An Iterative Approach towards Development of Ensemble Visualization Techniques for High-Impact Winter Weather Hazards.
Part 1: Product Development

Jacob T. Radford¹,², Gary M. Lackmann¹, Jean Goodwin³, James Correia Jr.⁴,⁵, and Kirstin Harnos⁴

¹ Department of Marine, Earth, and Atmospheric Sciences, North Carolina State University, Raleigh, North Carolina 27607
² Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado 80521
³ Department of Communication, North Carolina State University, Raleigh, North Carolina 27607
⁴ National Oceanic and Atmospheric Administration/National Weather Service/National Center for Environmental Prediction/Weather Prediction Center, College Park, Maryland 20740
⁵ Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado 80309

For submission as an article to Bulletin of the American Meteorological Society

14 September 2022

Corresponding Author: Jacob Radford, Jacob.Radford@noaa.gov

Early Online Release: This preliminary version has been accepted for publication in Bulletin of the American Meteorological Society, may be fully cited, and has been assigned DOI 10.1175/BAMS-D-22-0192.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2023 American Meteorological Society. This is an Author Accepted Manuscript distributed under the terms of the default AMS reuse license. For information regarding reuse and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
Capsule

We developed nine interactive tools to visualize probabilistic winter weather hazards based upon operational forecaster input. Five of the products depicting temporal aspects of snowstorms were selected for evaluation.
Abstract

We applied social science research principles to develop a suite of probabilistic winter weather forecasting visualizations for High-Resolution Ensemble Forecast (HREF) system output. This was achieved through an iterative, dialogic process with US National Weather Service (NWS) forecasters to design nine new web-based, interactive products aimed towards improving visualizations of winter weather event magnitudes, characteristics, and timing. These products were influenced by feedback from a preliminary focus group, which emphasized the importance of product credibility, contextualization, and scalability. In a follow-up discussion, winter weather forecasting experts found the event timing products to have the greatest utility due to their association with impact decision support services (IDSS). Furthermore, forecasters assessed snowfall rates as the most impactful variable, rather than snowfall totals and radar reflectivity. The timing products include plots of probabilistic snowfall onset time and duration, rush hour intersection probabilities, and a combination meteogram. The onset and duration plots are extensions of Demuth et al. (2020) which visualize the ensemble average onset time and duration of a specified snowfall rate, but with the addition of uncertainty information by visualizing the earliest, most likely, and latest potential onset times as well as the shortest, most likely, and longest potential durations. The rush hour product visualizes the probability of exceeding a specified snowfall rate during local commutes, and the combination meteogram allows rapid identification of high-impact periods by encoding probabilities of precipitation, precipitation-type probabilities, and average rates into one graphical tool. Examples of these interactive products are maintained on our companion website: https://www.visweather.com/bams2023.
1. Introduction

Over the past two decades there have been two paradigm shifts in operational meteorology that underpin this work. First, computational advances have allowed for the development of regional high-resolution, convection-allowing numerical weather prediction (NWP) models (CAMs; Benjamin et al. 2019) such as the High-Resolution Rapid Refresh model (HRRR; Benjamin et al. 2016). With grid-spacing of order 4 km or less, CAMs are capable of explicitly resolving convective-scale processes that are fundamental to mesoscale phenomena, such as quasi-linear convective systems (QLCS) and snowbands, and they often improve precipitation forecast skill (Mass et al. 2002). Second, operational emphasis has progressed from deterministic to probabilistic NWP forecasts based upon ensemble systems (Benjamin et al. 2019). Rather than relying on an individual solution, ensembles provide a spectrum of possible outcomes and allow forecasters to evaluate the most likely scenario, edge cases, and overall uncertainty.

More recently, these two developments have intersected into the implementation of convection-allowing ensemble systems. The first of these systems to be widely adopted within the US operational forecasting community is the High-Resolution Ensemble Forecast (HREF) system. The HREF is an ensemble of opportunity, meaning it combines output from existing CAMs rather than initializing the ensemble as an independent entity with varied initial conditions or physics parameterizations (Ebert 2001). With its introduction into AWIPS in 2017, the HREF became the first operational ensemble CAM available to US National Weather Service forecasters. However, cloud-based NWP products have gained prominence, and forecasters widely access HREF data through the Storm Prediction Center’s (SPC) HREF Viewer (Roberts et al. 2019). At the time of writing, the first true high-resolution ensemble, the Rapid Refresh Forecast System (RRFS), is set to become operational within the year.

CAM ensembles offer tremendous forecasting potential, but also present new challenges. Deterministic CAMs already produce more output than forecasters can reasonably analyze in an operational setting (Roebber et al. 2004). Extending CAMs to an ensemble system further increases output volume. Despite these advancements in NWP, progress in ensemble NWP post-processing and visualization has lagged (Hirschberg et al. 2011). Data volume growth has
outpaced effective analysis and visualization capacity. Visualization systems and strategies need to be developed to help forecasters extract meaningful information from an abundance of data and communicate this information to core partners (Hirschberg et al. 2011; NWS 2019). Rautenhaus et al. (2018), Roberts et al. (2019) and Wang et al. (2019) review many popular approaches to probabilistic visualization, including stamp maps, spaghetti plots, probabilistic heat maps, ensemble mean and standard deviation, and ensemble clustering methods.

There is no “silver bullet” solution for communicating probabilistic information, with various audiences demanding different development considerations. For example, professional weather forecasters have much greater technical understanding of ensemble systems than emergency managers, who in turn have greater familiarity than road workers (Fundel et al. 2019). Even within these groups there is diversity in background that informs interpretation and decision-making. However, risk communication research also offers guiding principles for developers of probabilistic information products. Fundel et al. (2019) suggest that developers “encode quantitative information in a way that fosters accurate decoding, explains probabilities, prevents deterministic misinterpretations of forecast uncertainty, puts rare but severe events into perspective, and chooses the level of detail in accordance with what can be reasonably predicted.”

