Meteorologists' Interpretations of Storm-Scale Ensemble-Based Forecast Guidance

KATIE A. WILSON

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

PAMELA L. HEINSELMAN

NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

PATRICK S. SKINNER, JESSICA J. CHOATE, AND KIM E. KLOCKOW-MCCLAIN

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 7 August 2018, in final form 30 October 2018)

ABSTRACT

During the 2017 Spring Forecasting Experiment in NOAA's Hazardous Weather Testbed, 62 meteorologists completed a survey designed to test their understanding of forecast uncertainty. Survey questions were based on probabilistic forecast guidance provided by the NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e). A mix of 20 multiple-choice and open-ended questions required participants to explain basic probability and percentile concepts, extract information using graphical representations of uncertainty, and determine what type of weather scenario the graphics depicted. Multiple-choice questions were analyzed using frequency counts, and open-ended questions were analyzed using thematic coding methods. Of the 18 questions that could be scored, 60%–96% of the participants' responses aligned with the researchers' intended response. Some of the most challenging questions proved to be those requiring qualitative explanations, such as to explain what the 70th-percentile value of accumulated rainfall represents in an ensemble-based probabilistic forecast. Additionally, participants providing answers not aligning with the intended response oftentimes appeared to consider the given information with a deterministic rather than probabilistic mindset. Applications of a deterministic mindset resulted in tendencies to focus on the worst-case scenario and to modify understanding of probabilistic concepts when presented with different variables. The findings from this survey support the need for improved basic and applied training for the development, interpretation, and use of probabilistic ensemble forecast guidance. Future work should collect data for a larger sample size to examine the knowledge gaps across specific user groups and to guide development of probabilistic forecast training tools.

1. Introduction

Uncertainty is inherent in forecasts of any natural system, including the weather. The limited predictability of the atmosphere and the resulting initial value problem thus calls for an ensemble of numerical weather predictions that can provide probabilistic forecast information (Bauer et al. 2015). Advancements in scientific understanding, computing resources, and observations have led to the development of operational numerical weather prediction systems that span the temporal and spatial scales of seasonal global forecasts to daily regional forecasts. Improvements in the skill of these forecast systems have been observed over the past several decades (Magnusson and Källén 2013). Since the mid-2000s, the development of convection-allowing models (CAMs) has provided short-term forecast guidance on the timing, intensity, location, and coverage of storms (e.g., Done et al. 2004; Sobash et al. 2011; Sobash and Kain 2017). An initial evaluation on the use of CAMs during the NOAA Hazardous Weather Testbed Spring Forecasting Experiment (SFE) in 2004 found that this guidance added value during the production of human forecasts for severe weather (Kain et al. 2006). The evaluation of CAMs has since expanded to a larger number of experimental ensemble systems in subsequent

Corresponding author: Katie Wilson, katie.wilson@noaa.gov

DOI: 10.1175/WCAS-D-18-0084.1

SFEs, motivating the development of the Community Leveraged Unified Ensemble in 2016 (Clark et al. 2018). Demonstrations of CAM benefits to the forecast process have resulted in several CAM systems becoming operationalized, namely, the High Resolution Rapid Refresh (HRRR; Benjamin et al. 2016) and the High Resolution Ensemble Forecast System version 2 (HREFv2; Jirak et al. 2018).

A more recent addition to the SFE activities is the testing and evaluation of the NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e). Developed as the prototype CAM ensemble system for the NOAA Warn-on-Forecast project (Stensrud et al. 2009, 2013), NEWS-e, and other prototype Warn-on-Forecast systems, provides real-time and short-term probabilistic forecast guidance for a variety of weather threats within the watch-to-warning timeframe (0–6 h) over a regional-scale forecast domain (e.g., Wheatley et al. 2015; Jones et al. 2016; Skinner et al. 2016; Yussouf et al. 2016; Skinner et al. 2018). In 2017, SFE participants used NEWS-e forecast guidance to issue two 1-h total severe outlooks (i.e., combined threats of severe hail, straight-line winds, and tornadoes). While this hands-on activity gave participants exposure to NEWS-e forecast guidance, the researchers did not expect for the group-driven end products to reveal the ways each participant understood, extracted, and applied the probabilistic forecast guidance. Given that CAM ensemble-based forecast guidance is becoming more widely available to forecasters, it is crucial to know what current knowledge and interpretive skills meteorologists employ when assessing such uncertainty information.

Research has shown that accurate interpretation of uncertainty information can be challenging for meteorologists, which in part stems from difficulties in correctly identifying and understanding the ways in which probabilities are calculated and the reference class they represent. For example, probabilistic information can be calculated using approaches based on frequency, subjectivity, or climatology (de Elía and Laprise 2005; AMS 2008). Additionally, probabilities have the potential to represent different aspects of a weather event, such as the frequency of occurrence, the areal extent in which it occurs, or its timing. Numerous studies have shown that incorrect attribution of uncertainty information to one of these aspects is a common cause for misinterpretation of probability of precipitation (PoP) forecasts among the U.S. public (Murphy and Winkler 1971, 1974; Gigerenzer et al. 2005; Joslyn et al. 2009) and among meteorologists (Stewart et al. 2016). Inaccurate interpretations of uncertainty information have the potential to result in inconsistent messaging of risk information, which in turn can present difficulties for end users when making sense of forecast products.

Despite the known challenges associated with understanding uncertainty information, over the past decade, numerous reports have outlined the potential benefits of quantifying uncertainty to both forecasters' and special end users' (e.g., broadcast meteorologists and emergency managers) decision-making (e.g., NRC 2006; Joslyn et al. 2007; AMS 2008; Hirschberg et al. 2011; Demuth et al. 2009). Additionally, in a survey investigating laypeople's use and understanding of forecast information, most respondents indicated a willingness or preference for receiving probabilistic forecasts (Morss et al. 2008). The movement toward the greater use and dissemination of uncertainty information is also supported in the Forecasting a Continuum of Environmental Threats (FACETs) concept (e.g., Rothfusz et al. 2014; Karstens et al. 2015; Rothfusz et al. 2018), where a continuum of probabilistic information is expected to drive weather-related decisions and modernize the current NWS watch and warning system. To achieve the FACETs vision and ensure the potential utility and benefits of probabilistic guidance is maximized, assessing forecaster knowledge and training needs is essential. In surveying NWS forecasters' understanding of uncertainty information, Novak et al. (2008) identified a need for improved education and training in this area. To make use of the guidance information within the forecast process, forecasters expressed a need to not only learn more about how ensemble prediction systems are constructed and how the guidance is derived, but to also have access to basic interpretation and application training of ensemble output (Novak et al. 2008).

Given this need for improved knowledge of forecasters' understanding of probabilistic guidance, the study presented herein was designed to examine the 2017 SFE participants' understanding and interpretation of NEWS-e forecast guidance prior to using it during the hands-on activity. The survey questions were developed to learn about participants' understanding of basic uncertainty concepts and to assess what mental models participants used when presented with NEWS-e probabilistic forecast guidance. To the authors' knowledge, this study is the first of its kind to examine meteorologists' understanding of probability concepts beyond PoP forecasts. The overall survey approach is described in section 2, and participants' responses to each of the questions are described in section 3. These results help to highlight the extent to which participants were able to successfully answer questions focused on various aspects of uncertainty information, the types of topics or tasks that were particularly challenging to participants, and the specific training needs required to enable effective assessments of NEWS-e (and other) probabilistic forecast guidance in the future.

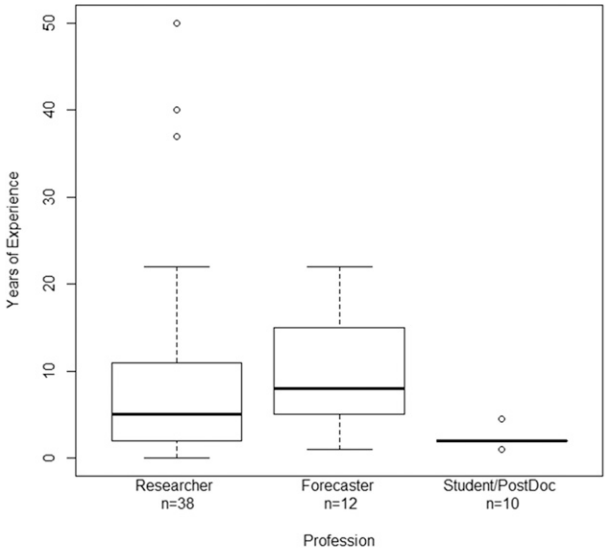

FIG. 1. Participants' years of experience in their reported professions, including research, forecasting, and student/postdoc.

2. Survey design and analysis

The NEWS-e survey was administered during the 2017 SFE, which is an annual NOAA Hazardous Weather Testbed experiment that evaluates new concepts and technologies for improving severe weather prediction (e.g., Kain et al. 2003; Clark et al. 2012; Gallo et al. 2017). Over the course of 5 weeks, 62 meteorologists participated in this survey prior to their initial exposure to NEWS-e. The survey took approximately 30 min to complete. The 62 meteorologists varied in terms of their occupation and years of experience (Fig. 1), with over half of them working in research (61%) and a smaller portion working in operational meteorology (19%) or working toward a graduate degree/ postdoc (16%). Participants were predominantly male (84%), and the majority of participants held a postgraduate degree (85%). Two respondents did not disclose their occupation, gender, or education.

The survey administered to participants consisted of a series of multiple-choice and open-ended questions designed to assess their understanding and interpretation of probabilistic and percentile concepts as they relate to ensemble-based probabilistic forecasts (see appendix). Most importantly, participants needed to understand how probability of exceedance (Fig. 2a) and percentile (Fig. 2b) values are derived from probability and cumulative distribution functions, respectively. Although historically percentile concepts have not been applied to forecasting in the way probability concepts have, the increasing availability of ensemble forecasts supports users' abilities to incorporate personal risk tolerance into forecasts and identify potential worst case scenarios (Novak et al. 2014). The use of percentiles to assess and communicate weather risk is thus being explored and tested for a variety of weather phenomena. The survey questions required participants to explain, interpret, and extract information from visualizations for a range of weather scenarios. Surveys were completed independently on computers, and all responses were saved to a database. While completing the survey, participants were able to seek clarification on questions from assisting researchers. Fewer than 10% of participants sought this clarification.

