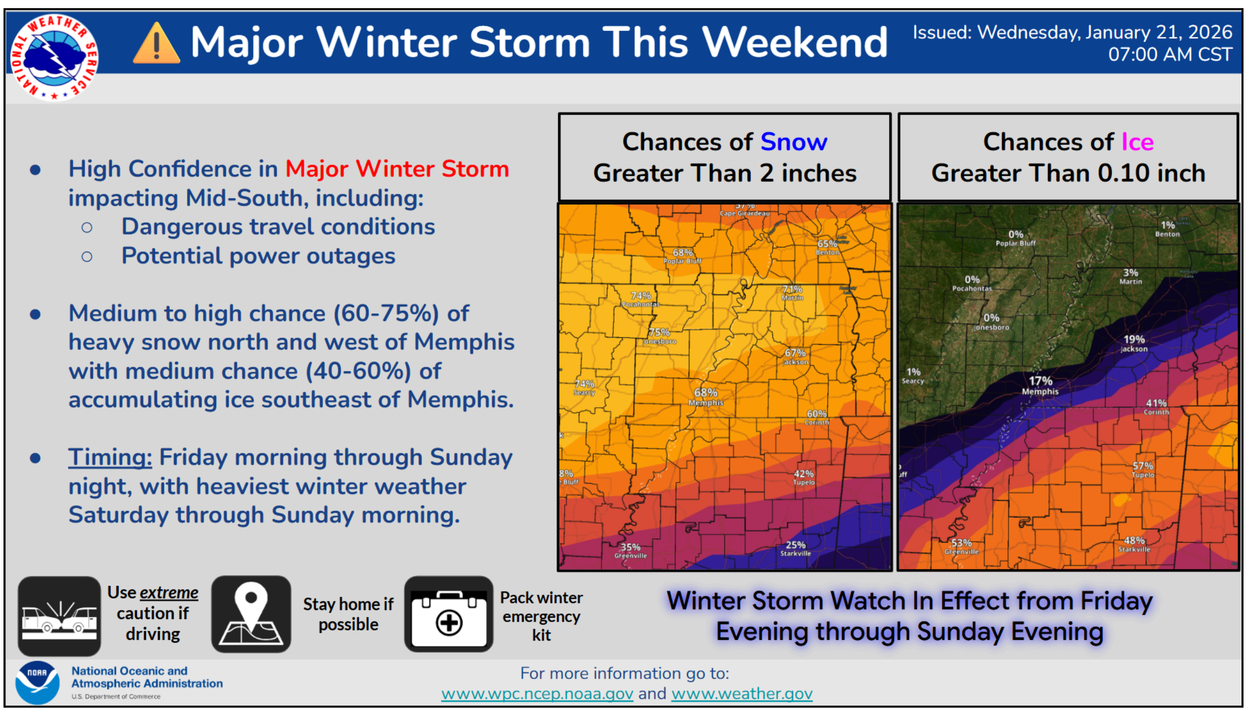

Figure 11. *Example Day 3 Key Message slide created for the Memphis, TN (MEG) CWA during the Afternoon Group session.*

b. Event 2 | 00 UTC February 23, 2026 to 00 UTC February 24, 2026

The second event analyzed during the testbed experiment occurred a month after Event 1 as a high-impact Nor'easter that affected much of New England (NOAA satellite loop shown by Figure 12). The events were similar in that both exhibited low predictability characteristics, but the features that made both cases challenging were quite different. Very subtle changes in storm track resulted in significant changes to snowfall totals, and subsequent travel impacts. Figure 13 gives a sense of the scope of high snowfall. This event was also notable in that it exhibited very high forecast uncertainty even within a 24-h lead time, where the February 22, 2026 11 UTC NBM snowfall seen through the Probabilistic Precipitation Portal still predicted a 10th to 90th percentile snowfall range of 4” to 29” for New York City, NY between 12 UTC February 22 and 12 UTC February 25, 2026. In addition to the high uncertainty at short lead times, model trends were erratic from cycle to cycle. The 75th percentile 24-h snowfall valid ending 00 UTC February 24, 2026 from the 100-member superensemble of GEFS, CMCE, and ECMWF members goes from 2” of snow from the 00 UTC February 20 initialization to 8” of snow for New York City from the 12 UTC February 20 initialization. Even the 12 UTC February 20 cycle ensemble envelope underforecasted the snow totals, with the LaGuardia airport reporting 22.5” of snow on the morning of February 24 and Warwick, Rhode Island shattering their state snow record with a whopping 37.9” snowfall report.

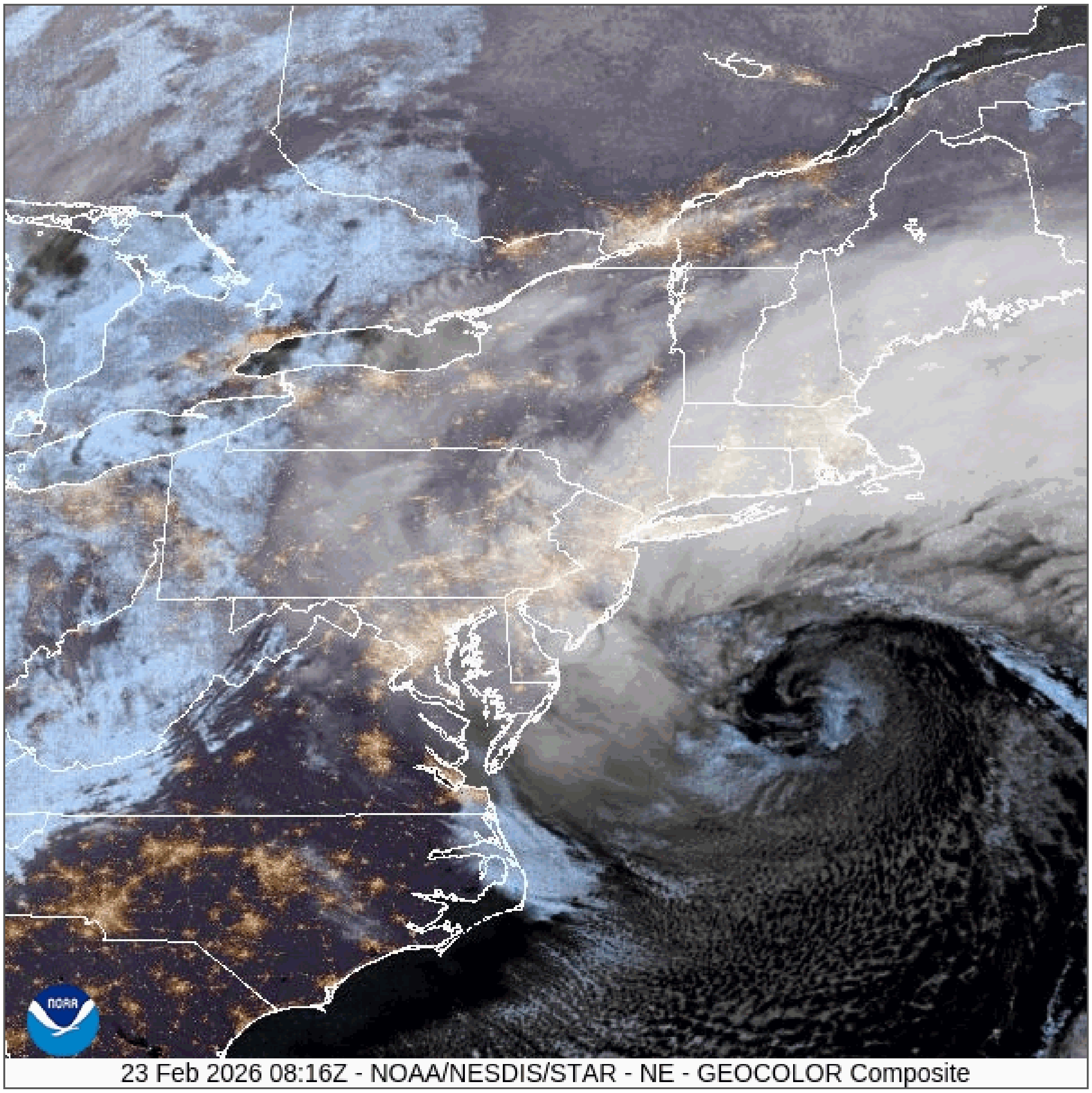

Figure 12. *Geostationary Operational Environmental Satellite 19 (GOES-19) loop through February 23, 2026. Image credit to the WPC Post-Event Review of this Nor'easter.*

