Exploring NWS Forecasters' Assessment of AI Guidance Trustworthiness

MARIANA G. CAINS,[a,b] CHRISTOPHER D. WIRZ,[a,b] JULIE L. DEMUTH,[a,b] ANN BOSTROM,[b,c] DAVID JOHN GAGNE II,[a,b] AMY MCGOVERN,[b,d] RYAN A. SOBASH,[a] AND DEIANNA MADLAMBAYAN[b,c]

[a] *National Center for Atmospheric Research, Boulder, Colorado*
[b] *NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography, Norman, Oklahoma*
[c] *University of Washington, Seattle, Washington*
[d] *University of Oklahoma, Norman, Oklahoma*

(Manuscript received 1 November 2023, in final form 7 May 2024, accepted 7 June 2024)

ABSTRACT: As artificial intelligence (AI) methods are increasingly used to develop new guidance intended for operational use by forecasters, it is critical to evaluate whether forecasters deem the guidance trustworthy. Past trust-related AI research suggests that certain attributes (e.g., understanding how the AI was trained, interactivity, and performance) contribute to users perceiving the AI as trustworthy. However, little research has been done to examine the role of these and other attributes for weather forecasters. In this study, we conducted 16 online interviews with National Weather Service (NWS) forecasters to examine (i) how they make guidance use decisions and (ii) how the AI model technique used, training, input variables, performance, and developers as well as interacting with the model output influenced their assessments of trustworthiness of new guidance. The interviews pertained to either a random forest model predicting the probability of severe hail or a 2D convolutional neural network model predicting the probability of storm mode. When taken as a whole, our findings illustrate how forecasters' assessment of AI guidance trustworthiness is a process that occurs over time rather than automatically or at first introduction. We recommend developers center end users when creating new AI guidance tools, making end users integral to their thinking and efforts. This approach is essential for the development of useful and *used* tools. The details of these findings can help AI developers understand how forecasters perceive AI guidance and inform AI development and refinement efforts.

SIGNIFICANCE STATEMENT: We used a mixed-methods quantitative and qualitative approach to understand how National Weather Service (NWS) forecasters 1) make guidance use decisions within their operational forecasting process and 2) assess the trustworthiness of prototype guidance developed using artificial intelligence (AI). When taken as a whole, our findings illustrate that forecasters' assessment of AI guidance trustworthiness is a process that occurs over time rather than automatically and suggest that developers must center the end user when creating new AI guidance tools to ensure that the developed tools are useful and *used*.

KEYWORDS: Social Science; Forecasting; Model evaluation/performance; Decision-making; Artificial intelligence; Machine learning

1. Introduction

National Weather Service (NWS) forecasters engage with a diverse set of observations and guidance[1] when making critical decisions that can affect the well-being of many people. Forecasters access and synthesize these myriad sources of information when forecasting for high-impact, severe weather events (Daipha 2015; Hoffman et al. 2006; Henderson et al. 2023). In recent years, artificial intelligence (AI) techniques have increasingly been used to produce new guidance tools with the goal of aiding weather forecasting (Chase et al. 2022; Roebber 2022), including for severe weather (e.g., Gagne et al. 2017; Burke et al. 2020; Lagerquist et al. 2020; Flora et al. 2021; Hill et al. 2023; Sobash et al. 2023). The growth of AI to produce experimental and operational guidance has driven increased emphasis on developing AI that is *trustworthy*. This emphasis on trustworthy AI exists across weather research and operational forecasting (McGovern et al. 2022; Roebber and Smith 2023), and it also has much broader national and international resonance. For instance, trustworthy AI is a focus of the European Commission High-Level Expert Group on Artificial Intelligence and the U.S. National Science and Technology Council's National Artificial Intelligence Research and Development Strategic Plan (European Commission 2019; National Science and Technology Council 2023).

[1] Consistent with Novak et al. (2008), we use the term "guidance" to broadly refer to any kind of predictive model—including dynamical, statistical, or AI models—or other tool that a forecaster might use to assess the meteorological situation and make a determination about the forecast they issue.

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-23-0180.s1.

Corresponding author: Mariana G. Cains, mgcains@ucar.edu

DOI: 10.1175/WAF-D-23-0180.1

Past research pertaining to AI trustworthiness and trust is vast, covering numerous concepts and empirical studies from across a wide range of fields and domains (Saßmannshausen et al. 2021; National Academies of Sciences, Engineering, and Medicine 2022; Bostrom et al. 2024; C. D. Wirz et al. 2024, unpublished manuscript). Within the literature are many efforts to define and measure trustworthy AI and end-user trust in AI (e.g., Hoffman et al. 2018b; Jacovi et al. 2020; Varshney 2021; Stanton and Jensen 2021; Bostrom et al. 2024). A few aspects are particularly relevant to the research presented here. The literature suggests that attributes such as being able to interact with the AI model inputs and outputs, strong performance of the AI model, and improved human–machine performance all contribute to users perceiving it as trustworthy (Hoffman et al. 2018a; Mueller et al. 2019; Kaplan et al. 2021). Researchers in the fields of cognitive engineering and decision-making have studied how the type and extent of explanations about the function, output, and performance of an AI model [i.e., explainable AI (XAI)] can influence users' trust in and reliance on it (Hoffman et al. 2018a,c). These and other aspects represent concepts that are central to AI, including those of AI model explainability, interpretability, and transparency. *Explainability* is the degree to which AI functionality can be understood with post hoc methods (AI2ES 2022a). *Interpretability* is the extent to which a person can understand the AI model functionality without supplementary techniques (AI2ES 2022b). *Transparency*[2] is providing the user with relevant details about the data, processing, and algorithms used within an AI system, so that the user can evaluate the strengths and weaknesses of the AI system for their use case. We draw upon these ideas in our study of AI trustworthiness in the weather forecasting domain.

Despite the recent increase in the provision of AI guidance in meteorology alongside efforts to develop trustworthy AI, there has been little research to date that examines how forecasters evaluate such guidance. This lack of research includes whether the AI guidance is trustworthy, what attributes might influence this, and how forecasters might use said guidance. This knowledge gap represents research located at the intersection of trust and trustworthiness of new technologies, domain expert decision-making and sense-making, and the development of new guidance for weather prediction. Working at this complex intersection requires the type of intentional, deep collaboration among social, atmospheric, and AI scientists that was designed as part of the National Science Foundation AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (e.g., AI2ES; McGovern et al. 2022).

Here, we present research that begins to fill this knowledge gap with a focus on the context of severe convective weather forecasting. Our focus is severe weather, one of several hazards our research team is researching as part of AI2ES (McGovern et al. 2022). Our first research question aims to understand how forecasters decide what model guidance and other tools to use in their operational forecasting process (RQ1). This knowledge provides a foundation for contextualizing how they evaluate AI guidance and for informing how AI guidance can be developed and provided in a way that better integrates into their forecasting process.

Our second research question aims to understand how different descriptive and performance attributes of prototype model guidance developed using AI influence forecasters' assessment of the guidance trustworthiness (RQ2). More specifically, we explore how initial background information about AI guidance affects forecasters' assessment of its trustworthiness (RQ2A). We then examine how the AI technique used, training of the AI model, AI model input variables, performance of the AI model, developer of the AI model, and interactivity with the AI model output each influence forecasters' assessments of trustworthiness (RQ2B). These guidance attributes were selected based on their relevance to forecasters' informational needs (Novak et al. 2008; Demuth et al. 2020) and on meta-analyses and systematic reviews of empirical research about how human, technological, and contextual factors influence trust in automation (Hoff and Bashir 2015; Schaefer et al. 2016; Glikson and Woolley 2020) from across multiple domains (e.g., social media, customer service, transportation, health care, military operations). Last, we elicit from forecasters what additional AI guidance attributes, if any, they want to further assess the trustworthiness of the AI guidance (RQ2C).

To address our research questions, we conducted preinterview surveys and in-depth, structured interviews with NWS forecasters focused on severe weather. Our research approach was informed by and contributes to decades of research focused on understanding the diversity of decisions made by and informational needs of operational forecasters and their core partners (e.g., Stewart et al. 1997; Doswell 2004; Stuart et al. 2006; Daipha 2015; Morss et al. 2015; Hoffman et al. 2017; Demuth et al. 2020). Here, we extend such research by valuing and eliciting forecasters' expertise, perceptions, and needs, with the ultimate goal of informing efforts to develop and provide AI guidance that is more trustworthy, trusted, and useful to them.

Below, we describe our methods, the rich set of results our research yielded, and a summary and discussion that highlights cross-cutting ideas and offers motivation for continued user-centered research that is inclusive of forecasters as domain expert collaborators.

2. Methods

a. *Research design, preinterview survey, and interview protocol*

We employed a mixed-methods research approach (Fig. 1). We recruited forecasters for the primary purpose of being interviewed, and we asked them to complete a short, web-based survey prior to the interview, both described below. As part of our analysis, we wanted to explore whether forecasters' assessments of the guidance we presented to them (RQ2) differed if we explicitly did or did not label that new guidance as

[2] Authors adapted this definition from algorithmic transparency defined by Molnar (2023).

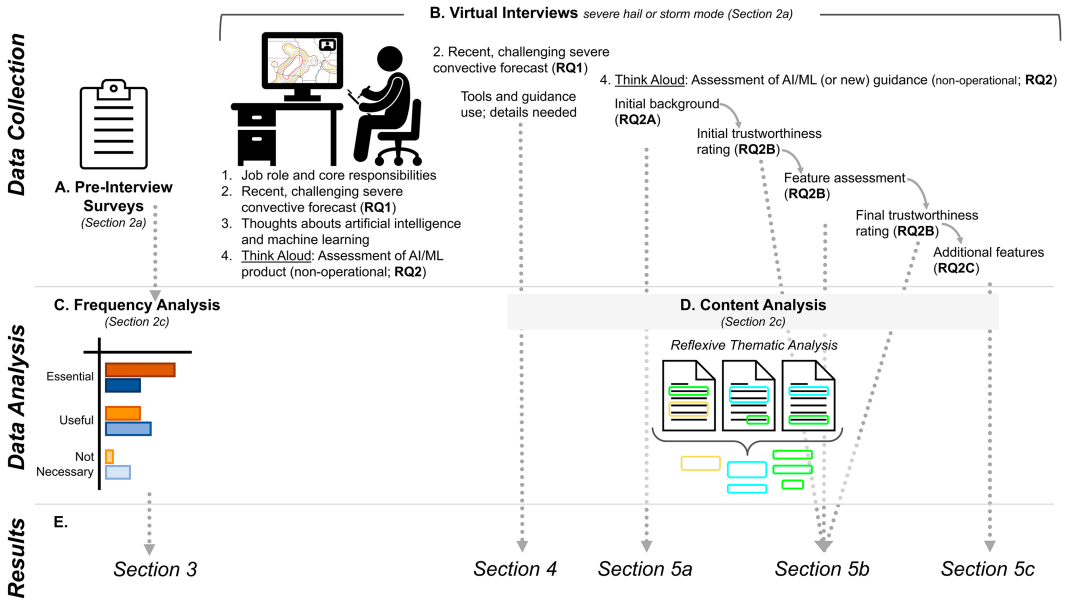

FIG. 1. Overview of the research process, data, and analysis approaches used to address research questions, along with the section of this paper where the results are located. (a),(b) Mixed-methods data collection, (c),(d) data analysis, and (e) results to understand whether, how, and why forecasters trust and use AI (or new) guidance and what attributes could increase or decrease guidance trustworthiness.

AI; below, we note how we incorporated the AI- versus no-AI-labeled versions into the survey and interview design.

1) PREINTERVIEW SURVEY

The preinterview survey had two purposes (Fig. 1a; Demuth et al. 2024). First, we asked the forecaster to think about a recent severe convective event that was challenging to forecast. This prompted them to recall a concrete forecasting situation to help them draw upon specific aspects of that experience when they responded to the survey questions. We also informed them that we would ask about this same challenging forecast event during the interview. Asking the forecaster to think about an actual event during the preinterview survey and then again during the interview was designed to trigger an episodic memory that more concretely anchored the survey and interview in specific professional practice experiences. This approach takes inspiration from critical incident technique, which results in more detailed and contextually relevant answers (Shattuck and Woods 1994).

Second, we asked the forecaster to rate a set of 20 forecast guidance attributes (Table 1). The 20 items were inspired by prior research about AI explainability, interpretability, transparency, and performance (Hoffman et al. 2018c,b; Heinrichs and Eickhoff 2020), which we tailored and expanded for our focus on severe weather forecasting and based on our team's prior related research experience (Bostrom et al. 2016; Gagne et al. 2019; McGovern et al. 2019; Demuth et al. 2020). Most of the survey items pertained to forecast guidance generally, whether derived from numerical weather prediction (NWP) models or AI (or other statistical) techniques (Table 1, items 1–12). Some items were designed to be specific to AI; for these, we explicitly mentioned "artificial intelligence/machine learning" or "AI/ML" in the item for the AI version of the survey and more generally mentioned "new guidance" in the item for the no-AI version of the survey[3] (Table 1, items 13–20). We asked the forecaster to rate each attribute as "essential," "useful, but not essential," or "not necessary" (Hoffman et al. 2018b) while keeping their challenging severe convective forecast experience in mind. We asked them to rate each attribute twice, once in the context of getting familiar with or training on new guidance and once in the context of using guidance operationally. Thus, each forecaster made 40 ratings overall.

