NWS Operations Proving Ground

Experiment Plan, Expectations, and Instructions

Probabilistic IDSS Evaluation

An OPG Virtual Experiment Evaluating the DESI and WSUP Probabilistic Data Viewers

NWS Operations Proving Ground
7220 NW 101st Terrace
Kansas City, MO 64153

1. Background and Evaluation Design

In March 2023, the Operations Proving Ground (OPG) conducted a two day virtual evaluation with 66 participants focusing on the use of experimental data visualization tools to produce probabilistic IDSS information. The Probabilistic IDSS Evaluation (PIE) objectives originated in 2021 when the Analyze, Forecast and Support office (AFS) asked the OPG to test the impact of probabilistic information on end user decision making. In particular, AFS wanted to know if forecasters could adequately apply recommendations from a NOAA funded project by the University of Oklahoma[1] that studied optimal strategies for communicating uncertainty information.

Then, in 2022, the Office of Science and Technology Integration (OSTI) asked the OPG to evaluate the experimental Dynamic Ensemble-based Scenario for IDSS (DESI) platform and the Whole Story Uncertainty and Probabilities Viewer (WSUP) to determine their operational readiness. OSTI also asked the OPG to evaluate probabilistic messaging to core partners and the value of cluster analysis in the forecasting messaging process.

In the PIE project plan for this evaluation, the OPG set out to answer several questions:

1. Should the DESI and WSUP viewers become operational?
2. How do forecasters feel about using the DESI and WSUP viewers as part of their decision making process?
    a. What do forecasters like or dislike about each tool, and what would they recommend to the tool developers?
3. Can we improve forecast efficiency by reducing data overload in the forecast process?
    a. What are the impacts on the probabilistic forecast and messaging processes if forecasters are limited to using three diagnostic tools (DESI, WSUP, and the Extreme Forecast Index (EFI)) to answer common IDSS requests from core partners?
4. How well do forecasters leverage cluster analysis techniques in their decision making process?
5. Can forecasters reasonably understand and apply probabilistic messaging recommendations from Dr. Joseph Ripberger and Dr. Makenzie Krocak from the University of Oklahoma[2]?
6. How do core partners respond to probabilistic IDSS information across a variety of hazard scenarios and do the DESI and WSUP viewers help core partners make more informed response decisions.

However, when the OPG began designing this Probabilistic IDSS Evaluation, we encounter two barriers we needed to resolve:

1. To generate probabilistic information, forecasters must first apply a probabilistic forecast process using probabilistic data. After specific interactions with forecasters, the OPG found evidence

[1] The research project was led by Dr. Joseph Ripberger and Dr. Mackenzie Krocek. Results from the research can be found here - https://crcm.shinyapps.io/probcom/#section-practical-recommendations
[2] https://crcm.shinyapps.io/probcom/#section-practical-recommendations

suggesting relatively few forecasters have received training on a probabilistic forecast process[3] or currently apply a probabilistic forecast process in daily operations.

2. To provide probabilistic IDSS, forecasters need a delivery mechanism and probabilistic forecast format. However, there are no current standards describing how to deliver probabilistic information, how to produce proper messaging, or how to design probabilistic visualizations.

To remain objective in our evaluations, the OPG could not be the group who designs a probabilistic forecast process, delivery mechanism, and data format and then tests the quality of those designs. So, in late 2022, the OPG hosted a "Probabilistic Think Tank" where subject matter experts designed a conceptual probabilistic forecast workflow. The OPG then leveraged the results of the think tank as design elements for our Probabilistic IDSS Evaluation.

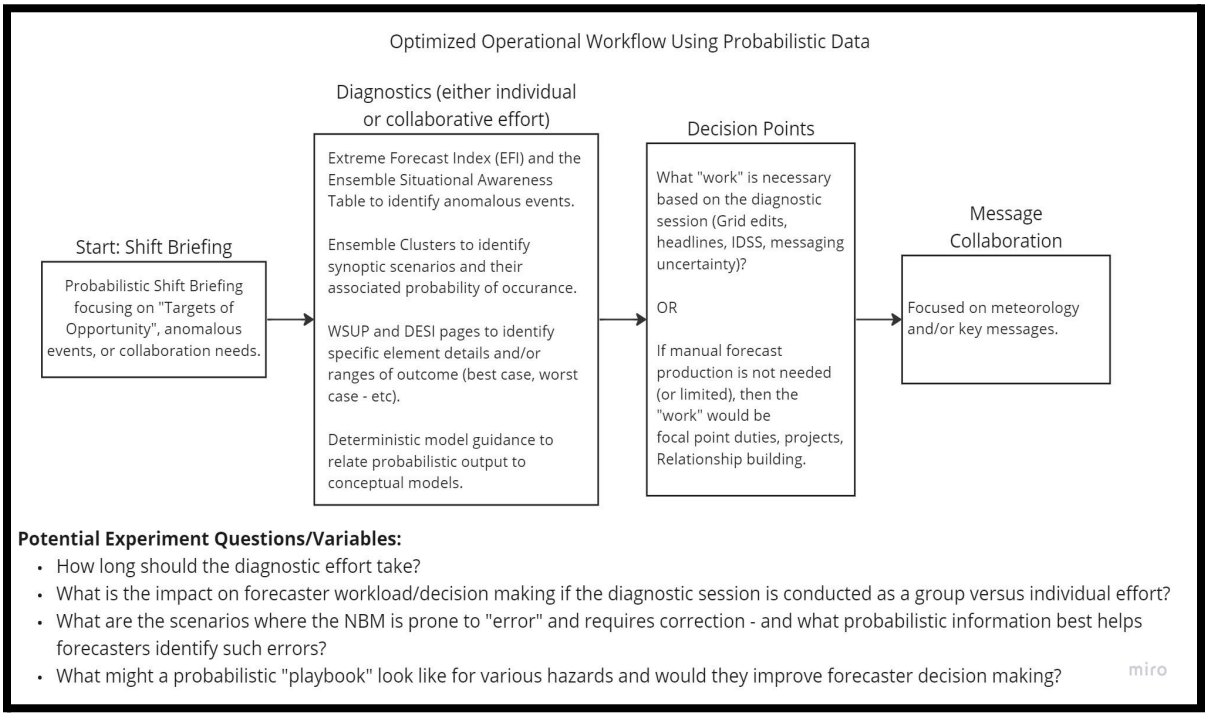

Image 1: Digitized version of a probabilistic workflow option designed by OPG Probabilistic Think Tank participants.

The Subject Matter Experts (SMEs) who participated in OPGs Probabilistic Think Tank designed several workflows for a probabilistic forecast and messaging process (see image above). A common theme from each workflow involved using the National Blend of Models (NBM) as a common starting point, then leveraging probabilistic visualization tools or diagnostic methods to identify forecast uncertainty and IDSS messaging opportunities. Specifically, these tools included the DESI and WSUP viewers along with other NWS data sources or viewers such as EFI and even third party web pages. The OPG then applied these common themes in our Probabilistic IDSS Evaluation design.

