Exploring the Usefulness of Machine Learning Severe Weather Guidance in the Warn-on-Forecast System: Results from the 2022 NOAA Hazardous Weather Testbed Spring Forecasting Experiment

MONTGOMERY L. FLORA,[a,b] BURKELY GALLO,[a,c] COREY K. POTVIN,[b,d] ADAM J. CLARK,[b,d] AND KATIE WILSON[a,b]

[a] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[c] *Storm Prediction Center, Norman, Oklahoma*
[d] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 22 February 2024, in final form 15 April 2024, accepted 7 May 2024)

ABSTRACT: Artificial intelligence (AI) is gaining popularity for severe weather forecasting. Recently, the authors developed an AI system using machine learning (ML) to produce probabilistic guidance for severe weather hazards, including tornadoes, large hail, and severe winds, using the National Severe Storms Laboratory's (NSSL) Warn-on-Forecast System (WoFS) as input. Known as WoFS-ML-Severe, it performed well in retrospective cases, but its operational usefulness had yet to be determined. To examine the potential usefulness of the ML guidance, we conducted a control and treatment (experimental) group experiment during the 2022 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (HWT-SFE). The control group had full access to WoFS, while the experimental group had access to WoFS and ML products. Explainability graphics were also integrated into the WoFS web viewer. Both groups issued 1-h convective outlooks for each hazard. After issuing their forecasts, we surveyed participants on their confidence, the number of products viewed, and the usefulness of the ML guidance. We found the ML-based outlooks outperformed non-ML-based outlooks for multiple verification metrics for all three hazards and were rated subjectively higher by the participants. However, the difference in confidence between the two groups was not significant, and the experimental group self-reported viewing more products than the control group. Participants had mixed sentiments toward explainability products as it improved their understanding of the input/output relationships, but viewing them added to their workload. Although the experiment demonstrated the usefulness of ML guidance for severe weather forecasting, there are avenues to improve upon the ML guidance, and more training and exposure are needed to exploit its benefits fully.

SIGNIFICANCE STATEMENT: We developed an artificial intelligence (AI) system to predict tornadoes, large hail, and damaging straight-line winds. The AI system was leveraged in real time during the 2022 NOAA Hazardous Weather Testbed Spring Forecasting Experiment. This study reveals that forecasters using AI guidance produced more reliable and spatially accurate outlooks than those without. While AI and complementary explainability products did not reduce forecaster workload, both demonstrated great potential for improving severe weather forecasting. This research also highlights the importance of user feedback in refining AI tools for severe weather forecasting.

KEYWORDS: Social Science; Severe storms; Forecast verification/skill; Short-range prediction; Machine learning; Model interpretation and visualization

1. Introduction

Improved predictions of severe thunderstorm hazards can reduce their impacts. A promising approach for improved severe weather predictions is leveraging machine learning (ML) methods, which analyze complex patterns from large datasets. The use of ML in severe weather forecasting has grown in the last few years (Lagerquist et al. 2017; Gagne et al. 2017, 2019; Hill et al. 2020; Loken et al. 2020; Sobash et al. 2020; Lagerquist et al. 2020; Cintineo et al. 2020; Flora et al. 2021; Hill et al. 2023; McGovern et al. 2023). For example, Storm Prediction Center (SPC) forecasters leverage ML-generated probabilistic guidance for the next day (24–36 h) time frame (Loken et al. 2020) and extended ranges (days 3–8; Hill et al. 2020, 2023) while National Weather Service (NWS) forecasters leverage ML-generated guidance for 0–1-h lead times ("ProbSevere"; Cintineo et al. 2020). The NWS is eager to adopt AI/ML guidance as it may accelerate the paradigm shift from forecasters generating forecast output to providing more weather decision support services (Roebber and Smith 2023). For example, translating ML guidance into usable information supporting impact-based decision support services (IDSS) is a top NWS goal (Roebber and Smith 2023).

A promising avenue for new forecast guidance development is user-centered design, where the final product matches the end user's need because it is developed iteratively in collaboration with the end user (Abras et al. 2004;

Gallo's current affiliation: U.S. Air Force, 557th Weather Wing, Offutt Air Force Base, Nebraska.
Wilson's current affiliation: Rand Corporation, Santa Monica, California.