2) INTERVIEW PROTOCOL

We conducted interviews, a qualitative research method, as our primary data collection method (Fig. 1b). Qualitative research aims to "come to terms with the meaning, not the frequency, of . . . phenomena in the social world" (Van Maanen 1979, p. 520). We chose this approach to obtain rich, detailed data to investigate our research questions.

Our interdisciplinary research team collaboratively developed the structured interview protocol, published in full in DesignSafe (Cains et al. 2024a,b). The interview included four sections. The first section included questions about the

[3] We used both the terms artificial intelligence and machine learning with the acronym AI/ML in the survey and interview questions. Both terms are used in the manuscript when reporting how a question was asked. Otherwise, only "AI" is used for simplicity; we recognize that some aspects discussed technically are ML, whereas others are more general AI.

TABLE 1. Preinterview survey items measuring whether forecasters deem different attributes of forecast guidance essential, useful but not essential, or not necessary.

Being able to interact with the guidance via a web-based tool or in AWIPS
Being able to sample the product and see the inputs that yield the output
Being able to see the evolution in the guidance, either as $d(\text{prog})/dt$ for a given event or over multiple events
Being able to examine the guidance predictions for past (archived) cases
Being able to assess the variability in the guidance output across a range of cases
Being able to compare the forecast from this guidance to a similar type of guidance that is valid at the same or similar time
Being able to compare the guidance with observational data
Knowing how the guidance output is derived
Knowing how the guidance verifies
Knowing what the failure modes of the guidance are
Knowing why the guidance works or fails
Knowing how the guidance verifies compared to other guidance
Knowing the person or group of researchers who developed the [AI/ML/new] guidance
Knowing the [AI/ML] techniques that were used to develop the [new] guidance, such as random forest techniques and neural networks
Knowing what modeled and/or observational data were used to train and calibrate the [AI/ML/new] guidance
Knowing what the input variables are in the [AI/ML/new] guidance
Knowing how each input variable is weighted and/or how they affect the outcome
Knowing what the [AI/ML/new] guidance deemed as the most important information to calculate the output (e.g., using ranking of input variables and attribution heatmaps)
Having the [AI/ML/new] guidance developed and provided in a way that resembles existing guidance products
Having spatially smoothed [AI/ML/new] guidance output, given that some [AI/ML/new] guidance has been developed that lacks spatial coherence and looks noisy at fine spatial scales

forecaster's current job role and core responsibilities and about their prior related work experience.

In the second section, we asked the forecaster about the recent, challenging severe convective forecast that they thought of while taking the survey. We asked them to briefly describe the event and their process of forecasting for it, including what data and products they used, what key decisions they made, and what forecast uncertainties were key. We further asked why they choose what tools and guidance they use and to what extent they need to know the details of how a product was derived to consider using it operationally. Data from these questions informed our analysis for RQ1.

In the third section—for the AI version only—we asked what came to mind when hearing the terms "artificial intelligence" and "machine learning" and how they felt about AI and ML in forecasting. For both interview versions, we then asked about the forecaster's interpretations of the terms "trustworthiness," "explainable," and "interpretable," including how they pertain to forecast guidance. Results from these questions will be reported in future work.

The fourth section comprised the bulk of the interview and elicited the forecaster's assessments of the trustworthiness of a specific AI product that was presented to them (RQ2), one of two nonoperational severe convective weather AI products under development by members[4] of the research team. One product uses a convolutional neural network (CNN) to predict the probabilities of storm objects being a supercell, quasi-linear convective system (QLCS), or disorganized (Sobash et al. 2023). The other product uses a random forest classification and regression to predict the probability of severe hail (≥2.54 cm; Gagne et al. 2017; Burke et al. 2020). Both products forecast up to a 36-h lead time and were brand new to the forecasters. The interview content and corresponding questions, further described below, were approximately parallel for these two products and investigated the same type of guidance attributes for the two different AI techniques and products.

Further, we designed and used parallel interview versions that either explicitly labeled the guidance as AI/ML or not. The no-AI version contained all the same technical details as the AI version—for example, it indicated that the storm mode probability product was developed using a CNN—but the product was referred to as new guidance or using a "new technique" rather than as "AI/ML guidance" or using an "AI/ML technique." The intention of this was to see whether or what effect explicitly labeling the product as AI/ML would have on forecasters' perceptions and responses, not to obscure the use of AI/ML techniques. At the end of the no-AI version of the interview, we disclosed that the new guidance the forecaster had just explored was AI/ML and asked whether this knowledge changed how they thought about it. We then followed up by asking what came to mind when hearing the terms artificial intelligence and machine learning, as well as how they felt about AI and ML in forecasting.

To assess how forecasters perceived the trustworthiness of the AI guidance they reviewed, we developed a Google Slides deck that presented information about the guidance and included an interactive virtual information board with which forecasters could learn about the different attributes. Full details of the slide deck content are published with the interview protocol in DesignSafe (Cains et al. 2024a,b). We asked the forecasters to think aloud and read the information while

[4] The coauthors who developed the severe hail and storm mode AI products did not participate in the data collection or analysis. Data collection and analysis were led by the risk communication team on the project, while developers provided the material and additional interpretation context through domain expertise.

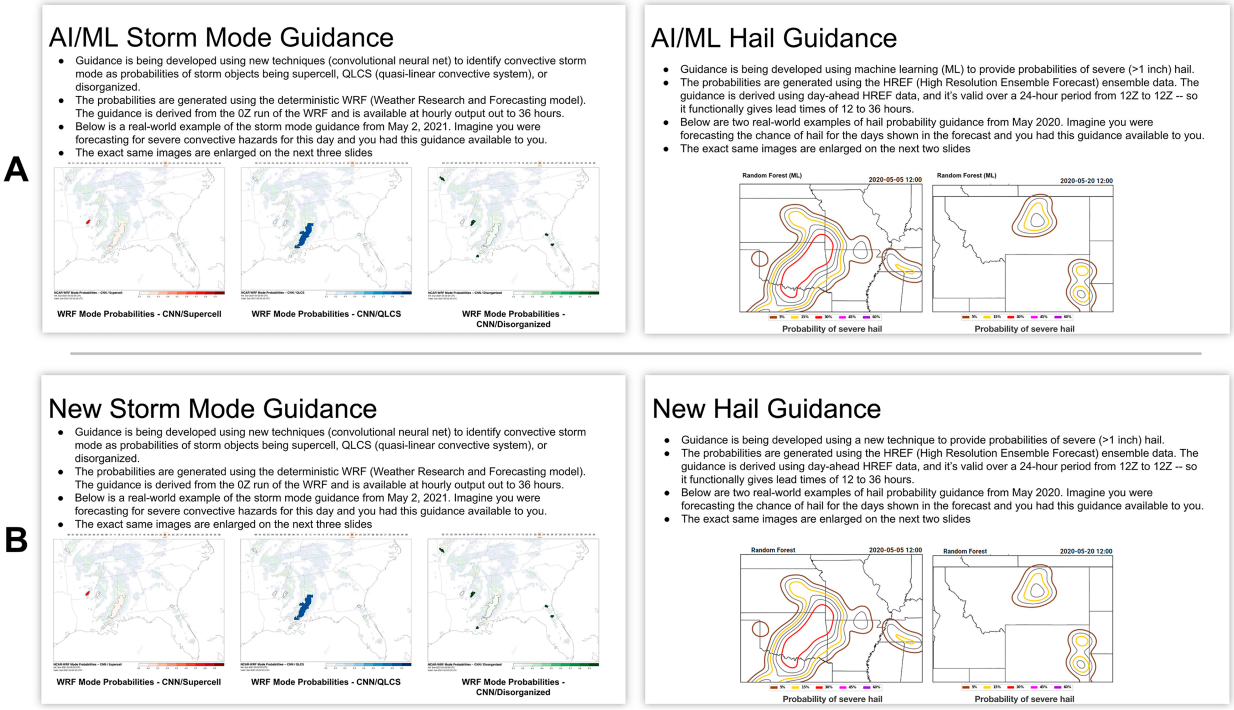

FIG. 2. Slides containing initial background information about (left) storm mode and (right) severe hail AI guidance that forecasters reviewed and provided initial thoughts and interpretations. (a) Slides clearly identifying the guidance as AI/ML and (b) slides with the no-AI version.

they performed the task of going through the slide deck. This think-aloud approach, as we explained to the forecaster participants, means to verbally share what is going through one's mind, without analyzing or justifying one's thoughts (Ericsson and Simon 1998; Charters 2003; Ericsson and Fox 2011; Schulte-Mecklenbeck et al. 2011).

The slide deck and corresponding think-aloud included multiple parts (Fig. 1b-4). We provided the forecaster with an initial slide that had basic background information about the guidance (Fig. 2). Then, we asked the forecaster their opinion about how trustworthy the guidance was on a scale from 0 to 10, where 0 meant "can't be trusted at all" and 10 meant "completely trustworthy," along with why they gave that rating. Next, we directed the forecaster to interact with the virtual information board while thinking aloud. The information board had six virtual sticky notes (Fig. 3a), each of which pertained to an attribute that we hypothesized might affect the forecaster's perceived trustworthiness of the guidance (Fig. 3c). Four of the attributes—the technique used, training, verification, and developers—were included in both the storm mode and severe hail versions. The other two attributes were specific to the product. For the severe hail product, these attributes included (i) details of the input variables and (ii) a comparison with the Storm Prediction Center's hail forecast. For the storm mode product, these attributes included (i) a web-based platform to interact with the guidance and (ii) the application of the algorithm trained from one NWP model to another. Each virtual sticky note linked to a slide with details about that attribute for the forecaster to read (Fig. 3b). For each sticky note, the forecaster was asked to think aloud as they read the information linked to the sticky note and placed the sticky note on a slide titled "Trustworthiness of [AI/ML] product" that included three panels labeled "Decrease," "No Impact," and "Increase," respectively (Fig. 3c). After working through all six guidance attributes, we reasked the forecaster's opinion about how trustworthy the guidance was on the 0–10 scale.

Finally, we asked the forecaster in two places about what else they would want to know or do with the product, if anything, before using it for forecasting. We first asked this question with the initial background slide and then again after they went through the full information board and all attribute details.

b. *Recruitment and sampling*

The sampling strategy for identifying and inviting forecasters to participate in this research was based on (i) the climatologies of storm mode and severe hail because those were what the AI guidance predicted and (ii) the forecaster's position in the NWS. Only one interviewee was invited from each Weather Forecast Office (WFO). For the forecaster's position, we recruited General Schedule (GS) 5–12 meteorologists, lead meteorologists, and Science and Operations Officers (SOOs). Details of how we developed our climatologically based sampling frame of WFOs and forecaster recruitment are in the online supplemental material.

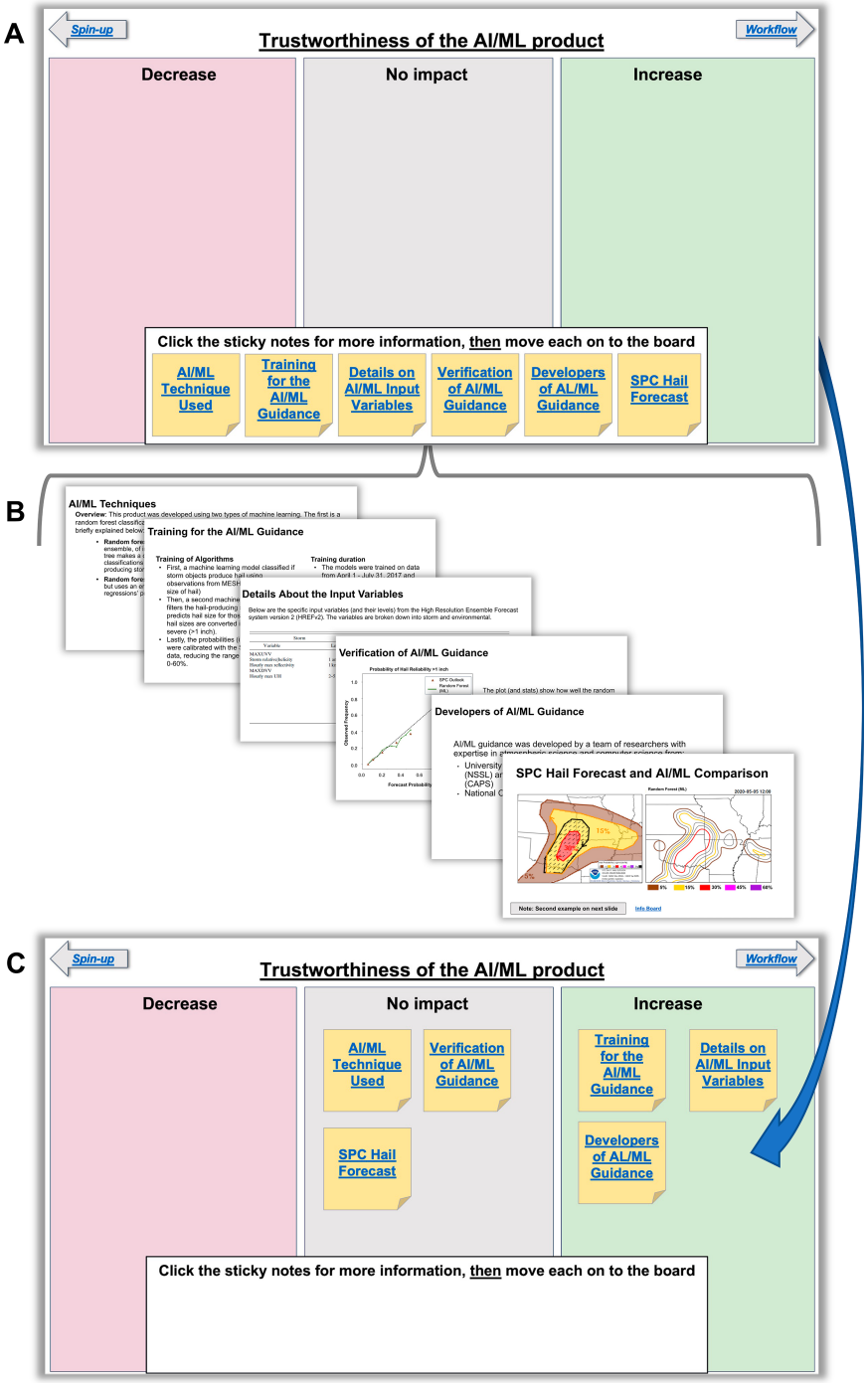

FIG. 3. Attribute assessment. (a) Information board with "sticky notes" for six attributes of the AI guidance. (b) Slides containing text and graphical information for the six attributes; severe hail shown. (c) Sticky notes sorted by a forecaster, while thinking aloud, as either decreasing, having no impact, or increasing the trustworthiness of the reviewed guidance based on the information provided.