[3] OCLO is currently developing an ensemble fluency training course that will address this challenge.

To evaluate the impact of DESI and the WSUP viewers on forecaster decision making, the OPG set up a control group and a test group. The control group was able to use any tool currently available in operations except DESI (because DESI was not available to all offices at the time of the evaluation) while the test group was restricted to using only DESI, WSUP, and EFI[4].

The OPG provided all the participants basic training on communicating uncertainty based on recommendations provided by Dr. Joseph Ripberger and Dr. Mackenzie Krocek from the University of Oklahoma. Dr. Ripberger and Dr. Krocek's NOAA funded research on communicating uncertainty resulted in 11 practical recommendations that can be applied by any forecaster in operations today. The OPG introduced these recommendations to our evaluation participants and provided training on how to apply the recommendations in real world situations.

The OPG also asked subject matter experts, Dr. Matthew Jeglum and Travis Wilson (developer of DESI), to provide training to our Test Group (the DESI/WSUP/EFI Group) on using DESI as part of a probabilistic forecast process. Finally, the OPG ensured all participants understood the evaluation expectations by practicing their actions as a group. After the practice session, the participants operated individually without collaborating with fellow participants or SMEs.

The table below more clearly highlights the similarities and differences between the two groups:

Group "A" the Test Group: 45 Forecasters	Group "B" the Control Group: 21 Forecasters
Tools Available: ● DESI ● WSUP ● EFI ● **NO Cloud AWIPS, NO other NWS or private web pages** Training Received ● How to apply Dr. Ripberger and Dr. Krocek's recommendations for communicating uncertainty ● How to complete the expected tasks during the evaluation ● How to use the DESI interface in the forecast process	Tools Available: ● Cloud AWIPS ● Any NWS hosted web pages (WSUP, EFI, NCEP Center Pages - etc) ● Any private vendor web pages (WeatherBell, Pivotal Weather etc.) ● **NO DESI** Training Received ● How to apply Dr. Ripberger and Dr. Krocek's recommendations for communicating uncertainty ● How to complete the expected tasks during the evaluation

1.1 Pivoting to a Modified Plan

Prior to executing the evaluation, the OPG determined that addressing the fourth objective in the project plan greatly increased the scope of the evaluation and exceeded OPG capabilities at the time. As such,

[4] The Extreme Forecast Index allows forecasters to identify anomalous events where the tail end of ensemble distributions are particularly - or anomalously - extreme.

the OPG decided to pivot and focus instead on the internal mechanisms of providing IDSS rather than the end user decision making process. The OPG intends to address end user decision making based on probabilistic IDSS in the future.

As it turned out, adjusting the evaluation to focus on internal processes and communication strategies proved highly informative. Participants in the evaluation produced a diverse set of responses to each IDSS request. It would have been very challenging for the OPG to select specific messages to share with core partners without interjecting bias.

2. Evaluation Workflow

To start each scenario, the OPG asked participants via a Google Meet session to, “Imagine you are writing a heads up email to a partner. Write one key message about the most important weather event you want your partners to know for the week ahead.” The participants were asked to keep the key message concise and attempt to use Dr. Ripberger's recommendations for communicating uncertainty.

Participants then used ArcGIS Online (AGOL) to draw a polygon within the evaluation domain and add their key message.

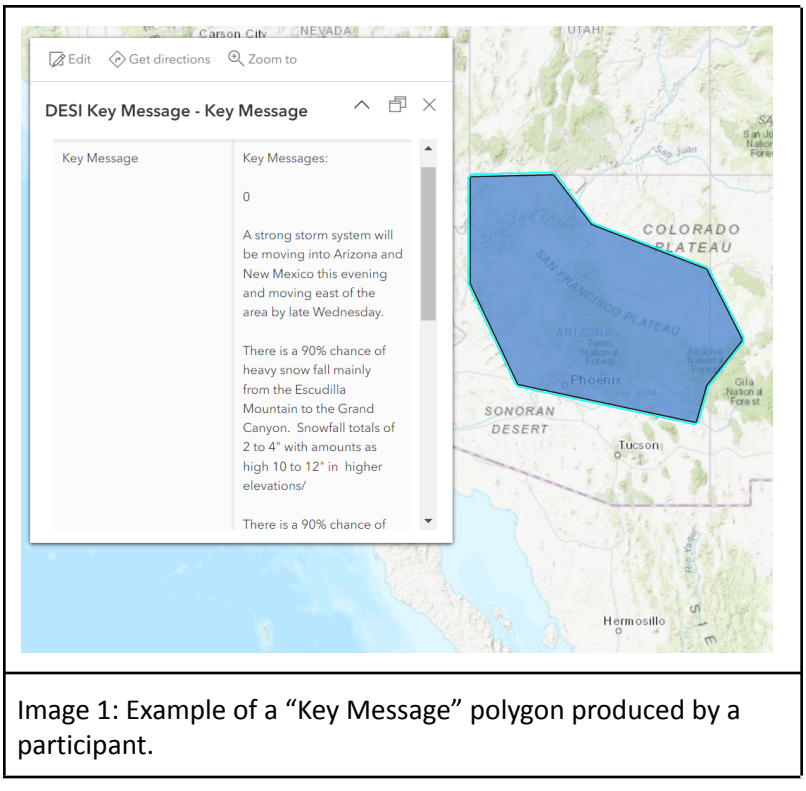

Image 1: Example of a “Key Message” polygon produced by a participant.

The OPG was able to monitor all of the participant entries in real time using a dashboard created in AGOL. Once 90% of participants provided their “Key Message” polygon, the OPG assumed the participants had conducted a thorough diagnostic evaluation for the 7-Day period and they were ready to answer more specific IDSS requests.

At that point, the OPG provided their second inject[5] which asked participants to draw a polygon where certain conditions were expected.

[5] An “inject” is simply a prompt asking the participants to complete some task.

In the example below (Image 2), the OPG asked participants to, “draw a polygon where they felt 2 inches or more of rain were expected ‘tomorrow’ (note our domain for the day as shown as a white square on the map roughly covering the state of Arkansas)”.

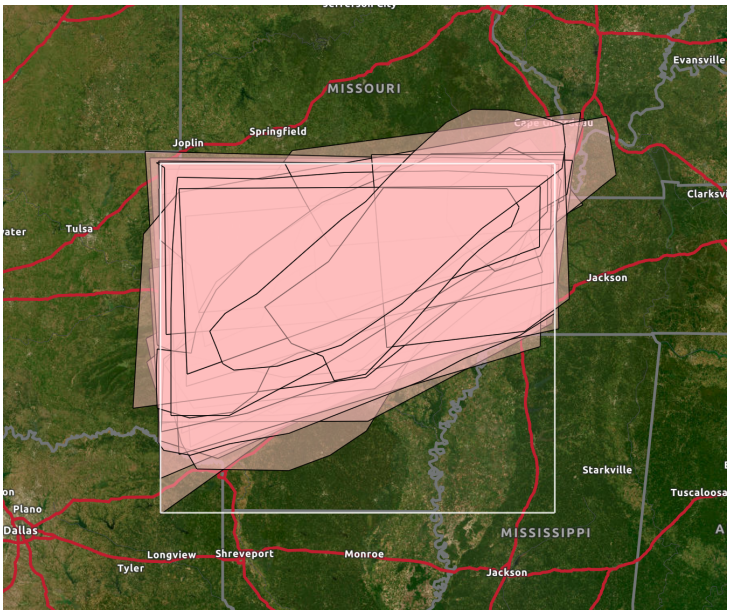

Image 2: All participant responses (red polygons) based on the inject, “Draw a polygon where you expect 2” or more of rain to fall on the following day.” Note the high level of consistency with these polygons showing a general SW to NE tilt and without any polygon in the SE portion of the domain.