Corresponding author: Montgomery Flora, monte.flora@noaa.gov

DOI: 10.1175/WAF-D-24-0038.1

Argyle et al. 2017). For example, leveraging forecaster feedback on the experimental Warn-on-Forecast System (WoFS) guidance (Wilson et al. 2019b) inspired the development of the event-based ML system discussed in this paper (Flora et al. 2019, 2021).

The authors have recently developed a novel ML system to generate probabilistic predictions for tornadoes, severe straight-line wind, and large hail using output from the experimental WoFS (Flora et al. 2021). Known as WoFS-ML-Severe, it was shown to outperform a baseline system that uses surrogate severe methods—which use a single convective-allowing model variable proxy for severe weather—to predict each hazard. The guidance helps to fill the watch-to-warning gap where numerical guidance is limited, and its storm-based design can support IDSS. To explore the operational utility of this novel ML guidance, a forecasting activity was developed in the annual NOAA Hazardous Weather Testbed Spring Forecasting Experiment (HWT-SFE; Clark et al. 2023; Gallo et al. 2017, 2022). The HWT-SFE brings together individuals from the operational and research sectors to assess and evaluate experimental forecast guidance.

In recent work exploring the impact of WoFS on issuing probabilistic forecasts, Gallo et al. (2024) used an experiment design with control and treatment (experimental) groups where both groups had access to non-WoFS data while the experimental group also had access to WoFS data. This framework can help determine new guidance's unique value and role in a data-rich environment. This study adopts a similar design where the control and experimental groups had access to the entire suite of WoFS products while the experimental group also had access to the experimental ML guidance. This framework also allows us to implicitly explore whether adding the ML guidance impacts forecaster overload. We hypothesize that alongside the full WoFS suite, the ML guidance offers concise information on severe weather hazard potential, with the WoFS suite allowing for deeper interrogation and situational awareness of ongoing severe weather threats.

For this study, participants were tasked with issuing 1-h probabilistic forecast outlooks valid for 2100–2200 UTC and 2200–2300 UTC for tornadoes, wind $\geq$ 50 kt (1 kt $\approx$ 0.51 m s^{-1}), and hail $\geq$ 1 in. Since the relationship between input and output is often unknown to ML users (McGovern et al. 2019), we also introduced a novel, interactive ML explainability[1] feature on the WoFS web viewer (see section 2b). Introducing simple explainability products is one step toward developing trustworthy AI, which is becoming a growing focus in environmental sciences (McGovern et al. 2022). We explored the following research questions:

1) Are the probabilistic outlooks generated by the experimental group (those with access to the WoFS-ML-Severe) more accurate than the control group?
2) Does WoFS-ML-Severe provide useful guidance beyond that already available on the WoFS web viewer?
3) Did the ML guidance affect participant confidence or significantly impact participants' forecast process (e.g., did it reduce forecaster workload)?
4) Were the explainability products useful, and did they improve confidence in the ML products?

To evaluate question 1, the 1-h forecast outlooks generated by participants were objectively and subjectively evaluated. For objective verification, probabilities were evaluated against storm reports. For subjective verification, HWT-SFE participants ranked the performance of the previous day's forecast on a scale of 1–5 using all available verification data (reports, radar-estimated hail size, and warning polygons). Regarding questions 2 and 3, automated guidance can reduce workloads (Karstens et al. 2015). We hypothesized that the ML guidance coupled with explainability products would improve forecast confidence and limit the number of WoFS products viewed by the experimental group compared to the control group. To test this, we surveyed participants about their confidence in their outlooks, the number of products they viewed, and their opinion on the usefulness of the ML guidance. The next day, participants who subjectively rated the outlooks were also asked questions about their confidence in using WoFS and the ML guidance in the future. Last, for question 4, we hypothesized that the explainability products would improve forecasters' confidence in the ML products.

2. Methods

a. *Description of the Warn-on-Forecast System data*

The WoFS is an experimental convection-allowing ensemble analysis and forecast system that rapidly updates real-time severe weather guidance by frequently assimilating ongoing convection. It comprises 36 analysis members and 18 forecast members with a 3-km grid spacing, using the Advanced Research version of the Weather Research and Forecasting (WRF-ARW; Skamarock et al. 2008) Model. It uses a limited-area domain whose location is updated daily according to the area of interest, usually one with some risk of severe weather. WoFS issues forecasts every 30 mins with 5-min output for up to 6 h. Additional details on the WoFS configuration have been covered in previous articles (Jones et al. 2016, 2020) and are omitted here for brevity.