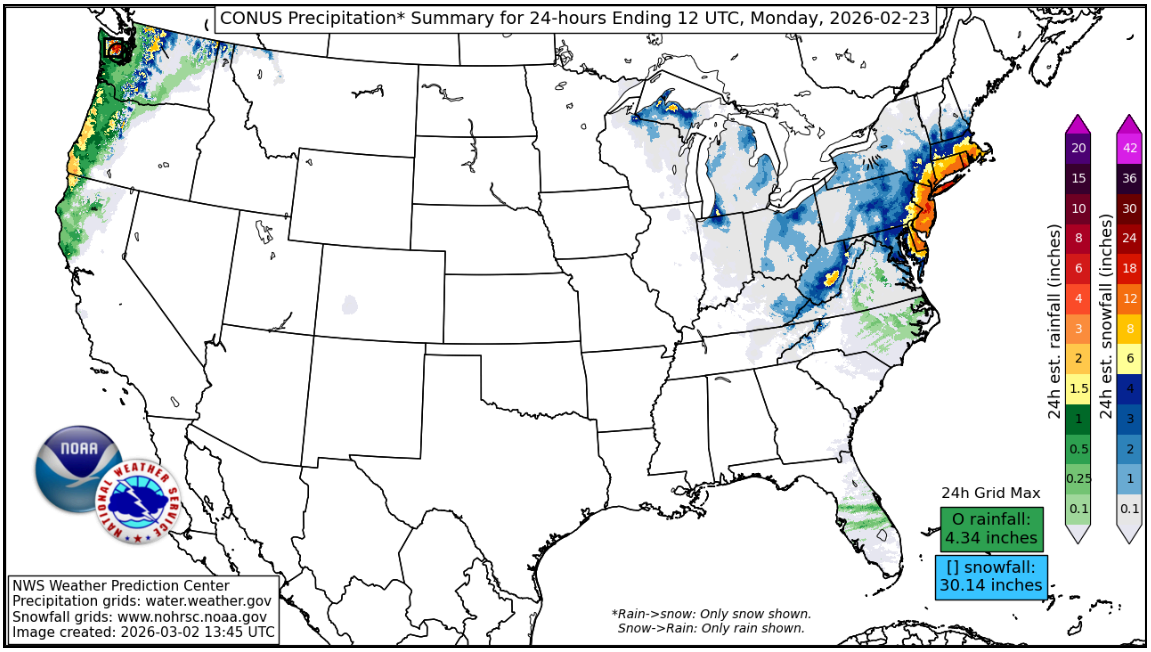

Figure 13. *As in Figure 2, except valid between 1200 UTC February 22 and 1200 UTC February 23, 2026 from the [WPC Winter Storm Archive](#).*

The substantial ensemble spread that persists well into shorter lead times combined with erratic model trends made this event a perfect candidate for performing data denial experiments and providing opportunities for participants to dive deeper into the ESA tools. Forecast activities for this case focused on generating forecast messaging for one of three assigned County Warning Areas (CWAs): Boston (BOX), New York (OKX), or Philadelphia (PHI). To understand the extent to which cluster and sensitivity variables used affected messaging, we restricted each CWA Group to one of three cluster variables. The Boston group could only use the WPC QPF Clusters and QPF PC-based response functions from the ESA graphics, the New York group was only able to use the WPC 500-hPa Height Clusters and related sensitivity graphics, and the Philadelphia group could only use the MSLP cluster and sensitivity fields. The use of DESI was prohibited for this event as well to keep from giving the New York group an advantage, as the DESI clusters are based on 500-hPa Heights. Similar to Event 1, Cody Snell and Mussie Kebede provided Day 6 and Day 3 forecast briefings for the Morning and Afternoon Groups, respectively.

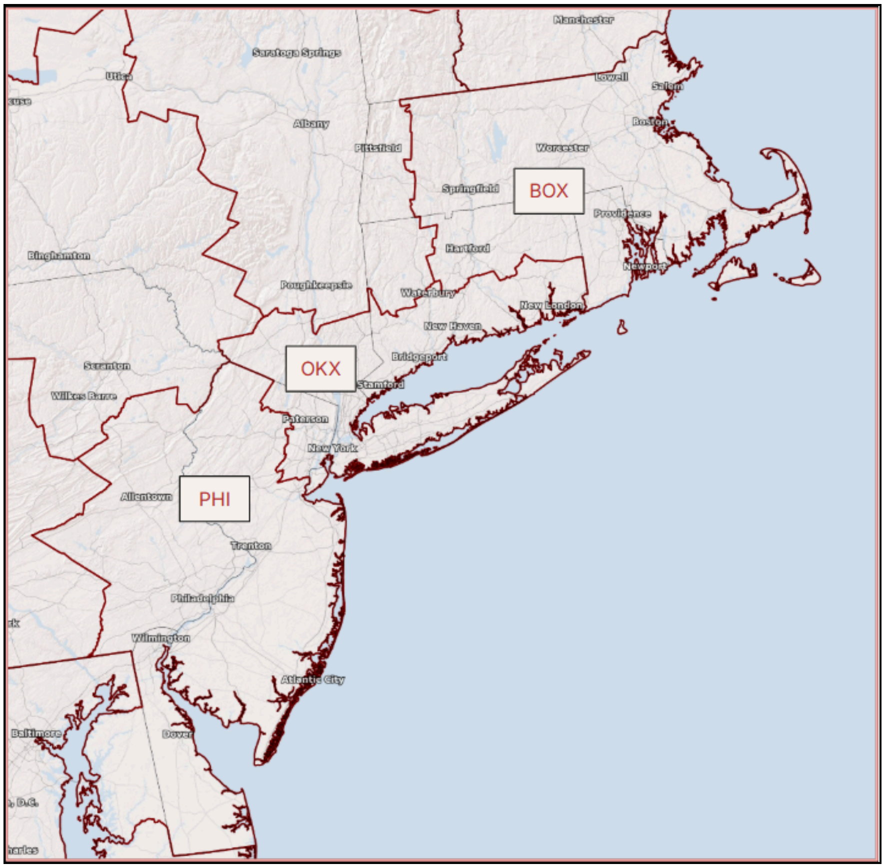

Figure 14. *Map describing the three CWAs used for Event 2 forecast activities. CWA boundaries are outlined in maroon, and abbreviated CWA titles are annotated in text boxes. The underlying geographic figure was generated with the Whole Story Uncertainty & Probabilities (WSUP) Viewer.*

I. Day 6 Forecast Activities

For the Day 6 lead time, there were data availability gaps in addition to the intentional data denial aspect of the experiment. Most of the WPC cluster data and a large chunk of WPC ESA data were missing from the 12 UTC February 17, 2026 initialization, leaving participants with 00 UTC February 17, 2026 data to focus on. To compensate for the data gap on Day 6, we allowed participants to use the WPC Ensemble Situational Awareness Tables (ESAT) and deterministic GFS and ECMWF ensemble mean data from Russ Schumacher's archive. The facilitators would like to show appreciation for the Colorado State University model archive that Russ Schumacher built and maintains, as it was very helpful for filling the data availability gap during this experiment.