Once again, after 90% of the participants responded to the polygon based inject, we moved on to point requests modeled after real world questions commonly asked by core partners[6] (see Appendix A for a complete list of the questions used during the evaluation). The OPG would announce to the participants on the Google Meet that a new inject was ready and showed a powerpoint slide with the inject information. At the same time, a red dot would pop up on the participant's AGOL display (Image 3) noting the location of the inject. The participants would then click on the red dot, read the request, and provide a text response all within AGOL.

[6] Field experts provided examples for the OPG to use during the evaluation.

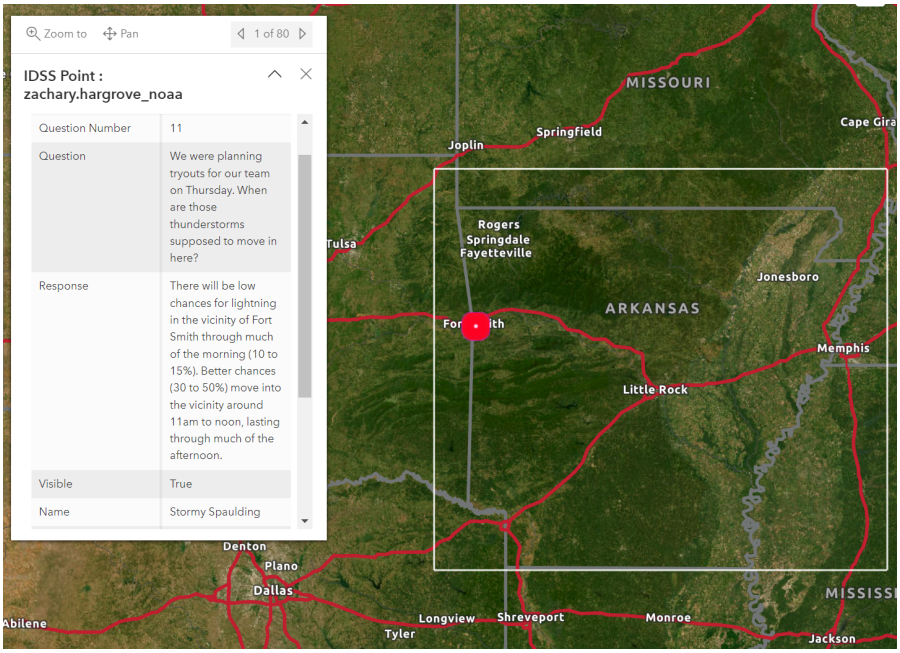

Image 3: Example showing a point request from a pretend Softball Coach named Stormy Spaulding who asked, “We were planning tryouts for our team Thursday. When are those thunderstorms supposed to move in here?” NOTE* We asked participants to respond to each question even though a softball coach would not be considered a “Core Partner”.

In total, the OPG asked participants to produce three key messages, three area polygons, and thirteen point responses across the three domains used during the evaluation. The first domain was set up over Arizona and New Mexico, then on day two, the participants operated over Arkansas in the morning, and then over New Hampshire and Maine in the afternoon.

3. Results

The OPG investigated several aspects of the participants responses:

1. Efficiency - how long did it take participants to respond to questions?
2. Length - how many words did participants use in their responses?
3. Scenarios - did forecasters properly apply cluster analysis tools when communicating scenarios?
4. Consistency - did participants produce similar probabilistic forecasts to one another?
5. Quality - did forecasters adequately apply Dr. Ripberger and Dr. Krocak’s recommendations?

3.1 Efficiency and Length

On average, it took our participants about 20 to 30 minutes to complete their diagnostic assessment and produce a Key Message for each domain. For the polygon and point injects, our participants responded in about 10 minutes on average.

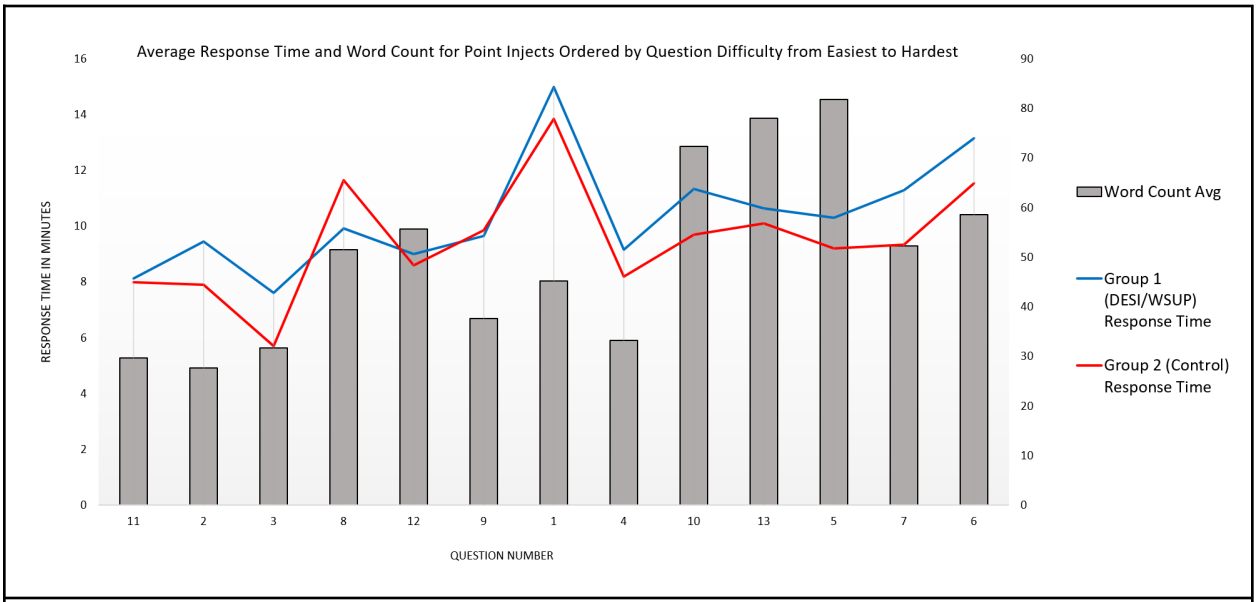

Figure 1: The blue (DESI/WSUP) and red lines (Control) represent the response time average for the point based injects referenced on the left hand Y-axis. The gray bars represent the word count average for each question referenced on the right hand Y-axis. The question numbers (x-axis) are in order of difficulty (easiest to hardest) based on participant rankings. Notice that for most questions the response times by both groups were within a minute of each other regardless of word count or question difficulty. Click the image to enlarge.