For the 2022 HWT-SFE, the WoFS was run daily from 2 May to 3 June 2022 (excluding weekends). The approximate location, maximum SPC risk, and the number of Storm Data reports during each date's total forecast period (2100–2300 UTC) are provided in Fig. 1. Multiple regions of the CONUS were sampled. The cases included days with multiple tornadoes (6, 11, 17, and 30 May), a mesoconvective vortex (11 May), a derecho event (12 May), and a series of cases that primarily produced damaging wind (3, 16, and 24 May). Damaging winds were the most frequent hazard (815 reports), while tornadoes were the least frequent (70 reports).

b. *WoFS-ML-Severe method*

A full description of the WoFS-ML-Severe products is provided in Flora et al. (2021), so we only give a brief account

[1] Explainability is the degree to which a person can understand an ML model using post hoc methods; see https://www.ai2es.org/products/education/glossary/explainability/.

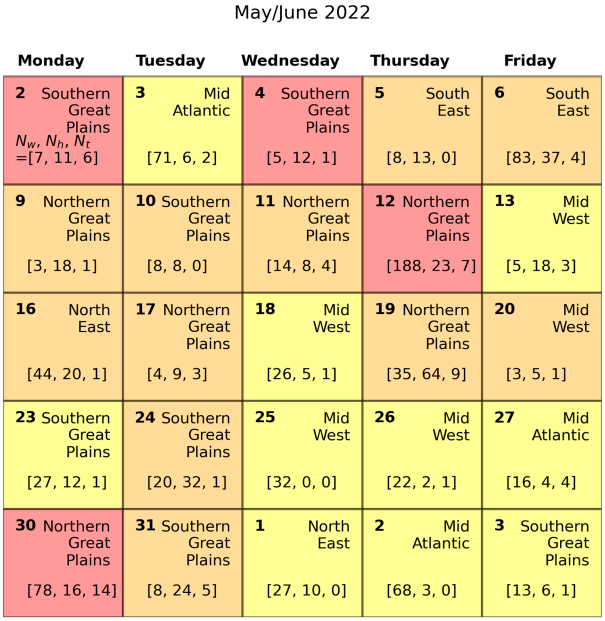

FIG. 1. Calendar covering the 2022 HWT-SFE period. The dates are color coded by the maximum categorical SPC risk issued within the WoFS domain at 1630 UTC (red = moderate, orange = enhanced, and yellow = slight). For each date, we provide the number of wind reports N_w, hail reports N_h, and tornado reports N_t within the WoFS domain during the 2100–2300 UTC period. The general location of the domain center is provided in the upper-right-hand corner for each date. A final count of each hazard is provided at the bottom of the figure.

here. A novel aspect of the guidance is that rather than producing spatial probabilities (predicting the likelihood that a given location or neighborhood will experience severe weather), the WoFS-ML-Severe guidance uses an object-based design that issues event probabilities (forecasting the likelihood that a given thunderstorm will produce severe weather; Flora et al. 2019). For the event probability framework, ensemble forecasts of future storm locations are aggregated over a 30-min period to produce "ensemble storm tracks" (an example is shown in Fig. 2a). The ML dataset is derived from these storm tracks by extracting ensemble statistics from intrastorm and environmental variables (see Table 1 from Flora et al. 2021) from points within the tracks. Additional variables characterizing the track morphology (e.g., area or major axis length) are also extracted. Environmental features are computed as spatial averages of the ensemble mean and standard deviation fields valid at the beginning of the storm track's forecast period (to mitigate sampling storm-modified environment fields). Intrastorm features consist of the spatial averages of the ensemble mean and standard deviation fields and the ensemble mean and standard deviation of the 90th percentile of each ensemble member (to capture storm intensity). The target data were based on whether an ensemble storm track contained a tornado, severe hail, or severe wind report, respectively.