Visualization and communication of probabilistic hazards present unique challenges, but share a foundation with science communication, a field with an abundance of literature. This foundation suggests that stakeholders are most likely to incorporate expert knowledge (or in this case, new visualization products), if the information is credible, salient, and legitimate (Cash et al. 2003). Credibility refers to the scientific soundness of the information, saliency to the relevance of the information to the stakeholders’ interests and decisions, and legitimacy to the unbiased nature of the information and incorporation of different viewpoints and backgrounds (Cash et al. 2003). In the absence of these three criteria, case studies have shown stakeholders to be critical of the science (Wynne 1992; White et al. 2010), despite being well-intentioned. Perhaps the most effective way to build credibility, salience, and legitimacy into science communication is through an iterative dialogue with the audience (Cash 2003; Feldman and Ingram 2009; Committee on the Science of Science Communication 2016), in which the researcher/developer draws upon the
experiences and expertise of the users to develop products that better fit their needs. A similar concept of user-centered design is presented by Abras et al. (2004), which suggests that end users are present in the development process in some capacity. This involvement can range from providing initial suggestions and guidelines to consistent, active participation in product design.

This dialogic approach is also suggested by Demuth et al. (2020), who state “instead of asking how to get forecasters to adopt new tools, the question that ought to be asked is how to effectively create and transition products that forecasters actually want, need, and can use.” Towards this end, Demuth et al. (2020) presented forecasters with three prototype products for feedback: A combination plot visualizing both deterministic and ensemble percentile guidance aiming to bridge the gap between deterministic and probabilistic output, an event onset product visualizing the ensemble mean onset time of a hazard, and an event duration product visualizing the ensemble mean duration of a hazard. The combination plot received mixed reviews due to perceived complexity, while the onset and duration plots were generally viewed favorably.

Our work builds upon that of Demuth et al. (2020), continuing development of novel ensemble visualization techniques from an operational perspective, specifically for winter weather events with an IDSS emphasis. In part one of this study, we conducted an initial focus group with WPC and NWS forecasters to establish a baseline of existing probabilistic winter forecasting tools and specific needs that could be solved with new products. More generally, this initial session sought to begin engagement with the operational community, thereby increasing forecaster investment and product credibility. The feedback received in this focus group was used to begin prototype product development to address the needs identified by forecasters. Following development, a second feedback session was conducted with a different group of winter weather forecasting experts to identify the prototypes with the greatest potential utility and to further refine product design. In part two of this work (Radford et al. 2023b), we evaluate potential benefits of and attitudes toward the new products through a series of forecasting experiments and follow-up discussions.

### 2. Methods
We chose to explore the perspectives of operational forecasters through an initial focus group to gauge the current state of winter weather ensemble visualization and an informal post-development feedback group to garner feedback and further refine designs. In addition to being an efficient way to gather responses to open-ended questions from a larger number of participants, focus groups allow exploration of shared group experiences and potentially decrease inhibitions that might limit criticisms (Tracy 2013). Product evaluation experiments were conducted in part two (Radford et al. 2023b). While co-designing products with users is an ideal development model, we sought to limit our ask of forecaster time commitments. We accomplished this by working with a different set of forecasters at each stage of work and by piggy-backing our work on existing NOAA testbed activities. IRB approval was received for each stage of the research involving human subjects, including this focus group, the following feedback group, and the evaluation groups detailed in part two (Radford et al. 2023b).

The focus group was conducted at the 2020 Weather Prediction Center’s (WPC) annual HydroMeteorological Testbed (HMT) Winter Weather Experiment (WWE). This group had seven total operational forecaster participants from the NWS WPC and Weather Forecast Offices (WFOs), with audio recording and transcription. This focus group had two key motivations. First, establishing communication with operational forecasters at an early stage of development in order to identify areas of need and increase chances that the new products would be used. This occurs not just through explicit forecaster feedback, but also through forecaster investment and interest in the design process, thereby increasing the credibility, salience, and legitimacy of our products (Cash et al. 2003). The second motivation was to gather baselines of the current state of ensemble visualization, identifying the tools that forecasters currently use, the challenges that forecasters face, and crafting objectives for new products. Focus group responses were transcribed, after which the lead author performed an inductive analysis to determine common themes (Braun and Clarke 2006).

Based on the results of the preliminary focus group, we began prototype development, seeking to address an array of forecast challenges, including snowfall rates and associated uncertainties, precipitation type uncertainties, snowfall event timing and impact (including onset and duration), and probabilities of mesoscale snowband development. At this stage a decision was made to
focus on web-based visualization using interactive Leaflet maps that allow users to pan and zoom on the output products. This decision was made based on feedback from forecasters that web-based tools are already their primary source of ensemble information and forecaster desire for greater customizability in products. Leaflet maps are inherently more customizable than static imagery as they are not locked to a particular region or scale (zoom level). A system was also put into place to produce all of the prototypes in real time for HREF output to prove the feasibility of incorporating the products into operations.

After approximately six months of development, prototype products were presented to the winter Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) group as the second phase to identify a subset of products for continued refinement and evaluation. This group consists of approximately a dozen developers, forecasters, and social scientists organized by the NWS and WPC, seeking to improve winter weather forecasting and decision-making (Perfater et al. 2023). Prototypes were roughly grouped into three categories based on visualization purpose – assessing event *magnitude*, *type*, or *timing*. After demonstrating the application of each visualization to the group, participants were asked to identify which goal and associated visualization should be pursued and refined. This feedback session was not formally analyzed. In part two (Radford et al. 2023b) of this work, we evaluate the utility of these products in a pseudo-operational setting.