The questions with the highest word counts (numbers 5, 10, and 13 above) resulted from injects that asked the participants to explain why there were large ranges for different forecasts. For example, question 10 asked, “We have to brief FEMA today about this snow in Maine for the weekend. Can you explain why the snow ranges are so large in Augusta over the weekend?”. (See section 3.2 for more detail).

The questions with some of the fewest words (number 11, 2, and 3 above) were framed in a deterministic manner. For example, question 11 asked, “We need a precipitation free 24 hour period to fill some potholes between Thursday and Sunday. What day looks best to do this work?”.

It is most important to note that there was not a strong correlation between average word count and average response time ($R^2 = 0.381$), or question difficulty and average response time ($R^2 = 0.544$). However, we did find a strong correlation, $R^2 = 0.872$, between the two groups average response times suggesting forecasters took a similar amount of time to answer questions regardless of the tools they had at their disposal.

3.2 Scenario

The three scenario questions were intended to stimulate participants to review clusters and craft scenario messages. Several of our participants described scenarios, and their associated probability of occurrence, when answering these questions.

For example, when asked, “I am seeing very different forecasts for rain amounts on Friday/Saturday from our local news stations. I am not really sure what to believe. Can you help explain why there is such a large range on Friday/Saturday?”, participant responded with the following examples:

> “*The rainfall on Friday into Saturday is tied to how quickly the cold front progresses through the area. There is still some spread in timing of this front in the guidance: there is about a 25% [chance] of the front moving through quicker (Friday evening) and about a 25% chance of it moving through slower (Saturday morning). The quicker frontal progression would yield heavier amounts, but we can refine this timing over the next day.*” Evaluation Participant

> “*There is a lot of uncertainty with regards to the precipitation totals on Friday/Saturday due to the different forecast scenarios in play. If a relatively narrow, but deep shortwave trough passes us Friday evening, as some solutions suggest, there is a 60% chance of accumulation totals between 1 and 3 inches on Friday/Saturday for the area. Other solutions have a shallow shortwave ridge passing by, which yields an 80% chance of accumulation totals between 0.5 inches and 4 inches. Finally, the third scenario offers a broader, deeper trough to the northwest on average, bringing with it an 60% chance of precipitation totals between 2 and 4 inches.*” Evaluation Participant

To be clear, the OPG would not recommend using as much meteorological jargon in IDSS responses (shortwave, trough, ridge - etc), but the answers did confirm that participants were leveraging clusters to form a conceptual model explaining various probabilistic values.

Participants also referenced the value of cluster analysis in the final survey stating:

> “*I used clusters to contextualize the range of ensemble solutions, or to answer the question ‘why is there so much spread in the forecast?*’” Evaluation Participant

> “*I found the clusters to be a good indicator of how uncertain a particular forecast was. If there were only two clusters, and the dominant cluster occupied greater than 50% of the membership, then I could tell confidence and model agreement was good. However, if there were five or more clusters, and the leading cluster comprised less than 30% of the total membership, I could tell it was a more uncertain forecast. It was also beneficial to see the difference from the grand ensemble overlaid onto the clusters to see if the magnitude of forecast disagreement was large or small*.” Evaluation Participant

> “*The clusters in DESI are far and away the most powerful clustering tool I have ever used. Being able to translate the clusters into their implications for numerous important forecast variables (anything from qpf/snow to 700 mb temperatures etc), allowed me to have a much more complete description of the scenarios at play. This allowed me to give way more context about the forecast in the cases where there was a large range of outcomes - something the social science tells us is extremely important for trust in the forecast*.” Evaluation Participant

Other participants felt the cluster analysis tools need more context to inform decision making:

> "*I tried to for the sake of the experiment (to use clusters), but they did not change my messaging. I struggle to know what is causing the differences between clusters. I don't know how to find out if it is timing, intensity, or location.*" Evaluation Participant

Finally, some forecasters felt the clusters were primarily valuable as a diagnostic tool for the extended range forecast:

> *"I only used this once in the practice run. I really like the idea of using clusters, but I feel these are better used for long range forecasts. Most of the time we were forecasting in the near/short term (generally within 4 days.)*" Evaluation Participant

> *"I envision cluster analysis having the greatest functionality during the long range forecast (especially days 5-7), since it helps break the forecast solutions and associated spread into similar patterns, and tie physical meaning to the associated uncertainty. This would be incredibly helpful when trying to message hazards in the long term AFD."* Evaluation Participant

Overall though, 85% of the participants who completed our final survey (42 people) stated the cluster analysis tools were moderately to extremely valuable in their decision making process. However, other probabilistic visualization tools were favored by forecasters including the Component Based Visualizations and Point Based Box and Whisker plots:

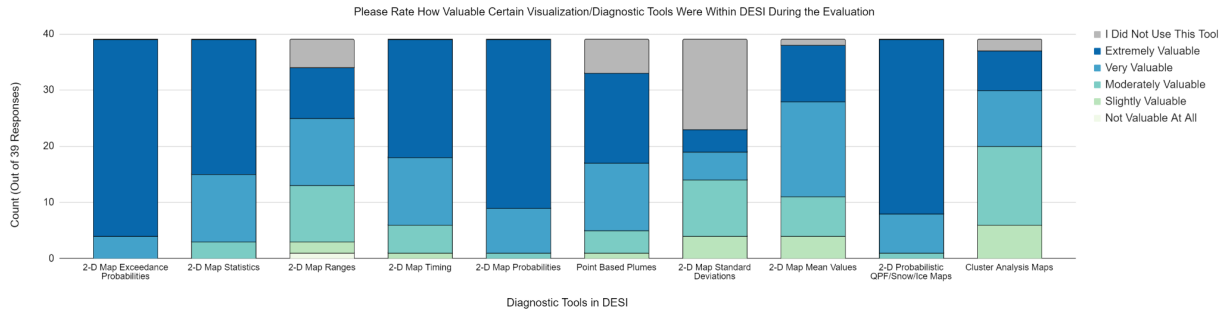

Figure 2: Results of our participant survey based on the DESI/WSUP group. Click on image to enlarge.