From the individual forecaster assessment surveys, participants' mini AFDs generally reflected a "high-ceiling, high-uncertainty" forecast related to the track and strength of the system in place along the Northeast coast on Day 6. AFDs also featured a much higher frequency of ensemble sensitivity incorporation into the forecast process, with slightly over half of AFDs (11 / 20) containing mentions of sensitivity analysis, either explicitly or through discussion of the upstream sources of later forecast uncertainty. Regardless of assigned atmospheric variables, the consensus from participants who used ESA tied the uncertainty of the system back to a parent trough located south of the Alaskan coast at the beginning of the period. Multiple participants used ESA to specifically highlight that greater forecast confidence could be expected on February 20 once the parent trough reached the west coast, as noted in the following example AFD:

"Ensembles suggest a strengthening ern-CONUS trough in the Day 5-6 timeframe. Among ensemble members, there remains quite a bit of uncertainty related to the orientation/strength of this trough, with some suggesting a slower, more-amplified, more-neutrally tilted structure, while others suggest a less-amplified, faster, and more-positively tilted structure.

The eventual strength/amplitude/timing of this system is tied to the amplitude/position of a strong low pressure system approaching the Pacific NW. If this system develops stronger and modestly farther to the south, the ern-CONUS system could acquire greater strength and pose potential for a much greater wintry impact, as cold air is drawn sewd by the associated strengthing coastal sfc low.

Hiding within a washed-out ensemble mean H5 height field is a solution set that contains a much-stronger, closed H5 low that develops across the nern CONUS in the Day 5-6 timeframe. The QPF signal within ESAT does not appear exceptionally strong related to climatology, but there are modestly anomalous wind/height features associated with this system evident within medium-range ensembles, so it is not out of the question that a notable coastal system could bring substantial impacts to the I-95 corridor in a late-winter nor'easter. This muted signal could be the result of a low-mode, long-tail distribution, however. This is not yet a high-confidence forecast given the strength/timing/amplitude uncertainties, but there is a potential signal for a stronger coastal system hiding within a broad, uncertain forecast distribution..

As this forecast could take one of two key general tracks at this time, we expect to have more detail about the more likely strength/amplitude of the Pacific NW system in the next few days and will be able to provide greater detail about the eventual nern-CONUS system by Feb 20."

Once participants returned to their CWA Groups and began developing their Day 6 Key Message slides, they all mentioned difficulty in selecting a graphic to use in their messaging. There were not obvious "more likely" and "less likely" scenarios, with all possibilities on the table. While all groups across the Morning and Afternoon Groups agreed that there was signal for an East Coast trough at Day 6, slight shifts in storm track could move all impacts entirely off shore. High amplitude uncertainty with the Day 6 trough further enhanced the difficulty of messaging potential impacts.

CWA Groups across both sessions seemed to rely on the PC Phase Space diagrams more often than during the first event to better contextualize the ensemble spread for their own understanding. Participants from the Boston CWA Morning Group were quick to point out that the ECMWF ensemble subsystem solutions were all confined to a tiny area in the QPF Cluster PC Phase Space, suggesting extreme underdispersion. This apparent collapse of the ECMWF ensemble on a narrow precipitation solution space gave Morning Group participants pause, and the Boston CWA Group gave less credence to the ECMWF-heavy QPF Cluster 1 in their messaging, despite the stark 63% of superensemble membership contained in Cluster 1. In the Afternoon Group, one participant emphasized how far displaced the deterministic GFS sat relative to the GEFS members in 500-hPa Height PC Phase Space. Since the GFS featured a high-end forecast solution and people would be looking at it, the challenge for the CWA Group became one of "getting ahead of the narrative," in which forecasters acknowledge the GFS outlier in their messaging but contextualize it as one of many possibilities.

In generating Key Message graphics, many CWA Groups struggled without the ability to use DESI for scenario-based messaging graphics. They further mentioned wishing that there was a way to combine cluster scenarios and build simple, public-friendly graphics from ensemble guidance. The Philadelphia CWA Group in both Morning and Afternoon sessions decided to forego showing MSLP Cluster graphics entirely in favor of a conceptual low track scenario graphic (example from the Morning Group shown by Figure 15), possibly inspired by discussion of a recent UK Met Office scenario-based messaging graphic.

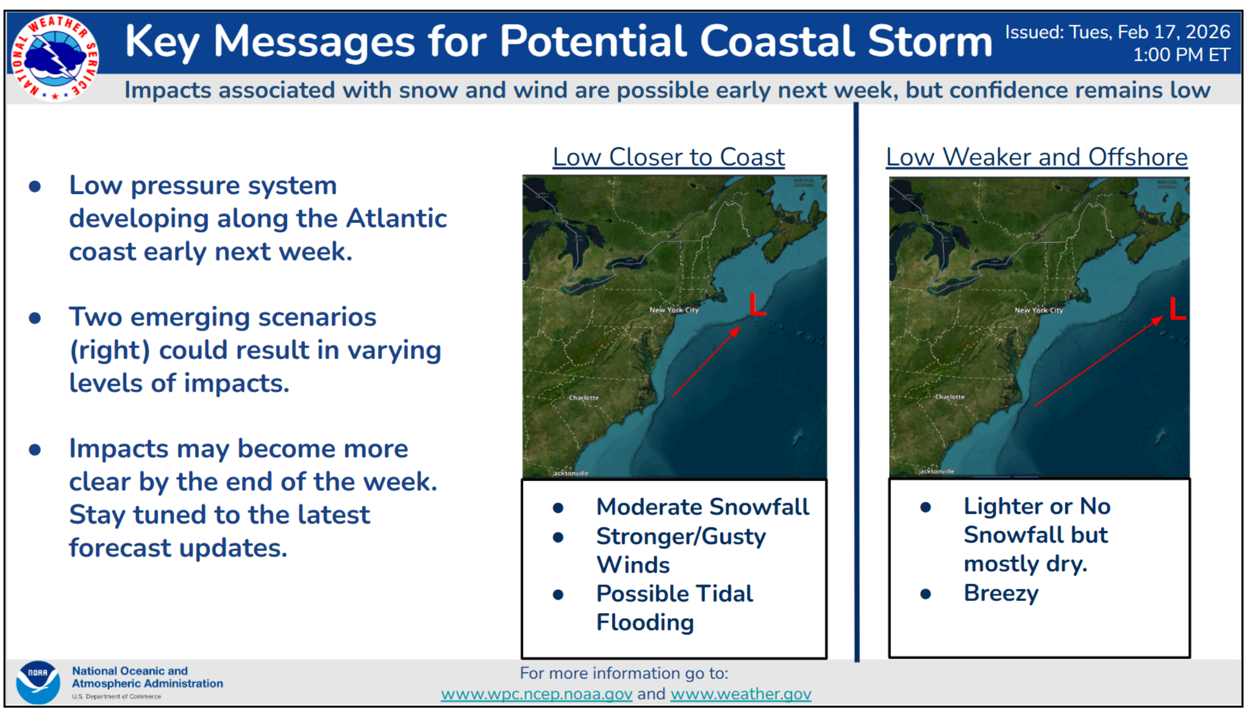

Figure 15. *Example Day 6 Key Message slide created for the Philadelphia, PA (PHI) CWA during the Morning Group session.*

II. Day 3 Forecast Activities

By the Day 3 lead time, confidence in a major Nor'easter had solidified as the ensemble envelope trended towards the coast. However, higher synoptic confidence gave way to low mesoscale confidence and sharp forecast precipitation gradients, where subtle differences in the placement of the surface low had massive implications for precipitation totals and impacts. Discussions depicted participants creating their own forecast scenarios by consolidating clusters with similar impacts to tell a coherent weather story. Many participants also noted a particular focus on conveying more confidence along the coast where impactful significant precipitation was near-certain regardless of the exact magnitudes. Across all CWAs, participants mentioned increased uncertainty for snow totals further inland. This example AFD focuses on the subtle differences between forecast scenarios and implications for the I-95 and the coast:

"Ensembles and associated clusters still show a slight positional difference but all are in good agreement in a rapidly strengthening low off the NE/Mid Atlc coast. Clusters 1/4 shows the deepest and further west while clusters 2 and 3 are further west. This is very important for snowfall across the area. The further west solutions would cause significant snowfall for the entire area including I-95 corridor while the east solutions would pin the heaviest snow at the coast. Messaging needs to indicate that the greatest likelihood of significant snowfall more than a foot will be at the coast along with blizzard conditions due to strong onshore flow. There is still a good chance for more than a foot of snow inland, but more uncertainty. Coastal flooding seems like a good chance given the strong NE flow in all of the guidance."