TABLE 2. Characteristics of the 16 NWS forecasters interviewed.

Characteristic	Interview sample
NWS experience: median (range)	16 years (1–29)
NWS job position	
GS 5–12 meteorologists	7
Lead meteorologist	4
SOO	5
Type of AI product reviewed[a]	
Probability of storm mode	9
Probability of severe hail	7

[a] Forecasters only reviewed one of the two AI products.

We planned for and ultimately conducted interviews with a sample of 16 forecasters[5] to balance capturing a diversity of attitudes and experiences while not overburdening the forecaster community. By the last interview, few to no new notable ideas emerged from the interviews, indicating that we were reaching "saturation" (Merriam and Tisdell 2015). The characteristics of the sample are provided in Table 2.

The interviews were conducted online between October 2021 and May 2022 using Google Meet. Two researchers (the first and second authors) participated in each interview: One led the interview, while the other observed and took notes. The audio and video of each interview were recorded for transcription and data analysis purposes. The median interview was 89 min long (mean: 91; range: 76–110 min).

The research design was reviewed and approved by NCAR's Human Subjects Committee. All interviewed forecasters consented to participate and to the audio and visual recording of the interview. Per human subject regulations and our research ethic, the forecasters' identities (name and WFO) are de-identified, meaning we have redacted any language that could reveal their identity (e.g., if they discuss a local landmark). Such an approach allows interviewees to be honest and open when expressing their thoughts and feelings (Scott 2005).

c. *Data analysis*

For the preinterview survey, we analyzed item response frequencies (Fig. 1c). The forecaster interview transcriptions cumulatively produced hundreds of pages of rich and expressive textual data, which we analyzed with both qualitative and quantitative content analysis methods (Fig. 1d). Content analysis is a flexible method (Cavanagh 1997; Hsieh and Shannon 2005) where the "researcher is the primary instrument" for analysis (Merriam and Tisdell 2015, p. 16).

Our primary analytical focus was qualitative content analysis, specifically reflexive thematic analysis using the constant comparative method, which emphasizes the researcher's "reflexive engagement with theory, data, and interpretation" (Braun and Clarke 2021a). This approach involves the researcher iteratively analyzing the data to identify and define codes, synthesize codes into themes, and connect relevant, cross-cutting ideas (see Braun and Clarke 2006, 2019, 2021b). Figure 4 is an annotated example from our interview data of reflexive thematic analysis to illustrate how thematic codes are identified in the data, which are further synthesized into themes.

Last, we quantitatively analyzed forecasters' ratings of trustworthiness on the scale from 0 to 10, including their pre- and postinformation board ratings and the change in their ratings.

3. Results: Preinterview survey about attributes of forecast guidance

In Fig. 5, we show the attributes that at least 8 (≥50%) of the forecasters deemed as essential in the context of getting familiar with or training on the new guidance (familiarization) and in an operational context for using the guidance (operational). Seven attributes were deemed "essential" by half of forecasters in both contexts. One attribute is being able to compare the guidance with observations (item A). Three attributes pertain to verification, including generally knowing what the failure modes are (item B), how the guidance verifies (item C), and how it verifies compared to other guidance (item D). The three other attributes pertain to being able to interact with the guidance (item E), being able to sample the product (a form of interacting) to see the inputs that yield the output (item F), and knowing what the input variables are (item G).

Specific to familiarization and training, all forecasters indicated it is essential or useful to know why the guidance works or fails (item H). All forecasters also deemed it essential or useful to know how the output was derived and what data were used to train and calibrate the AI guidance (items I and J). Additionally, most forecasters deemed it essential to examine AI predictions for past cases (item K).

Specific to operational use, two attributes were deemed essential or useful by most forecasters: being able to compare the forecast from AI guidance to that of similar guidance valid at the same time (item L) and being able to examine the evolution in the guidance over time (item M). These likely reflect forecasters assessing model consistency as a way of shaping their forecast confidence (Demuth et al. 2020; Henderson et al. 2023). Interestingly, neither attribute was deemed important for familiarization or training.

4. Results: Understanding forecasters' guidance use decisions

To investigate how forecasters make guidance use decisions (RQ1), we analyzed the forecasters' responses to questions about why they use the guidance that they do and the extent to which they need to know the details of how a product was derived to use it for forecasting given that guidance can sometimes seem like a "black box," all in the context of their recent challenging forecast event.

[5] Prior to the 16 interviews that the dataset analyzed here is composed of, we conducted pretest interviews with six colleagues with relevant forecasting expertise whom we directly contacted. The pretest interviews informed changes to the interview structure, wording, organization, and length. None of the data from the pretest interviews are included in the analysis reported here.

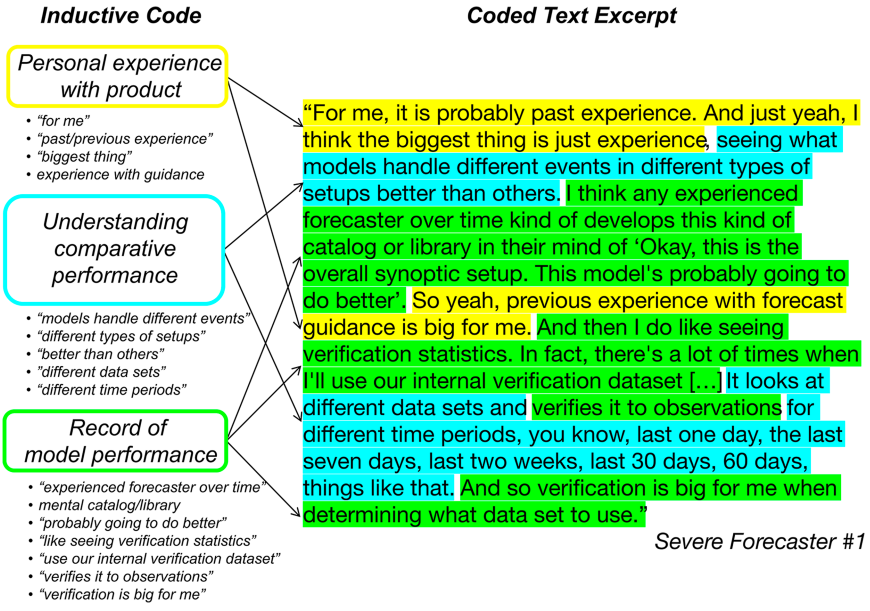

FIG. 4. Annotated reflexive thematic analysis of Forecaster 1 explaining how they decide what guidance, tools, and information to use for forecasting.

We synthesized three interrelated themes from our reflexive thematic analysis related to 1) forecasters' desire to understand guidance, 2) their interactive process for exploring and evaluating guidance, and 3) how they prioritize information when time is limited.

Theme 1.1: *Forecasters want to understand guidance to use it, but what constitutes "understanding" is complex and multifaceted*. Per the forecasters' quotes, understanding includes learning the functionality of a model,[6] its strengths and weaknesses, how it performs under different scenarios, and how it performs compared to other models (Table 3: 1.1A–C). Forecasters do this by developing a "catalog or library in their mind" as to which model is "probably going to do better" in different situations (Table 3: 1.1C; Fig. 4). These aspects of understanding models are central to how forecasters use them for given atmospheric conditions and weather events and allow forecasters to mentally bias correct as needed. However, forecasters' need to understand guidance is more complex and varied when it comes to knowing the details of how a model was derived. Some forecasters said the specifics of model inputs and derivation of outputs are not needed if the model performs well (Table 3: 1.1D). Others emphasized that understanding the model inner workings is important when initially using it, "but once [they] trust it, then it's not as important" other than if it is particularly sensitive to certain conditions (Table 3: 1.1E). Forecaster 2 described how black boxes can impede their trust in models and that they need to "see it in action" to believe it (Table 3: 1.1F); this sentiment links this theme about understanding to the next theme about forecaster experience.

Theme 1.2: *Forecasters go through an iterative process of exploring and evaluating guidance to determine whether, when, and how it should be used during operations*. Multiple forecasters described how familiarity with models, using a given tool repeatedly, and having direct experience using guidance are important, illustrating the iterative evaluative process they go through for guidance use. One forecaster noted they do not trust guidance at face value nor use it solely based on others' recommendations, but rather that they need to use it themselves repeatedly to "believe it" (Table 3: 1.1F, 1.2A). As the quotes in Table 3 illustrate, repeated personal experience with a model allows forecasters to develop familiarity with it, to learn how it performs for their geographic area and forecasting needs, and to know how to adjust from what the model output is if needed, all of which influence forecasters' trust and/or confidence, both of which were terms explicitly mentioned by forecasters. Connecting this theme to Theme 1.1, iterative evaluation to develop familiarity and assess performance appears to be a key mechanism by which forecasters develop their desired deep understanding of model guidance, which in turn can influence their trust or confidence in it and ultimately use of it. Importantly, the "personal" aspect of the process stands out; as Forecaster 12 noted, "We're all doing the same job, but we all have our own specific processes to come to our ultimate decisions" (Table 3: 1.2E).

Theme 1.3: *Forecasters are overloaded with information but limited on time, and they manage this tension by prioritizing using guidance and other resources that best help them do their job*. Multiple forecasters expressed that there is so much information available to them that they must "pick and choose"

[6] "Model" is used in section 4 to mean any type of model (e.g., NWP and AI).