3.2 Consistency and Quality

To calculate message consistency, the OPG focused on the predicted values rather than the words used on conveying the message[7]. The OPG did try to compare the entire content of the messages using a Cosine Similarity Matrix approach but the results showed very little consistency among participants regardless of question context or group. The OPG believes that the variability in individual communication styles masked potential consistency in the primary forecast message. Therefore the OPG extracted the predicted values from the threshold based questions to assess consistency. There were four questions that provided a specific threshold and request for the probability of exceedance. The table below lists the questions and the values the OPG extracted from each response:

Question Text	Data Used in the Consistency Analysis
"What time do you expect the strongest winds on Wednesday? Any chance we get higher than 50 mph? We are supporting a community fair with a bunch of vendors in tents. The tents are rated for 50 mph winds."	Probability of exceeding 50 mph wind speeds
"We are concerned about high profile vehicles blowing over on I-25. What is the chance we see gusts higher than 60 mph on Wednesday afternoon?"	Probability of exceeding 60 mph wind speeds
"We are hosting a major biker rally on Beale Street that runs from Saturday through Tuesday. These bikers hate the cold weather and if we have high temperatures less than 60 degrees then I have to set up outdoor heaters for their comfort. We don't want a replay of last year. What are the chances we have high temperatures below 60 any time Saturday through Tuesday?"	Probability of maximum temperatures remaining below 60 degrees
"It sounds like we could see a lot of snow over the weekend. We are considering school delays for Monday morning if the roads are bad. How much snow are you thinking for this weekend in Bangor?"	Probability of exceeding various snow amounts

After extracting the numeric predictions, the OPG plotted the results using Box and Whisker plots separated by group (blue boxes were for Group 1 who used DESI/WSUP, and red boxes for our control group):

[7] The OPG used ChatGPT to generate python and javascript code to run the Cosine Similarity Matrix calculations. However, when we ran an independent analysis we found the resulting Cosine Similarity values were not the same as the results provided by the ChatGPT generated code. Regardless, the values produced by the independent analysis and the ChatGPT generated code were similar in showing very little consistency in message text. Thus this report will not show the Cosine Similarity Matrix values, but instead note that the messages, generally speaking, were inconsistent.

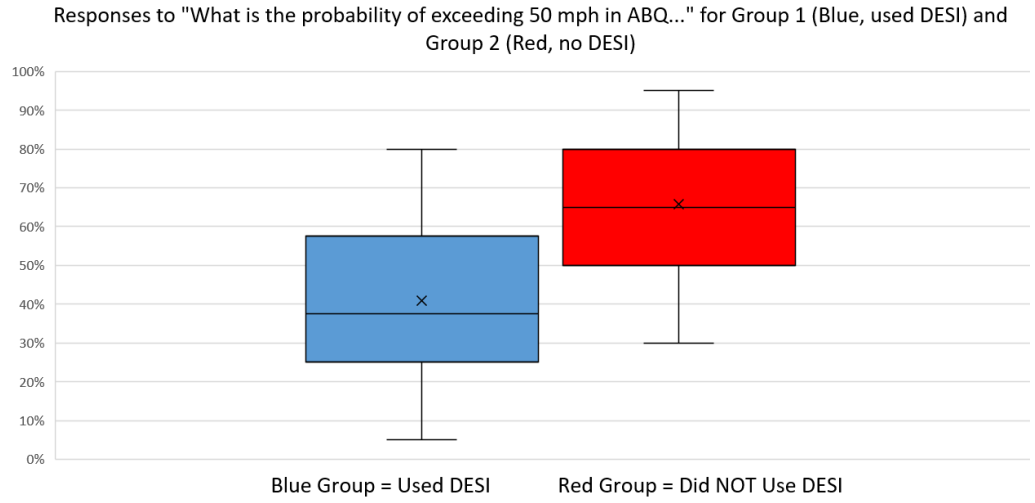

Figure 3: Participant responses to the question, “What are the chances we exceed 50 mph in Albuquerque, NM…”.

The wind related questions produced the most interesting output. In this Albuquerque case, both groups produced responses with large interquartile ranges although the DESI/WSUP group indicated lower overall chances of exceeding 50 mph wind speeds. The OPG also asked participants to determine the potential of exceeding 60 mph winds in Ilfield, NM.

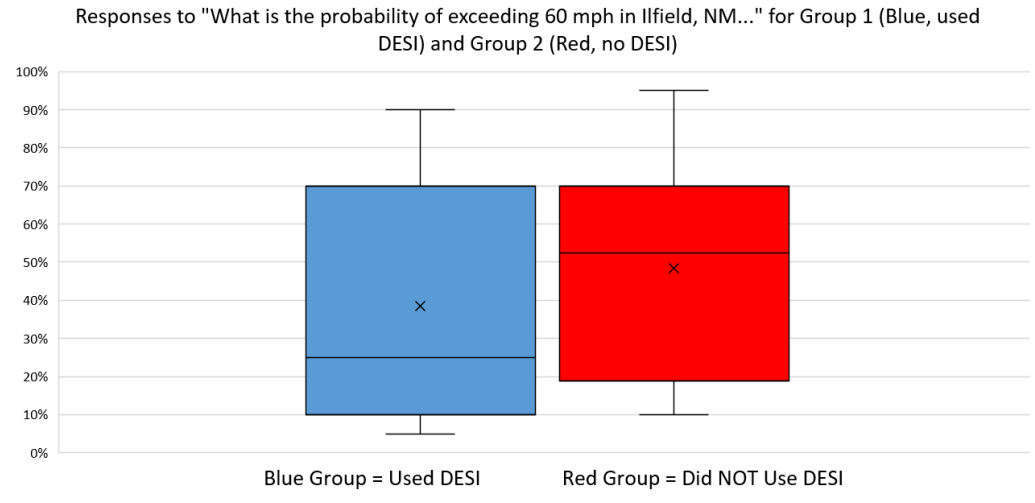

Figure 4: Participant responses to the question, “What are the chances we exceed 60 mph in Ilfield, NM…”. Note that both groups exhibited a large interquartile range but the mean values were much lower again for the DESI/WSUP group..

Interestingly, participants produced a much larger range of responses for this Ilfield scenario. When attempting to understand why the ranges were so large, one participant suggested the ranges stemmed from differences in the guidance:

> "*Tool note, discrepancy between HREF (14% chance of exceeding 60 mph) vs NBM (60-80% chance. This definitely caused me to investigate more. Ended up deciding to take a blended approach to the probability I answered above. However I should note that DESI made it easy to identify and assess the difference between the two pieces of guidance.*" Evaluation Participant

In addition to wind speed related questions, participants also had to address a nuanced temperature forecast. In this case, the simulated Emergency Manager of Memphis, TN wanted to know if maximum temperatures would remain below 60 ℉ between Saturday and Tuesday.