Back in their CWA Groups, participants using the QPF Clusters were surprised that the ECMWF ensemble system was still overconfident and encompassed a narrow corridor in the solution space. The Morning Group Boston CWA kept the ECMWF solutions out of their messaging, due to skepticism around this forecast behavior from the ECMWF. The Morning Philadelphia CWA Group called the coast a "slam dunk", but felt that the uncertainty increased exponentially further inland. The New York CWA Group across both sessions struggled to message the impacts at the Day 3 lead time while only using 500-hPa Height Clusters and Sensitivity Fields, since synoptic-scale differences aloft were very subtle but derived QPF clusters from the 500-hPa Height Cluster membership still revealed large QPF differences. Furthermore, sharp precipitation gradients were difficult to interrogate without a way to zoom in or show snowfall totals. One participant noted that having actual Snow-to-Liquid Ratio (SLR) values would have helped substantially in this case. They expanded that the NBM would have been useful here given that it accounts for the snow ratios.

Participant CWA Groups exhibited more variety with their messaging strategies at the Day 3 lead time for this event, likely at least partially owed to the data denial aspect of the forecast activities herein and the spatial variability of ensemble spread. The Philadelphia CWA Group in the Afternoon session opted to take an Outlook-style approach to annotating areas of enhanced certainty for impacts (Figure 16). There seemed to be consensus amongst participants that communicating the higher coastal confidence for more impactful snow and the lower confidence for impactful snow further inland was a primary challenge at this lead time.

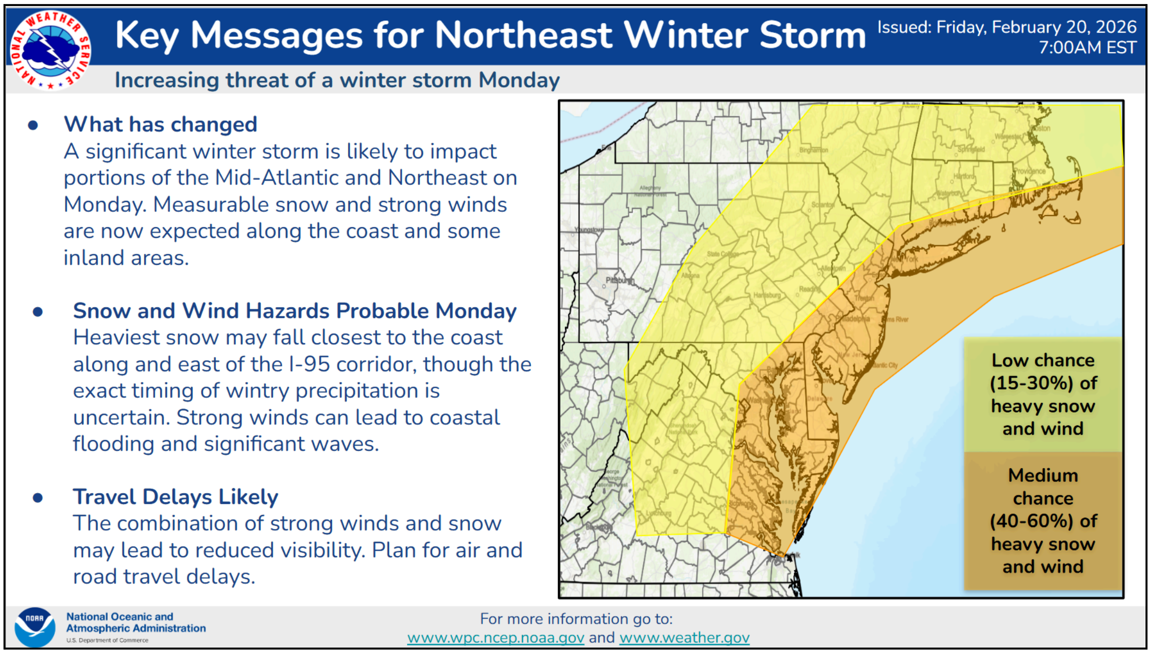

Figure 16. *Example Day 3 Key Message slide created for the Philadelphia, PA (PHI) CWA during the Afternoon Group session.*

III. Ensemble Sensitivity Analysis Activity

The last forecasting activity of the week was arguably the most challenging. The activity began with an ESA Demo in which facilitators (Austin Coleman for the Morning Group and Brian Colle for the Afternoon Group) briefly went through interactive examples of Event 2 ESA and Cluster trend interpretations over various lead times. The goal was to provide guidance on how to interpret and combine sensitivity and cluster tools before setting participants loose to develop their own ESA and Cluster Trend briefing slides. Following the demonstration, participants regrouped with their original Event 2 CWA Groups with the goal of highlighting:

1. relevant trends in [assigned variable] cluster forecasts
2. which features the [assigned variable] forecasts are most sensitive to
3. any trends in those sensitive features

Different groups addressed these goals in a variety of ways. The Morning Boston CWA Group noted that the NBM QMD members used for WPC QPF Clustering behave fundamentally differently than the raw global ensemble QPF used in ESA response functions, which can lead to a disconnect that reduces cross-dataset interpretation. Tyler Wixtrom from the Meteorological Development Laboratory (MDL) confirmed in the

Afternoon Group that the two datasets need to be treated separately, and added that NBM QMD can face bias correction issues in the transition and shoulder seasons. Other groups brought up the importance of knowing ESA limitations as interpretation guardrails. Given that coastal lows such as that seen in Event 2 can evolve nonlinearly, it is important to keep in mind that ESA only shows a linear estimation of some of these extremely complex interactions in the atmospheric flow. Finally, a few Afternoon Group participants focused on the evolution of Day 6 WPC 500-hPa Height Cluster 4 (outlier deep, slow trough scenario) to a near-ensemble mean scenario once we reach the Day 3 lead time. Mark Klein's post-event verification overview revealed that Day 6 WPC 500-hPa Height Cluster 4 was by far the best forecast scenario for 500-hPa heights, with a 24 meter MAE relative to other scenarios with MAEs between 40 and 65 meters. Participants noted that model trends and cluster trends were fairly reliable for this forecast situation, but solely messaging the 500-hPa height cluster scenarios with larger membership on Day 6 would have led the forecast messaging astray.

Hosting Steve Willington's seminar pertaining to ensemble exploitation strategies at the UK Met Office immediately before these ESA Activity discussions additionally sparked a lot of fascinating conversation. One of the key seminar points centered around "demonstrating value," and that these cultural shifts of moving from deterministic forecast guidance to ensemble forecast guidance cannot happen simply by telling forecasters to use these tools. The value of ensemble tools over deterministic tools needs to be proven for forecasters. In the context of this testbed, the utility of ESA organically revealed itself throughout the week. After the ESA activity, one of the Afternoon Group participants pointed out that ESA helps with credibility. Forecasters shrugging off high forecast uncertainty can look bad, but saying specifically why the forecast is uncertain and when we expect to know more helps build trust. The facilitators were somewhat surprised by this use case for ESA, given its technicality, but participants seemed more confident using ESA "under the hood" in a sense and messaging the key takeaways from the guidance without going too far into the weeds.