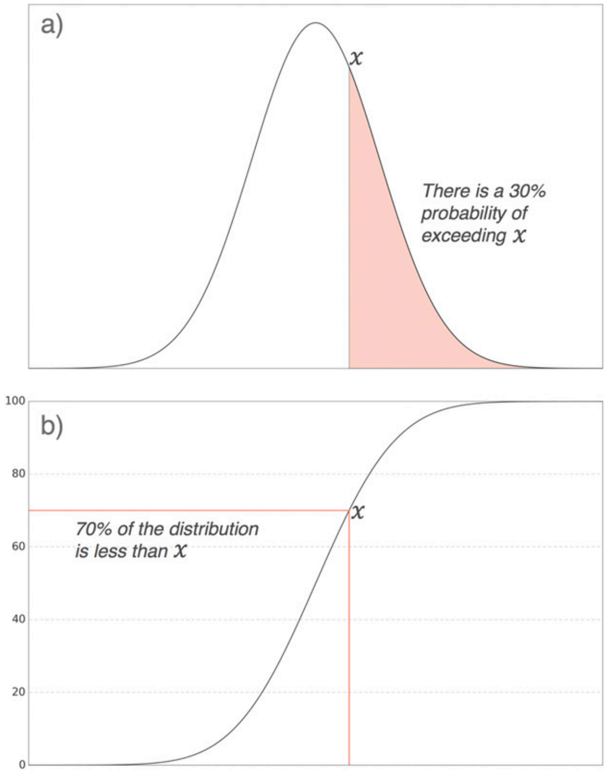
a)
x
There is a 30% probability of exceeding x
b)
100
80
60
40
20
0
x
70% of the distribution is less than x

FIG. 2. Schematic showing (a) the probability of (or percent of ensemble members) exceeding a threshold value X and (b) the corresponding percentile at which value X occurs for a Gaussian distribution.

All survey responses were combined for each question and analyzed. Multiple-choice questions or questions requiring a simple response (e.g., reporting of a probability value) were analyzed using a frequency count. Nine open-ended questions that yielded more elaborate responses were examined in greater depth. After identifying recurring themes within the responses, the five authors met regularly to ensure that they agreed on the meaning and interpretation of these different themes. Once consensus was established, two research team members were assigned to each of these open-ended questions and coded all of the qualitative responses. The

TABLE 1. Concepts tested in the survey.

Probability concepts	Percentile concepts	Combined probability and percentile concepts
Making sense of and extracting information from visualizations using probability values and exceedance thresholds.	Explaining percentile representations of accumulated rainfall.	Using percentile representations of accumulated rainfall to determine the probability of exceeding a given value of rainfall.
Relating probability values to ensemble members.	Comparing percentile and PoP representations of accumulated rainfall.	Combining probability and percentile representations of accumulated rainfall and 2–5-km UH to decipher whether the graphic depicts a high-/low-probability event with high/low consequence.
Understanding why probability values decline in swaths of accumulated rainfall and UH.	Using percentile representations of accumulated rainfall to extract the least, greatest, and range of expected rainfall amounts.	

coding process required the researchers to consider the meaning of each response and attribute it to the relevant theme(s) identified previously [see Saldaña (2016) for further details on qualitative coding]. The pairing of researchers varied for most of the nine questions. To ensure sufficient consistency between researchers, interrater reliability was calculated for each of these coded questions using Cohen's kappa. Unlike the more simplistic percentage agreement statistic, the Cohen's kappa takes into account the possibility of chance agreement between two coders (Cohen 1960; Fleiss 1981; McHugh 2012). The kappa statistic ranges from −1 to +1, with larger positive values representing greater interrater reliability. Using the *Kappa.test* function in R and the thresholds recommended by Landis and Koch (1977), almost perfect interrater agreement was established (i.e., kappa ≥ 0.8) for eight of the nine questions, and substantial interrater agreement (i.e., $0.6 \leq \text{kappa} < 0.8$) was established for the remaining question. In the following sections, results from these coded responses and those from the multiple-choice questions are shared. Questions are grouped depending on whether they assessed participants' knowledge of probability concepts, percentile concepts, or a combination of probability and percentile concepts, which are further described in Table 1 and the appendix. The results are organized and discussed according to these three groups, and participants' overall success in answering each question is summarized in Fig. 3.

3. Results

a. Probability concepts

The survey began with a question that required participants to extract and make sense of probability values for accumulated rainfall exceeding 0.01 in. (Fig. 4). Participants were first asked to describe the type of weather event that the graphic depicted (Q1). The researchers' intended response was that a widespread area, with some isolated regions, has a greater than 90% probability of exceeding 0.01 in. of rainfall between 0000 and 0130 UTC. Of the responses, $n = 13$ participants provided most if not all elements of the intended response (e.g., ''a relatively large area characterized with a high probability of precipitation accumulation exceeding 0.01 in. threshold during an hour and a half period''), while $n = 17$ participants provided responses focused on the definition of the product (e.g., ''probability of measurable precipitation based on how many ensemble members depict QPF over 0.01″ in a grid point''). Of the remaining participants, $n = 13$ reported that the graphic depicted heavy rainfall or severe storms, while $n = 19$ either simply described the widespread nature of the event or speculated on the storm mode and/or forcing mechanisms responsible for the event. While some of these answers captured elements of the intended response, about a third of participants misinterpreted the graphic and/or question, which was near the maximum percentage of misunderstanding demonstrated in any question (Fig. 3). Notably, the most common error involved an inference of severity that went beyond the presented probability information. The design of the graphics and a deterministic construal error are possible explanations for this inference. The color code used in the graphic resulted in a widespread shade of red in Fig. 4, which is often associated with danger (Ash et al. 2014). Additionally, the apparent oversimplification of probabilistic forecasts resulted in deterministic responses. While participants may have reduced their cognitive loads to provide such responses, the given uncertainty information was not used to inform their understanding of the event (Savelli and Joslyn 2013).

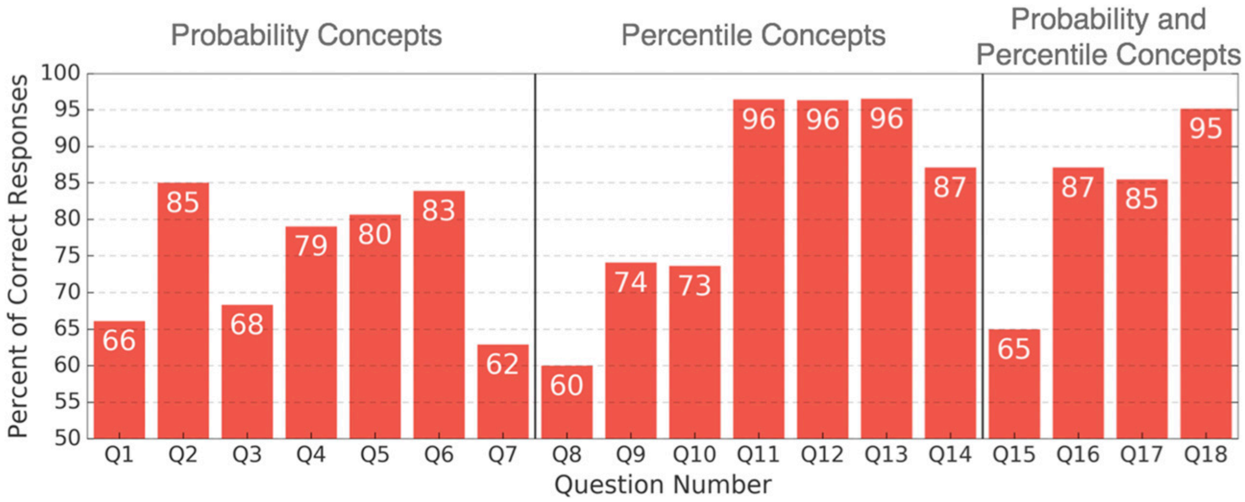

FIG. 3. The percent of participants providing correct responses for Q1–Q18.

Using this same graphic, participants were then asked to identify the maximum amount of rainfall possible within boxes A and B (Q2 and Q3). The intended response was that the maximum could not be determined from the given information, but that the graphic shows a 90% probability at A (Q2) and a 10%–20% probability at B (Q3) of exceeding 0.01 in. of rainfall. In both questions, participants' correct responses most frequently stated they could not determine the maximum value (n = 43 in Q2 and n = 39 in Q3) and/or there was some chance of receiving greater than 0.01 in. of rainfall (n = 32 in Q2 and n = 26 in Q3). Some participants gave both of these responses (e.g., ''this information isn't sufficient to convey a maximum amount of precipitation, only the likelihood that any measurable precipitation will fall''), while others elaborated further by indicating the high/low likelihood of receiving greater than 0.01 in. of rainfall (n = 12 in Q2 and n = 19 in Q3). Additionally, in Q3, a subset of participants (n = 10) reported the relative likelihood of exceeding the same amount of rainfall at location B compared to location A (i.e., B < A with strong/moderate confidence or B = A with lower probability).

A subset of participants in Q2 and Q3 either did not grasp that a definitive answer could not be provided using the information presented in Fig. 4, or were influenced by perceived demand characteristics (e.g., Orne 1962) and inferred that an answer must exist given the question was asked. These participants instead reported a specific maximum value of accumulated rainfall ranging between 0.01 and 5 in. Examples of responses include ''with such little information I would assume around an inch of rain for box A,'' and ''around one quarter inch, assuming the CAMs are correct with placement of convection.'' The most popular values given were 0.01 and 1 in. Participants giving the former value did not demonstrate an understanding of the exceedance threshold, while participants giving the latter may have misinterpreted the color bar and assumed the colors represented accumulated rainfall amounts rather than probabilities. There is little explanation for why the other values were given (e.g., 0.03, 0.25, and 5 in.). Despite the similarity in question style and use of the same

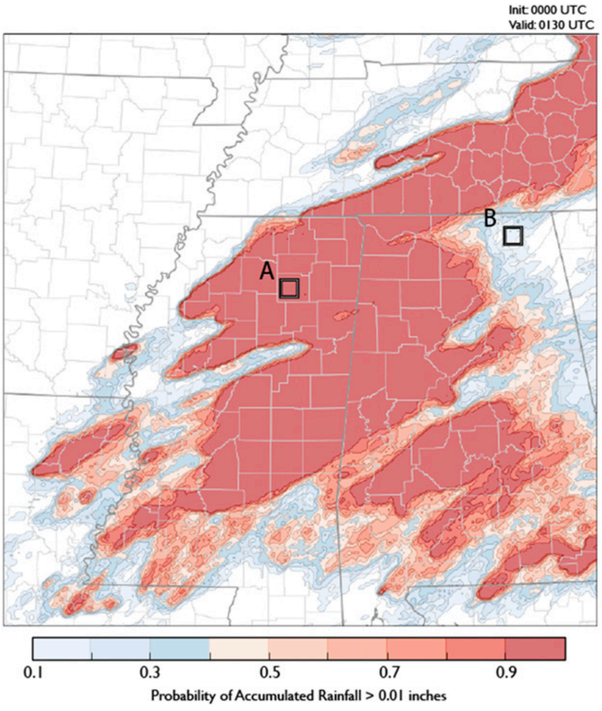

FIG. 4. Probability of accumulated rainfall exceeding 0.01 in. (referred to in Q1–Q3).

graphic in Q2 and Q3, more than twice as many participants did not provide the intended response in Q3 ($n = 19$) compared to Q2 ($n = 9$; Fig. 3). Therefore, some participants were unable to apply their conceptual understanding of the problem consistently across locations A and B.