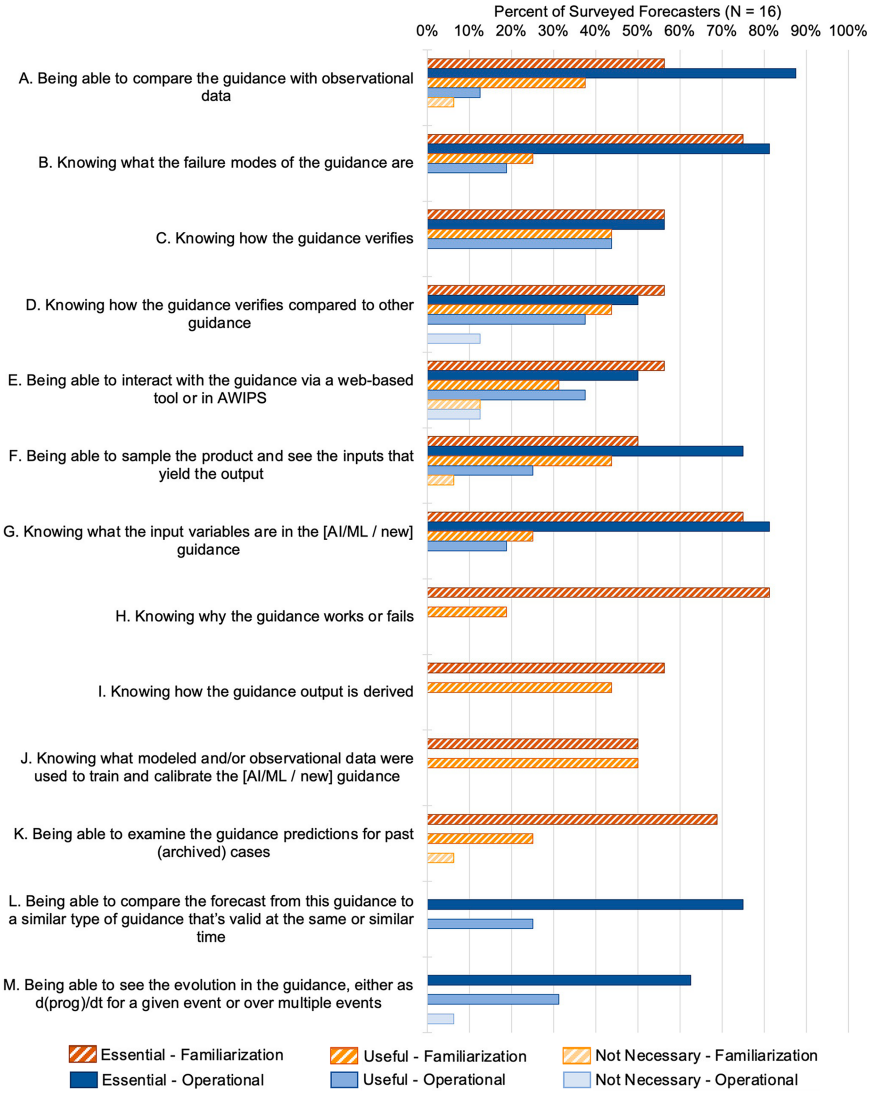

FIG. 5. Preinterview survey items that at least 8 (50%) forecasters deemed essential for the familiarity/training phase or operational phase.

what to use (Table 3: 1.3A). As Forecaster 7 explained, the challenge of using new guidance in addition to existing tools is "where it starts to pile on" (Table 3: 1.3B). Forecasters do not have time, however, to dig into the inner workings of every tool they use because of the fast-paced environment they work in, especially when working on high-impact events (Table 3: 1.3C). Therefore, they have developed strategies to manage the information overload, such as looking at a tool postevent to determine whether it might have helped them in the moment (Table 3: 1.3A) and, in particular, using the tools that best help them do their job. There are multiple reasons why a tool can be especially helpful, such as a tool that curates and centralizes key pieces of information on a website. Forecaster 13 indicates the Storm Prediction Center's (SPC) mesoanalysis web page as an exemplar noting that it is a "really good one-stop shop . . . where you can look at all the different parameters in a very quick time frame" (Table 3: 1.3D). Theme 2.3 presented below (section 5a) describes another reason why forecasters find certain tools especially helpful.

5. Results: Assessing the trustworthiness of AI forecast guidance

To investigate RQ2, we analyzed the think-aloud portion of the interview, including the initial background information provided about the AI forecast guidance (RQ2A; section 5a); the effect of different guidance attributes on forecasters' perceived trustworthiness (RQ2B; section 5b); and additional attributes not provided that forecasters mentioned as important (RQ2C; section 5c). We did not find meaningful differences between the AI and no-AI versions in the analyses we present here, for which reason we combine reporting of them in all of the results discussed below. Note that this only applies to the analyses presented here of the think-aloud section of the

TABLE 3. Thematic codes and example quotes supporting Themes 1.1, 1.2, and 1.3 regarding how forecasters make guidance use decisions (RQ1). Formatted as **Identifier** | ***Thematic code***: Example quote.

Theme 1.1: Forecasters want to understand guidance to use it, but what constitutes "understanding" is complex and multifaceted	Theme 1.2: Forecasters go through an iterative process of exploring and evaluating guidance to determine whether, when, and how it should be used during operations	Theme 1.3: Forecasters are overloaded with information but limited on time, and they manage this tension by prioritizing using guidance and other resources that best help them do their job
1.1A | ***Model functionality (e.g., how and why it works)***: "You've gotta know how things are put together and how they're made to have a deeper understanding of what it's giving you and why it's giving you what it's doing. So I think [understanding how a model works is] essential. There are probably degrees of how much you need." Forecaster 11	**1.2A** | ***Guidance should not be taken at face value***: "Somebody can send me a white paper, 'Hey, check it. Look at this video' and all that. That's great. [The model] worked great for wherever the study was done. But until I actually try it, for a couple of times myself personally, I do not necessarily believe it." Forecaster 2	**1.3A** | ***Not enough time to look at all the potentially relevant information***: "You have your standby [products] that you know, and how do I kind of sprinkle this [new guidance]? Sometimes we do a post analysis using some of these tools and say, 'Would this have helped us?' Again if we don't have time . . . There's so much out there now that you gotta pick and choose." Forecaster 4
1.1B | ***Understanding model's strengths and weaknesses for proper use***: "I think the main thing that's critical is just understanding the limitations–strengths and limitations – with whatever it is that's being derived, not necessarily how it was derived." Forecaster 7	**1.2B** | ***Explore product prior to operational use***: "I think a lot of [why I use what I use is] is familiarity. Even from the time I started in the Weather Service you had people that had been in for a longer time training me on what they used, what they like to use." Forecaster 16	**1.3B** | ***Amount of information and number of resources can become overwhelming***: "I try to stay involved with some of the latest and greatest research that's coming out. I'll be trying to utilize a lot of these new products and new model suites and new stuff. I try to hold on to some of the old stuff too. So that's where it starts to pile on. You have a lot to look at." Forecaster 7
1.1C | ***Understanding comparative performance (e.g., other models, scenarios)***: "Seeing what models handle different events in different types of setups better than others. I think any experienced forecaster over time kind of develops this catalog or library in their mind of, 'Okay, this is the overall synoptic setup. This model's probably going to do better.'" Forecaster 1	**1.2C** | ***Record of model performance***: "I probably have a favorite model that I go to all the time, but it's consistent. I know its weaknesses. I know what it's good at, and I can adjust to that." Forecaster 14	**1.3C** | ***Limited time to dig into inner workings***: "There are times when I probably wish I knew a little bit more about what goes into specific products so that I had a better feel for what I'm analyzing. I just think that forecasting is becoming such a fast-paced environment sometimes; especially in high impact events where you don't really have enough chance or time to focus on a lot of those sorts of specifics." Forecaster 12
1.1D | ***Model (e.g., black box) specifics are not immediately needed if model performs well***: "If I see [the model] performing well, I'm going to latch on to it right away. . . . If it does well, I think since I'm in operations, I don't need to know what was in it or how it was deriving things." Forecaster 3	**1.2D** | ***Model performance validated with case studies***: "You have trust in what you see – I'm going to use the word confidence – but you have confidence in what you're looking at. And the reason I do personally is because I validated all of this. It's not just looking at model data, forecast data. It is that, but it's more tied into doing case studies, research projects. They don't have to be large ones. They're just small ones dealing with these summertime convective events." Forecaster 11	**1.3D** | ***Time-saving and efficient source of information***: "So I guess [the mesoanalysis data] is what we've become comfortable with using over the last year or two. We incorporate, in terms of the mesoanalysis data, the SPC[a] mesoanalysis site. That's a really good one-stop shop because it has a variety of information available. I think that's very important from a mesoanalyst role or a severe weather forecasting role is to have some sort of quick one-stop shop to go to where you can look at all the different parameters in a very quick time frame, especially when things are fast breaking." Forecaster 13

TABLE 3. (*Continued*)

Theme 1.1: Forecasters want to understand guidance to use it, but what constitutes "understanding" is complex and multifaceted	Theme 1.2: Forecasters go through an iterative process of exploring and evaluating guidance to determine whether, when, and how it should be used during operations	Theme 1.3: Forecasters are overloaded with information but limited on time, and they manage this tension by prioritizing using guidance and other resources that best help them do their job
1.1E | ***Understanding how a model is derived is important during the initial or training phase with a new tool***: "I want to know initially, there's no question, what goes into [the model] to make it work. But once I trust [the model], then it's not as important other than to know if something is very sensitive to dewpoint, which all CAMs[b] are basically." Forecaster 2	**1.2E** | ***Personal experience with products***: "I think [why I use what I use is] familiarity and that's probably the biggest thing. And also helping with giving me more confidence because it's the products that I look at routinely. I just know kind of what to look for. . . . We're all doing the same job, but we all have our own specific processes to come to our ultimate decisions." Forecaster 12	
1.1F | ***Black boxes can impede forecaster trust in tool/model without first-hand understanding and use***: "I don't trust black boxes. Just if somebody says, 'This is great.' Okay, I'll take a look at it. But I do not believe it until I understand it and see it in action." Forecaster 2		

[a] Storm Prediction Center.
[b] Convection-allowing model.

interview protocol.[7] We are analyzing other sections of the interview separately and will report in future papers on any differences we find between the AI and no-AI versions.

a. *RQ2A*: *Impact of initial background information on forecasters' assessment of AI (or new) guidance*

Three complementary themes were synthesized from the analysis of forecasters' think-alouds as they explored initial background information: forecasters' 1) use of conceptual models, 2) desire for AI model explanations, and 3) favorable assessment of guidance that meets an unmet need.

Theme 2.1: *Forecasters apply their conceptual models of the atmosphere when initially making sense of new guidance.* Forecasters have expert knowledge about how the atmosphere is structured and how it behaves, and they applied their conceptual models in multiple ways when looking at the initial information about the new guidance to make sense of and evaluate it. Several forecasters explicitly noted that the guidance made sense to them given that it matched their conceptual models of severe convective weather. For instance, Forecaster 10 noted that the storm mode objects evolved in ways they would expect (Table 4: 2.1A). Further, some forecasters who explored the storm mode guidance interpreted the information by comparing the probabilities of the different modes at a given forecast valid time and thinking about the feasibility of the specific mixed-mode scenarios (Table 4: 2.1B). The probabilistic storm mode guidance included composite reflectivity as an underlay (see Fig. 2), which the forecasters also examined and mentioned as they applied their conceptual models. Having the new storm mode guidance coupled with radar data reflects forecasters' real-world operational environment in which they have access to and consult multiple sources of information to make sense of atmospheric conditions. Although the forecasters' use of their conceptual models typically resulted in them being better able to make sense of the new guidance, some forecasters were unsure how to interpret the disorganized storm mode (Table 4: 2.1C).

Theme 2.2: *Forecasters want additional, baseline information about AI model development to better understand it, particularly about the AI techniques used, the model inputs and how they are defined, and how the model outputs are derived.* As discussed in Theme 1.1, most forecasters want some basic understanding of guidance to use it, including some understanding of model functionality. We identified that desire in a few specific ways when showing forecasters the initial background information of the storm mode and severe hail guidance. Several forecasters stated that they were unfamiliar either with AI in general or with the specific CNN or random forest technique used, and they wanted to learn more about it (Table 4: 2.2A). Forecasters also wanted to know different things about the model inputs, including how a storm object was defined for the storm mode predictions and whether and

[7] In a later part of the no-AI interview after the think-aloud, forecasters were asked questions about AI (data not analyzed in the present study), and two forecasters did mention that the term "CNN" made them think of AI. The first said AI was what came to mind when they read CNN, and the second understood that the use of a CNN meant the product they were reviewing was AI.

TABLE 4. Thematic codes and example quotes supporting Themes 2.1, 2.2, and 2.3 regarding the impact of providing the initial background information on forecasters' assessment of the AI (or new) guidance (RQ2A). Formatted as **Identifier** | ***Thematic code***: Example quote.

Theme 2.1: Forecasters apply their conceptual models of the atmosphere when initially making sense of new guidance	Theme 2.2: Forecasters want additional, baseline information about AI model development to better understand it, particularly about the AI techniques used, the model inputs and how they are defined, and how the model outputs are derived	Theme 2.3: Forecasters favorably respond to guidance that fills an important, unmet operational need, in this case, storm mode prediction
2.1A | ***Guidance makes sense to forecasters given that it matches their conceptual model***: "This [visual] makes sense to me. This is what I would expect from a conceptual model of how I would expect [the storm objects] to evolve. This first image is something I would think would be initially trustworthy based on what I've seen. There's nothing here that would set off any alarm bells that would make me think 'Oh, this literally doesn't have a clue what's going to happen.'" Forecaster 10	**2.2A** | ***Forecasters are unfamiliar with either AI or specific technique and want more information***: "[The guidance] is something I'm definitely interested in based on the description given here. At this point my interest is piqued, and I'm definitely looking to take a closer look at the guidance. I guess my only question is I don't know what a convolution neural net is. That's something I haven't heard of yet." Forecaster 10 "I'd probably have to familiarize myself with exactly what the random forest process is; the methodology there. I think that's probably the key: knowing what's all involved with the random forest." Forecaster 6	**2.3A** | ***Forecasters see immediate utility in storm mode guidance (e.g., threat and hazard assessment, screening tool, messaging)***: "[…] this guidance is going to help show probabilities, which I love. Probabilities of storm objects being supercell, QLCS, or disorganized. I love it because if I'm thinking supercells, I'm thinking this set of hazards. If I'm thinking QLCS, I'm thinking this set of hazards. If I'm thinking disorganized, I'm not really thinking of hazards other than pulse or very, very infrequent in time and space and not of great magnitude." Forecaster 15
2.1B | ***Forecasters interpret and make sense of different storm modes by comparing across modes***: "When looking at the QLCS[a] probabilities, those are very, very high in that same area from Alabama to Eastern or Northeastern Louisiana, very high probabilities. So to me, this suggests that the predominant mode would be QLCS and maybe you could have an embedded Supercell in that line here or there. Which is certainly a scenario that is not really uncommon. And if ultimately that's what happened, that would be pretty impressive, that it was able to show that in a probabilistic sense." Forecaster 1	**2.2B** | ***Forecasters want to know how AI model was trained to identify storm objects and classify storm mode***: "How is [the AI model] defining what the [storm] object is? Because I can see [from radar underlay] that there's areas of convection that are not defined as an object, so I want to know what goes into defining some areas of objects." Forecaster 16	**2.3B** | ***Storm mode is a challenge and guidance would fill an operational need***: "'Guidance is being developed using techniques, convolution neural net to identify convective storm mode as probabilities of storm objects being supercell, quasilinear convective system or disorganized.' Okay. Very good. That's good to know. That could be very useful operationally. Trying to use new techniques to try to figure out convective storm mode as we talked about earlier. So that's kind of exciting to see." Forecaster 13
2.1C | ***Forecasters are unsure how to interpret disorganized storm mode***: "There's some other disorganized looking stuff that doesn't show up. So again guessing it has to do something with the background field. I guess I'm not seeing the utility of this one as much. Kind of struggling to interpret this disorganized. I think the other two are fairly straightforward in what they're trying to depict. This one I'm not so sure about, or the utility." Forecaster 4	**2.2C** | ***Forecasters want to know how multiple inputs combine with each other to produce AI model outputs***: "This is a supercell where I wouldn't generally expect it kind of behind this leading line of storms. If I know what the inputs are, I could go look at those individual inputs, like, 'Okay, is it like this because of an updraft helicity track.' I could go look at the updraft helicity track fields or something like that." Forecaster 5	**2.3C** | ***Forecasters think storm mode guidance would be useful (e.g., hazards) when paired with other sources of information***: "The supercell you're dealing maybe with [is] more of a hail threat; [it] gets a little bit more maximized there. So that's something where some sort of hail or hazard guidance would be very useful as well." Forecaster 10