3. Pre-Experiment to Post-Experiment Evolution

The testbed experiment fostered a stronger cultural shift in the participants than expected. Most participants walked away with a much deeper understanding of how to use ensemble clusters beyond solely interpreting the basic cluster mean visualizations. They reported a significant increase in their conceptual understanding of EOFs and the PC Phase Space, stating that the activities successfully "illuminated more of the how and why" behind these diagnostic tools. This technical growth enabled forecasters to identify "red flags" in the data, such as high intracluster spread or scenarios dominated by a single ensemble subsystem, which allowed them to move beyond a blind trust in cluster means. The DESI platform was central for many of these "red flag" diagnoses, since it allowed participants to interrogate cluster spread from both a plan-view perspective as well as in the form of ensemble point soundings. Retaining this degree of flexibility in ensemble cluster visualizations will be crucial to the utility of cluster tools in any future NWS Visualization platforms.

The experiment also facilitated a cultural shift away from the common, tempting practice of “picking-a-winner”, in which forecasters tailor their messaging to the cluster scenario with the highest ensemble membership. This practice, while potentially worthwhile at shorter lead times, prematurely turns a probabilistic forecast tool back into a deterministic forecast tool when applied to medium-range solutions. While pre-experiment responses reflected a tendency to gravitate toward either a single scenario deemed “most likely” or a situationally-dependent messaging strategy that included a “most likely” scenario and an alternate scenario (Figure 17), post-experiment feedback indicated a consensus to embrace messaging multiple scenarios. Along these lines, participant feedback during the Pre-Experiment Survey indicated that many of them felt comfortable with using the ensemble clusters, but not with translating them into actionable messaging for the general public or core partners. Participants felt that they lacked the resources to connect the clusters back to ensemble sensitivity analysis tools in order to better communicate the “why”. The following survey response to the question: “Are there any specific topics you would like us to cover during the experiment?” summarizes this common sentiment nicely:

“I’d especially like more coverage on the “last mile” of Ensemble Clustering and Sensitivity Analysis: how to turn the results into clear, practical messaging for partners. Since the output can be very technical, I’d like to learn best practices for communicating uncertainty and alternate scenarios in a simple, jargon-free way.”

This sentiment came up often throughout the week, with participants highlighting the need to adopt a more holistic probabilistic framework that centers adding valuable context to forecast uncertainty over providing the most accurate deterministic forecast grids. One participant from the Morning Group described this framework with a “railyard” analogy, where they compared the current strategy of creating deterministic forecast grids at longer lead times to an attempt to pick which rail a train will end up on several days in advance. They went on to describe the need for a cultural shift in which the forecaster’s goal becomes giving stakeholders and the general public a sense of the layout of the railyard itself rather than picking a particular track. The deliverable would be to describe the railyard in a way that is actionable. In the analogy, ensemble sensitivity analysis would be used to identify when the track splits in the railyard and when to expect further information as the bifurcation in the forecast approaches. In this way, the lead time from the event plays a lesser role in driving messaging strategy than the predictability characteristics of the event. Defining the railyard allows forecasters to move away from leadtime-based messaging strategies to predictability-centric messaging strategies specific to the uncertainty characteristics of the event itself. Participants from both groups generally echoed these sentiments.

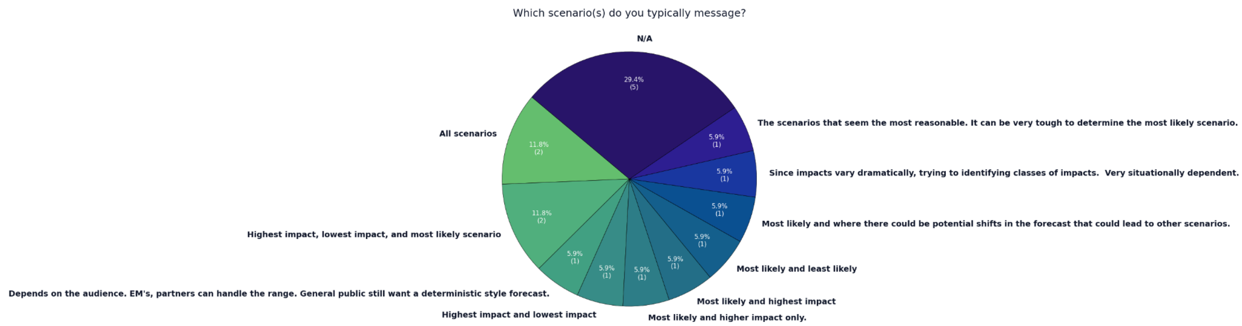

Figure 17. *Pre-Experiment responses to the question: "Which scenario(s) do you typically message?"*

The starkest differences between pre- and post-experiment survey responses were found in the feedback regarding ESA. Participants swung from seeing it as a research-centric diagnostic to an operational tool that identifies exactly when to expect an increase in forecast confidence. Sharpest increases in familiarity and understanding of ESA products were found with participants who came in with low familiarity of ESA (Figure 18). This theme was reflected by several participants who, independently across both groups, exhibited "lightbulb moments" during the week. For example, while forecasting the late February Nor'easter event, one participant enjoyed using ESA for the event at the Day 6 lead time. They elaborated that ESA made it easy for them to say: "Of course this winter storm is uncertain! The trough responsible for it is still south of Alaska!" Furthermore, they learned from ESA to expect greater clarity in the Nor'easter forecast once the parent system approached the west coast around the Day 3 lead time. Representative examples from the testbed Jam Board session regarding the use of ensemble sensitivity in forecast messaging are shown by Table 2.

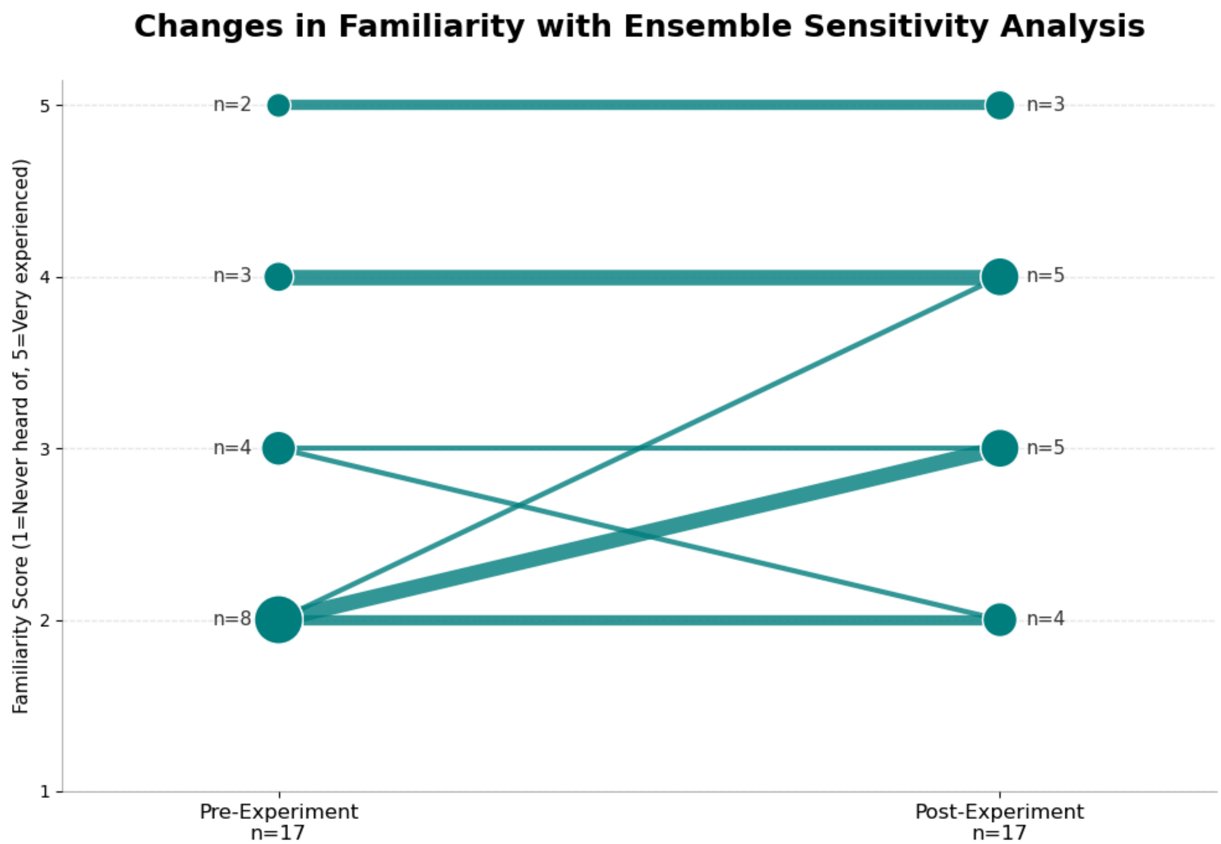

Figure 18. *Evolution of participants' responses to the question: "How familiar are you with Ensemble Sensitivity Analysis?" before the testbed experiment and after the testbed experiment on a 1 to 5 scale. A score of 1 indicates that the participant has never heard of ESA whereas a score of 5 indicates the participant is very experienced with ESA. Thickness of the lines between familiarity scores corresponds to the number of participants who changed from one pre-experiment score to another post-experiment score.*