In Flora et al. (2021), three machine learning models per hazard were trained: random forest, logistic regression, and gradient-boosted trees [extreme gradient boosting (XGBoost); Chen and Guestrin 2016]. For the HWT-SFE, we only used logistic regression and random forest models due to technical limitations with the XGBoost models. As described in Flora et al. (2021), different models were trained for different lead times. The original lead times were separated into two groups: lead times ending at 60–90 min (first hour) and lead times between 90 and 150 min (second hour). However, for the HWT-SFE, ML guidance was produced for up to 4 h of lead time. Due to time constraints, we could not train additional models for these later lead times and instead used the second hour models for all times greater than 90 min. It is unknown if introducing additional models for these lead times would have produced more skillful guidance, but we defer this to future work. However, for the lead times that participants were issuing outlooks—<2–3 h—they were not heavily relying on forecasts > 2 h. To summarize the ML guidance, additional time-composite products for 1- and 4-h time spans are produced on the web viewer (Figs. 2c,d).

Since ML models' input/output relationship is often complex or unknown, we introduced a novel interactive explainability feature on the WoFS web viewer (see Fig. 2b). Participants could click on a particular storm track, and a product like that shown in Fig. 2b would appear. For a given hazard, the product shows five predictors and where the value of a given predictor lies within the training distribution of storms previously matched to a report. Adding this context allows a user to quickly determine how the input compares with previous severe storms. For consistency, the same set of features for a given hazard was shown for each track (and for both the logistic regression and the random forest). The methods developed in Flora et al. (2024) determined these top five predictors by combining multiple feature ranking methods and manual selection based on subject-matter expertise. However, we did ask participants whether they preferred a consistent set of predictors (global explainability), predictors that were specific to a given track (local explainability), or some combination of both (section 3e).

c. *SFE experiment design*

1) PARTICIPANTS

Every spring, the NOAA HWT-SFE gathers professionals from operational and research sectors to evaluate experimental forecast guidance (Gallo et al. 2017, 2022; Clark et al. 2023). Participants include NWS forecasters and individuals from academia and the public and private sectors. During the winter before HWT-SFE, participant solicitations are sent to NWS Weather Forecast Offices (WFOs). Forecasters apply to participate through these solicitations and are selected by their respective regional Science Services Division (SSD) chiefs. Information provided by forecasters includes name, position, region, WFO, week of participation preference, and an interest statement. Participation is considered a part of their regular duties as federal employees. For the 2022 HWT-SFE, the NWS forecasters came from 25 WFOs. The breakdown by NWS region is as follows: western (4), central (11),