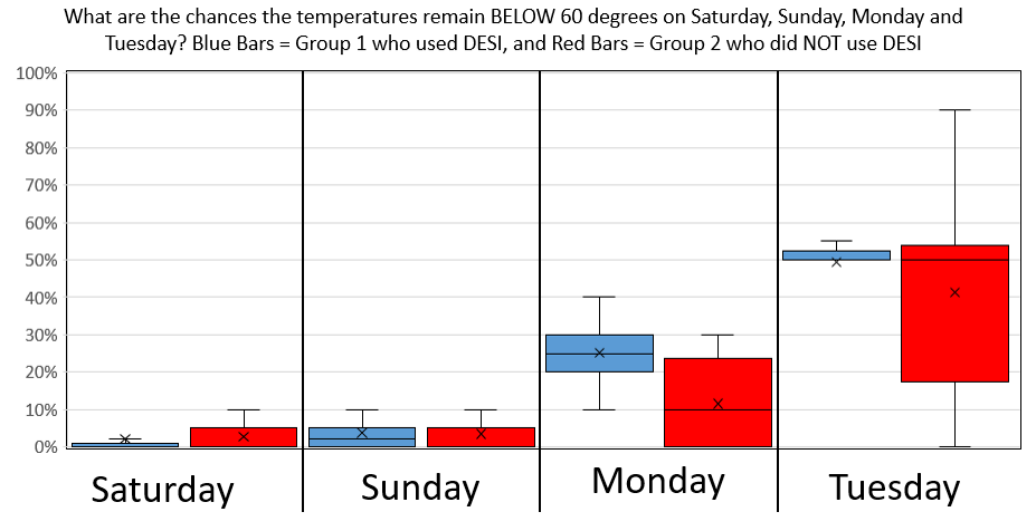

Figure 5: Participant responses to the question, "What are the chances we remain below 60 ℉ in Memphis, TN from Saturday through Tuesday…". Note that the control group exhibited a very large interquartile range on Tuesday, but otherwise the ranges and predictions were closely aligned.

Overall, the DESI/WSUP group produced more consistent responses to our temperature based question than the Control group. It is unclear why the Control group produced such a large range of responses for the Tuesday temperature forecast, but OPG believes it is likely related to the tools available to answer such a non-standard question. With a tool like DESI, participants can set the threshold and then calculate the probabilities. In this case, participants could tell DESI to produce the probability of remaining below 60 degrees on each day of the forecast. Assuming the forecaster did not change the output significantly, then the resulting predictions should be consistent. Without DESI however, forecasters have to infer a probability from the NBM or various ensemble systems.

Finally, the OPG also asked threshold based questions relating to snowfall amount for an event roughly four to five days in the future. The question presented to the forecasters was deterministic in nature, "How much snow are you thinking for this weekend in Bangor, Maine?". With the training and tools provided for this evaluation, participants responded with probabilistic information including a "most likely" snowfall and a "worst case scenario".

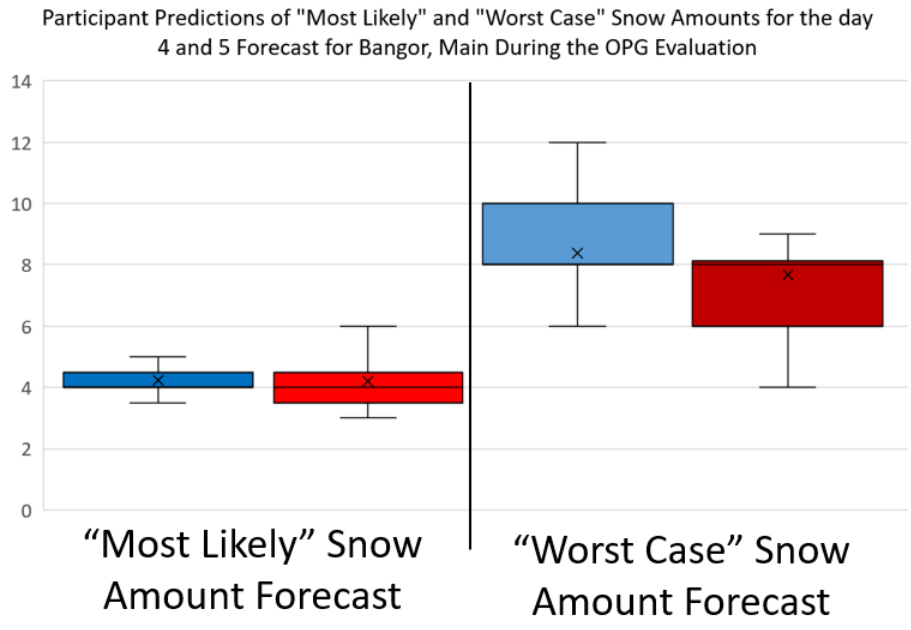

Figure 6: Participant responses to the question, "How much snow are we going to get in Bangor, Maine over the weekend?". The blue plots represent Group 1 (who had access to DESI), and Group 2 (control group). Notice that both groups showed a most likely snow amount of around 4 inches, but differed slightly on the worst case scenario.

The participants were highly consistent in their "most likely" snowfall forecast, but the DESI/WSUP group did predict higher amounts for the worst case. As it turned out, Bangor received between 3" and 4" of snow while northern Maine observed around 10".

After reviewing the data from the probability of exceedance based questions the OPG could not make any definitive statements about which group produce more consistent responses, or who produced more accurate forecasts, however a few interesting themes emerged:

1. The wind related questions generated the largest spread of responses. The OPG found evidence that these large ranges resulted from differences between raw ensemble guidance (HREF) and post processed (NBM) guidance.
2. Similar to point #1, neither DESI nor WSUP solve the "model picking" forecast methodology that results in inconsistent forecasts or IDSS messages. As such, forecasters will generate different forecasts for the same event based on their choice of model, or ensemble system, or percentile.
3. Using the DESI platform allowed participants to plug in any threshold value for various time periods and generate a probability of exceedance value.
4. Participants, including those with little to no northeast winter weather forecast experience, were able to produce a representative "most likely" and "worst case scenario" forecast for a snowfall event 4 to 5 days into the future.

5. Findings and Recommendations

The OPG was asked to evaluate topics for several entities (STI, AFS, GSL, MDL and WPC). As such, when appropriate, the recommendations below each finding provide suggestions for each of these groups.

Finding 1: Given the right set of tools, clear expectations, and some training, forecasters CAN provide **reasonable probabilistic IDSS** in real time for forecast periods between 12 hours and Day 7.

> To unpack this finding, we need to provide context, definitions, and examples. First, we provided our DESI/WSUP group participants with tools, such as DESI, WSUP, and EFI, and some training by subject matter experts (SMEs) to properly interpret probabilistic data from these platforms. We then provided participants from both groups with training on communicating uncertainty based on Dr. Ripberger and Dr. Krocak's work. We also worked through a real time training case with all participants to ensure they understood their expectations and technical processes for completing tasks.
>
> After the evaluation, Dr. Ripberger and Dr. Krocak reviewed participant responses to various injects and noted the responses were an improvement over current operational messages. Even though core partners did not participate in this evaluation, Dr. Ripberger and Dr. Krocak's comments support our first finding.

Recommendation 1a: As much as possible, forecasters should apply Dr. Ripberger and Dr. Krocak's recommendations when crafting messages as part of IDSS or legacy NWS products.

Recommendation 1b: OCLO should incorporate Dr. Ripberger and Dr. Krocak's recommendations into existing training courses, such as the IDSS Bootcamp or Advanced Hurricane Messaging Course.

Recommendation 1c: Each AFS program should review existing directives and consider updating content to eliminate use of words of estimative probability and instead apply Dr. Ripberger and Dr. Krocak's recommendations.

Recommendation 1d: GSL and MDL should consider adding a “message crafting” tool that takes real time probabilistic information and uses a template to produce messages aligned with Dr. Ripberger and Dr. Krocak's recommendations.