Selected Participant Responses to the Jam Board Question: "How can ensemble sensitivity influence your forecast messaging?"	
Morning Group	**Afternoon Group**
It can result in more informed confidence communication and introduce a timing component to when things may become more clear.	*Not only do I have a better understanding of WHY there are differences in output from the grand ensemble, but now I can get an idea of WHEN I may have better clarity. This is forecast gold to the partners!*
It helps me understand the source(s) of uncertainties that grow with time, and helps inform when I might know better when the goal posts can be narrowed around my uncertain time periods of the forecast (e.g. what do I need to see to know which of many tracks the forecast is	*It can help drive the messaging of uncertainty. Frequently uncertainty word is used in isolation without any real context. Ensemble sensitivity influence has the ability to help explain what is uncertain about a given pattern which could become an integral (if not prominent) facet of*

Selected Participant Responses to the Jam Board Question: “How can ensemble sensitivity influence your forecast messaging?”	
on)	*future IDSS.*
It allows for a better understanding of where/when that source of uncertainty is located at, and when I may be able to provide better information. Perhaps not the complete answer, but it allows me to narrow down the potential answers (goalposts)	*We can message what features we are watching to put the forecast into perspective and build trust/transparency.*

Table 2. *Selected remarks in response to the question: “How can ensemble sensitivity influence your forecast messaging?” from the Jam Board brainstorming session.*

By the end of the week, the vast majority of respondents expressed the intent to use ESA more frequently to explain the "why" behind forecast variability (Figure 19). One SOO participant even mentioned a plan to give a seminar to their WFO about using ESA in operations after the experiment. This feedback was particularly encouraging given that the participant reported an ESA Familiarity Score of 2 / 5 before the experiment.

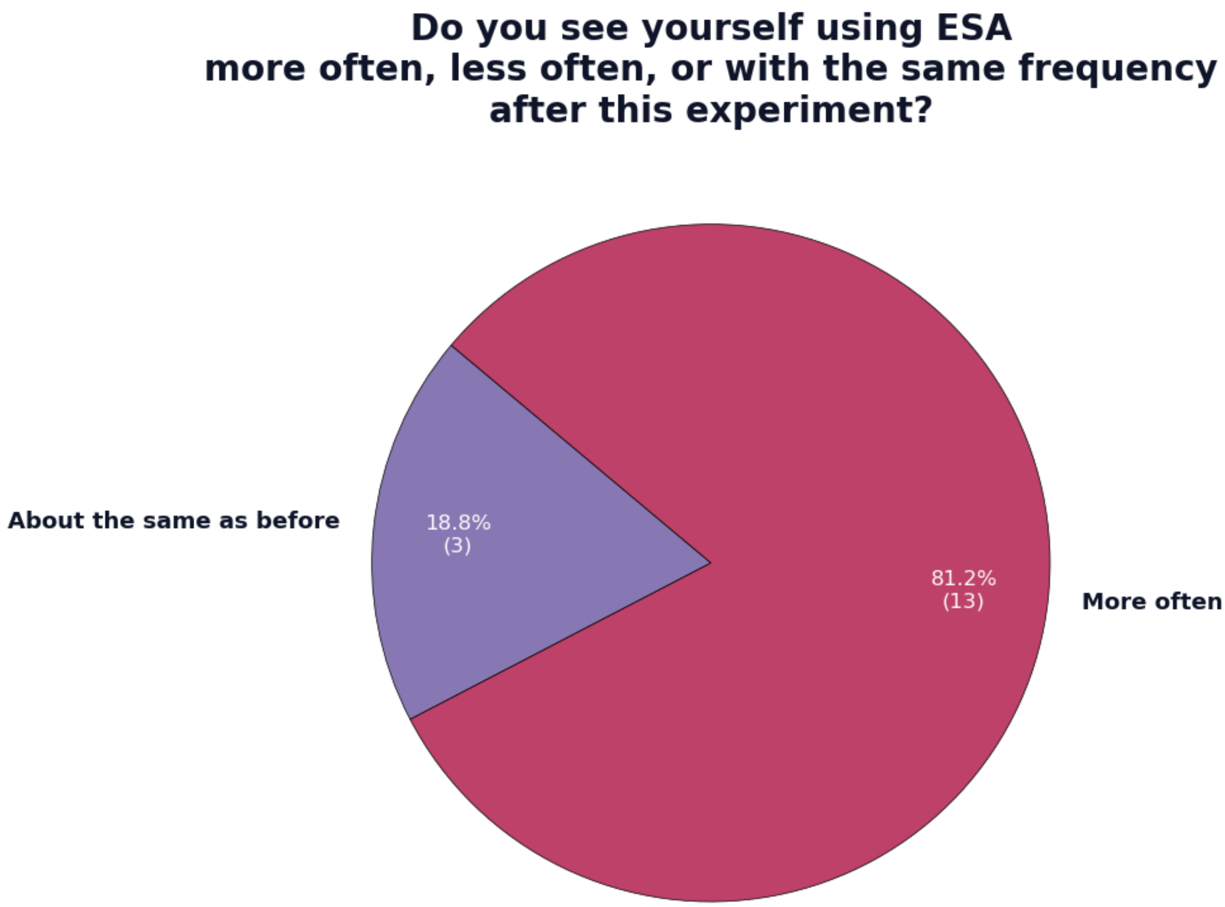

Figure 19. *Post-experiment survey responses to the question: "Do you see yourself using ESA more often, less often, or with the same frequency after this experiment? (You can be brutally honest with us, here!)".*

4. Forecasting Best Practices

a. Using Ensemble Clusters

i. PC Phase Space can be very helpful for analyzing dispersion characteristics of the super ensemble. For example, if membership from an ensemble subsystem (e.g., GEFS or ECMWF) is clustered in one quadrant of the phase space at long lead times, it can signal to the user that the ensemble subsystem may be underdispersive or untrustworthy for that event.

ii. The WPC 500-hPa height, WPC MSLP, and DESI clusters all have their number of cluster scenarios preset to four to optimize cluster mean verification metrics. In practice, however, there may be more or less than 4 meteorologically distinct clusters. When in doubt, interrogate intracluster spread to determine if multiple physically-distinct scenarios are being averaged into a single cluster. This analysis may be accomplished by examining the PC Phase Space more closely or by examining ensemble spread statistics of the clusters through DESI.