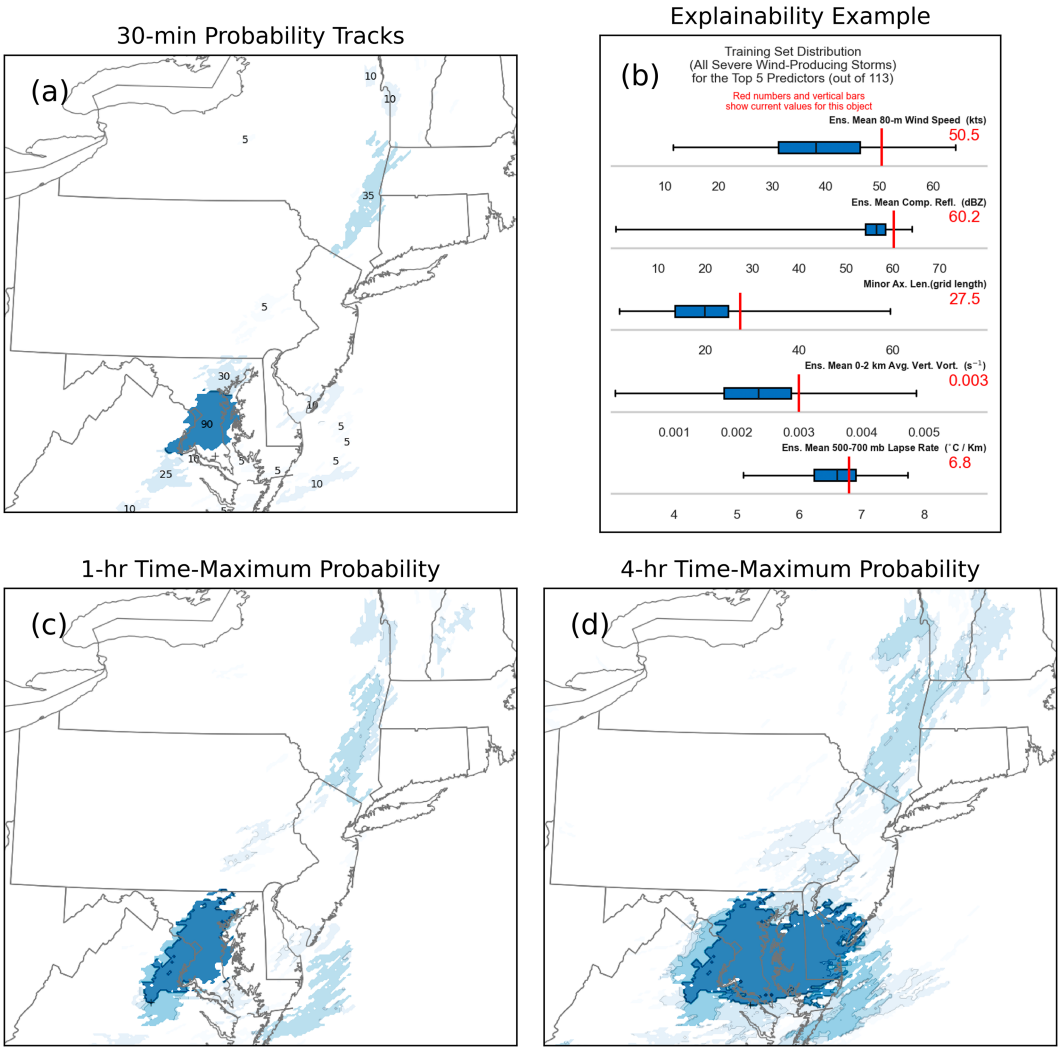

FIG. 2. Examples of the available ML products on the real-time WoFS web viewer. These include (a) 30-min guidance, (b) the interactive ML explainability product, (c) 1-h guidance, and (d) the 4-h guidance. For the 30-min guidance in (a), probabilities rounded to the nearest 5% are overlaid. The 1- and 4-h products are time-maximum composites of the 30-min guidance. These graphics come from the cloud-based WoFS web page (https://cbwofs.nssl.noaa.gov/).

southern (5), and eastern (10). The participants' experience levels generally range from novice to decades of experience. We did not explicitly collect the participants' experience level information. The HWT-SFE development team similarly solicits non-NWS participants, and their respective funding sources cover their compensation. Forecasters and other participants eager to adopt cutting-edge guidance will likely apply to the SFE. This does limit the generalizability of the results to all NWS forecasters. Most participants only attend the HWT-SFE for a single week (Monday–Friday). In total, the 2022 HWT-SFE had 260 participants. The HWT-SFE, including this study, was approved under expedited review by the University of Oklahoma's Office of Human Research Participant Protection Institutional Review Board under project 13320.

2) PROCEDURE

For the HWT-SFE activity, we performed a two-group experiment in which a control group had access to the standard WoFS product suite only, while the treatment (experimental) group had access to the WoFS suite plus the ML guidance. Each group was assigned two NWS forecasters (known as "expert" forecasters) and 4–6 non-NWS participants (known as "nonexpert" forecasters). Though nonexperts are not NWS forecasters, many still have extensive knowledge of forecasting and severe weather (e.g., researchers and model developers) and could still be considered experts in the field. Participants were randomly assigned to a group and could not opt into the experimental group. Throughout the week, participants experienced both groups/conditions. For example, a

participant may be in the control group Monday–Wednesday and the experimental group Thursday and Friday. Though a participant spent multiple days in the experimental group, the period was short enough that the learning bias was minimal. Due to COVID-19 restrictions, participants attended the 2022 HWT-SFE virtually. Participants in the control and experimental groups were in separate Google Meet sessions, and there was no communication between the two groups. The facilitators guided participants in the control and experimental Google Meet sessions. The control and experimental groups accessed the same web page; only the honor system was in place for control group participants not to use the ML products. Due to time constraints, we could not implement a method for restricting specific users from accessing the ML guidance. Facilitators frequently reminded participants to avoid the ML guidance if they were in the control group. Participants from both groups could also access any other online weather information (e.g., SPC mesoanalysis) at their discretion.