3. Results

a. Preliminary Focus Group

A preliminary focus group was conducted with four WPC forecasters and three WFO forecasters to establish a baseline of the current state of ensemble visualization in forecast operations. We are careful to note that the conclusions made based upon this small group may not generalize to the broader forecaster population. Forecasters were first asked about their perspectives on the availability of ensemble products within their current operational workflow. Participants indicated that current ensemble capabilities in operations are extremely limited, with AWIPS offering only very basic ensemble products such as flattened probabilities (products that have
been stripped of information on the individual members or model evolution) and spaghetti plots. One forecaster lamented that there has been little ensemble AWIPS development because not all NWP ensemble data are available through AWIPS and getting the data into AWIPS hasn’t been a priority due to the lack of visualization development, creating a “chicken or the egg” problem. Forecasters emphasized that web-based tools such as the SPC HREF viewer are used (or abused, as one forecaster put it) over regularity.

For the example of high-impact winter weather phenomena, forecasters were asked specifically about their process for forecasting mesoscale snowbands and the role that probabilistic information plays in that process. Discussion of mesoscale banding dynamics dominated much of the focus group. As might be expected, forecasters described investigating the forecast frontogenesis and stability fields, which have been identified in previous studies to be important ingredients in the formation of bands (e.g., Thorpe and Emmanuel 1985; Nicosia and Grumm 1999; Novak et al. 2004; Baxter and Schumacher 2017). More interesting was how forecasters accessed this information from ensembles. Ensemble products highlighting frontogenesis, stability, and cross-sections of these variables are uncommon, so one forecaster described either looking at postage stamps (if the ensemble is relatively small) or cycling through the different solutions (if the ensemble is larger) until they feel they have a grasp on the solution space. While the small membership of current ensembles makes these processes feasible, potential future ensembles with hundreds or thousands of members will require strategies that scale with ensemble size. In addition, diagnostic probabilistic information is uncommon, and so this same forecaster compared his process to an old car that usually gets the job done but occasionally fails, a sentiment that seemed to resonate with others in the group. Though the prevailing opinion of the group was that better probabilistic tools are needed, one forecaster noted that, with the vast amount of information at their fingertips, more products is not always better and a simpler approach may at times be favorable. Another way ensemble data is utilized is as a check or bound on deterministic output. While forecasters appreciate the realism and interpretability offered by deterministic models, the presence or lack of the same signal in the ensemble allows them to calibrate their confidence. Quotes pertaining to current ensemble visualization tools and how forecasters interrogate probabilistic data are shown in Table 1.
Table 1: Forecaster quotes pertaining to current ensemble visualization tools and methods for interrogating ensemble data. Forecasters are anonymously identified with letters A through G and the quote is attributed to the first forecaster listed, with following forecasters reiterating or verbally agreeing with the quote.

<table>
<thead>
<tr>
<th>Forecaster Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ensembles at the WFO level are only represented through some flattened threshold probabilities. There’s some old school spaghetti with not even all of the HREF members because we don’t even get all of the HREF members yet...part of it’s that there’s been no AWIPS2 development on ‘how do you use ensemble data’ because we haven’t had the data in order to develop technically. So it’s a chicken egg problem...The post-processed stuff like the probability matched means is in [AWIPs], but that was just recently, within the last six months. We get summarized fields. The web browsers are abused all the time trying to find [more information]. So the HREF viewer at SPC, all that stuff is heavily leveraged. A lot of it is outside.</strong> [Forecaster A, Forecaster D, Forecaster G]</td>
</tr>
<tr>
<td>![We mostly have access to] threshold probabilities, percentile, or spaghetti diagrams and see how much overlap there is of the members...paintball plots for reflectivity or precip. There’s probably a little matching stuff that’s available nowadays and they have utility. <strong>[Forecaster B, Forecaster A, Forecaster C, Forecaster G; (all interjected)]</strong></td>
</tr>
<tr>
<td>We have a lot of engineered solutions and some of them are very poorly put together vehicles, but they get us to and from work, right? But once in a while you’re calling AAA. <strong>[Forecaster A, Forecaster C, Forecaster G]</strong></td>
</tr>
<tr>
<td>Sometimes simplifying the problem is exactly what you need, right? We’re really good now with a lot of tools to make it harder than it needs to be to let the complexity get you. We think we have a lot of information, and we do, but we don’t all speak that language all the time. <strong>[Forecaster C]</strong></td>
</tr>
<tr>
<td>The threat scores for snowfall for individual models are actually better than the individual ensemble members and so where the ensemble members are really helpful is you look at ‘is the axis of orientation within the ensemble members the same as the deterministic forecast?’ And that might tell you where there’s uncertainty and is the magnitude of the event [as forecast by deterministic models] duplicated within the ensemble? <strong>[Forecaster B]</strong></td>
</tr>
<tr>
<td>The deterministic model keeps you tied to all the physics of the atmosphere, so its greatest value is it meets you where you are as a physical scientist...I don’t think any of us trust a mean value from an ensemble, or even the bulk spread of an ensemble. <strong>[Forecaster A, Forecaster B, Forecaster C]</strong></td>
</tr>
<tr>
<td>You use your deterministic model as your control member and then you look at the ensemble for hints of things you might find in any random control member like snowband orientation, the amount, when the snow rates are the highest during that band, how does that snowfall come about? Do they all come about in the same way? <strong>[Forecaster C, Forecaster B]</strong></td>
</tr>
<tr>
<td>We’re kind of doing some caveman approaches to try to get at information implicitly by looking at a whole bunch of things in either rapid fashion or side by side with postage stamps...if you just animate through the ensemble members at the same time step you see the relative changes amongst each solution. It’s almost like you sit and stare at it as it loops through and you kind of get a feel for ‘now I kind of know what’s controlling this.’ <strong>[Forecaster A, Forecaster F]</strong></td>
</tr>
</tbody>
</table>