Participants were presented with a new graphic in Q4 that tested their understanding of how probabilities relate to ensemble members (Fig. 5). In this question, participants needed to correctly identify which panel represented the exceedance threshold of 0.01 in. of rainfall (top), extract the color and corresponding probability value at location B, and use those values to estimate the number of ensemble members (out of 20) predicting at least 0.01 in. of rainfall. The intended response was two, three, or four ensemble members (i.e., 10%–20% of ensemble members). Most participants ($n = 45$) provided a response within this range, with $n = 31$ reporting two members, $n = 11$ reporting two or three members, and $n = 1$ each reporting three members, two to four members, or four members. Additionally, $n = 4$ participants reported a percentage of members [10% ($n = 1$), 10%–20% ($n = 2$), and 20% ($n = 1$)], and although not incorrect, these participants did not successfully translate the percentage of members into a quantity. The remaining $n = 13$ participants gave incorrect responses, including one member ($n = 6$), one or two members ($n = 2$), 18 members ($n = 2$), zero members ($n = 1$), zero or one members ($n = 1$), and "information is not given" ($n = 1$).

Next, using this same graphic (Fig. 5), Q5 asked participants to find the probability of rainfall exceeding 0.5 in. in box A. In this question, participants needed to extract the color located at box A in the middle panel that represented the 0.5 in. of rainfall exceedance (Fig. 5). As in Q4, this color represented probability values of 0.1–0.2, and the intended answer was therefore a 10%–20% chance of box A exceeding 0.5 in. of rainfall. The majority of answers ($n = 50$) fell within this range, with participants reporting 10% ($n = 37$), 10%–20% ($n = 8$), 20% ($n = 3$), 15%, ($n = 1$), and 10%–19% ($n = 1$). Some participants continued to think about the problem in terms of ensemble members, reporting two members ($n = 4$) and two to four members ($n = 1$). The remaining participants ($n = 7$) did not provide the intended response and instead answered 0% ($n = 2$), 0%–20% ($n = 1$), 5%–10% ($n = 1$), 90% ($n = 1$), 18–20 members ($n = 1$), and "low but A is higher than B" ($n = 1$). Four of these seven participants also did not provide the intended response in the previous Q4.

The final two probability concept questions investigated participants' understanding of why probability values decline in swaths of (Q6) accumulated rainfall exceeding a

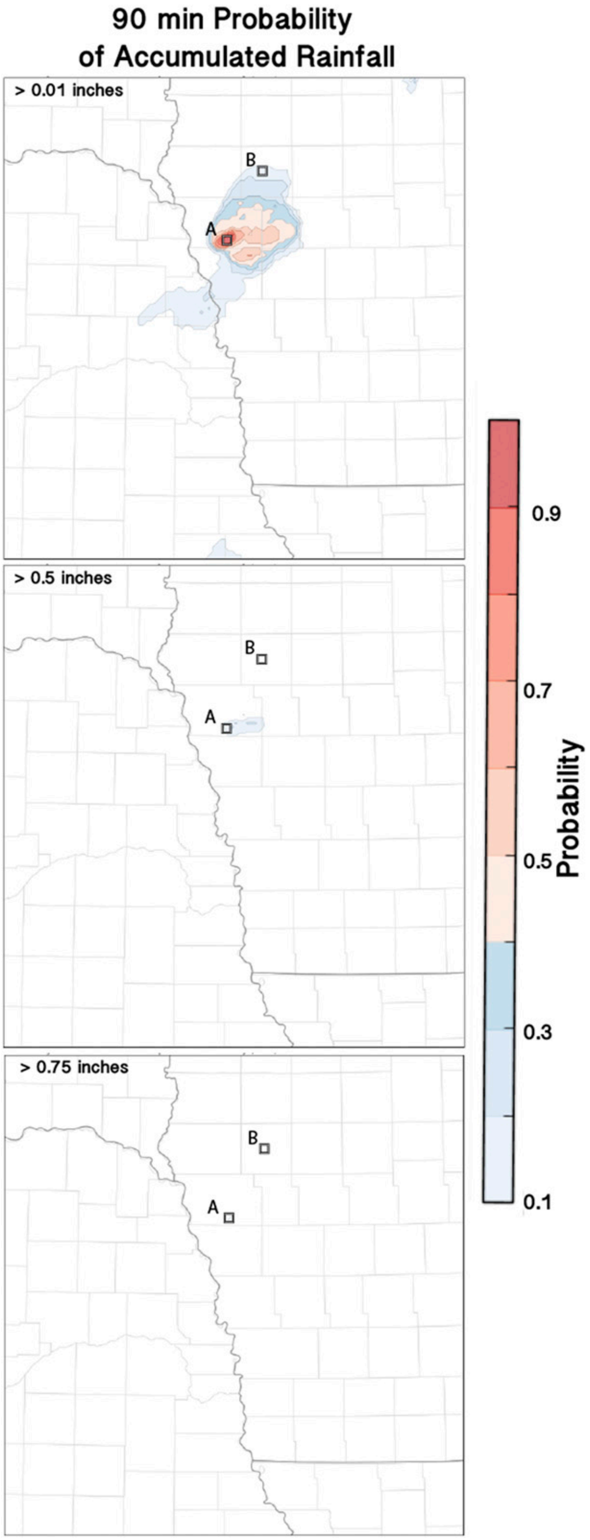

FIG. 5. The 90-min probability of accumulated rainfall exceeding (top) 0.01, (middle) 0.5, and (bottom) 0.75 in. (referred to in Q4 and Q5).

threshold of 0.25 in. and (Q7) updraft helicity (UH) exceeding a threshold of $20\,\mathrm{m}^2\,\mathrm{s}^{-2}$ (Kain et al. 2008; Fig. 6). Presented with four possible answers, participants were asked to select which answers most intuitively explained this effect (appendix). Regardless of the weather scenario presented, an underlying driver for the declining probability values is the *increased trajectory uncertainty toward the end of the swath*; this answer was therefore expected to dominate participants' responses in both questions. For the accumulated rainfall example, most of the participants ($n = 52$) selected the trajectory uncertainty answer (Q6). However, participants ($n = 42$) tended to choose more than one answer and therefore also selected that the *storm is expected to decrease in intensity/strength* ($n = 32$) and/or that *there is less time for the rainfall to accumulate at that point during the forecast period* ($n = 33$). The least popular answer was also the least likely to be true: *the storm is growing in size and rainfall rate is forecasted to decrease* ($n = 5$). Unlike in Q6, the most popular answer for the updraft helicity example in Q7 was that *the storm is expected to decrease in intensity/strength* ($n = 47$). Fewer participants selected that the *trajectory uncertainty is greater toward the end of the swath* ($n = 39$), most of whom also selected this answer in Q6 ($n = 35$). Overall, participants' selections were more concentrated to one or two choices in Q7, so a smaller subset of participants (compared to Q6) selected the less likely answers that *there is less time to experience maximum updraft helicity during the forecast period* ($n = 14$) and that *the storm is growing in size and rotation is broadening* ($n = 6$). The results from Q6 and Q7 show that for some participants, their interpretation of declining probability values depended on the underlying meteorological variable of the probability swath.

b. Percentile concepts

To assess participants' knowledge and understanding of percentile concepts, Q8 asked participants to explain what the 70th-percentile value of accumulated rainfall from an ensemble-based probabilistic forecast represents (Fig. 7). The qualitative responses ($n = 60$) were thematically coded and sorted into responses that demonstrated a clear understanding of this concept ($n = 36$; Table 2) and responses that demonstrated misunderstanding or ambiguity ($n = 24$). Most participants demonstrating a clear understanding explained that 70% of the ensemble members had a value less than what was shown for the 70th percentile ($n = 29$). It is possible that the phrasing of the question biased participants' responses to focusing on the 70th percentile. A smaller portion of participants explained that 30%, or a minority, of ensemble members had a value more than what was shown ($n = 10$). Of these participants, most responses ($n = 8$) had also been coded for the previous theme. A smaller number of participants ($n = 8$) qualitatively described the 70th percentile as being a high-end

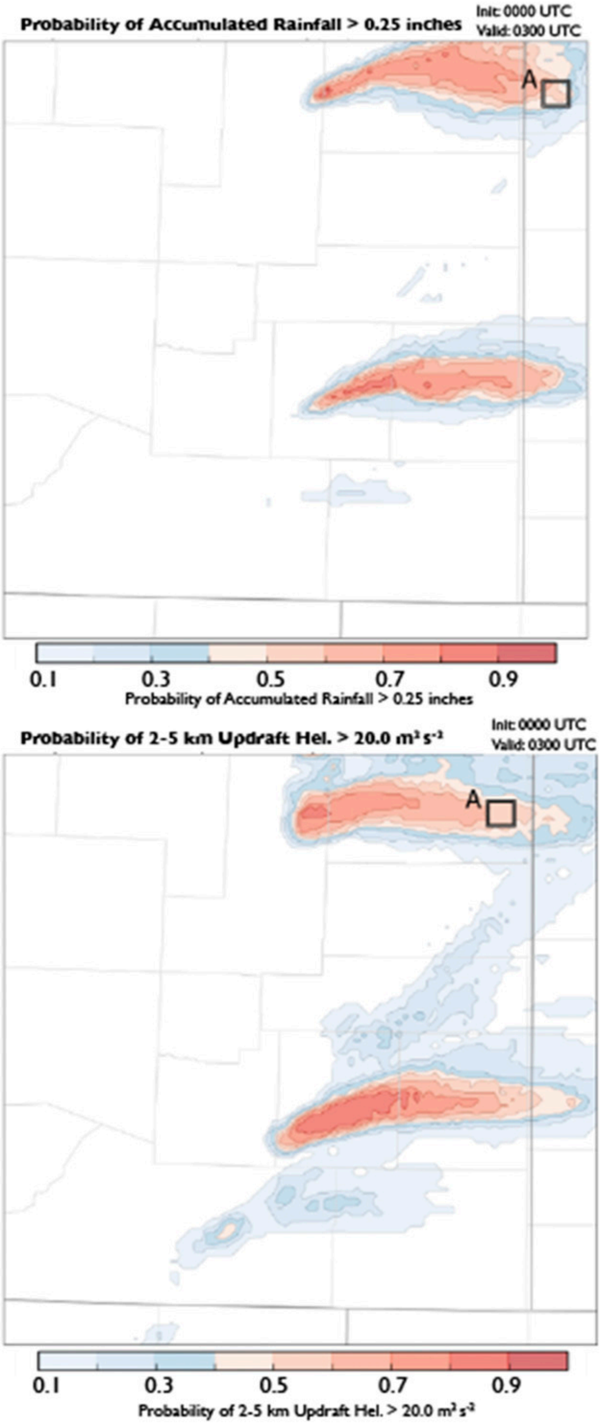
Probability of Accumulated Rainfall > 0.25 inches
Init: 0000 UTC
Valid: 0300 UTC
A
0.1
0.3
0.5
0.7
0.9
Probability of Accumulated Rainfall > 0.25 inches
Probability of 2-5 km Updraft Hel. > 20.0 m^2 s^{-2}
Init: 0000 UTC
Valid: 0300 UTC
A
0.1
0.3
0.5
0.7
0.9
Probability of 2-5 km Updraft Hel. > 20.0 m^2 s^{-2}

FIG. 6. Swaths of maximum ensemble probability of (top) accumulated rainfall exceeding 0.25 in. and (bottom) 2–5-km UH exceeding $20\,\mathrm{m}^2\,\mathrm{s}^{-2}$ for two storms (referred to in Q6 and Q7).