- An example of this phenomenon can be found in the Event 1 Day 5 discussion above.

iii. Having equitable representation from each ensemble subsystem within a given cluster can provide confidence in a particular scenario. Conversely, cluster scenarios dominated by membership from a single ensemble subsystem can indicate potential underdispersion.

iv. This feature is currently specific to WPC Ensemble Clusters and may be worth integrating into future releases of DESI or NWS Viz. Viewing deterministic global model forecasts in the context of the PC Phase Space with their global ensemble counterparts can be a good indicator of how representative a deterministic forecast is relative to its corresponding ensemble subsystem.

v. Analyzing model trends (d[model]/dt) via clusters is more efficient than interrogating individual members or ensemble means, allowing for a rapid assessment of how the ensemble envelope is shifting over successive initializations.

vi. If a meteogram depicts high ensemble spread in a given variable, that suggests that a deeper dive into the spread via ensemble clusters may be warranted. In situations with low ensemble spread, clusters may not be necessary in the forecast process for the given event.

b. Using Ensemble Sensitivity Analysis

i. The primary use of ESA is to explain the underlying mechanisms of forecast event uncertainty and identify which features (e.g., a Pacific trough) are driving downstream spread. It also identifies when higher confidence can be expected, specifically when the responsible feature enters a dense observation network.

ii. The customizable ESA floater domain application is preferred to gain specific, focused understanding of uncertainty contributors for localized high-impact weather events. [Side anecdote - forecaster interest in the floater domain application has increased significantly after the completion of this testbed experiment.]

c. Scenario-based Messaging

i. It is important to keep in mind that just because one cluster has the largest membership, that doesn't necessarily mean that it is the "Most Likely" given the reliance on an uncalibrated ensemble of opportunity. Ensemble cluster products based on a calibrated ensemble system will theoretically not possess this limitation. In situations where a "most likely" scenario is not obvious, it is best to mention 2 or 3 plausible scenarios along with their associated impacts.

ii. Relatedly, it is best to avoid messaging overly-specific percentages with the scenarios for cluster products based on uncalibrated ensemble systems, as the percentages may not be representative of the true likelihood of occurrence.

iii. Key messages with simple, annotated graphics that focus on potential impacts rather than specific numeric amounts were perceived as the most effective. Removing any technical jargon (e.g., replacing "system" with "storm") can go a long way in improving comprehension.

iv. For core partners and decision makers, communicating the sources of forecast uncertainty and when to expect more clarity through ESA can build trust and transparency.

5. Future Directions

a. Near Term Technical To-Dos for Cluster and Sensitivity Products

- **Unified WPC Ensemble Cluster and Sensitivity Analysis Platform**
  Participants really liked the ECSA Testbed website, since it combined all the cluster applications into a one-stop shop. Adapting this page for realtime applications and adding the d[prog]/dt feature would be an easy win for participants who overwhelmingly mentioned the need for one unified application.

- **Consistent EOF Pattern Signs**
  Ryan Torn mentioned that EOFs for Atmospheric River (AR) Reconnaissance missions are normalized such that the signs are consistent from initialization to subsequent initialization. Forcing EOF

patterns to exhibit greater consistency would drastically improve interpretability of EOFs and ESA plots based on PC response functions.

- **Event Driven Time Windows**
  Participants mentioned that the 00 UTC - 00 UTC 24-h average window often temporally sliced an event in half, making it very inconvenient for incorporation into forecast messaging without using DESI. Participants in the Morning and Afternoon Groups alike indicated a preference for a 12 UTC - 12 UTC 24-h window for the cluster and sensitivity tools. They further mentioned a preference for event-driven time windows that capture the duration of a high-impact weather event.

- **Third EOF / PC Toggle for Advanced Users**
  Several of the advanced cluster users lamented that the first two EOFs and PCs often did not tell the whole forecast uncertainty story. This aspect is most obvious with the QPF Clusters, in which many nuanced uncertainty modes exist leading to lower explained variance in the leading two EOFs relative to other cluster variables. Whether a cluster application that relies on three PCs instead of two is likely a future science question that would require user input as well. Parallel cluster applications relying on different numbers of PC dimensions would constitute an obstacle in consistent scenario-based messaging.

- **Basic Response Function Options in ESA Application**
  While the facilitators hypothesized that there would be a relatively steep learning curve for newer ESA users to begin applying it in their forecast messaging, the use of PC-based response functions in the ESA products proved to be a larger blockade to understanding than initially thought. Many participants less familiar with using the EOF patterns found themselves struggling to mentally match EOF signs to sensitivity signal signs, and if the EOF sign flipped with the subsequent ensemble initialization, they often had to restart their interpretation. Basic response functions (e.g., QPF Max in a response function region, Coverage of High 10-m Wind Speed in a response function region, etc.) can more directly connect the meteorology to the ESA product for newer users, providing a more accessible entrypoint for less experienced users in addition to insightful complementary sensitivity products for experienced users.

- **Day 2 and Day 1 ESA Products**
  Several participants mentioned that the added situational awareness of subtle upstream features important to the details of the forecast even at shorter lead times would be beneficial to the forecast process. This addition would be fairly straightforward given that short lead time options already exist for the floater domain application.

b. Future Ensemble Product Wishlist

- **Convection-allowing Ensemble Clusters and Sensitivity Analysis**
  Many participants brought up the need for more high-resolution cluster and sensitivity products, as well as customizable domains and forecast windows for these applications.

- **Expansions to DESI or NWS Viz**
  - Participants want to be able to interrogate the additional WPC Cluster Variables on the DESI platform and/or NWS Viz.
    - For instance, a participant who used the MSLP Clusters to forecast for Event 2 explained that they would have liked to view snowfall diagnostics from the MSLP Clusters specifically given that uncertainty in surface low track was the primary driver of snowfall totals.

  - Multiple participants also expressed a desire for the incorporation of ensemble sensitivity analysis products into future iterations of DESI and/or NWS Viz.

- **Expanded WPC Cluster Variables**
  - Precipitable Water (PWAT) and Integrated Vapor Transport (IVT) were mentioned frequently as potentially valuable cluster variables, especially for Atmospheric River (AR) prediction.

  - Dewpoints or Relative Humidity (RH) for boundary uncertainty quantification or fire weather prediction.

  - Snowfall, SLR, Freezing Rain clusters would all be useful in diagnosing snowfall amount or precipitation type uncertainty.

  - Cluster variables that quantify impacts over specific atmospheric conditions. Ensemble clustering within the probabilistic Winter Storm Severity Index (WSSI) comes to mind here.

- **Incorporate intracluster spread to determine the optimal number of clusters**
  One theme that emerged throughout the week was that keeping the number of clusters preset to four in order to maximize cluster mean verification metrics could be situationally-misleading. Participants expressed a desire for a cluster application that employs a varying

number of clusters and relies on ensemble spread in order to determine the optimal number of clusters for a given atmospheric event.

c. Training Needs

- **Show deterministic guidance in ensemble context to demonstrate value of a probabilistic framework**
  One SOO participant mentioned that she uses the PC Phase Space diagrams to demonstrate the utility of an ensemble framework for new forecasters, given that the deterministic guidance often falls outside of its ensemble counterpart envelope in the PC Phase Space. Deterministic guidance should be plotted in phase space for each of the cluster applications and training modules should implement this strategy to more effectively convey the need for probabilistic guidance.

- **Training for identifying “red flags” in cluster output**
  There is a need for training that moves beyond teaching the cluster fundamentals and into the details of identifying when cluster output is untrustworthy, as discussed further in the Forecasting Best Practices section above.

- **Ensemble sensitivity analysis training for different experience levels**
  There is a great deal of variability in users’ experience levels with ESA, and future training modules should reflect that. One participant suggested developing short videos to demonstrate using ESA for everyday forecasting. These videos should be developed for an array of different experience levels. Some participants from the Afternoon Group discussed the need to acknowledge that not every forecaster needs to be an ESA expert. There is a need for a “bench” of subject matter experts at the regional level or national level with a strong grasp of these concepts, with fewer meteorologists at the local office level needing as deep of an ESA understanding.

- **Cross-WFO and WFO-NCEP collaboration training**
  One question that remained open-ended involved asking participants about the kinds of WFO-WFO or National Centers for Environmental Prediction NCEP-WFO collaboration needed for messaging a variety of atmospheric situations. Participants had many different ideas on how to establish best practices for scenario-based messaging collaboration, some of which are summarized by Table 3. There did seem to be some consensus that national centers should provide broader guidance on trends and overarching synoptic scenarios moving forward, with WFOs refining the messaging for local impacts. WFO-WFO collaboration

strategies were less clear-cut. Workshops incorporating further social science perspectives would likely be needed to identify the best way forward for these interactions, particularly in the broader context of NWS Transformation.