The forecasters in the group identified several key challenges to ensemble visualization to be addressed in prototype products. Perhaps the largest point of emphasis by forecasters in the preliminary focus group was increasing the transparency, interactivity, and interrogability of ensemble visualizations. Said one forecaster, “There’s a lot of stuff that runs straight to the user.
We can’t reverse engineer this stuff. [It] kind of skips the forecaster and the forecast channel and I don’t really know how to communicate about that product.” This is particularly true of web-based information, where forecasters are dependent on the developers’ ideas of what forecasters need. Given a product that emphasizes a high-impact event, forecasters need both background on how the product was produced and the capability to investigate the model perturbations and forcings to be able to communicate and add expert value to the product. This result somewhat contrasts findings by Evans et al. (2014), who found that forecasters generally used ensemble information to identify a “most likely” scenario and did not further investigate member solutions. However, this investigation of individual solutions is contextual, as forecasters may not have the time or need to access that level of detail. For example, one forecaster noted that there is a time pressure environment in forecasting, while another remarked that their analysis varies depending on what their partners need. Along those same lines, multiple participants expressed the need for tying verification data into ensemble tools. If the true solution often doesn’t fall within the ensemble solution space, the visualization likely has limited utility. The forecaster needs to know the circumstances in which they can trust the model and the visualization (i.e. the model climatology) and if this has changed over time. As a final point, even if forecasters have the time and capability to access all of the available model data and verification for a particular phenomenon, it may not be possible to produce a satisfactory graphic for all use cases. This is especially true for small-scale and low probability but high impact events, where the at-risk area is large but the phenomenon itself is small. If possible, the forecasters feel more comfortable verbalizing and contextualizing their thoughts to partners rather than presenting a graphic. Quotes related to ensemble visualization and communication challenges are shown in Table 2.

Table 2: Forecaster quotes pertaining to current ensemble visualization and communication challenges. Forecasters are anonymously identified with letters A through G and the quote is attributed to the first forecaster listed, with following forecasters reiterating or verbally agreeing with the quote.

<table>
<thead>
<tr>
<th>Forecaster Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>We can’t reverse engineer all of this stuff. [It] kind of skips the forecaster and the forecast channel and I don’t really know how to communicate about that product. Like a flat probability space – I have no idea why the space looks the way it does because I don’t have a tool of diagnosis. The inputs add a meaningful level as an expert to make an expert physical interpretation… I think the forecasters have kind of been missing in some of the applications, especially with an ensemble</td>
</tr>
</tbody>
</table>
frame...If our job is to be expert interpreters of probabilistic information to enable decision makers, you can’t wash that [out]...that’s a really challenging paradigm. You find that sweet spot – you can’t boil it down too much. [Forecaster A, Forecaster B]

We can look at the output from all different guidance - where did that come from? What initialization was changed? In order to make that member be different, especially in a single model ensemble-based system where you’re shaking the initial conditions around in a way that makes numerical sense...but does that perturbation make physical sense? [Forecaster D, Forecaster A, Forecaster C]

Every operation is a summary operation. How do you pull out from that? It’s really difficult. [Forecaster C, Forecaster B]

That time pressure environment makes you choose how much, how deep can you go? And look, I’m a slow person in parts of my life, so there are times when it’s the answer because I can’t do it fast enough. [Forecaster C]

You don’t have all the time to just go through ‘yep, nope, yep, nope,’ so it’s about how do you visualize the data in a way that you can kind of get yourself there without destroying the information that you need in order to actually go through that exercise? [Forecaster A]

The amount of details that I want to know from ensembles is gonna be different depending on what type of detail my partner needs. Like if there’s an appropriate time to look at hourly stuff or get quantitative, and then there’s sometimes where qualitative visualization is more important...there’s obviously need for both. [Forecaster G]

I think one of the biggest challenges is you need to know the model climate a little bit to know ‘this model is good at doing X, Y, and Z but it sucks at these other things. [Forecaster C, Forecaster G]

You need to know the climatology of the model but we don’t have that climatology because the model changes a lot. [Forecaster E, Forecaster A, Forecaster G]

From a national perspective, if you have a band being realized and you’ve got a bunch of data that’s somewhere in [this area], if you average that all together you squashed, flattened the space in there, and yeah, somewhere it’ll be higher, and a lot of the area won’t even precipitate potentially. We try to put a bound on it, but it’s because of banding that it will often exceed those bounds, particularly if it’s the heavy rate and longer timeframe, there’s gonna be a boom in there somewhere, that it’s unreasonable to say this is the maximum amount for this whole area. It would be nice to say ‘this is your reasonable forecast. Locally it’s going to be higher. [Forecaster D]

There’s a lot of information content that you can’t just plug in as part of your final answer on a gridded solution, but that doesn’t mean that there’s not really good information content there that you can communicate. You might have to verbalize it, but it’s still intel that as a forecaster helps you out or might even be able to help the end user out. You know, uncertainty, timing, and things like that. Or if it holds off three more hours, holy hell, this is gonna be really good today. [Forecaster A]

You know there’s going to be a corridor of severe weather, but your probabilities do not reflect anything that looks like that corridor because you don’t know where it’s going to be, but the entire area is favorable, so you have to cover it...the one super cell is the corridor and good luck figuring out what that one’s gonna start. [Forecaster C, Forecaster A]

It might be a brilliant forecast, but you don’t have the ability to actually put it down on a piece of paper in a way that’s satisfactory to anybody, but it’s great in your head. [Forecaster A]

Event timing was also brought up as a concern, echoing Demuth et al. (2020), with three sub-topics. First is the idea that atmospheric processes are inherently Lagrangian, with features evolving in both space and time, simultaneously. In particular, ensemble systems either represent time by selecting a single location or represent space by selecting a single timeframe.