TABLE 4. (*Continued*)

Theme 2.1: Forecasters apply their conceptual models of the atmosphere when initially making sense of new guidance	Theme 2.2: Forecasters want additional, baseline information about AI model development to better understand it, particularly about the AI techniques used, the model inputs and how they are defined, and how the model outputs are derived	Theme 2.3: Forecasters favorably respond to guidance that fills an important, unmet operational need, in this case, storm mode prediction
	"I guess further understanding [the AI model] to make sure I have the idea of how [the output is] being derived. What factors is [the AI model] taking into account so I can know like, 'Okay, I know this product is taking Shear multiplied by CAPE[b] plus, et cetera, et cetera.' It would make my processing a bit more efficient because if I know this is how [the output] is derived then I don't need to check all the other parameters." Forecaster 9 **2.2D** | ***Forecasters want additional training and reference material for new guidance***: "I think the biggest thing is I would like some training or explanation of how [the AI model] works and how these probabilities are derived. Not necessarily for something for me to reference the day of a severe event but for training, like preseason training." Forecaster 1	

[a] Quasi-linear convective system.
[b] Convective available potential energy.

what storm-based fields (e.g., updraft helicity) or environmental parameters (e.g., shear) "the model is taking into account" (Table 4: 2.2B,C). This desire to know the model inputs reflects Theme 2.1 about forecasters' conceptual models of the atmosphere and how they are making sense of the guidance, including when the model produces output that differs from what a forecaster might expect. Forecaster 5 expressed such desire when they noted that the supercell probability was in a different location than expected and if they "[knew] what the inputs are, [they] could go look at those individual inputs." This represents an opportunity where explainable AI could potentially enhance forecasters' understanding of what input variables (e.g., composite reflectivity and updraft helicity) influenced the classification of a storm object or what variables yielded higher supercell probabilities. Finally, forecasters expressed a desire to know how the storm mode and severe hail probabilities were derived, but as Forecaster 1 explained, they would want such in-depth information as part of "preseason training" and "not necessarily . . . to reference the day of a severe event" (Table 4: 2.2D). This desire to know more during preseason training about how guidance is derived is also supported by the training/familiarity-only findings from the preinterview survey (Fig. 5, items H, I, and J). Overall, these results show that when initially introduced to the guidance, forecasters were interested in and essentially seeking information about attributes that, unbeknownst to them at the time, would be introduced later in the interview as part of the information board. This suggests that the AI attributes presented (see section 5b) aligned well with forecasters' needs.

Theme 2.3: *Forecasters respond favorably to guidance that fills an important, unmet operational need, in this case, storm mode prediction.* Forecasters saw immediate utility in the storm mode guidance, given that storm mode is a forecasting challenge, and they indicated that the guidance would fill a need and "could be very useful operationally" (Table 4: 2.3A,B). The forecasters also noted that the storm mode guidance would also be useful for assessing threats of specific hazards (Table 4: 2.3A) and when paired with other sources of information such as "hail or hazard guidance" (Table 4: 2.3C). Forecasters expressed immediate, favorable reactions only for the storm mode guidance as filling a critical informational gap, whereas they perceived the severe hail guidance as being very similar to an existing SPC product (as it was designed to be). This does not mean that the severe hail guidance was not useful but rather that guidance that is critical to forecasters doing their job and that does not already exist, like the storm mode probabilities, can be especially useful (see also Theme 1.3).

b. *RQ2B*: *Trustworthiness assessment of AI (or new) guidance and attributes*

We categorized the initial and final trustworthiness ratings of the guidance into low, medium, and high trustworthiness. No new themes are presented in this section, rather our interpretations of the rating justifications and guidance attributes

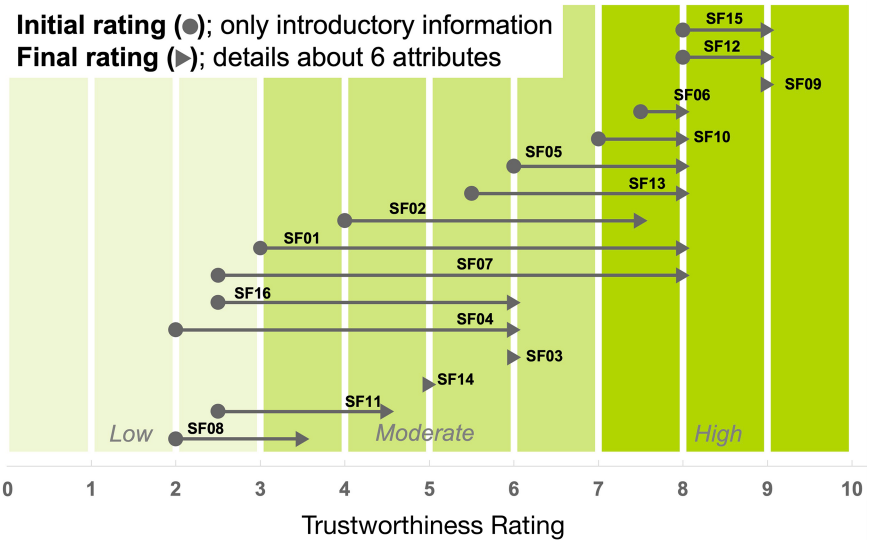

FIG. 6. NWS forecasters' perceived trustworthiness rating of severe weather guidance developed using AI techniques (SF). Initial ratings (circle) were given before the think-aloud and based on initial background information, and final ratings (arrowhead) were given after the think-aloud containing details about the guidance. SF03, 09, and 14 did not provide an initial rating. No pattern was found when ratings were categorized based on years of forecasting experience.

think-alouds yielded applied and specific examples (i.e., concretized manifestations) of the more generalized themes synthesized in sections 4 and 5a.

1) PHASE 1: INITIAL TRUSTWORTHINESS RATING

The initial trustworthiness ratings of the storm mode and severe hail AI guidance ranged from 2 to 8 (Fig. 6). Forecasters' reasoning about ratings reflected several of the themes discussed in section 4b (Table A1; see the appendix). Forecasters who rated the trustworthiness of the guidance low (rating ≤ 3) noted they did not know enough about the guidance or its performance. Additionally, three forecasters [severe forecasters (SFs) 03, 09, and 14] did not even provide an initial trustworthiness rating for the severe hail product because they did not feel they had enough information to make such a judgment. Forecasters who rated the trustworthiness of the AI guidance as moderate ($4 \geq$ rating ≤ 6) stated that the AI models matched their conceptual models and that the subsequent product output looked realistic (supports Theme 2.1). Forecasters who rated the trustworthiness of the AI guidance as high (rating ≥ 7) said the AI models matched their conceptual models and filled an operational need (supports Themes 2.1 and 2.3).

2) PHASE 2: ASSESSMENTS OF HOW GUIDANCE ATTRIBUTES INFLUENCE TRUST

Tables 5 and 6 are thematically synthesized representations of the forecasters' thoughts during the think-alouds while they examined and assessed the guidance attributes. Overall, each of the attributes shown tended to increase forecasters' assessed trustworthiness of the guidance, but there were instances in which some attributes decreased or had no impact on trustworthiness. Although Tables 5 and 6 are a curated synthesis, we caution that these should not be interpreted as a metrics checklist.

(i) AI/ML model technique

Learning more information about the AI model techniques and inputs increased trustworthiness for many forecasters, which further supports multiple themes discussed in previous sections about forecasters wanting to *understand* more (Themes 1.1 and 2.2). Several forecasters responded positively to learning that humans were involved in the AI model training process for the storm mode guidance (Table 5: 1C, 2C). However, the human hand labeling of storm mode images was also noted as a source of uncertainty by the forecasters given their knowledge about the difficulty of categorizing storm mode and thus the subjectivity and differing interpretations in doing so. Forecaster 16 conveyed this as "differences in opinion and some lack of consistency" of how to classify storm modes. Comparatively, Forecaster 7 acknowledged that humans were a source of uncertainty but, "assuming the human knows what they're doing, I think [hand-labeling] would be an increase in trustworthiness."

(ii) Training of AI/ML model

The training of the AI model had a mixed effect on trustworthiness (Table 5: 2A–C). Several forecasters said their familiarity with the sources of information [e.g., Weather Research and Forecasting (WRF) and High-Resolution Ensemble Forecast (HREF)] increased the trustworthiness of the AI guidance (Table 5: 2C), which is likely due to the forecasters having *personal experience* with those sources and building a mental *record of performance* (Table 3: 1.2F,C). Forecasters noted that a larger and longer storm mode dataset was needed to account for potential seasonal or climatological phenomena (Table 5: 2A); they also wanted to know whether the training datasets were geographically representative of the CONUS or a specific region, echoing results found by Demuth et al. (2020). Thus, when thinking about improving

TABLE 5. Part 1: Thematic codes developed from the analysis of forecasters' thoughts shared while thinking aloud as they decided whether the text and graphics provided for each guidance attribute decreased, had no impact, or increased the trustworthiness of the storm mode ($n = 9$) and severe hail ($n = 7$) AI guidance (RQ2B).

Guidance attribute	A. Decreases trustworthiness	B. No impact	C. Increases trustworthiness
1. AI/ML model technique	—	3 forecasters Intriguing but "just telling" about AI model Additional information is needed about the specifics of the technique	13 forecasters Human involvement in AI model training (storm mode only) Information about technique inputs and process Explanation and use of ensembles (severe hail)
2. Training of AI/ML model	1 forecaster Larger and longer dataset needed to account for potential seasonal or climatological phenomena (storm mode)	2 forecasters Similar information to technique slide Not enough information Content raises additional questions	11 forecasters Useful details about data, inputs, and processes Human hand-labeling storm mode images Familiarity with referenced sources of information Duration and timeframe of training data match weather seasonality (severe hail)
3. Verification metrics	—	2 forecasters Not enough information to correctly interpret statistics (e.g., AUC[a] and BSS[b])	14 forecasters Reliability diagram, false alarm rate, probability of detection Verification is valuable and "always important" to forecasters Impressed by verification given forecasting difficulty (storm mode) Comparable verification to SPC[c] (severe hail)
4. Developers	—	5 forecasters Does not affect use of guidance "It doesn't matter" Does not track the developers of all the guidance they use Taken for granted that developers are experts	11 forecasters Familiarity with institutions and positive reputations Inclusion of operational expertise, not just academia Multidomain expertise and experience Collaborative effort from multiple groups

[a] Area under the receiver operating characteristic curve.
[b] Brier skill score.
[c] Storm Prediction Center.

the utility of guidance, developers should consider broadening *where* and for *what* types of storms and scenarios the training data represent and not just *when* the AI model is trained for. The importance of the *where* and *what* does not diminish the need for temporal representation (Table 5: 2A); however, the absence of more geographically representative training datasets could cause decreased trustworthiness due to model outputs that conflict with the *forecaster's knowledge and conceptual model* of the geographic region (Table 4: Theme 2.1).