Selected Participant Responses to the Jam Board Question: "In the context of messaging forecast scenarios, how should cross-CWA collaboration look? How about NCEP-WFO collaboration?"	
Morning Group	**Afternoon Group**
Collaboration of messaging is a lot more complex than just trying to come to agreement on a deterministic forecast outcome. We have to agree on the relevant features, the types and sources/origins of uncertainties, and where to place goalposts around plausible impactful outcomes. And on top of that, we have to then agree on strategies for conveying that information in meaningful ways. In some ways it's easier in that we don't have to have entirely consistent agreement on potential outcomes because there is room in the probabilistic space for multiple messages. There's no one right way to message a particular event (although there may be wrong ways), and focusing in on the story we want to tell is key. If we can agree on that part, how we tell it may matter less.	*There seems to be a misconception that people are collaborating on picking a "favorite" cluster or something, but rather we are collaborating on (1) which scenarios to message, and (2) how to graphically represent the impacts (e.g., long range is the map of where the low tracks, short range is the prob exceedance, near term is the most likely and high end). NCEP can advise on trends, provide broader context, and ensure that the timing of the messaging is consistent across future sectors, time zones, etc. Y'all deal with a lot!*
I think that it would make more sense to have messaging done on the level of scale. Have an ERS at the larger scale take care of FEMA regions, have state level ERS have messaging at a state level for all state partners, and WFO at a county level. Tailoring the message to the level that is needed for the correct audience. The message for something as large as FEMA is going to be different than a county level type partner, EMA, DOT, or general public. Not having too many hands in the cookie jar so to speak. I still don't like the idea of WFOs being these arbitrary boundaries that mean nothing to the general public or partners and wish they were better matched geographically (state boundaries or by tv coverage areas, or some other boundary that isn't some political or otherwise made up boundary from 30 years ago). That being said I think having a platform	*I'm uncertain. However, I think the coordination between the I-95 offices could be a model for that type of future WFO-WFO (and to some degree CWSU/RFC). Training and tools will be necessary.* *NCEP to WFO collaboration may largely remain unchanged? However, these tools/techniques can be infused into that collaboration, but it'll need to be seamless/flexible.*

Selected Participant Responses to the Jam Board Question: "In the context of messaging forecast scenarios, how should cross-CWA collaboration look? How about NCEP-WFO collaboration?"	
where we are grabbing images all from the same platform will help in collaborative messaging.	
This would most easily be accomplished by having something similar to the 'Expert Hub' that the UKMET folks have. A way for a regionalized/sub-regionalized scenario assessment was passed down to WFOs to fine-tune impact communication within those scenarios.	*NCEP-WFO collaboration can look like, for example, WPC providing ESA of upstream synoptic features and how they influence downstream impacts illustrated by the clusters. Whereas, WFOs can share what clusters are showing at the regional level and how that influences their local messaging.*

Table 3. *Selected remarks in response to the question: "In the context of messaging forecast scenarios, how should cross-CWA collaboration look? How about NCEP-WFO collaboration?" from the Jam Board brainstorming session.*

6. Concluding Remarks

The first 2026 Ensemble Clustering and Sensitivity Analysis (ECSA) Testbed successfully demonstrated that the routine integration of advanced ensemble diagnostic tools into the operational forecast process can help facilitate not only enhancements to the forecast process and trust with core partners, but a cultural shift in the NWS forecast process as well. By engaging researchers, SOOs, and forecasters in a collaborative virtual environment, the experiment revealed a way to transition from a more deterministic messaging mindset to a scenario-based probabilistic framework. This path forward is exemplified most acutely with the use of a "railyard" analogy to describe the future of forecast messaging. Participants concluded that attempting to pick a single deterministic "rail" several days in advance is increasingly counterproductive. Rather, the most value can be added to the forecast by defining the "layout of the railyard". In other words, forecasters can comprehensively explain the various tracks a storm may take and incorporate ESA to identify bifurcations in the forecast (when the "track" splits) and when greater forecast confidence can be expected. Adopting this framework would allow the predictability characteristics of the specific event, instead of the lead time, to drive the messaging strategy.

While initially viewed as a technical research tool, ESA surprisingly became co-opted as a communication and IDSS tool by the end of the week. Forecasters discovered that by tracing downstream forecast uncertainty back to specific upstream features, such as the trough south of Alaska identified for Event 2, they could provide users with the "why" behind the uncertainty along with a timeline of when to expect greater clarity. This type of utility was even described at one point as "forecast gold". This diagnostic empowers the forecaster to do meteorological "detective work", which will be increasingly valuable as the role of the forecaster evolves toward adding context and providing IDSS to partners. The experiment further established "red flags" for cluster interpretation, and developed a set of future ensemble visualization and training needs for the forecasting community at both the medium- and short-range lead times. Collaboratively, the testbed suggested a new model for NCEP-WFO interactions surrounding scenario-based messaging strategies. They recommended a tiered approach where National Centers provide the broad synoptic scenarios to message along with their respective sources of uncertainty via sensitivity, while WFOs refine those scenarios for localized impacts.

Insights from the 2026 ECSA Testbed Experiment have provided a clear path forward for advancing probabilistic forecasting and messaging strategies. The findings herein outline a blueprint for near-term specific product updates and long-term development and training initiatives. As the agency continues implementing NWS Transformation, these tools and the cultural shift that they foster will be essential for translating complex forecast knowledge into simple, actionable messaging for NWS core partners and the general public.

7. Acknowledgements

Lastly, we would like to acknowledge the incredible contributions by each participant of this testbed experiment. We were fortunate to learn from such an insightful, engaging group of forecasters and researchers, and are exceedingly grateful for each of them. The facilitators are particularly appreciative of the invited seminar speakers who provided critical context during the experiment week, Ryan Torn, Steve Willington, and Brian Ancell. Thanks are owed to Travis Wilson for his contributions in implementing ensemble clusters onto the publicly-available DESI platform in time for the experiment. We appreciate Massey Bartolini for his thoughtful suggestions that improved the final report. Finally, we would like to thank James Correia Jr. and Sarah Trojniak of the NOAA HMT for their valuable guidance in the design of this experiment.

Research Participants

1. Brian Ancell
2. Ryan Torn
3. Dana Tobin
4. Leticia Williams
5. Steve Willington
6. Kim Hoogewind
7. Bryan Guarente
8. Leif Swenson
9. Diana Stovern
10. Tyler Wixtrom

NWS SOO Participants

11. Matthew Kramar
12. Donal Harrigan
13. Michael Dutter
14. Greg Mann
15. Jared Klein
16. Stephen Harrison

17. Lamont Bain
18. Bob Setzenfand
19. Trent Davis
20. Samuel Shea
21. Katie Magee

NWS Forecaster Participants

22. Ashley Novak
23. Derek Hodges

WPC Forecaster Participants

24. Cody Snell
25. Mussie Kebede

WPC DTB Helpers

26. Bruce Veenhuis
27. Kirstin Harnos
28. Massey Bartolini

Facilitators

29. Austin Coleman
30. Mark Klein
31. Jim Nelson
32. Brian Colle

<u>Demo Day Participants</u>

The facilitators would also like to acknowledge the participants who took part in the experiment before the experiment (Demo Day - February 2, 2026), where research questions and experiment activities were refined ahead of the full-length testbed.

1. Stephanie Avey —— NWS SOO
2. Russell Danielson — NWS Forecaster
3. Matthew Kramar —— NWS SOO
4. Katie Magee ———— NWS SOO
5. Bill Rasch ————— NWS SOO
6. Paul Schlatter ——— NWS SOO
7. Josh Weiss ———— WPC Forecaster