Comparison of solutions at individual time stamps have limited utility to forecasters because
they are not able to convey when members have the same solution but unfold at different rates. This is particularly difficult when ensemble forecasts are clustered, as the pathway that each member takes to a similar outcome may be very different. One forecaster suggested that for some fine scale features it may even be better not to provide a map background and instead highlight the intrinsic properties of the feature, such as the intensity or shape, because it is unlikely we will know where and when they will occur. Second, ensemble systems also tend to integrate hourly data into time windows (such as 6-hr accumulations), which may or may not be appropriate for the feature of interest. For example, one particularly poignant comment foreshadowed feedback from the follow-up group and the evaluation focus groups: “it would be nicer to get into things like rates. At particular time steps, what would the rate be [associated with a mesoscale snowband]?” On the other hand, it may not be feasible to look at all of the individual hourly frames or the effect of integrating these higher rates through time may be desirable in order to highlight long-duration impacts. Systems that allow greater customizability in this regard may be beneficial. Lastly, multiple forecasters noted the importance of the time evolution of ensemble solutions, or dProg/dt. Tools or strategies for investigating the time evolution of deterministic solutions are common, but strategies for ensembles may also be valuable in evaluating the stability of signals. How to do this most effectively is an open question. For example, is it better to track the tendency of individual members and aggregate these tendencies, or to track the tendency of summary statistics, such as the ensemble mean or median? Quotes pertaining to temporal components of ensemble data are presented in Table 3.

Table 3: Forecaster quotes pertaining to temporal components of ensemble data. Forecasters are anonymously identified with letters A through G and the quote is attributed to the first forecaster listed, with following forecasters reiterating or verbally agreeing with the quote.

<table>
<thead>
<tr>
<th>Forecaster Quote</th>
<th>Forecaster</th>
</tr>
</thead>
<tbody>
<tr>
<td>It’s usually the things that you cannot objectively identify [that present a forecast challenge]. Orientation of a snowband – you can get after that. When does it happen? Sometimes you’re forced into time windows that don’t match the phenomenon that you’re trying to convey something about, so you try to get it as fine of detail as you can. Like hourly output, for example, but then forecasters might say ‘well, I don’t have time for hourly output from 37 models that I look at.’ You’re gonna have to summarize it in some way. [Forecaster C]</td>
<td>Forecaster C</td>
</tr>
<tr>
<td>How do you interpret rapidly changing things…if you look at hourly snowfall it might not look all that different, but man, when you integrate it over six hours, that’s a way different result. How do you assemble data in a way that the human can make an intuitive decision? [Forecaster A]</td>
<td>Forecaster A</td>
</tr>
</tbody>
</table>
It’d be nice to get into things like rates. At particular time steps what that rate would be. If it just happens that you’re pivoting along it, then you’re gonna have successive high rates and you get the particularly disruptive snow. [Forecaster D]

The biggest challenge that we have in general, and it doesn’t matter if it’s cold season or warm season, is the atmosphere behaves Lagrangian and we have no way of visualizing, but that’s how the processes operate. Our interpretation has to be at the system relative scale…we see this all the time. One model started two hours before the other model and so you get completely different evolution because the background environment had a chance to evolve two more hours and they might have the same forcing going on, but the initiation happened at a different time…it’s ridiculously exciting because that’s why we do this from a meteorological perspective, but it’s also horribly scary when you try to communicate that to an end user. [Forecaster A, Forecaster F, Forecaster G]

You’re actually gonna get the right answer for the wrong reasons, which is your ensemble has variability, but really the variability occurred in time. [Forecaster C, Forecaster D]

We go toward what we visualize as a cluster. Particularly for a day three forecast, we know two models got to a similar answer in a different way. You wouldn’t use that blend on day two, but suddenly on day three [you would]. [Forecaster D]

We either pick time or space to divide things up and it turns out we really need spacetime. We really need these things that coexist together…At the FO level it turns out that our most valuable aspect is that you keep the dimensionality probability space and then eliminate the space space because your users are at a pseudo point. So you can collapse the data on a timeline and maintain the integrity of the distribution so that you can show the context of the probability space. When you force the two dimensions to be X and Y, you’re now forced to pick a threshold probability or a couple…and then it makes the interpretation difficult because you can’t see the inner relationship amongst the probability thresholds…It’s a really hard problem at the national level. [Forecaster A]

You could just do one of those things where you draw how wide you think the band is gonna be. You put no map background – you just say it is gonna be a two mile wide band and it’s gonna be this long and don’t pay attention to the background map because that’s not where it’s gonna be located. [Forecaster E, Forecaster A, Forecaster C]

I think what we don’t witness here [at the winter weather experiment] so much and use all the time is time tendencies of solution space, or dProg/dt for deterministic models. But then you can kind of do the same thing with an ensemble, like how stable is this solution? [Forecaster A]

Something like plotting a time-lagged difference of of angle of a snowband from each model run so you could kind of see the variability in how the band might set up and how it’s changed over time [might be useful]. [Forecaster E]

b. Prototype Visualization Products

Given the preliminary group’s appreciation for flexibility, we decided that an interactive system would offer the greatest opportunity for product customizability and interrogability. We chose to do this with raster tiles on web-based Leaflet maps, which offer total control in spatial scale and smooth temporal transitions. While we used the focus group as inspiration for our products, there is not necessarily a one-to-one relationship between every prototype product and forecaster input. This is partially due to the fact that while forecasters generally agreed that there was need
for improved probabilistic tools, they did not necessarily have specific ideas on what these tools might look like or challenges they might address (other than timing). Instead, we prototyped products that cover a broad swath of winter forecast challenges and use additional subject matter experts to distill these to those with the greatest potential utility. The prototypes broadly address three forecasting challenges. The first set of products sought to visualize the magnitude of snowfall events and associated uncertainty in terms of snowfall accumulations and rates. The second set of products depict the type of event, emphasizing precipitation type probabilities or snowbanding potential. The third set of products visualizes event timing, including the range of onset times and durations of the event and the intersection of the event with local rush hour. This temporal set of products is most aligned with previous work by Demuth et al. (2020) and the feedback from the initial focus group. There is certainly overlap between categories, but products were assigned based on their primary forecasting goal. We now present examples of products in each category for a snowstorm in the Northeastern U.S. on Dec. 17th, 2020. Interactive web-based (Leaflet) versions of these figures will be maintained on https://www.visweather.com/bams2023 indefinitely so that readers can fully investigate and interact with examples of each product.