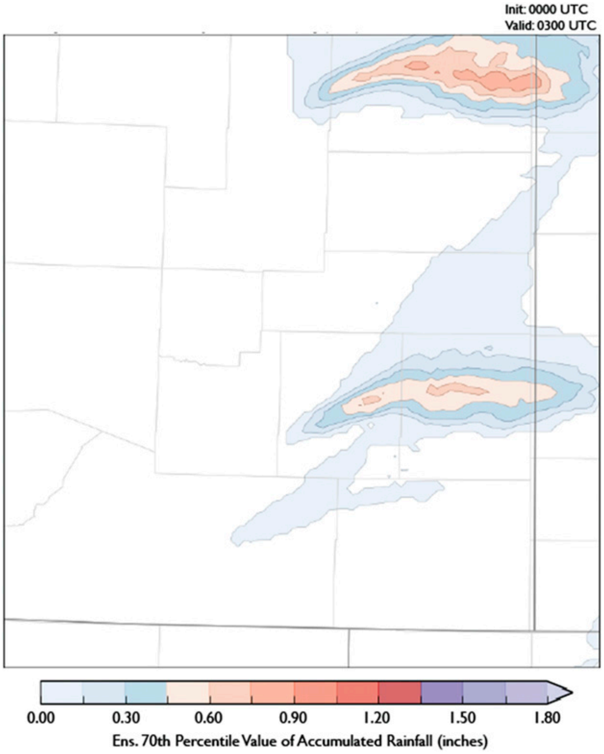

FIG. 7. The ensemble 70th-percentile swath of accumulated rainfall (referred to in Q8–Q9).

possibility or that the values showed something akin to the maximum. Half of these participants' responses were also coded for the first theme. Finally, a subset of participants made reference to the probability distribution function (pdf) concept in their responses. In addition to fitting this pdf theme, all but one of these participants' responses were coded for at least one of the other three themes. Responses indicating misunderstanding or ambiguity included explanations such as, ''accumulated rainfall that at least 70% of the members agree on,'' ''the regions shaded represent the union of at least 70% of ensemble members,'' and ''the accumulated rainfall amount that 70% of the members are producing.'' Rather than recognizing that the 70th percentile represented a rank within a distribution, these examples suggest that some participants applied deterministic ideas to understand ensemble output. It is also possible that this misinterpretation stemmed from participants confusing the concept of percentile value with that of the probability of exceedance.

Participants' understanding of percentile representations of forecast uncertainty was further assessed in Q9, where they were asked to explain how it compares to the PoP that is used today. As in Q8, the qualitative responses ($n = 58$) were thematically coded, and in comparison, more participants demonstrated a clear understanding ($n = 43$), and fewer demonstrated misunderstanding and ambiguity ($n = 15$). Most participants demonstrating a clear understanding explained that the percentile representation of rainfall gives information on the amount of rainfall, whereas PoP does not ($n = 31$). The next most frequent theme was that unlike PoP, the percentile representation does not tell you the incidence/probability of rainfall ($n = 16$). Most of these participants' responses were also coded for the first theme ($n = 12$). A smaller number of participants ($n = 8$) noted differences in the coverage and/or timing information provided in the percentile and PoP representations of rainfall, and, as in Q8, a group of (different) participants ($n = 12$) referenced the pdf concept in their responses. Participants demonstrating misunderstanding or ambiguity gave varying responses, including that the percentile and PoP representations of rainfall were ''roughly equivalent,'' that ''I believe it will show overall higher probability values,'' and that ''much of the model uncertainty (in areal coverage) has been removed, showing where the greatest agreement among ensemble members remain.'' These responses suggest that some participants have not yet grasped the meaning of and differences between percentile and PoP representations of rainfall. In particular, 10 participants struggled to convey a clear understanding of these concepts in both Q8 and Q9.

The remaining percentile concept questions (Q10–Q14) required participants to examine graphics of the 10th, 50th, and 90th percentiles of accumulated rainfall and extract the least and greatest amounts for boxes A and B. In each of these questions, four to seven participants (often the same people) did not provide a response. For Q10, participants were asked to find the least amount of rainfall expected in box A. Participants first needed to recognize that ''least'' corresponded to the 10th-percentile panel (Fig. 8) and then use the color bar to find what value in the circle of the inset for box A was represented. Most participants reported an accumulated rainfall value corresponding to pink/red colors, with greatest consensus ($n = 28$) being between 0.75 and 1.05 in. (Fig. 9a). Eight or fewer participants each reported a value corresponding to one of the other shades of pink. Additionally, $n = 15$ participants gave values that were either at the lower end (blues) or higher end (purples) of the scale (Fig. 9a). Based on a handful of queries from participants when taking the survey, we believe that some of these outlier values were selected due to confusion over whether to report the color in just the circle or the entire inset box. If assessing the entire box, blue colors are evident, which may also be construed as purple depending on participants' sensitivity to the two colors.

TABLE 2. Examples of themes identified in participants' responses to Q8 and Q9.

Question 8: In an ensemble-based probabilistic forecast, what do you think the 70th percentile value of accumulated rainfall represents?

Themes	70% of members have a value less than this/70th percentile ($n = 29$)	30% (or at least a minority) of members have a value more than this ($n = 10$)	High-end possibility/showing something akin to the max ($n = 8$)	PDF concept ($n = 12$)
Response examples	"That 70% of solutions will be within this amount."	"That there is about a 1-in-3 chance of at least this amount of rain."	"This would capture some of the higher rainfall amounts expected from storms."	"If there are 10 ensemble members, the QPF from the 7th-highest member at any given point."
	"It means that 70% of ensemble members have this much precipitation or less over the accumulated time period."	"The 70th percentile value represents the upper 30 percent of membership values accumulated rainfall."	"The upper-end of the likely precipitation range. . ."	"Accounting for all the accumulated rainfall values produced by the ensemble members, the values are aggregated and ordered by magnitude. The 70th percentile is as near to the 70th highest value as is possible to obtain."
	"70% of the ensemble members produced a rainfall accumulation smaller than this value (at each grid point)."	"70% of the ensemble members have less than the amount shown /30% of the members have more."	". . . Getting towards a 'maximum reasonable' precipitation estimate from the raw ensemble distribution."	

Question 9: How does this representation of forecast uncertainty compare to the PoP used today?

Themes	Amount of rain ($n = 31$)	Incidence/probability of rain ($n = 16$)	Coverage and/or timing ($n = 8$)	Ensemble-based/pdf concept ($n = 12$)
Response examples	"It provides a better display of rainfall amount compared with PoP."	"This is different to PoP because it is not a probability."	"It provides a better representation of the spatial distribution of the magnitude of the values. I think it may connect better to the dynamics of the system intuitively."	"It provides more nuanced information as to the spread in the ensemble members, whereas PoP is more of a deterministic yes/no metric."
	"It gives an idea of how much rainfall could occur, I think this could be good for communicating flash flood risk. I think people view PoP as either yes or no today rather than a chance, and I think this may be a good tool to convey the severity of rainfall."	"PoP corresponds to the probability of exceeding an amount. Which is decidedly different from a percentile, which is not related to exceedance."	"PoP would also take into account coverage. This does not take into account areal coverage and instead focuses on grid-point values from each individual ensemble member."	"The PoPs used today convey the likelihood of any measurable precipitation at a given point for a certain forecast period. The information regarding the percentile values actually show a portion of the underlying PDF amount."

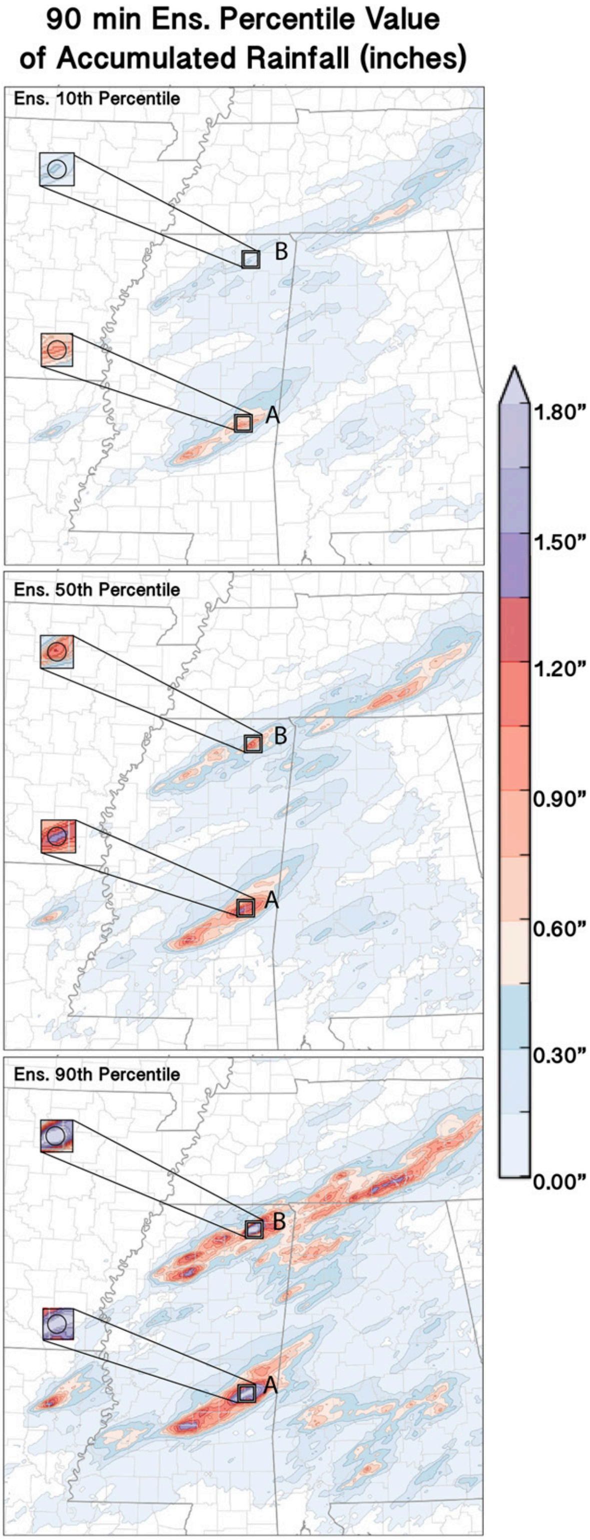

FIG. 8. The 90-min ensemble (top) 10th-, (middle) 50th-, and (bottom) 90th-percentile values for accumulated rainfall (referred to in Q10–Q16).