(iii) Verification metrics

The verification information for both the storm mode and severe hail increased trustworthiness of the guidance for nearly all forecasters (Table 5: 3C). Several forecasters noted that verification information is valuable and always important for them; some forecasters specifically cited the inclusion of the reliability diagram, false alarm rate, and probability of detection. However, a few forecasters noted that there was not enough information provided to properly interpret the verification statistics, which in those instances was associated with a no-impact-on-trustworthiness decision (Table 5: 3B). This further supports that forecasters want to understand guidance but that *understanding is multifaceted* (Theme 1.1; Table 3). Even though the verification information provided basic definitions and interpretations of the shown statistics, this was not sufficient or the right type of information to influence some forecasters' AI guidance trustworthiness judgments.

(iv) Developers

Information about the AI model developers had no impact on the trustworthiness of the guidance for one-third of forecasters (Table 5: 4B). However, it did increase trustworthiness for two-thirds of forecasters, in part due to the inclusion of operational expertise (Table 5: 4C). Forecaster 11 emphasized the need for such expertise when saying, "The fact that [the developers] actually put operational forecasters in the team is very important. That's the end user of this kind of guidance so they need to have some kind of input, and that's important."

TABLE 6. Part 2: Thematic codes developed from the analysis of forecasters' thoughts shared while thinking aloud as they decided whether the text and graphics provided for each guidance attribute decreased, had no impact, or increased the trustworthiness of the storm mode ($n = 9$) and severe hail ($n = 7$) AI guidance (RQ2B).

Guidance attribute	A. Decreases trustworthiness	B. No impact	C. Increases trustworthiness
1. Interactive output (storm mode)	—	—	9 forecasters Ability to evaluate dynamic evolution of fields, storm objects Location and time of storm modes match expectations from conceptual model Predictions are internally consistent and make meteorological sense Familiar graphical user interface Operational utility of guidance
2. Input information (severe hail)	—	2 forecasters Do not know AI model's sensitivity to the inputs Seeing the actual product would be more helpful	5 forecasters Knowing what is going into the AI model Inputs are reasonable, comparable to what forecasters would use Familiarity of variables
3. WRF and HRRR comparison (storm mode)	2 forecasters Cautious since CNN[a] was not trained on HRRR[b] data but was applied to it	3 forecasters CNN was not trained on HRRR data but was applied to it "Can't really say . . . doesn't decrease"	4 forecasters Additional tools and information to build forecast Location of HRRR storm objects makes sense given HRRR output CNN did well with HRRR even with WRF[c] differences
4. Random forest and SPC comparison (severe hail)	—	3 forecasters Did not affect how they felt No history with AI severe hail product to determine trustworthiness	4 forecasters Additional tools and information to build forecast "Good agreement" to SPC[d], a familiar and used resource Finer resolution than SPC[d]

[a] Convolutional neural network.
[b] High-resolution rapid refresh.
[c] Weather Research and Forecasting.
[d] Storm Prediction Center.

(v) Interactive output (storm mode)

Forecasters' ability to interact with the output of the storm mode guidance via a web-based interface and to evaluate the dynamic evolution of the parameter fields and storm objects increased trustworthiness, partly because it helped forecasters evaluate the guidance by comparison with their conceptual models (Table 6: 1C). The location and timing of the different modes matched forecasters' expectations, aided by the radar underlay that was a part of the guidance visualization, which is an applied example of Theme 2.1. The desire for interactivity for sense-making of black box–esque tools is not unique to AI guidance; similar needs were found by researchers investigating forecasters' assessment of uncertainty using ensemble prediction systems (Novak et al. 2008).

(vi) Input information (severe hail)

Forecasters who reviewed the severe hail guidance said the familiarity of the input variables for the AI model increased their trustworthiness of it (Table 6: 2C). As Forecaster 9 said (which also supports Theme 1.1), "If I understand how I use these variables, and I'm already seeing that in this product, then [I'm] very happy with [the AI guidance]." For some forecasters, the input information had no impact on trustworthiness because they still did not know the model's sensitivity to those inputs (Table 6: 2B).

(vii) WRF and HRRR model comparison (storm mode)

The output comparison of the WRF input versus High-Resolution Rapid Refresh (HRRR) input for the storm mode CNN model produced the most mixed reactions about trustworthiness (Table 6: 3A–C). The comparison decreased trustworthiness for the forecasters who thought that guidance trained with the WRF and then applied to the HRRR was cause for pause. This supports Theme 1.1 wherein understanding guidance includes knowing the appropriate use of the model used to develop the guidance (Table 3: 1.1B). Conversely, this cross application increased the trustworthiness of storm mode for forecasters who drew on their experience with the HRRR to make sense of the location of the predicted HRRR storm objects. This is a concrete example of how *personal experience with products* (Table 3: 1.2F), like the WRF and HRRR, contributes to forecasters developing an *understanding of when and how a new piece of guidance should be used* (Table 3: Theme 1.2).

(viii) Random forest and SPC comparison (severe hail)

The output comparison of the severe hail AI guidance versus the SPC hail outlook either had no impact on the assessed trustworthiness of the guidance or increased its trustworthiness (Table 6: 4B,C). Some forecasters noted that the comparison did not affect how they felt about the guidance, whereas others said they had no experience with the AI product to evaluate its trustworthiness (Table 6: 4B). The forecasters who said the comparison increased their trustworthiness noted that it highlighted that the AI guidance was in "good agreement" with the SPC, which is a familiar and commonly used resource, and thus, they deemed that the AI guidance was a useful, additional tool for forecasting (Table 6: 4C).

3) PHASE 3: FINAL TRUSTWORTHINESS RATINGS

After the attribute assessment, the final trustworthiness ratings of the storm mode and severe hail AI guidance ranged from 3.5 to 9 (Fig. 6). The majority of the forecasters rerated the trustworthiness of the AI guidance as high (rating ≥ 7), crediting the additional background information provided via the sticky notes, the forecasters' familiarity with input data sources (e.g., WRF and HRRR), and comparison to conventional resources (e.g., SPC). Familiarity with the operational and domain expertise and reputations of the developers' institutions was also credited as contributing to higher trustworthiness.

The increases in the trustworthiness ratings did not mean that forecasters were ready to immediately adopt the new guidance for forecasting. Forecaster 16 said they "want to be able to see more events and . . . a real-world example with actual radar" to investigate the storm mode AI model's predictability. Additionally, although the severe hail guidance matched Forecaster 2's conceptual model of the atmosphere, and they rerated the guidance as an 8 (initial rating was 2), they said, "I'll just say I wouldn't completely trust it; I know that machine learning can get spurious results. . . . I would need my own personal experience with [the guidance] to trust it further" (Table A2; supports Theme 1.2).

c. *RQ2C*: *Additional important attributes for assessing guidance trustworthiness*

Two themes were synthesized regarding additional attributes that forecasters indicated were important beyond those presented in the information board: 1) Forecasters want specific information to learn and understand their tools, and 2) they want to know what value is added by the new AI guidance.

Theme 3.1: *Forecasters are continually learning; they not only seek training and reference material to understand guidance but also require hands-on experience with guidance.* Forecasters wanting training and reference material to understand the AI guidance is resonant of Theme 1.1, where forecasters want to understand guidance to use it (Table 3). The informational slide deck provided basic information about certain attributes of the AI model, but, as underscored by the forecasters, initial exposure to and learning about new guidance are only part of the understanding process (see Theme 1.1). Forecasters want training about AI in general. Although some forecasters have had "basic training" about AI, they still feel that "something like a training exercise would probably be useful for staff" (Table 7: 3.1A). Such training exercises and reference materials also need to extend to specific guidance applications, such as the AI applications for the presented storm mode and severe hail guidance. One forecaster noted that having "a very solid description" of the new guidance from the developers would increase the trustworthiness of it (Table 7: 3.1B). An interviewed SOOs said that the content of the informational slide deck was "certainly enough for [them] to integrate this and to get staff brought in and trained up" (Table 7: 3.1B).

Training and reference material can only go so far, however, in helping forecasters understand the functionality and utility of new guidance. Illustrating this, Forecaster 14 pointed out that using new guidance once or twice does not increase trustworthiness, "it's got to be multiple cases" (Table 7: 3.1C). Some forecasters also want to learn as much as possible about new guidance and get into the "nitty gritty" to understand the scientific background and not just its operational applications. Forecaster 9 illustrated this when they said, "As a scientist, I believe very strongly in trying to understand as much as I possibly can" (Table 7: 3.1D). The desire to fully understand their tools and respective nuances was illustrated by forecasters wanting to see specific information and additional analyses performed on the new AI guidance (supports Themes 1.1 and 2.2). For example, Forecaster 8 was not sure whether they were interpreting the provided verification information correctly for the severe hail guidance, which they were. Models that perform well with predicting low probability events (e.g., severe hail AI guidance) are not common, and thus, the forecaster wanted more information beyond the slide deck to make sure they were fully understanding the AI model's performance capabilities (Table 7: 3.1D).

Theme 3.2: *Forecasters want to know how AI guidance and products will improve weather forecasting, i.e., what is the value added and what need is being met.* Given the numerous sources of information available to forecasters, they want to know what sets the AI guidance, or any new piece of guidance, apart from what they already know, use, and trust (Table 7: 3.2). When the forecasters are not centered in product development, undue burden is placed on them to determine whether and how new products *could* be useful. Answering this question of value added requires developers to work with and codevelop guidance and products *with* users such that the output is informed by users' needs in order to be actually useful (as also found by Demuth et al. 2020).

6. Discussion

In this study, we collected data from 16 NWS forecasters via a short web-based survey and online, structured interviews. The survey asked forecasters to rate how essential different guidance aspects were when they were familiarizing with or training on new guidance as well as when they were using guidance operationally. The interview included questions about a recent, challenging convective forecast scenario and the data and guidance the forecaster used during that event. The interview also introduced two new AI guidance products (storm mode probability or severe hail probability)

TABLE 7. Thematic codes and example quotes supporting Themes 3.1 and 3.2 regarding additional attributes that forecasters report being important for assessing the trustworthiness of new AI guidance (RQ2C).

Theme 3.1: Forecasters are continually learning; they not only seek training and reference material to understand guidance but also require hands-on experience with guidance
Identifier | ***Thematic code***: Example quote
3.1A | ***Forecasters want general training on AI (technique) in addition to specific applications for forecasting***: "We've done some basic training on what is AI and this sort of stuff […] but prior to doing a tool like this, basic training on AI like some of the questions you asked: What is AI? What's the difference between AI and Machine Learning? What do these terms mean? Something like a training exercise would probably be useful for staff." Forecaster 4
3.1B | ***Forecasters want a short training course on new guidance as both training and reference materials for future use (e.g., content of informational slide deck)***: "I think we would need some training on [the new guidance], as forecasters, before we really started using it. If the developers could put together a short training module just talking about it in detail, describe what we're looking at in detail and talk about it. All these things we're looking at and maybe just something that a forecaster could go to for a refresher you know. I think that would increase their trustworthiness of it too, just having a very solid description of it. I think that could be useful." Forecaster 13
"You're never going to appease everybody, but I think what you presented here is certainly enough for me to integrate this and to get staff brought in and trained up on using it." Forecaster 15
"I think the way [the guidance] was presented in those slides was good, where you say: this is where it came from, this is what the inputs were, it was trained against this. Especially if it's something where it's human assisted or human guided like that." Forecaster 5
3.1C | ***Forecasters want hands-on experience with new guidance to give a full impression; however, this does not guarantee trustworthiness***: "I have found one of the best ways to try these new products is to just make them available to forecasters, where we can look at them. One honest complaint I've had sometimes is [researchers and developers] will either hold stuff back or something, which I understand because some people [are] looking at certain things. You don't want to mislead the forecasters if it's like, 'Oh, it's a new product' and it actually turns out it doesn't work very well." Forecaster 5
"But even using [new guidance] once or twice […] doesn't increase that. It's got to be multiple cases that I've used myself to get my trustworthiness above whatever I call the 'standard' or whatever my starting point was." Forecaster 14
3.1D | ***Some forecasters want to learn as much as possible about new guidance and get into the "nitty gritty" to understand the scientific background***: "I'm really into the science side of things. And I personally really do like to know the background. I like to read scientific papers about whatever it is I'm using to make sure I'm understanding it as well as possible, knowing its strengths and limitations." Forecaster 7
"As a scientist I believe very strongly in trying to understand as much as I possibly can. So I'd be more than happy to learn as much as I can but obviously not in a warning scenario." Forecaster 9
"So, the way I interpreted it, [the severe hail AI model] did better with lower probability events, which in my mind is good, but that's not historically how model guidance works. Usually model guidance doesn't do well with low probability events. So I'm not sure if I was interpreting that correctly or not. I want more information there." Forecaster 8
Theme 3.2 Forecasters want to know how AI guidance and products will improve weather forecasting, i.e., what is the value added and need is being met
Example Quotes
"Is what we are seeing now operationally a result of AI/ML integration, or is this [new approach] something totally groundbreaking that we need to look at, that you're looking into to improve our forecasts?" Forecaster 8
"What's the need? What need is this [new guidance going] to match? What's the goal?" Forecaster 4

via an interactive informational board about different guidance attributes, with forecasters thinking aloud as they explored the board. Our analysis provides formative knowledge about how NWS forecasters make guidance use decisions (RQ1) and how information about select descriptive and performance attributes of AI guidance influences forecasters' assessment of guidance trustworthiness (RQ2). We synthesized multiple quantitative and qualitative themes from the data. From these results, three key, cross-cutting findings and implications were identified that support and/or triangulate results from previous forecaster studies.

a. NWS forecasters' development of trust in AI guidance is a deliberative, dynamic process

The first key, cross-cutting finding is that *NWS forecasters' development of trust in AI guidance is a deliberative, dynamic process. Assessments of AI guidance trustworthiness result from iterative, intentional engagements with the guidance*. Hoffman (2017) refers to this process as progressive trusting, which is seen here as forecasters requiring experience using a model over time to evaluate it as valid or true. Forecasters' process-based approach to evaluating new guidance is not specific to AI tools; rather, it occurs for most, if not all, new tools and guidance made available to them (e.g., Stuart et al. 2007; Novak et al. 2008; Evans et al. 2014; Demuth et al. 2020). This is because forecasters understand there are meteorological predictability limitations and corresponding model guidance errors, which they encounter in their forecasting roles. Because of the high-impact mission of their jobs to protect life and property and enhance the national economy, forecasters must constantly consider these potential prediction limitations and errors, which they do in part through the process of evaluating and reevaluating new guidance over time.