To the extent possible, we developed these products with the principles of risk communication in mind. Of particular importance to forecasters was contextualizing products to increase transparency and scientific soundness. This included explaining probabilities and preventing deterministic interpretations of probabilistic information, as suggested by Fundel et al. (2019). We did this first by providing background information to forecasters via a presentation of the products, the concepts behind them, and a sense of how they are produced. Once interacting with the products, a text bubble outlining similar information was available for rapid user reference. In addition to providing the basics of each product, this bubble also contextualized probabilistic information. For example, in addition to providing the 10th, 50th, and 90th percentiles for event onset time, we referred to these levels as “near the earliest possible onset,” “the most likely onset,” and “near the latest possible onset.” We also attempted to “encode information to foster accurate decoding” (Fundel et al. 2019) through the use of limited color intervals and color maps shown to improve readability, such as “viridis” and “plasma” (Smith and van der Walt 2015). Despite this attempt, data encoding would be the subject of much constructive criticism and has
since been improved to incorporate feedback from our evaluation focus groups, such as the incorporation of a mouse sampling tool.

i. Event Magnitude and Uncertainty

Snowfall Rate Percentile Plots

Percentile plots are popular for visualizing ranges of ensemble solutions. At each grid point, the solutions from each ensemble member are sorted and a percentile threshold is selected, providing viewer perspective on the low (10%), middle (50%), or high (90%) end HREF outcomes. In Figure 1 we do the same applied to snowfall rates, similar to Novak et al. (2014). Novak et al. (2014) note that higher resolution output (such as the HREF) would enable better visualization of snowfall accumulation spatial gradients. Second, Novak et al. developed snowfall accumulation plots rather than snowfall rates. Finally, the Leaflet and JavaScript delivery platform offers some advantages over existing platforms: A slider allows the user to cycle through percentile thresholds seamlessly. This is in contrast to existing products, which either place different percentile graphics side by side or require loading a new URL for a different percentile which may contribute to “change blindness,” a phenomenon in which the viewer loses information by switching focus between images (Simons and Levin 1997). Our maps are now also hoverable, allowing users to “mouse over” a location and simultaneously display the 10th, 50th, and 90th percentile solutions at that location. The products presented here use a 10:1 snow to liquid ratio (SLR). The inclusion of different SLR algorithms and the ability to toggle between each of these scenarios is a feature that would be beneficial to add in further iterations.
Fig. 1: 12-h HREF snowfall rate percentile plot valid 1200 UTC 17 December 2020. This product visualizes the snowfall rate percentile at the chosen forecast hour. The percentile can be toggled with the vertical slider to view the range of the ensemble distribution (https://www.visweather.com/bams2023 Figure 1).

**Snowfall Rate Attribute Blocks**

Portraying both the magnitude of an event and the associated uncertainty within the same image while avoiding information overload is challenging. One common method uses filled contours to display the variable mean or median, and overlaid contours to depict some measure of uncertainty. This places greater responsibility on the viewer to combine the information from the baseline (filled contours) and information from the uncertainty (unfilled contours) into a cohesive interpretation of the event. Attribute blocks (Miller 2007) are an effective means to quickly communicate different event solutions in a single graphic. The output grid is divided into blocks of $n \times n$ pixels, with each pixel in the block controlled by the value of one ensemble member. This preserves the magnitudes of the ensemble member solutions (at the cost of some spatial resolution) and ensures representation of each member solution. For example, in Figure 2, the snowfall rate output is broken up into blocks of 3 x 3 pixels, with each of the ten HREF members controlling one pixel (one member is randomly omitted in each block). Should all ensemble members have the same solution, the net result is a smooth, natural looking image.
Greater divergence between members increases the color contrast, resulting in a blockier, pixelated image while retaining a sense of the precipitation feature’s structure. A limitation of this product is that we are making an implicit assumption that neighboring 3-km pixels should have similar color hues and thus this may not be ideal for situations where you might expect very sharp gradients in a solution.

Fig. 2: As in Fig. 1, but showing attribute blocks, which partition the output grid into square blocks of pixels, with each pixel being controlled by the solution from one ensemble member. Here, attribute blocks are applied to hourly snowfall rate (https://www.visweather.com/bams2023 Figure 2). The full image shown here is an amalgam of 3 x 3 pixel blocks, as indicated by the inset grid.

**ii. Event Type**

*Ternary Precipitation Type Plot*

One of the most difficult forecasting tasks for winter weather is determining precipitation types, particularly for events featuring mixed precipitation. Generally, precipitation type visualization is accomplished by looking at the individual probabilities of each of the four main precipitation categories (rain, snow, freezing rain, and ice pellets) in four different images. These can either be
juxtaposed in a gridded format or toggled. Keeping track of four individual probabilities for a particular location in this manner is extremely challenging. The question becomes, how can we merge the individual probabilities into one graphic? We can accomplish this by making one simplification – combining the freezing rain and ice pellet categories into one “ice” category, a necessary sacrifice to reduce data dimensionality. This leaves three categories which can be represented by three color channels (red-green-blue; RGB) in a ternary plot (Figure 3). Uncertain precipitation types can thus be represented by proportional mixtures of these three colors. For example, a 50% probability of snow and 50% probability of rain would be represented with an equal mixture of green and blue, or teal. The same principle applies to mixtures of snow and ice, rain and ice, or mixtures of all three. This ternary plot effectively visualizes precipitation type transition zones in winter events. We envision this to be a product that is most valuable in support of the rate products or temporal products. For example, Demuth et al.’s (2020) onset time product (which we expand upon in section 3.b.iii) applied to snowfall rates may paint an incomplete picture, as the onset time is implicitly conditioned upon there being snow at a given location. The product does not communicate how many ensemble members predicted snow in the first place, as opposed to rain, ice pellets, or freezing rain, a piece of information that may be provided by such a p-type product.
**Snowband Contours and Probabilities**