Finding 2: There are still cultural biases in place that inhibit forecasters' ability to embrace a probabilistic forecast mindset. These biases include:

- A desire to pick “winners and losers” among a set of guidance in an effort to improve the forecast.
    - “*I think a barrier would be that some folks aren't comfortable tweaking probabilities from what is in the NBM, even if we know the NBM is off…*” Evaluation Participant

- A lack of trust in the numbers that fall out of ensemble systems. Forecasters still perceive ensemble guidance as a "black box" and as such struggle to believe the probabilistic data is valid or accurate.
    - *"It is hard to know if the data we are given verifies well enough and is providing the service that is necessary and accepted."* Evaluation Participant
    - "*The main barrier is building trust in the tools. Sometimes forecasters disagree with the probabilistic numbers being given.*" Evaluation Participant
- A belief that end users, including core partners and the public, do not want or can not properly understand probabilistic information.
    - "*What I foresee being a greater challenge is getting our partners, many of whom still think with a binary yes/no mindset, to adjust to probabilistic messaging*." Evaluation Participant
    - "*Some of my coworkers and I are concerned the general public doesn't have firm understanding of probabilities and may prefer a deterministic forecast (i.e. what we feel is the most-likely outcome based on our analysis)"* Evaluation Participant
    - "*Perhaps culture more than anything. Despite the findings, there does seem to be an opinion that partners nor the public understand percentages.*" Evaluation Participant
- Prior training that advised against using probabilistic information when communicating uncertainty.
    - "*Most times I struggled to incorporate the recommendations simply because it didn't match how I would typically write to a partner or in a discussion, that is to say it's a taught, cultural bias of sorts. There's certainly going to be a learning curve on the parts of both forecasters and customers to write and interpret more probabilistic text, but once I had become used to using that sort of language, it came much more naturally.*

Recommendation 2a: The "Ken's 10 Probabilistic IDSS Team" should recognize the barriers posed by cultural biases in the Probabilistic IDSS Roadmap and suggest methods our agency can apply to overcome such barriers.

Recommendation 2b: The NWS, through OSTI (including test beds and proving grounds) and the SOO/DOH community, should consider formally evaluating various forecast methods (e.g. picking winners and losers, applying rules of thumb, or manually correcting perceived model biases) to identify skillful methods that can translate to future probabilistic forecasting practices. In short, the NWS needs to determine a probabilistic forecast process that all forecasters can apply that either increases skill in correcting errors to a starting point, and/or informs IDSS messaging.

Recommendation 2c: OCLO should consider producing short, FAQ style content that addresses forecaster perceptions regarding probabilistic guidance (e.g. ensembles as a black box, end user understanding of probabilistic data, or the use of numbers versus words of estimative probability to convey uncertainty).

Finding 3: Forecasters are using non-operational tools to evaluate probabilistic information because our existing operational tool, AWIPS, has limited probabilistic capabilities.

- "*The biggest takeaway from this evaluation is that my mindset to think in probabilistic information is not engrained yet. I also realized the tools we have to get this information is very limited. How we interpret, create, and message that information also was not fully part of my forecast process. For somebody who has forecasted for almost 20 years, that is a hard thing to reprogram without much more practice.*" Evaluation Participant

Recommendation 3: While the OPG does not advocate for any particular solution based on this evaluation alone, it is clear that forecasters need an operational solution for probabilistic data visualization, diagnostics, and dissemination. Thus, the OPG recommends that the NWS consider prioritizing, possibly under the Probabilistic IDSS "Ken's 10" initiative, providing an operational probabilistic data visualization solution to meet forecasters probabilistic forecasting needs.

6. Conclusion and Thank You

On behalf of the OPG, I would like to thank all the participants who gave their time and expertise to help us understand the aspects of the probabilistic forecast and messaging process. You all performed exceptionally well. I also want to say a special thank you to the “control group” who were not able to use DESI during the evaluation. We learned a great deal from both groups and enjoyed working with all of you!

Thank you to all the managers who allowed their staff to participate in this evaluation. We greatly appreciate your sacrifice and hope you continue to see value in OPG evaluations.

Thank you to GSL and MDL who developed DESI and the WSUP browsers. We appreciated your help and guidance prior to, during, and after our evaluation. We hope you find the results beneficial to your future development efforts.

Finally, thank you to my incredible OPG staff who continue to find creative solutions to challenging experimental design and execution problems. Your work is important to the agency and to the public.

This was a rather interesting experience where we attempted to compare forecast methods. We found it challenging to control all the variables involved in forecasting, yet still identified several important findings that should inform future probabilistic forecasting efforts.

If there was just one message that we hope readers of this report “hear” loud and clear, it is that using words of estimative probability (instead of numbers) decreases end user decision making quality and therefore reduces the value of our information. These new forecasting tools are incredible and help forecasters assess uncertainty and form conceptual models regarding potential outcomes, yet they can not prevent forecasters from using words like “possible, may, should, or could” when conveying uncertainty.

We showed in this evaluation that simply exposing forecasters to Dr. Ripberger's recommendations, and providing opportunities to practice probabilistic messaging, greatly reduces use of words of estimative probability. These are techniques that can be applied today by forecasters but will require policy changes in cases where words of estimative probability are built into products by design.

Thank you for reading this report.

Respectfully,

John J. Brost

Director, Operations Proving Ground

Appendix A: Questions Asked During the Evaluation

1. What time do you expect the strongest winds on Wednesday? Any chance we get higher than 50 mph? We are supporting a community fair with a bunch of vendors in tents. The tents are rated for 50 mph winds.
2. We have a county fair starting at 4 pm on Wednesday. Will the rain be out of the area by then?
3. We have a crew pouring concrete at our new recreation center on Wednesday. When is our best 6 hour window to avoid rain?
4. We are concerned about high profile vehicles blowing over on I-25. What is the chance we see gusts higher than 60 mph on Wednesday afternoon?
5. I am seeing very different forecasts for rain amounts on Friday/Saturday from our local news stations. I am not really sure what to believe. Can you help explain why there is such a large range on Friday/Saturday?
6. We saw something on the news about possible severe weather on Friday is that true? If we get severe weather during the afternoon we will consider going to an early release. What can you tell me at this point?
7. We are providing support for an air show on Friday. My weather app said rain was expected. I wanted to get your opinion on it. What time do you expect rain on Friday?
8. We are hosting a major biker rally on Beale Street that runs from Saturday through Tuesday. These bikers hate the cold weather and if we have high temperatures less than 60 degrees then I have to set up outdoor heaters for their comfort. We don't want a replay of last year. What are the chances we have high temperatures below 60 any time Saturday through Tuesday?
9. We were planning tryouts for our team on Thursday. When are those thunderstorms supposed to move in here?
10. We have to brief FEMA today about this snow in Maine for the weekend. Can you explain why the snow ranges are so large in Augusta over the weekend? How likely is it they exceed 8 inches of snow?
11. We need a precipitation free 24 hour period to fill some potholes between Thursday and Sunday. What day looks best to do this work?
12. It sounds like we could see a lot of snow over the weekend. We are considering school delays for Monday morning if the roads are bad. How much snow are you thinking for this weekend in Bangor?
13. The guy on TV said we are getting snow Saturday but the app on my phone says it's all going to be rain here. Can you explain what is going on with this precipitation forecast?