Participants next reported the greatest amount of rainfall in box A (Q11). Almost all of the participants correctly referred to the 90th percentile and reported values corresponding to a shade of purple, with greatest consensus ($n = 44$) corresponding to the highest contour (>1.8 in.) of rainfall (Fig. 9b). A subset of participants ($n = 10$) reported values corresponding to other shades of purple representing lower amounts of accumulated rainfall. Two participants reported a value corresponding to pink, possibly due to confusion from the presence of this color located outside of the circle in the box A inset (Fig. 9b).

The same questions were asked in Q12 and Q13 but for box B. In the first of these two questions, all but two participants successfully identified the least amount of accumulated rainfall in box B as corresponding to one of the blue contours (Fig. 9c). Participants' responses were also in strong consensus in Q13 for the greatest amount at this location, with again all but two (different) participants reporting values corresponding to a shade of purple (Fig. 9d). These two participants instead reported values corresponding to pink, possibly due to this color appearing outside of the circle in the box B inset. For Q10–Q13, participants occasionally included modifiers with their reported values, with the greater-than modifier (>) being used most frequently. This modifier was used predominantly in Q11 and Q13, likely due to these questions asking for the greatest value and possibly influenced by the arrow located at the top of the color bar indicating values in excess of what was shown.

Finally, Q14 asked participants to identify whether location A or B had the greatest range of potential rainfall. Most participants ($n = 54$) were able to integrate the information provided in the different percentile graphics to determine the correct answer of location B. Of those participants answering incorrectly, half also provided at least one incorrect/unanswered response in Q10–Q13, while the other half provided correct responses for those questions but were unable to assess the potential rainfall ranges successfully.

c. *Combined probability and percentile concepts*

Approximately one-third of the survey questions used both probability and percentile concepts to test whether participants could combine their understanding of the two concepts to provide correct responses. The first two questions, Q15 and Q16, required participants to use the percentile representations of accumulated rainfall from the previous question (Fig. 8) to determine the probability of exceeding a given value of rainfall. In Q15, over half of the participants ($n = 35$) gave the intended response of 50% for the probability of exceeding 1.35 in. of rainfall in box A. These participants were able to correctly extract the purple contour on the color bar

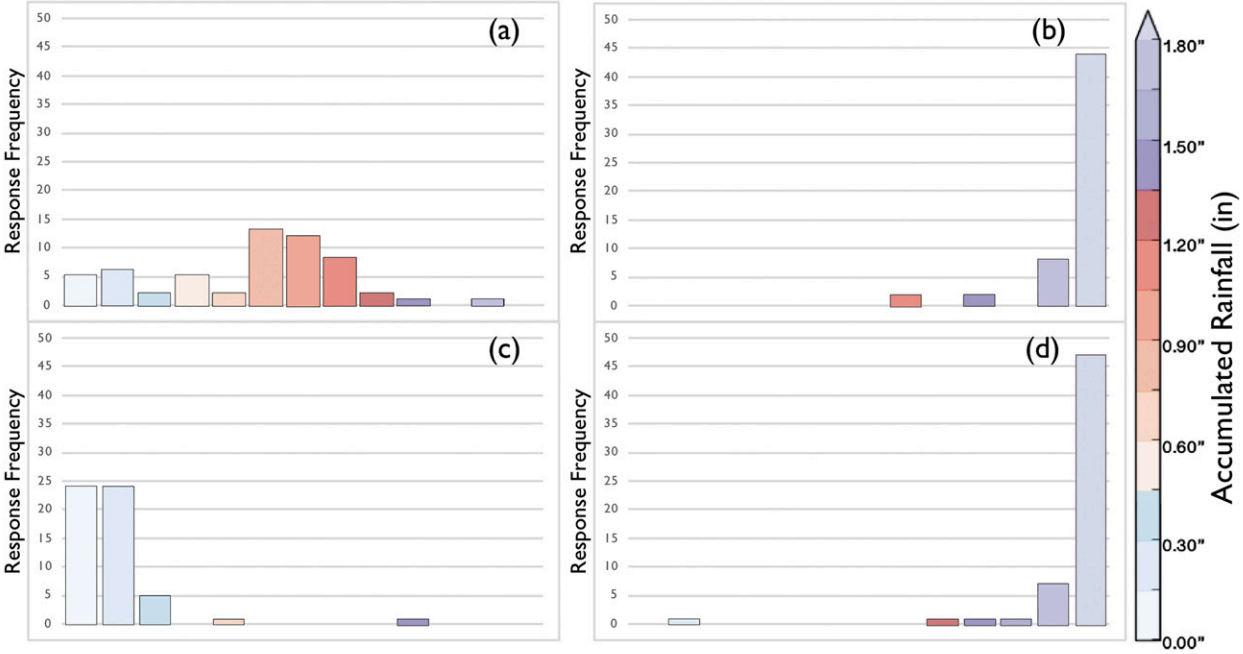

FIG. 9. Responses for the (a) least and (b) greatest amount of expected accumulated rainfall at location A (Q10 and Q11) and the (c) least and (d) greatest amount of expected accumulated rainfall at location B (Q12 and Q13).

and see that this value for box A first appears in the 50th-percentile graphic (Fig. 8). Given that this percentile is the median of the distribution, these participants understood that 50% of the ensemble members predicted accumulated rainfall values greater than 1.35 in. Furthermore, $n = 4$ participants drew information from both the 10th- and 50th-percentile graphics to report values of 10%–50%. Of the remaining participants, a handful reported 90% probability of exceedance ($n = 7$), while a few reported 50%–90% ($n = 3$) and 70% ($n = 3$). Additionally, one or two participants each responded with 10%, 55%, 60%, 100%, 60%–70%, or ''high.''

Next, Q16 was framed as in Q15; however, participants were asked to report probability values for exceeding 1.2 in. of accumulated rainfall in box B. This time, a higher number of participants provided the intended response of 50% ($n = 48$), and a small subset reported 10%–50% ($n = 6$). Of the remaining participants, two answered 10%, and one participant each answered 30%, 90%, 100%, 30%–40%, 50%–90%, and ''less than A.'' Although it is unclear why each of the incorrect responses in Q15 and Q16 were given, one possible reason for some of the incorrect responses is that participants referred to the wrong percentile graphic. This reason could be especially true in Q15, where participants may have mistaken the light blue in the inset box at location A for purple, resulting in an answer of 90%. Additionally, although answers including a range of values could not be ascertained with the limited information provided in the three graphics, participants offering these responses appeared to apply knowledge of the pdf concept in both Q15 and Q16 and therefore demonstrated an awareness of the potential plausible values that could hold true.

Question 17 presented a graphic that showed both percentile and probability representations of 2–5-km updraft helicity (Fig. 10). Participants were asked to assess this information and indicate whether the graphics depicted a high- or low-probability event with a high or low consequence. The majority of participants ($n = 53$) selected the intended answer that the graphics showed a *low-probability high-consequence* event. Two of these participants also selected an alternative correct answer (*high-probability low-consequence* event). Additionally, a subset of participants ($n = 9$) only selected one of these incorrect answers. Showing similar products in Q18, participants were asked to combine the information and select one of four answers that provided the best comparison between the two storms (Fig. 11). One of two intended answers was that *Storm A could become more intense than storm B, but both storms are about as likely to reach UH (updraft helicity) values of 120* $m^2 s^{-2}$, which most participants chose ($n = 47$). Fewer participants ($n = 12$) selected the other plausible but perhaps less obvious answer, which was that *Storms A and B could become equally intense, but A is more likely to reach that peak intensity than B.*

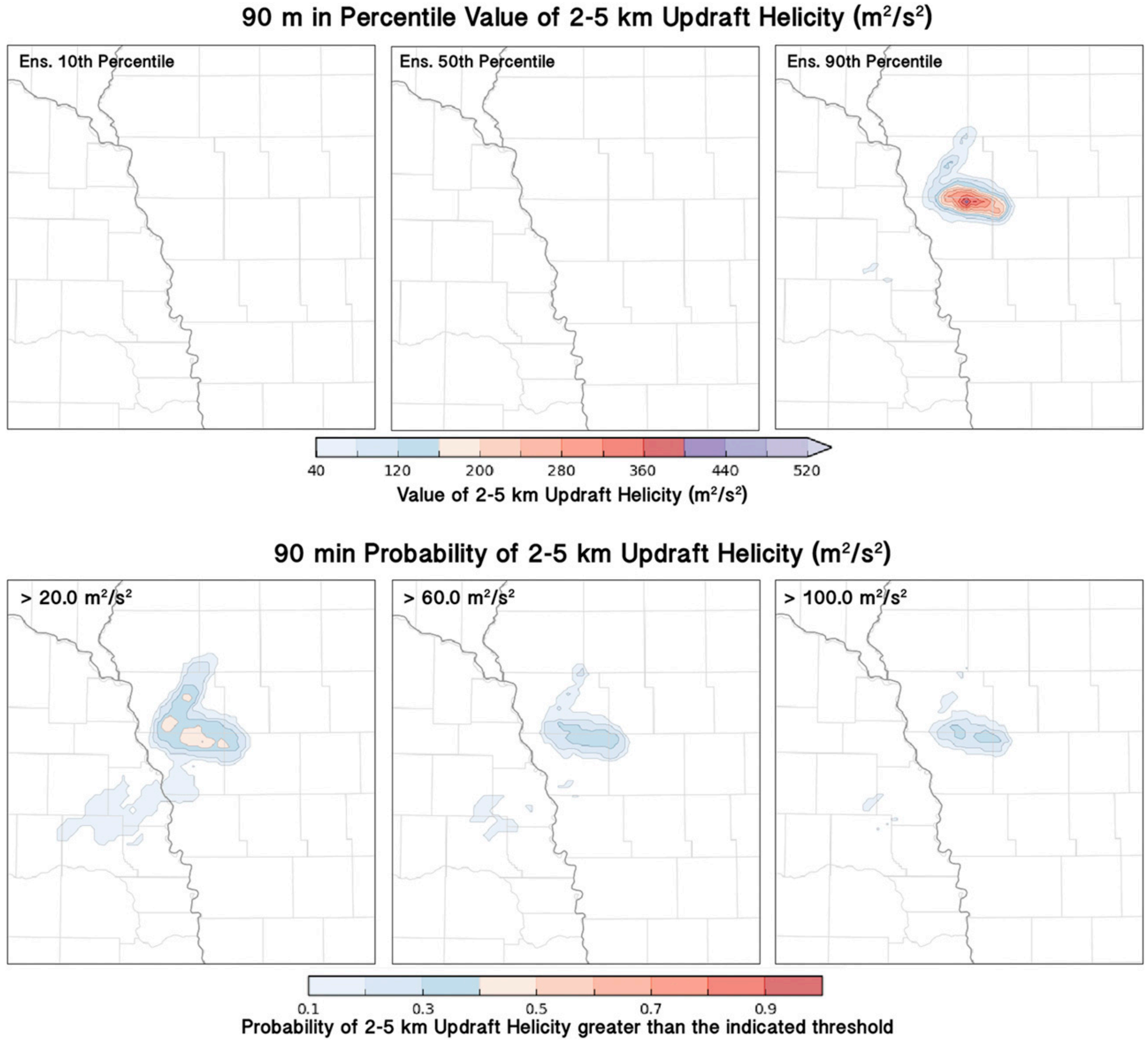
90 m in Percentile Value of 2-5 km Updraft Helicity (m²/s²)
Ens. 10th Percentile
Ens. 50th Percentile
Ens. 90th Percentile
40
120
200
280
360
440
520
Value of 2-5 km Updraft Helicity (m²/s²)
90 min Probability of 2-5 km Updraft Helicity (m²/s²)
> 20.0 m²/s²
> 60.0 m²/s²
> 100.0 m²/s²
0.1
0.3
0.5
0.7
0.9
Probability of 2-5 km Updraft Helicity greater than the indicated threshold