Common among strong cool season mid-latitude cyclones are mesoscale snowbands, narrow regions of locally enhanced snowfall rates (Novak et al. 2004). Snowbands have limited predictability, even for CAMs like the HRRR (Radford et al. 2019). However, that is not to say there is no forecast skill for CAMs, and CAM ensembles could yield further improvements to snowband predictability. Visualizing times and locations with heightened potential for banding could be a useful tool for forecasters attempting to pinpoint where white-out conditions and traffic disruptions could occur. Radford et al (2019) developed a snowband detection algorithm based on simulated 1000-m reflectivity and 2-m temperature or categorical p-types. Here we present two ways to visualize this information for the HREF. The first method applies the detection algorithm to each HREF member and overlays these contours in a paintball-type plot over the mean simulated reflectivity (Fig. 4a). This provides a larger synoptic context for potential banding development. The second method uses the overlap between detected snowbands to produce a probabilistic heatmap, highlighting the most likely locations for band development (Fig. 4b).
Fig. 4: As in previous figures but showing (a) snowbands (detected according to Radford et al. 2019), then contoured and overlaid onto the HREF mean simulated 1000m reflectivity (https://www.visweather.com/bams2023 Figure 4a); (b) snowbands detected via the same method, with overlap between detected bands used to generate a probabilistic heatmap (https://www.visweather.com/bams2023 Figure 4b).

iii. Event Timing
Snowfall Rate Onset

Event onset visualizations were proposed by Demuth et al. (2020) and received positive feedback from forecasters. Demuth’s product chose a precipitation rate threshold, identified the first time that this threshold was exceeded in each ensemble member, and then visualized the mean of these onset times. We continue development of this product in three ways. First, Demuth stopped short of demonstrating the feasibility of this product using real-world data and in an operational or pseudo-operational setting, which we test through real-time implementation on https://www.visweather.com and an evaluation stage in Radford et al. (2022b). Second, rather than using the ensemble mean onset time, we apply the same procedure as the percentile plot to identify the 10th, 50th, and 90th percentile onset times. This approach (Fig. 5) seeks to reconfigure Demuth et al.’s (2020) product into a probabilistic product, more aligned with Fundel et al.’s (2019) recommendation to avoid deterministic solutions to probabilistic problems. The percentile approach captures a greater amount of the ensemble solution distribution to provide forecasters with a sense of the earliest, most likely, and latest potential arrival times. Finally, we provide a threshold slider to allow forecasters to choose the intensity threshold of interest, rather than restricting the visualization to a single threshold.
Fig. 5: As before but showing onset time plot adapted from Demuth et al. (2020), showing the first time of arrival of the chosen hourly snowfall rate, which can be toggled with the horizontal slider. The vertical slider can toggle between the “earliest” (10th percentile), “most likely” (50th percentile), and “latest” (90th percentile) onset times (https://www.visweather.com/bams2023 Figure 5).

Snowfall Rate Duration

The basis for a snowfall rate duration product also originates with Demuth et al. (2020). Here, we define a forecast period and identify a precipitation rate, then sum the number of hours that each ensemble member exceeds this rate within the period. These individual member durations are then averaged and displayed. We take the same approach here as with the onset graphic, visualizing the 10th, 50th, and 90th percentiles rather than the mean and providing a variable threshold slider to evaluate the duration of different intensities (Fig. 6).

Fig. 6: As in Figs. 1-5, except showing the total duration of the chosen hourly snowfall rate, which can be toggled with the horizontal slider. The vertical slider can toggle between the “shortest” (10th percentile), “most likely” (50th percentile), and “longest” (90th percentile) durations (https://www.visweather.com/bams2023 Figure 6).

Rush Hour Rate Probabilities
Given the NWS emphasis on IDSS, we sought to develop a product that not only assisted forecasters with prediction and communication of event timing but that also emphasizes time periods at higher risk for significant societal impacts. The most obvious of these periods is “rush hour”, roughly from 7AM to 9AM and 4PM to 6PM local time in many urban areas. We hypothesized that it would be beneficial to forecasters and their core partners to visualize the probability of exceeding a chosen precipitation intensity threshold at any point during these time periods. This product (Figure 7) displays the proportion of ensemble members that exceed a chosen threshold during rush hour, while adjusting for local time zones. We believe this could be a particularly useful tool for forecasters to communicate traffic risk with both emergency managers and commuters. A complication to this product is that rush hour may be location-dependent, a consideration that we address in part two (Radford et al. 2023b).

Fig. 7: As in previous figures but showing the probability of exceeding the chosen snowfall rate during the local rush hour (7AM to 9AM or 4PM to 6PM). The snowfall rate threshold can be toggled with the horizontal slider (https://www.visweather.com/bams2023 Figure 7).

Combination Meteogram

Evans et al. (2014) found that spatial representations of ensemble output were generally favored for QPF forecasting over meteograms. However, our final prototype attempts to identify time
periods with the potential for significant impacts using a graphical rather than a traditional contour approaches. Our goal when developing this product was to quickly draw the viewer’s attention to periods with low ensemble spread and high precipitation rates. We accomplish this with an interactive bar chart using D3 (Figure 8; Bostock et al. 2011). The height of the bars represents the probabilities of precipitation at each forecast time, while the width of the bars represents the ensemble mean precipitation rate. Thus, the time periods with the highest confidence in high precipitation rates will have bars with the largest area, drawing the viewer’s eye. Periods of low confidence but high average rates (low probability, high risk) are represented by short, thick bars. In addition, the bars are colored proportionally to indicate the conditional probabilities of each precipitation type. This chart is interactive, so hovering over a bar displays the probability of precipitation, average precipitation rate, and the precipitation type probabilities. These meteograms are produced at every point on a 0.25° x 0.25° grid across the CONUS.