Appendix B: Participant Suggestions to Improve the DESI and WSUP Viewers

The following themes emerged from our participants when we asked them what specific tools they used in the DESI and WSUP browsers and what suggestions they have to improve the DESI and WSUP browsers. These suggestions are unedited (other than spelling corrections).

WSUP

Theme 1: Suggested Features to Add or Adjust

- Would be nice to be able to select exceedance thresholds, at least by the inch, especially for snow. Different partners have different thresholds.
- I'm always interested in adding more elements, especially for convection. Perhaps shear or SRH?
- NBM 6-hourly snowfall and ice accumulation data for at least the first four days of the forecast cycle.
- A user guide including context on how to use the standard deviation guidance correctly.
- Increase city density and labeling. Otherwise, it functions well.
- Making it more suited to create graphics for social media/partners.
- Add the ability to create the Pie Charts at custom points and display the actual probabilities for each portion of the Pie Chart.
- Customizable color table options.
- It'd be nice if the wind probability information was also included in the "surface" menu instead of just "fire weather", but that may be a menu size issue.
- Having time steps in local times instead of zulu in the upper righthand corner.
- A two panel or 4 panel layout would be awesome to be able to see all precip types on the screen at once.
- Being able to see max and min of NBM values in DESI was helpful instead of stopping at 5th and 95th percentiles in WSUP.
- Using a map tack to limit the area WSUP loads NBM data might make it load faster.
- The ability to have custom sample points for the 1-D viewer.
- Would like more options for timing of precip onset.

Theme 2: Minor “Bugs”

- A lot of times when I change the forecast hour, the POE threshold reverts back. It would be nice if the threshold stayed selected without having to double click.
- When going forward and backward in time via the left and right arrow keys, the slider on the left reverts to the lowest exceedance threshold, so you keep having to use the down arrow to go to the threshold of interest. For example, if looking at probabilities of QPF > 4 inches, it would be nice for that to stay there when going back and forth in time.

- More variable zoom levels in WSUP would be nice when we want to export maps to use in messaging. At least for my WFO, there is a zoom issue where either we zoom in just a little too close and cut off a portion of the CWA or we zoom out one level and include way too much area outside of what we want.
- I would like to see a data status message. Sometimes it is obvious that all NBM members are not ingested for whatever reason. It'd be nice to see what is being included and a message of what is not included because of technical issues

DESI

Theme 1: Request for Additions/Changes

- More severe parameters (several similar comments like the following):
    - Would love to see probabilistic indices in box and whisker format for severe weather indices showing individual member values and means. Also, I would like to see soundings based on ensemble clusters and be able to 4 panel to see if there are significant differences.
- Point sounding (several similar comments like the following):
    - Remake BUFKIT in DESI - anything resembling NSHARP, especially if you can have the analogs in there. Time height cross-sections. Momentum transfer winds.
- It would also be nice as an option to have the 25-75th percentile range highlighted for both temp/Td, like the EC-Ens sounding viewer does. That way it's a quick look at where the spread is.
- Instead of one thermodynamic profile that is an average, being able to plot all ten members at once and then being able to pick and choose which members to take out/ include and then compare would be tremendous.
- It would be nice to see probability distributions of the parameter space, similar to the 10/25/50/75/90th percentiles etc.
- Wet Bulb Temp trace, Highlighted DGZ, parameters useful for winter forecasting as well as convection.
- QPF based clustering (the 500mb heights can be less relevant to sensible weather at low latitudes in the summer for example).

Appendix C: Participant Suggestions for OCLO

The following themes from our participants relate to training needs for probabilistic forecasting.

Value of Task Based Training

- I felt like this entire experiment was a great example of how training should be set up. Given this question, how would you respond? Things like that would be very helpful to make people more comfortable speaking in probabilistic terms.
- Having drills/training sessions on how to use DESI/WSUP for an operational setting would be great. This OPG event was great because it would actually force you to take a deep dive into the information.

Formal Course for Probabilistic Forecasting

- PAC (Probabilistic Applications Course). Required training. Cover Dr. Ripberger's recommendations. Practice, practice, practice using the viewers, adding probabilistic data and contextualizing it. If we don't use it consistently and often, we resort back to what we always have done and to what we are comfortable with.
- Deep dive understanding of statistics (as part of a formal course)

Contextualizing or Communicating Probabilistic Information

- Training that discusses the use of likely to extreme events and how to communicate this effectively with EM or core partners.
- Training on how to communicate the probabilistic information in a way that is useful to the general public would be helpful because I frequently find myself in the statistical weeds and am unable to think of how to communicate the probabilities have changed without using words like percentile or standard deviation.
- Best practices or suggestions on how to handle messaging when different probabilistic datasets are telling you different stories.
- Advice on how to best message a forecast when your "official" forecast falls at the tail ends of the probabilistic distribution.
- How to better provide context to the probabilities being provided.
- I would like more training on how to create visualizations in weather graphics from probabilistic information.
- Specific guidance on what words to NOT use. For example, there is a lot of hesitance to use low/medium/high because these words can be both ambiguous and subjective.

Understanding and Applying Different Types of Probabilistic Data

- I would like more info on the different data sets. What are the differences between NBM, NBM QMD, and HREF. How do we leverage each data set to put out the best possible forecast?

- How to interpret and convey standard deviation guidance correctly.
- More training or resources on how the probabilistic data are calculated for different fields.
- Taking advantage of clustering analyses to better understand and subsequently communicate the physical reasoning behind high ensemble spread.
- I would like to see more training on model bias and more social science training for (communicating) with partners and the public.
- Perhaps some basic training on how NBM probabilities are derived (seems like a black box to most of the field), as well as how to interpret it, adjust if necessary, and how to word responses to different types of questions from partners, much like we did in the evaluation.
- Guidance on how to most appropriately stray from/tweak NBM probabilities when we do not think they accurately portray the range of possible outcomes.
- Training on how to leverage probabilities that are more difficult for folks to interpret such as severe weather probabilities from SPC.
- Right now in the NWS, there is no established work flow for using and critically thinking about probabilistic data sets. We are currently just handed the data in neat web viewers, and told to try this to quantify/qualify your deterministic forecast's uncertainty. Once the forecasters have a better understanding on how to apply probabilistic data to their forecast work flow, then we can transition this messaging.

Miscellaneous

- Using large language models to incorporate probabilistic data into our IDSS messaging would be really interesting. We may be a ways away from this, but perhaps not too far off with the advance of OpenAI's ChatGPT.
- Examples of how this has been applied in the field. This could be done by recorded webinars of people talking about how they used the WSUP Viewer, HREF Probs, etc. during webinars, DSS packets, in verbal briefings, and in-person meetings.
- We need to learn how to educate partners on the increased use of probabilities.
- Training on using the visualization tools like DESI/WSUP.