FIG. 10. (top) Percentile and (bottom) probability swaths of 2–5 km UH ($m^2 s^{-2}$) for storms A and B (referred to in Q17).

The remaining $n = 3$ participants incorrectly chose that *Storms A and B could become equally intense, but B is more likely to reach that peak intensity than A*. Finally, no participants chose the remaining incorrect answer that *Storm B could become more intense than storm A, but both storms are about as likely to reach UH values of 120 $m^2 s^{-2}$*.

The final two questions of the survey gave participants a side-by-side view of a probabilistic representation of composite reflectivity and a percentile representation of 2–5-km updraft helicity (Fig. 12). Used together, these two forecast products provide measures of both mesocyclone likelihood and severity, which can be used to rank the storms in terms of potential impacts. To compare participants' perceptions of storm severity using these two sources of information, they were asked to list which of the six storms would be of primary concern (Q19) and least concern (Q20). In response to Q19, almost all participants ($n = 60$) selected storm E as being of greatest concern, and half also selected storm B ($n = 28$) and/or storm D ($n = 31$). In Q20, participants were in strong consensus that storm A ($n = 45$) and storm C ($n = 50$) were of least concern (with two different anomalous participants selecting storm A and selecting storm C as a primary concern in Q19). Generally, participants looking to encompass most, if not all, of the storms in their answers also selected storm F. These participants were fairly split between whether this storm was

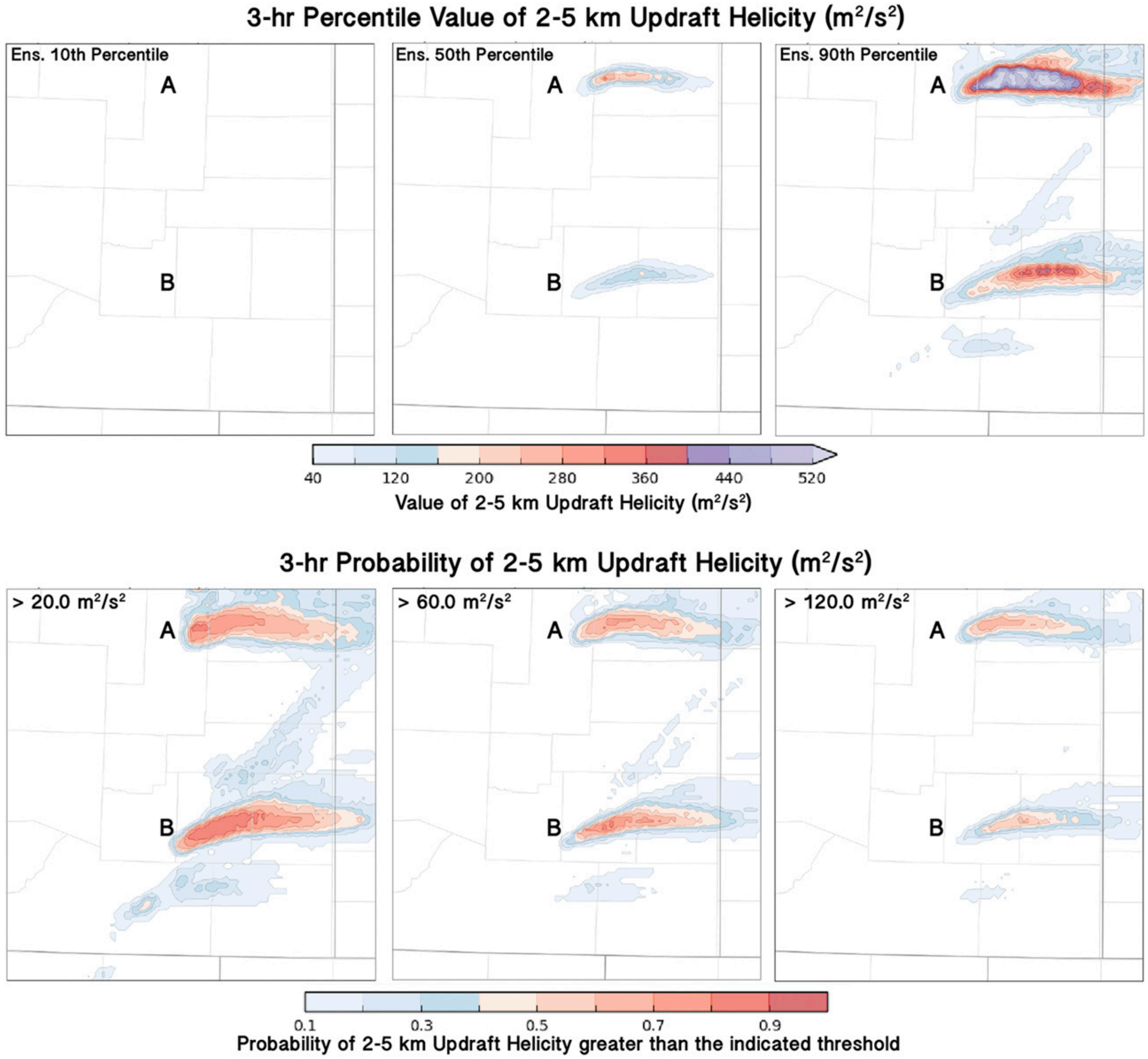
3-hr Percentile Value of 2-5 km Updraft Helicity (m^2/s^2)
Ens. 10th Percentile
Ens. 50th Percentile
Ens. 90th Percentile
A
B
40
120
200
280
360
440
520
Value of 2-5 km Updraft Helicity (m^2/s^2)
3-hr Probability of 2-5 km Updraft Helicity (m^2/s^2)
> 20.0 m^2/s^2
> 60.0 m^2/s^2
> 120.0 m^2/s^2
0.1
0.3
0.5
0.7
0.9
Probability of 2-5 km Updraft Helicity greater than the indicated threshold

FIG. 11. As in Fig. 10, but for storms referred to in Q18.

of primary concern ($n = 8$) or least concern ($n = 6$), likely due to its more marginal representation of severity relative to the other storms. While approximately half of the participants preferred to focus on a single storm for each answer, the other half included most if not all of the storms in their two responses. Overall, participants' responses to this question show that their perceptions of storm severity corresponded well when using both probabilistic and percentile forecast guidance.

4. Discussion

The survey presented in this paper provides insight into meteorologists' understanding and interpretation of probabilistic and percentile ensemble forecast products during the 2017 SFE. Traditionally, the majority of weather information that meteorologists use and communicate is deterministic, meaning that event occurrence is often treated dichotomously (i.e., it either is or is not going to occur). In some survey questions, we found substantial variability in the extent to which participants could think beyond this deterministic mindset and successfully synthesize, extract, and apply uncertainty guidance information. This varying ability was evident in both qualitative and numerical responses. For example, although most participants recognized that only limited conclusions could be drawn for the maximum amount of accumulated rainfall from the information given in both Q2

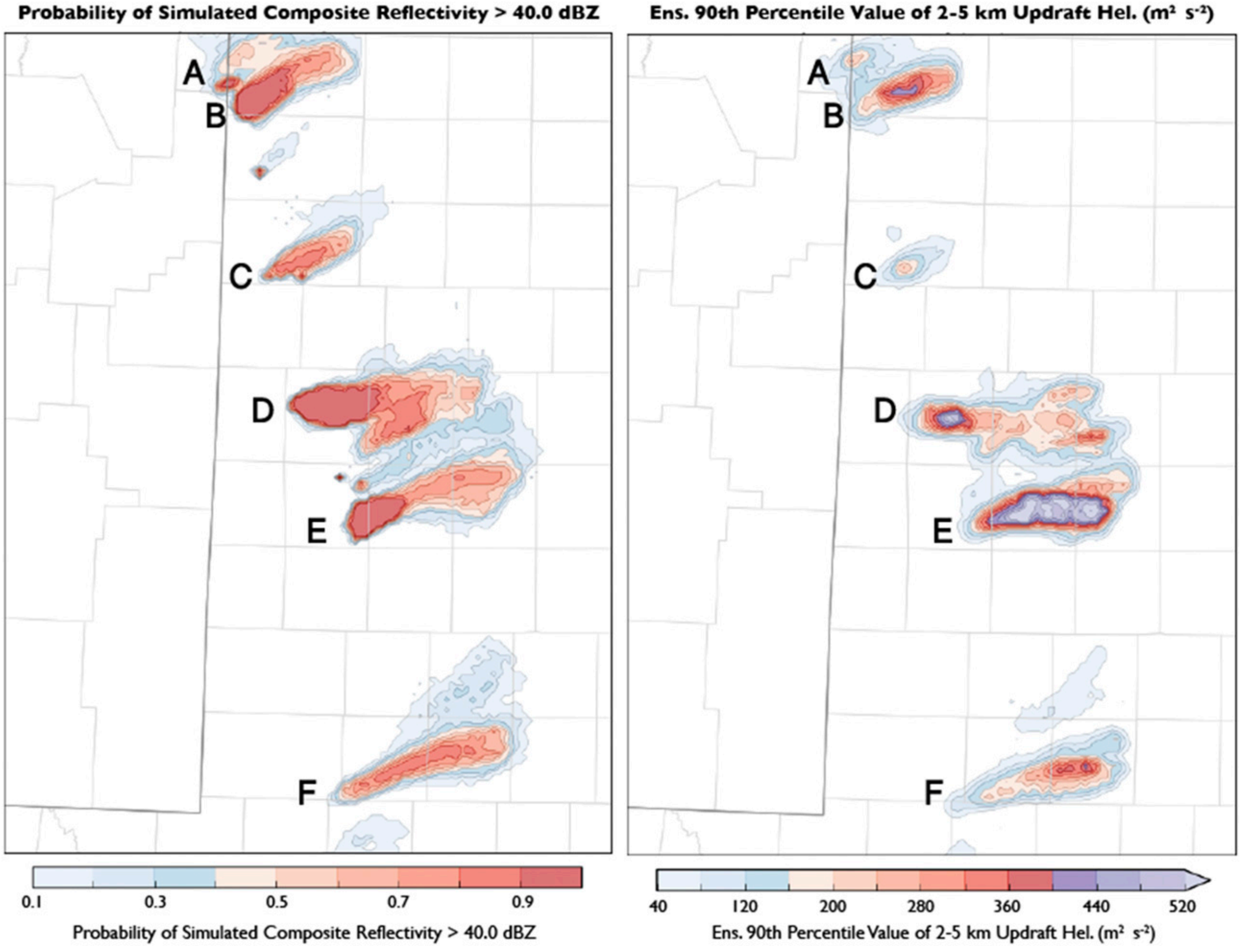
Probability of Simulated Composite Reflectivity > 40.0 dBZ
Ens. 90th Percentile Value of 2-5 km Updraft Hel. (m² s⁻²)
A
B
C
D
E
F
A
B
C
D
E
F
0.1
0.3
0.5
0.7
0.9
Probability of Simulated Composite Reflectivity > 40.0 dBZ
40
120
200
280
360
440
520
Ens. 90th Percentile Value of 2-5 km Updraft Hel. (m² s⁻²)