Through our data collection and analysis, we saw at least three phases[8] in which forecasters evaluate new guidance in different ways: 1) initial exposure and orientation to new guidance, 2) further familiarization with new guidance through nonoperational information-seeking and interrogation, and 3) operational experience through real-time observation of guidance and potentially use of it for forecasting.

In the first evaluation phase, we see the forecaster's initial introduction to new guidance and interpretation of it along with the accompanying information that was curated and "pushed" to them (e.g., our informational slide deck). Forecasters have extensive meteorological expertise and are used to having many pieces of information to look at, and thus, they have experientially developed skills to quickly evaluate a new piece of guidance to discern utility. Even when looking at guidance output in isolation, as with the spinup slide we provided, some forecasters applied their meteorological knowledge to begin evaluating whether the predictions matched their conceptual model and therefore made sense on the surface. The extent to which they could do this depended on how much initial information was made available. Forecasters have sophisticated mental models of the atmosphere for different hazards and different hazard scenarios, and they can immediately interpret new guidance—even without any a priori meteorological context—to start forming and updating their conceptual models.

The guidance attributes we asked forecasters to step through provided additional pieces of information for them to further develop their initial interpretation of the information. On balance, each of the attributes increased forecasters' perceived trustworthiness of the guidance. In particular, learning about the AI/ML model technique, training of the AI/ML model, model verification, and input information all had predominantly positive effects on forecasters' perceptions of AI guidance trustworthiness. This is largely because the attribute information jointly helped them understand the guidance better and further evaluate whether it matched their conceptual model. Additionally, being able to interact with the model output had a positive effect for all forecasters who had this option (for storm mode only) because it facilitated deeper sense-making. It did so by allowing forecasters to evaluate the evolution of the predictions at hourly forecast valid times and to interpret the model predictions coupled with radar data. Although providing this information about these attributes was useful, all forecasters' final trustworthiness ratings were below the maximum, "completely trustworthy" option offered, and some ratings remained far lower. These ratings suggest that this initial introductory phase is useful but only part of the forecasters' process of developing trust in new guidance.

The second evaluation phase includes forecasters becoming more familiar with new guidance through their own initiative by seeking more information and exploring it outside of their operational forecasting duties. Assessing how the guidance verifies is a big part of this second phase, in which forecasters aim to learn failure modes, why guidance works or fails, and how it compares to other guidance. Further understanding how the guidance functions is another big part of this phase, in which forecasters want to learn how it was derived, what the input variables are, and how it was trained and calibrated. In doing so, they want to interact with the guidance, including by sampling to see which inputs yielded a given output. Forecasters also want to examine guidance predictions for past cases to further explore verification and guidance functionality.

In the third evaluation phase, forecasters develop operational experience with guidance by observing how it functions during real-world situations and potentially by integrating the guidance as part of their forecast process. Forecasters assess new guidance in multiple ways, e.g., comparing it with observations, comparing it against similar guidance valid at the same time, and assessing how it evolves over time for different initialization and valid times. This repeated personal experience with guidance allows forecasters to more deeply understand how it functions and performs for the geographic area they forecast for and across a range of meteorological scenarios.

Based on this first key finding, *AI* (*and other*) *developers would be well served by understanding and appreciating that forecasters' assessment of AI guidance trustworthiness is a process that occurs over time and by consequently developing mechanisms to facilitate the forecaster's evaluation process.* Developers can do so by offering accessible baseline information and other training information about a new tool, particularly about the technique, inputs, training, and verification. Further, developers can facilitate forecaster interaction with guidance, particularly to evaluate the sensitivity of inputs, how the predictions evolve over space and time, and how the guidance compares or contrasts with other guidance or observations.

b. Influence of attributes on perceptions of trustworthiness varies across attributes and forecasters

The second key, cross-cutting finding is that *the specific reasons for the influence of various attributes on perceptions of trustworthiness vary across attributes and forecasters.* In other words, the determinants of trustworthiness for each of the six AI guidance attributes varied across forecasters. For example, several of the forecasters who evaluated the storm mode guidance said that the AI modeling technique increased trustworthiness because the model used human-labeled training data, while other forecasters noted more generally that *having information* about the AI modeling techniques and inputs increased trustworthiness. A coarse view of the results would be that the AI modeling techniques increase trustworthiness. However, although some simply cite the availability of that attribute information, a more granular view reveals that other forecasters cite the specific characteristics of the attribute (e.g., human-labeled training data). Similarly, learning about the AI model training increased most forecasters' perceptions of trustworthiness, but the specific reasoning varied from the general availability of the training information down to being familiar with the specific sources of training data referenced. The diversity of determinants suggests there is more nuance to what shapes forecasters' trustworthiness perceptions than the broad classification of attributes (e.g., technique, training, and developers) applied in parts of this study. Additional

[8] We note these phases may overlap and are not necessarily a linear progression, but rather are three different key parts of the process forecasters use to assess trustworthiness.

research with forecasters as expert collaborators would be helpful to further unpack these determinants (see Tables 5 and 6), as well as to discover other yet identified determinants.

c. User-centered approach to developing new AI guidance

The third key, cross-cutting finding is that *developers who make a user-centered approach to developing new AI guidance integral to their thinking and efforts stand to better meet forecasters' needs and produce AI guidance that forecasters perceive as trustworthy*. This key finding emerged both directly by forecasters expressing this desire and indirectly by forecasters favorably responding to the storm mode predictions, which filled an unmet, important operational need. Centering the user requires developers to consider the utility, value added, and operational need that can be filled from the initial, early stages of development throughout the development life cycle to the testing phase and operationalization of the guidance. Crucially, it also requires recognizing forecasters' deep expertise and developing ways to integrate that knowledge into the design and development of new guidance. Three complementary mechanisms for further centering forecasters in AI guidance development include explicitly involving forecasters in the development process as collaborative partners, further integrating systematic social science research with forecasters into the development process, and establishing "visitor" programs that allow developers to shadow forecasters in their naturalistic operational environment. These efforts yield deliberate development of products that are not just useful but needed. This notion of deliberate development was also identified in an interdisciplinary research effort to understand how convective-allowing model ensembles are used to inform forecasting decisions (Demuth et al. 2020). Further, deliberate development is also central to achieving priority weather research recommendations as identified by the NOAA Science Advisory Board (e.g., multisector and interdisciplinary collaboration for diagnostic and guidance product development; NOAA Science Advisory Board 2021). In support of forecasters being directly involved as partners and/or participants in social science research, several forecasters expressed appreciation for their inclusion in the presented work, and Forecaster 16 causally linked their inclusion to the trustworthiness of guidance by saying

> I appreciate you guys getting in touch with all these WFO forecasters and I think that's probably one of the things that's going to make stuff more trustworthy too. It's always nice when people working at national centers reach out to us. I think that makes it more trustworthy when a forecaster is using a product if they know that 'Hey other WFO people have given their input on this.'

7. Conclusion

In closing, this research produced rich and expressive data regarding how forecasters assess the trustworthiness of new AI guidance and consider incorporating it into their operational forecast processes. Integrating atmospheric science and social science research enables robust, systematic evaluation of weather forecast guidance and product development through relevant feedback from end users. Yet, this work is only a first step in further understanding the complexity of forecasters' perspectives and practices and how these intersect with the breadth and complexity of AI guidance. The analyses presented here did not reveal obvious inconsistencies between stated preferences on the surveys and those observed in the think-aloud task. Next steps could evaluate how the results of this multimethod study compare to real-world actions and behaviors of forecasters to better understand how they assess trustworthiness and use AI forecast guidance during real training sessions or active operations (i.e., the intention–behavior gap; e.g., McKnight et al. 2002; Norberg et al. 2007). Studies could also examine how forecasters interpret and use AI products when the inner workings of the model as explained by XAI do not match their physics-based conceptual model of the atmosphere (e.g., Lapuschkin et al. 2019). Another application of the intention–behavior gap could be providing a heat map of the forecast uncertainties within AI model outputs to see when and how forecasters use less certain guidance, as well as what guidance they use instead. There are multiple, high-priority research-to-operational efforts that should be undertaken, topically and methodologically (Bostrom et al. 2024). Through deeply collaborative, user-driven research like that presented in this paper, AI guidance can be developed and refined in ways that make it not only useful but actually usable and used by forecasters.

Acknowledgments. The authors sincerely thank all of the NWS forecasters for their time and their thoughtful and detailed responses, Imme Ebert-Uphoff for her collaborative conceptualization of the research approach, and Amanda Burke for the severe hail product output. This research is supported by the National Science Foundation under Grant ICER-2019758. Additional support was provided by NOAA OAR Grant NA17OAR4590114. The NSF National Center for Atmospheric Research is sponsored by the U.S. National Science Foundation. The authors have no conflicts of interest to declare.

Data availability statement. Anonymized survey data are available via DesignSafe (Cains et al. 2024c). Interview data are not available in accordance with NSF NCAR Human Subjects Committee approval.

APPENDIX

Trustworthiness Ratings and Justification Interpretations

Tables A1 and A2 detail the results of the reflexive thematic analysis of the reasonings/justifications stated by forecasters for their trustworthiness ratings.

TABLE A1. Reflexive thematic analysis of reasoning/justification stated by forecasters for their initial trustworthiness rating. Rating was made before the think-aloud assessment, based on initial background information.

Initial trustworthiness rating and justification interpretation	Exemplar quote
Low: 0–3 Do not know enough about AI model No experience with AI guidance No model verification	"So I'll give it, as far as I can see, maybe a 2 at this point. I haven't seen any numerical plots. Of course, you get the [map] legend there, but we're not sure what the legend actually means with the gray lines. I haven't compared it to any other model data, to observational data. There's a whole slew of things that I'd want to look at before I give this model any value." Forecaster 8
Medium: 4–6 Made sense climatologically AI model matches conceptual model Output looks realistic	"I don't know what went into the model necessarily, what the synoptic conditions are, or haven't had any experience with it up until this point either. So it'd be great guidance, but I would want to verify it myself that it made sense with my conceptual model of what's going on." Forecaster 2
High: 7–10 Meteorologically valid AI models used well-known resources Operational utility of storm mode guidance	"I would give it a solid 8. It meshes well with my conceptual model of what the storm modes would be with this system. That's not to say if it didn't mesh it would be wrong. But at first blush it seems to have a pretty good handle on how things are evolving with its forecast. That makes it seem pretty trustworthy to me." Forecaster 10

TABLE A2. Reflexive thematic analysis of reasoning/justification stated by forecasters for their final trustworthiness rating. Ratings were given after the think-aloud assessment of guidance attributes.

Final trustworthiness rating and justification interpretation		Exemplar quote
Low: 0–3	—	—
Medium: 4–6 and High: 7–10	Additional background information Familiarity with input data sources (e.g., WRF and HREF) Comparison to conventional resources (e.g., SPC) Developers' operational and domain expertise Reputations of the developers' institutions	"After seeing the detail, I think [the trustworthiness rating] is probably more like a 6. […] I want to be able to see more events, and then I'd want to be able to see a real-world example with actual radar to be able to tell if it looks like what [the AI model] says it was going to like." Forecaster 16 "I would go 8 now. I'll just say I wouldn't completely trust it; I know that machine learning can get spurious results. But since I've only looked at basically two cases here, I don't have that long background experience information to bump it up higher. So knowing what's behind it and some of the verification. . . . Definitely somewhere between 7 and 8. But then I would need my own personal experience with it to trust it further." Forecaster 2

REFERENCES

AI2ES, 2022a: Explainability. Glossary, https://www.ai2es.org/products/education/glossary/explainability/.

——, 2022b: Interpretability. Glossary, https://www.ai2es.org/products/education/glossary/interpretability/.

Bostrom, A., R. E. Morss, J. K. Lazo, J. L. Demuth, H. Lazrus, and R. Hudson, 2016: A mental models study of hurricane forecast and warning production, communication, and decision-making. *Wea. Climate Soc.*, **8**, 111–129, https://doi.org/10.1175/WCAS-D-15-0033.1.