Fig. 8: HREF precipitation-type meteogram showing the evolution of precipitation over time for a 41.25°N, 73.25°W for the event on Dec. 17th, 2020. Probability of precipitation is shown on the ordinate, while the width of bars is controlled by the ensemble mean precipitation rate. The color of the bar represents the conditional probability of different precipitation types (green for rain, blue for snow, red for ice pellets, and pink for freezing rain). Hovering over a bar displays a
c. Refinement Group

Needing to narrow our focus for evaluation purposes, we presented each of these sets of products to the winter FACETs team to solicit feedback on which products should be further pursued. Echoing sentiments from the preliminary focus group, this group resoundingly envisioned the greatest value in the temporal visualization products due to their clear relationship with impact-decision support services. In addition, the group identified snowfall rates, rather than total snowfall accumulations or radar reflectivity as the most important model output variable to emphasize. Though not explicitly related to event timing, the snowband probability heatmap was also reviewed favorably as potentially valuable to IDSS and thus incorporated for evaluation in part two as well. The broader scope of the FACETs program may have led to a small bias towards IDSS products, but weather forecast office experts were also involved and voiced similar sentiments to the group as a whole.

4. Conclusions

Social science research espouses the benefits of dialogue between developers/researchers and end users, but researchers often default to the deficit model of science communication, in which new insights are simply conveyed to users without feedback opportunities. Well-intentioned visualizations and tools may gather dust because they either are not trusted by users, don’t meet specific needs of users, or don’t respect the diverse viewpoints and background of users. That said, we are cognizant of the fact that forecasters have limited bandwidth for participation in research activities. In this study we balanced this realistic understanding with the benefits of product co-development by using an iterative, dialogic approach to develop new visualizations for probabilistic winter weather threats. Different forecasters participated in each phase of the work, limiting per forecaster time to a maximum of two hours. An additional component contributing to the success of this iterative process was coordination through NOAA testbeds. The initial focus group was conducted at the HMT’s winter weather experiment, where forecasters were participating in testbed activities. The evaluation groups in part two of this work
(Radford et al. 2023b), though not strictly part of the winter weather experiment, were coordinated by the HMT and benefitted from the concurrent winter weather experiment and the relationships that this annual testbed has built. The protection and expansion of testbed activities and accommodation for forecasters to participate in these activities is vital for user-driver development of new forecast products.

The preliminary focus group with operational weather forecasters at the WPC’s annual HMT WWE established that NWSFO systems are not currently optimized to display ensemble data. Nonetheless, probabilistic information is fundamental to modern weather forecasts, and forecasters have turned their attention to the web for ensemble visualization solutions. Websites such as the SPC’s HREF Viewer are possible outlets for such information – while AWIPS is still a valuable tool, external viewers have a growing influence on official NWS forecasts (Radford et al. 2023b). However, this information comes with trade-offs. Forecasters are at the mercy of developers to produce the visualizations they need and there is less transparency in how the visualizations are constructed and capabilities for customizability and interaction.

The group brought up key considerations for probabilistic visualization development, including transparency in methods, contextualization, and a desire for more customizability. The issue of product scalability wasn’t discussed directly but was implicitly addressed in how forecasters interact with ensemble output. For example, one forecaster described their ensemble analysis as toggling between members until they attain a broader understanding of the ensemble solution space. While this is feasible for a small ensemble like the HREF, it would not be possible for larger ensembles. We sought to make our products flexible enough for application to any system, regardless of membership.

Using feedback from this group as inspiration, we then set out to develop prototype probabilistic winter weather visualization products that maintain credibility in the eyes of forecasters. In response to desire for customizability, products were developed with interactive, web-based Leaflet maps, allowing forecasters to transition between location, scale, and time without cycling through a series of static images. We also attempted to contextualize the products through a brief presentation and text “help” bubbles. Visualizations were developed with three different goals.
First, snowfall rate percentiles and snowfall rate attribute blocks convey both the magnitude and forecast spread of heavy snowfall events. Second, a ternary precipitation type plot and a snowband probability heatmap communicate the type of event to expect. Finally, percentile snowfall onset and duration plots, a rush hour probability plot, and a combination meteogram build upon work by Demuth et al. (2020) to communicate temporal attributes of heavy snowfall events. Our website, https://www.visweather.com demonstrates real-time application of these tools to HREF output. Interactive versions of the figures presented in this article are also available at https://www.visweather.com/bams2023.

We presented the nine prototype visualizations to a group of winter weather experts, asking for feedback and recommendations on which products to further pursue. The group had favorable impressions of all of the prototypes but resoundingly favored the temporal products for IDSS purposes. The snowband probability product was also well-received. Thus, the probabilistic snowfall onset time and duration, rush hour probability, combination meteogram, and snowband probability tools were selected to be evaluated for operational utility in part two of this work (Radford et al. 2023b).
Acknowledgements

Support for this research was provided by NOAA Grant NA19NWS4680001, awarded to North Carolina State University. We would like to thank the NOAA NWS Weather Prediction Center HMT for organizing and hosting our focus groups and we express our gratitude to the dozens of National Weather Service forecasters who took the time to participate and provide invaluable insights into NWP visualization. We would also like to thank the Storm Prediction Center, and Dr. Brett Roberts, specifically, for making an archive of HREF system data available. We thank Jeff Waldstreicher and the winter FACETs team for valuable suggestions early in this process and for facilitating contact and discussion with other NWS groups working on similar problems.

Data Availability Statement

 Archived HREF data is available through the Storm Prediction Center: https://data.nssl.noaa.gov/thredds/catalog/FRDD/HREF.html, while real-time HREF data is available through NOAA NOMADS: https://nomads.ncep.noaa.gov/. The de-identified transcript from the preliminary focus group is stored on a shared drive and is available upon request.
References