FIG. 12. Accumulated maximum probabilities of (left) simulated composite reflectivity (dBZ) and (right) the ensemble 90th-percentile values of 0–2-km UH ($m^2 s^{-2}$; referred to in Q19 and Q20).

and Q3 (i.e., using exceedance thresholds and probability values), a selection of participants disregarded the uncertainty depicted in the graphic and instead provided specific (deterministic) values for accumulated rainfall. Furthermore, in qualitative responses to Q8, just under half of the participants demonstrated misunderstanding or ambiguity in their explanations of the percentile concept. Oftentimes, these inaccurate explanations were deterministic in nature, such that participants described the 70th percentile in an aggregated sense rather than as a rank within a distribution of values. Most commonly, this type of aggregated explanation reflected a belief that the 70th percentile in an ensemble prediction system meant that 70% of the members agreed on the accumulated rainfall value.

In other questions, participants providing correct responses did not always interpret the graphical information equally; while some participants accepted the given information at face value and provided correct, single-value answers, others attempted to recognize the inherent uncertainty and interpreted the graphics a step further by providing all possible answers. For example, in responses to Q4 and Q5, while most participants looked at the discrete steps of the color bar and provided a single value answer (e.g., ‘‘2 ensemble members’’ or ‘‘10% probability’’), a smaller portion of participants interpolated between the color bar steps and provided a range of values (e.g., ‘‘2–4 members’’ or ‘‘10–20%’’). Similarly, in Q15 and Q16, while the majority of participants provided single probability value answers, a smaller portion of participants instead opted to interpolate between the discrete color bar steps to give a range of probability values that, although plausible, were not fully supported with the limited information provided. One possible explanation for these observed differences in participants' graphical interpretations is the extent to which they were willing or able to move beyond a deterministic mindset and instead consider the problem within probabilistic space. The reasons behind meteorologists' different applications of deterministic and probabilistic mindsets needs to be explored further to better understand why some meteorologists are able to transition between these two mindsets more easily than others.

Participants' understanding of probabilistic representations of uncertainty information was also found to be a function of the variable that the graphic depicted. When provided with swaths of probability values representing accumulated rainfall and updraft helicity (Q6 and Q7), participants were asked to identify the most intuitive reason(s) for why the probability values decline toward the end of the swath. In both examples, the primary underlying reason was attributable to increased trajectory uncertainty toward the end of the swath. However, more than twice as many participants ($n = 23$) did not select this answer for the updraft helicity example compared to the accumulated rainfall example ($n = 10$). This result suggests that a large proportion of the participating meteorologists did not apply their understanding of probabilistic concepts uniformly to the accumulated rainfall example and the updraft helicity example. Instead, participants shifted how they interpret uncertainty guidance depending on the variable they were examining. This result may have been influenced by differences in participants' overall familiarity with these two variables (with greater exposure expected to the accumulated rainfall variable), their level of understanding for the atmospheric science principals that explain these two variables, and the extent to which they have encountered probabilistic representations of these variables previously. This inconsistent application of knowledge supports the need for educating meteorologists on how various meteorological variables are treated in CAMs and how graphical representations of uncertainty information are subsequently created.

In addition to the meteorological variable influencing participants' understanding of probabilistic concepts, strong tendencies to focus on the worst-case scenario also influenced answer selections in Q17. When given percentile and probabilistic representations of 2–5-km updraft helicity, $n = 53$ participants reported that the graphics depicted a low-probability high-consequence event, of whom only $n = 2$ also gave the second correct answer of a high-probability low-consequence event. Therefore, most participants defaulted their attention to the scenario that had greatest potential impact without recognizing the range of potential outcomes. A similar outcome was observed in Q1 when forecasters inferred a worst-case-scenario interpretation into the probability of exceeding a measurable amount of rainfall. However, when encouraged to consider all scenarios in Q19 and Q20 to identify which storms were of greatest and least concern when using a combination of percentile and probabilistic forecast guidance, responses showed that participants' perceptions of storm severity were well aligned. This result suggests that probability and percentile information, when presented together, may improve overall understanding.

5. Conclusions

This survey was designed to explore professional meteorologists' current knowledge, understanding, and application of probabilistic guidance during the 2017 SFE. Overall, the results are encouraging, with 60%–96% of participants providing correct answers for each of the questions addressing probability concepts, percentile concepts, and a combination of probability and percentile concepts (Fig. 3). Participants providing correct responses varied in terms of their depth of understanding, levels of interpretation, and abilities to think beyond a deterministic mindset. Although many participants demonstrated a strong understanding of probabilistic and percentile concepts, those providing incorrect responses demonstrated that knowledge gaps relating to the use and interpretation of uncertainty information currently exist even for those actively working in the field of meteorology. Open-ended questions requiring participants to depend on their own knowledge to provide qualitative explanations of either a graphic or percentile concepts proved to be most challenging (e.g., Q1, Q8, and Q9). Additionally, in some instances, participants also struggled to apply a correct understanding of concepts consistently across questions that were similar in nature. That is, although the same knowledge and skills were required to answer multiple questions that were of the same style, inquiring about a different aspect of the same graphic or a different variable impacted participants' abilities to provide the intended response (e.g., Q2 and Q3; Q6 and Q7).

The findings from this survey therefore support the need for improved basic and applied training for the development, interpretation, and use of probabilistic ensemble forecast guidance as the meteorological community moves toward increased generation and communication of uncertainty information. This training need has been recognized in past comprehensive reports, with recommendations for revision of undergraduate and graduate education to include uncertainty training (Hirschberg et al. 2011), for improvements in training for operational use and applications of uncertainty information (NRC 2006; Novak et al. 2008; Hirschberg et al. 2011), and for educating the public on the meaning of products that communicate uncertainty and risk (NRC 2006). In addition to these recommendations, our survey findings suggest a need for training materials that are available to meteorologists already working in various sectors of the weather enterprise. This training material should focus on reviewing basic concepts of storm-scale probabilistic forecast guidance, educating meteorologists on the interpretation and application of this information for various scenarios, and ensuring that meteorologists understand how different probabilistic products are generated and how they can be used in a complementary manner to most effectively evaluate the weather scenario. Training should build upon the educational resources already available, such as the NOAA NWS/UCAR COMET program's distance learning lessons on forecast uncertainty and ensemble prediction systems, and be developed in coordination with other official training agencies, such as the NOAA NWS Warning Decision Training Division. Furthermore, rather than being associated with misinformation or a lack of knowledge, on occasion participants' incorrect survey responses appeared to relate to possible difficulties interpreting color schemes or their inattention to detail in the graphics. Further research exploring the visualization of weather information is therefore needed to maximize users' comprehension and use of probabilistic forecast guidance, such that the overall graphic design supports forecasters' intuitive understanding and approaches to assessing such information (e.g., Hegarty et al. 2010; Hogan Carr et al. 2016; Quinan and Meyer 2016).

Given the small sample size and relatively uniform demographics of the participants that completed the survey during the 2017 SFE, recommending specific training needs of meteorologists working in different professions (e.g., in research or operations) is imprudent. A major limitation of this study is that the majority of respondents identified themselves as working in the research community, leading to the operational and student/postdoc participants being underrepresented in the results. However, while not a focus of this survey, our results did indicate that inaccuracies in participants' understanding of probability and percentile concepts spanned each of these professional categories. This finding motivates the need for future research that will expand on these survey efforts by increasing the sample size of responses for each profession. This research would build on the findings presented herein and develop a more precise understanding of what knowledge gaps exist among meteorologists serving in different professions and how education and training could be tailored to meet the needs of specific user groups.

Acknowledgments. We are thankful to the 62 participants who graciously completed this survey and to the facilitators of the Spring Forecasting Experiment for their effort in organizing and leading the 2017 activities. We also appreciate manuscript reviews provided by Kodi Berry, Burkely Gallo, and three anonymous reviewers. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. The contents of this paper do not necessarily reflect the views or official position of any organization of the U.S. government.

APPENDIX

Survey Questions

Probability Concept

1) In 1–3 sentences, please describe what kind of event is depicted by this graphic.
2) According to the information provided, what is the maximum amount of rainfall possible within box A?
3) According to the information provided, what is the maximum amount of rainfall possible within box B?
4) Out of 20 ensemble members, how many predict at least 0.01 in. of rain will fall within box B?
5) Given the information presented, what is the probability of exceeding 0.5 in. of rainfall within box A?
6) Within box A, toward the end of the swath, the probability values (at the given level of intensity) decline. In your opinion, which of the following factor(s) most intuitively explain(s) what is happening? (Choose all that apply: the storm is expected to decrease in intensity/strength; there is less time for the rainfall to accumulate at that point during the forecast period; trajectory uncertainty is greater toward the end of the swath; the storm is growing in size and rainfall rate is forecasted to decrease.)
7) Within box A, toward the end of the swath, the probability values (at the given level of intensity) decline. In your opinion, which of the following factor(s) most intuitively explain(s) what is happening? (Choose all that apply: the storm is expected to decrease in intensity/strength; trajectory uncertainty is greater toward the end of the swath; the storm is growing in size and rotation is broadening; there is less time to experience maximum updraft helicity during the forecast period.)