——, and Coauthors, 2024: Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Anal.*, **44**, 1498–1513, https://doi.org/10.1111/risa.14245.

Braun, V., and V. Clarke, 2006: Using thematic analysis in psychology. *Qual. Res. Psychol.*, **3**, 77–101, https://doi.org/10.1191/1478088706qp063oa.

——, and ——, 2019: Reflecting on reflexive thematic analysis. *Qual. Res. Sport Exercise Health*, **11**, 589–597, https://doi.org/10.1080/2159676X.2019.1628806.

——, and ——, 2021a: One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qual. Res. Psychol.*, **18**, 328–352, https://doi.org/10.1080/14780887.2020.1769238.

——, and ——, 2021b: Can I use TA? Should I use TA? Should I *not* use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling Psychother. Res.*, **21**, 37–47, https://doi.org/10.1002/capr.12360.

Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning–based probabilistic hail predictions for operational forecasting. *Wea. Forecasting*, **35**, 149–168, https://doi.org/10.1175/WAF-D-19-0105.1.

Cains, M., and Coauthors, 2024a: Interviews with NWS Forecasters related to severe weather and new artificial intelligence/machine learning (AI/ML) guidance predicting severe hail and storm mode: Interview materials for "AI/ML" version. Accessed 20 March 2024, https://doi.org/10.17603/DS2-8MGD-2J44.

——, and Coauthors, 2024b: Interviews with NWS Forecasters related to severe weather and new artificial intelligence/

machine learning (AI/ML) guidance predicting severe hail and storm mode: Interview materials for "No AI/ML" version. Accessed 20 March 2024, https://doi.org/10.17603/DS2-NFGD-3G25.

——, C. Wirz, J. Demuth, and A. Bostrom, 2024c: Interviews with NWS Forecasters related to severe weather and new artificial intelligence/machine learning (AI/ML) guidance predicting severe hail and storm mode: Pre-interview survey data. Accessed 20 March 2024, https://doi.org/10.17603/DS2-11Y2-BG84.

Cavanagh, S., 1997: Content analysis: Concepts, methods and applications. *Nurse Res.*, **4**, 5–16, https://doi.org/10.7748/nr.4.3.5.s2.

Charters, E., 2003: The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Educ. J.*, **12**, https://doi.org/10.26522/brocked.v12i2.38.

Chase, R. J., D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern, 2022: A machine learning tutorial for operational meteorology. Part I: Traditional machine learning. *Wea. Forecasting*, **37**, 1509–1529, https://doi.org/10.1175/WAF-D-22-0070.1.

Daipha, P., 2015: From bricolage to collage: The making of decisions at a weather forecast office. *Sociol. Forum*, **30**, 787–808, https://doi.org/10.1111/socf.12192.

Demuth, J., A. Bostrom, C. Wirz, and M. Cains, 2024: Interviews with NWS Forecasters related to severe weather and new artificial intelligence/machine learning (AI/ML) guidance predicting severe hail and storm mode: Pre-interview survey. Accessed 20 March 2024, https://doi.org/10.17603/DS2-MR3Z-7947.

Demuth, J. L., and Coauthors, 2020: Recommendations for developing useful and usable convection-allowing model ensemble information for NWS forecasters. *Wea. Forecasting*, **35**, 1381–1406, https://doi.org/10.1175/WAF-D-19-0108.1.

Doswell, C. A., III, 2004: Weather forecasting by humans—Heuristics and decision making. *Wea. Forecasting*, **19**, 1115–1126, https://doi.org/10.1175/WAF-821.1.

Ericsson, K. A., and H. A. Simon, 1998: How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind Cult. Act.*, **5**, 178–186, https://doi.org/10.1207/s15327884mca0503_3.

——, and M. C. Fox, 2011: Thinking aloud is not a form of introspection but a qualitatively different methodology: Reply to Schooler (2011). *Psychol. Bull.*, **137**, 351–354, https://doi.org/10.1037/a0022388.

European Commission, 2019: Ethics guidelines for trustworthy AI: High-level expert group on artificial intelligence. EC Rep., 39 pp., https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Evans, C., D. F. Van Dyke, and T. Lericos, 2014: How do forecasters utilize output from a convection-permitting ensemble forecast system? Case study of a high-impact precipitation event. *Wea. Forecasting*, **29**, 466–486, https://doi.org/10.1175/WAF-D-13-00064.1.

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate storm-scale probabilistic guidance of severe weather hazards in the warn-on-forecast system. *Mon. Wea. Rev.*, **149**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

Gagne, D. J., II, A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

——, S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev.*, **147**, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1.

Glikson, E., and A. W. Woolley, 2020: Human trust in artificial intelligence: Review of empirical research. *Acad. Manage. Ann.*, **14**, 627–660, https://doi.org/10.5465/annals.2018.0057.

Heinrichs, B., and S. B. Eickhoff, 2020: Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Hum. Brain Mapp.*, **41**, 1435–1444, https://doi.org/10.1002/hbm.24886.

Henderson, J., J. Spinney, and J. L. Demuth, 2023: Conceptualizing confidence: A multisited qualitative analysis in a severe weather context. *Bull. Amer. Meteor. Soc.*, **104**, E459–E479, https://doi.org/10.1175/BAMS-D-22-0137.1.

Hill, A. J., R. S. Schumacher, and I. L. Jirak, 2023: A new paradigm for medium-range severe weather forecasts: Probabilistic random forest–based predictions. *Wea. Forecasting*, **38**, 251–272, https://doi.org/10.1175/WAF-D-22-0143.1.

Hoff, K. A., and M. Bashir, 2015: Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum. Factors*, **57**, 407–434, https://doi.org/10.1177/0018720814547570.

Hoffman, R. R., 2017: A taxonomy of emergent trusting in the human–machine relationship. *Cognitive Systems Engineering: The Future for a Changing World*, P. J. Smith and R. R. Hoffman, Eds., Expertise: Research and Applications Series, CRC Press, 137–163.

——, J. W. Coffey, K. M. Ford, and J. D. Novak, 2006: A method for eliciting, preserving, and sharing the knowledge of forecasters. *Wea. Forecasting*, **21**, 416–428, https://doi.org/10.1175/WAF927.1.

——, D. S. LaDue, H. M. Mogil, and P. J. Roebber, 2017: *Minding the Weather: How Expert Forecasters Think*. The MIT Press, 488 pp.

——, G. Klein, and S. T. Mueller, 2018a: Explaining explanation for "explainable AI." *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, **62**, 197–201, https://doi.org/10.1177/1541931218621047.

——, S. T. Mueller, G. Klein, and J. Litman, 2018b: Measuring trust in the XAI context. DARPA Explainable AI Program Tech. Rep., 26 pp.

——, ——, ——, and ——, 2018c: Metrics for explainable AI: Challenges and prospects. arXiv, 1812.04608v2, https://doi.org/10.48550/arXiv.1812.04608.

Hsieh, H.-F., and S. E. Shannon, 2005: Three approaches to qualitative content analysis. *Qual. Health Res.*, **15**, 1277–1288, https://doi.org/10.1177/1049732305276687.

Jacovi, A., A. Marasović, T. Miller, and Y. Goldberg, 2020: Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. arXiv, 2010.07487v3, https://doi.org/10.48550/arXiv.2010.07487.

Kaplan, A. D., T. T. Kessler, J. C. Brill, and P. A. Hancock, 2021: Trust in artificial intelligence: Meta-analytic findings. *Hum. Factors*, **65**, 337–359, https://doi.org/10.1177/00187208211013988.

Lagerquist, R., A. McGovern, C. R. Homeyer, D. J. Gagne II, and T. Smith, 2020: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Mon. Wea. Rev.*, **148**, 2837–2861, https://doi.org/10.1175/MWR-D-19-0372.1.

Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K. R. Müller, 2019: Unmasking Clever Hans predictors

and assessing what machines really learn. *Nat. Comm.*, **10**, 1096, https://doi.org/10.1038/s41467-019-08987-4.

McGovern, A., R. Lagerquist, D. J. Gagne II, G. Eli Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

——, and Coauthors, 2022: NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). *Bull. Amer. Meteor. Soc.*, **103**, E1658–E1668, https://doi.org/10.1175/BAMS-D-21-0020.1.

McKnight, D. H., V. Choudhury, and C. Kacmar, 2002: Developing and validating trust measures for e-Commerce: An integrative typology. *Inf. Syst. Res.*, **13**, 334–359, https://doi.org/10.1287/isre.13.3.334.81.

Merriam, S. B., and E. J. Tisdell, 2015: *Qualitative Research: A Guide to Design and Implementation*. John Wiley and Sons, 368 pp.

Molnar, C., 2023: *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. Christoph Molnar, 320 pp.

Morss, R. E., J. L. Demuth, A. Bostrom, J. K. Lazo, and H. Lazrus, 2015: Flash flood risks and warning decisions: A mental models study of forecasters, public officials, and media broadcasters in Boulder, Colorado. *Risk Anal.*, **35**, 2009–2028, https://doi.org/10.1111/risa.12403.

Mueller, S. T., R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein, 2019: Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv, 1902.01876v1, https://doi.org/10.48550/arXiv.1902.01876.

National Academies of Sciences, Engineering, and Medicine, 2022: *Human-AI Teaming: State-of-the-Art and Research Needs*. National Academies Press, 140 pp.

National Science and Technology Council, 2023: The National Artificial Intelligence Research and Development Strategic Plan: 2023 update. 56 pp., https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf.

NOAA Science Advisory Board, 2021: A report on priorities for weather research. NOAA Science Advisory Board Rep., 119 pp., https://sab.noaa.gov/wp-content/uploads/2021/12/PWR-Report_Final_12-9-21.pdf.

Norberg, P. A., D. R. Horne, and D. A. Horne, 2007: The privacy paradox: Personal information disclosure intentions versus behaviors. *J. Consum. Aff.*, **41**, 100–126, https://doi.org/10.1111/j.1745-6606.2006.00070.x.

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.

Roebber, P. J., 2022: A review of artificial intelligence and machine learning activity across the United States National Weather Service. NOAA Tech. Memo. NWS MDL 86, 25 pp., https://vlab.noaa.gov/documents/6609493/8249989/TechMemo86.pdf.

——, and S. Smith, 2023: Prospects for machine learning activity within the United States National Weather Service. *Bull. Amer. Meteor. Soc.*, **104**, E1333–E1344, https://doi.org/10.1175/BAMS-D-22-0181.1.

Saßmannshausen, T., P. Burggräf, J. Wagner, M. Hassenzahl, T. Heupel, and F. Steinberg, 2021: Trust in artificial intelligence within production management – An exploration of antecedents. *Ergonomics*, **64**, 1333–1350, https://doi.org/10.1080/00140139.2021.1909755.

Schaefer, K. E., J. Y. C. Chen, J. L. Szalma, and P. A. Hancock, 2016: A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Hum. Factors*, **58**, 377–400, https://doi.org/10.1177/0018720816634228.

Schulte-Mecklenbeck, M., A. Kühberger, and R. Ranyard, 2011: The role of process data in the development and testing of process models of judgment and decision making. *Judgment Decis. Making*, **6**, 733–739, https://doi.org/10.1017/S1930297500004162.

Scott, C. R., 2005: Anonymity in applied communication research: Tensions between IRBs, researchers, and human subjects. *J. Appl. Commun. Res.*, **33**, 242–257, https://doi.org/10.1080/00909880500149445.

Shattuck, L. G., and D. D. Woods, 1994: The critical incident technique: 40 years later. *Proc. Hum. Fact. Ergon. Soc. Annu. Meet.*, **38**, 1080–1084, https://doi.org/10.1177/154193129403801702.

Sobash, R. A., D. J. Gagne II, C. L. Becker, D. Ahijevych, G. N. Gantos, and C. S. Schwartz, 2023: Diagnosing storm mode with deep learning in convection-allowing models. *Mon. Wea. Rev.*, **151**, 2009–2027, https://doi.org/10.1175/MWR-D-22-0342.1.

Stanton, B., and T. Jensen, 2021: Trust and artificial intelligence. Draft NISTIR 8332, 30 pp., https://doi.org/10.6028/NIST.IR.8332-draft.

Stewart, T. R., P. J. Roebber, and L. F. Bosart, 1997: The importance of the task in analyzing expert judgment. *Organ. Behav. Hum. Decis. Processes*, **69**, 205–219, https://doi.org/10.1006/obhd.1997.2682.

Stuart, N. A., and Coauthors, 2006: The future of humans in an increasingly automated forecast process. *Bull. Amer. Meteor. Soc.*, **87**, 1497–1502, https://doi.org/10.1175/BAMS-87-11-1497.

——, D. M. Schultz, and G. Klein, 2007: Maintaining the role of humans in the forecast process: Analyzing the psyche of expert forecasters. *Bull. Amer. Meteor. Soc.*, **88**, 1893–1898, https://doi.org/10.1175/BAMS-88-12-1893.

Van Maanen, J., 1979: Reclaiming qualitative methods for organizational research: A preface. *Administrative Sci. Quart.*, **24**, 520–526, https://doi.org/10.2307/2392358.

Varshney, K. R., 2021: *Trustworthy Machine Learning*. Independently Published, 267 pp.