Percentile Concept

8) In an ensemble-based probabilistic forecast, what do you think the 70th percentile value of accumulated rainfall represents?

9) How does this representation of forecast uncertainty compare to the PoP used today?
10) Based on the information shown, within box A, what is the least amount of rainfall expected? Please refer to the insets on the left for a closer look at the areas of interest.
11) Based on the information shown, within box A, what is the greatest amount of rainfall expected?
12) Based on the information shown, within box B, what is the least amount of rainfall expected?
13) Based on the information shown, within box B, what is the greatest amount of rainfall expected?
14) Based on the information provided, at which location, A or B, was the range of potential rainfall greatest?

Probability and Percentile Concept Questions

15) Based on the information shown, what is the probability of exceeding 1.35 in. of rainfall within box A?
16) Based on the information shown, what is the probability of exceeding 1.2 in. of rainfall within box B?
17) Choose all responses that apply. Considered together, these graphics show a: high probability, low consequence event; high-probability, high consequence event; low-probability, low consequence event; low-probability, high-consequence event.
18) Considered together, these graphics show: storms A and B could become equally intense, but A is more likely to reach that peak intensity than B; storm A could become more intense than storm B, but both storms are about as likely to reach UH values of $120\,\mathrm{m}^2\,\mathrm{s}^{-2}$; storm A and B could become equally intense, but B is more likely to reach that peak intensity at A; storm B could become more intense than storm A, but both storms are about as likely to reach UH values of $120\,\mathrm{m}^2\,\mathrm{s}^{-2}$.
19) Based on the information given, which storm(s) would be your primary concern?
20) Based on the information given, which storm(s) would you be least concerned with?

REFERENCES

AMS, 2008: Enhancing weather information with probability forecasts. Amer. Meteor Soc., https://www.ametsoc.org/index.cfm/ams/about-ams/ams-statements/statements-of-the-ams-in-force/enhancing-weather-information-with-probability-forecasts/.

Ash, K. D., R. L. Schumann, and G. C. Bowser, 2014: Tornado warning trade-offs: Evaluating choices for visually communicating risk. *Wea. Climate Soc.*, **6**, 104–118, https://doi.org/10.1175/WCAS-D-13-00021.1.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Clark, A. J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, https://doi.org/10.1175/BAMS-D-11-00040.1.

——, and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448, https://doi.org/10.1175/BAMS-D-16-0309.1.

Cohen, J., 1960: A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, **20**, 37–46, https://doi.org/10.1177/001316446002000104.

de Elía, R., and R. Laprise, 2005: Diversity in interpretations of probability: Implications for weather forecasting. *Mon. Wea. Rev.*, **133**, 1129–1143, https://doi.org/10.1175/MWR2913.1.

Demuth, J., J. J. Lazo, and B. H. Morrow, 2009: Weather forecast uncertainty information: An exploratory study with broadcast meteorologists. *Bull. Amer. Meteor. Soc.*, **90**, 1614–1618, https://doi.org/10.1175/2009BAMS2787.1.

Done, J., C. A. David, and M. L. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) Model. *Atmos. Sci. Lett.*, **5**, 110–117, https://doi.org/10.1002/asl.72.

Fleiss, J. L., 1981: *Statistical Methods for Rates and Proportions*. 2nd ed. Wiley, 800 pp.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

Gigerenzer, G., R. Hertwig, E. van den Broek, F. Fasolo, and K. V. Katsikopoulos, 2005: "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Anal.*, **25**, 623–629, https://doi.org/10.1111/j.1539-6924.2005.00608.x.

Hegarty, M., M. S. Canham, and S. I. Fabrikant, 2010: Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *J. Exp. Psychol.*, **36**, 37–53.

Hirschberg, P. A., and Coauthors, 2011: A weather and climate enterprise strategic implementation plan for generating and communicating forecast uncertainty information. *Bull. Amer. Meteor. Soc.*, **92**, 1651–1666, https://doi.org/10.1175/BAMS-D-11-00073.1.

Hogan Carr, R., B. Montz, K. Maxfield, S. Hoekstra, K. Semmens, and E. Goldman, 2016: Effectively communicating risk and uncertainty to the public: Assessing the National Weather Service's flood forecast and warning tools. *Bull. Amer. Meteor. Soc.*, **97**, 1649–1665, https://doi.org/10.1175/BAMS-D-14-00248.1.

Jirak, I. L., A. J. Clark, B. Roberts, B. Gallo, and S. J. Weiss, 2018: Exploring the optimal configuration of the High Resolution Ensemble Forecast System. *29th Conf. on Weather Analysis and Forecasting/25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 14B.6, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html.

Jones, T. A., K. Knopfmeier, D. Wheatley, G. Creager, P. Minnis, and R. Palikonda, 2016: Storm-scale data assimilation and ensemble forecasting with the NSSL Experimental Warn-on-Forecast System. Part II: Combined radar and satellite data experiments. *Wea. Forecasting*, **31**, 297–327, https://doi.org/10.1175/WAF-D-15-0107.1.

Joslyn, S., K. Pak, D. Jones, J. Pyles, and E. Hunt, 2007: The effect of probabilistic information on threshold forecasts.

Wea. Forecasting, **22**, 804–812, https://doi.org/10.1175/WAF1020.1.

——, L. Nadav-Greenberg, and R. M. Nichols, 2009: Probability of precipitation: Assessment and enhancement of end-user understanding. *Bull. Amer. Meteor. Soc.*, **90**, 185–194, https://doi.org/10.1175/2008BAMS2509.1.

Kain, J. S., M. E. Baldwin, P. R. Janish, S. J. Weiss, M. P. Kay, and G. W. Carbin, 2003:Subjective verification of numerical models as a component of a broader interaction between research and operations. *Wea. Forecasting*, **18**, 847–860, https://doi.org/10.1175/1520-0434(2003)018<0847:SVONMA>2.0.CO;2.

——, S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF Model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, https://doi.org/10.1175/WAF906.1.

——, and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

Karstens, C., and Coauthors, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting*, **30**, 1551–1570, https://doi.org/10.1175/WAF-D-14-00163.1.

Landis, J. R., and G. G. Koch, 1977: The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174, https://doi.org/10.2307/2529310.

Magnusson, L., and E. Källén, 2013: Factors influencing skill improvements in the ECMWF forecasting system. *Mon. Wea. Rev.*, **141**, 3142–3153, https://doi.org/10.1175/MWR-D-12-00318.1.

McHugh, M. L., 2012: Interrater reliability: The kappa statistic. *Biochem. Med.*, **22**, 276–282, https://doi.org/10.11613/BM.2012.031.

Morss, R. E., J. L. Demuth, and J. K. Lazo, 2008: Communicating uncertainty in weather forecasts: A survey of the U.S. public. *Wea. Forecasting*, **23**, 974–991, https://doi.org/10.1175/2008WAF2007088.1.

Murphy, A. H., and R. L. Winkler, 1971: Forecasters and probability forecasts: Some current problems. *Bull. Amer. Meteor. Soc.*, **52**, 239–248, https://doi.org/10.1175/1520-0477(1971)052<0239:FAPFSC>2.0.CO;2.

——, and ——, 1974: Probability forecasts: A survey of National Weather Service forecasters. *Bull. Amer. Meteor. Soc.*, **55**, 1449–1453, https://doi.org/10.1175/1520-0477(1974)055<1449:PFASON>2.0.CO;2.

Novak, D. R., D. R. Bright, and M. J. Brennan, 2008: Operational forecaster uncertainty needs and future roles. *Wea. Forecasting*, **23**, 1069–1084, https://doi.org/10.1175/2008WAF2222142.1.

——, K. F. Brill, and W. A. Hogsett, 2014: Using percentiles to communicate snowfall uncertainty. *Wea. Forecasting*, **29**, 1259–1265, https://doi.org/10.1175/WAF-D-14-00019.1.

NRC, 2006: *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. National Academies Press, 124 pp.

Orne, M. T., 1962: On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *Amer. Psychol.*, **17**, 776–783, https://doi.org/10.1037/h0043424.

Quinan, P. S., and M. Meyer, 2016: Visually comparing weather features in forecasts. *IEEE Trans. Vis. Comput. Graph.*, **22**, 389–398, https://doi.org/10.1109/TVCG.2015.2467754.

Rothfusz, L. P., C. Karstens, and D. Hilderband, 2014: Next-generation severe weather forecasting and communication. *Eos, Trans. Amer. Geophys. Union*, **95**, 325–326, https://doi.org/10.1002/2014EO360001.

——, R. Schneider, D. Novak, K. E. Klockow-McClain, A. Gerard, C. Karstens, G. Stumpf, and T. Smith, 2018: FACETs: A proposed next-generation paradigm for high-impact weather forecasting. *Bull. Amer. Meteor. Soc.*, **99**, 2025–2043, https://doi.org/10.1175/BAMS-D-16-0100.1.

Saldaña, J., 2016: *The Coding Manual for Qualitative Researchers*. 3rd ed. SAGE Publications, 368 pp.

Savelli, S., and S. Joslyn, 2013: The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Appl. Cognit. Psychol.*, **27**, 527–541, https://doi.org/10.1002/acp.2932.

Skinner, P. S., L. J. Wicker, D. M. Wheatley, and K. H. Knopfmeier, 2016: Application of two spatial verification methods to ensemble forecasts of low-level rotation. *Wea. Forecasting*, **31**, 713–735, https://doi.org/10.1175/WAF-D-15-0129.1.

——, and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast system. *Wea. Forecasting*, **33**, 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

Sobash, R. A., and J. S. Kain, 2017: Seasonal variations in severe weather forecast skill in an experimental convection-allowing model. *Wea. Forecasting*, **32**, 1885–1902, https://doi.org/10.1175/WAF-D-17-0043.1.

——, ——, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

Stensrud, D., and Coauthors, 2009: Convective-scale Warn-on-Forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

——, and Coauthors, 2013: Progress and challenges with Warn-on-Forecast. *Atmos. Res.*, **123**, 2–16, https://doi.org/10.1016/j.atmosres.2012.04.004.

Stewart, A. E., C. A. Williams, M. D. Phan, A. L. Horst, E. D. Knox, and J. A. Knox, 2016: Through the eyes of the experts: Meteorologists' perceptions of the probability of precipitation. *Wea. Forecasting*, **31**, 5–17, https://doi.org/10.1175/WAF-D-15-0058.1.

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL Experimental Warn-on-Forecast System. Part I: Radar data experiments. *Wea. Forecasting*, **30**, 1795–1817, https://doi.org/10.1175/WAF-D-15-0043.1.

Yussouf, N., J. S. Kain, and A. J. Clark, 2016: Short-term probabilistic forecasts of the 31 May 2013 Oklahoma tornado and flash flood event using a continuous-update-cycle storm-scale ensemble system. *Wea. Forecasting*, **31**, 957–983, https://doi.org/10.1175/WAF-D-15-0160.1.