During the SFE, we rely on real-time weather conditions, distinguishing it from other experiments where one preselects past cases. Using real-time weather events eliminates potential bias from forecasters being familiar with historical events. As for forecaster experience or familiarity with regional climatology, we note that the WoFS domain shifts daily. For example, domains were centered over at least three geographic regions per week (Fig. 1). Though the control and experiment groups were determined randomly, which should mitigate the impact of experience and familiarity across both groups, we do explicitly control for regional familiarity in our experiment design.

Both groups generated hourly probabilistic outlooks for tornadoes, wind, and hail, valid from 2100–2200 UTC and 2200–2300 UTC. Participants had access to the standard SPC contours used for full-day convective outlook products,[2] with other intermediate contours also available. The participants drew initial outlooks between 1915 and 2015 UTC and then generated final, updated outlooks between 2015 and 2115 UTC based on updated WoFS output. Though the experiment is SPC-centric, Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) goals are such that there is a seamless flow of probabilistic information across scales. This forecasting experiment is in the watch-to-warning space and tests one potential iteration or realization of that FACETs-style guidance. WFO and SPC forecasters are not producing 1-h probabilistic outlooks, and this exploratory activity could benefit both types of forecasters.

Given that the experimental group varied daily, facilitators presented an overview of WoFS and the ML products at the beginning of each experimental group session. Facilitators also emphasized to both groups that the 1-h time window outlooks should not be treated as the SPC convective outlooks, valid over much longer periods. Given the short lead times, the forecasts should, in theory, be more accurate and precise, meaning highlighted areas should have higher probabilities and cover smaller areas than SPC's day 1 convective outlooks. An example set of forecasts for 1 June 2022 is shown in Fig. 3. All forecast outlooks generated during the 2022 HWT-SFE are available at https://hwt.nssl.noaa.gov/sfe/2022/.

After completing their outlooks, participants were asked to complete a survey (see Tables 1 and 2). Many questions were answered using a Likert scale or forced choice, but some were open ended. A Likert scale measures respondents' attitudes in questionnaires using various response options from one extreme to another (Likert 1932). These questions asked participants about their confidence in their forecast, how useful they found the ML guidance, and how useful the explainability products were. The experimental group was asked additional questions about the value of the ML guidance and the interactive explainability products.

To subjectively evaluate the previous day's forecasts, participants used observed storm reports, radar-based maximum estimated hail size (MESH; Witt et al. 1998), NWS warnings, and practically perfect forecasts (Hitchens et al. 2013; example shown in Fig. 3). The evaluation participants were often a combination of those who did and did not help generate the outlooks. Though this includes self-evaluation samples, the total sample size is primarily dominated by participants evaluating outlooks generated by other participants. This is not a typical social science experiment design, but these subjective evaluations provide useful information, especially when evaluating model guidance (Kain et al. 2003; Gallo et al. 2024). Participants evaluated the outlooks generated by the NWS forecasters (two outlooks) and a consensus outlook from the non-NWS participants from both the control and experimental groups. The consensus outlooks were created by converting the non-NWS participants' outlooks to continuous spatial probabilities using a method developed at SPC (Karstens et al. 2019) and then averaging them together. If nonexpert forecasters drew a significant severe contour, indicating a greater than 10% probability for significant severe weather, the consensus significant contour was drawn around points where at least half of the participants drew a significant severe contour. We note that consensus outlooks smooth the intensity of contours of the individual outlooks and tend to have less extreme values than the expert forecasters' outlooks.

The primary objective of this exercise was to quantify ML's value in the experimental outlooks by comparing those made with and without ML guidance. Participants categorized each outlook as "poor," "below average," "average," "above average," or "excellent." The previous categories were converted to a 1–5 rating scale to quantify the subjective forecast evaluations, where five corresponds to excellent. Practically perfect forecasts were tuned with a smaller standard deviation than is used for verifying SPC outlooks to increase amplitudes over smaller areas, and participants were reminded of this for each case. Then, the average ratings were computed for each hazard and forecast type (expert and nonexpert consensus) for both the control and experimental groups (four ratings per hazard).

d. Objective verification methods

To evaluate the performance of the 1-h forecast outlooks, we use the receiver operating characteristic (ROC; Metz 1978)

[2] Standard contour levels for hail and wind are 5%, 15%, 30%, 45%, and 60% and for tornado includes an additional 2% and 10%.

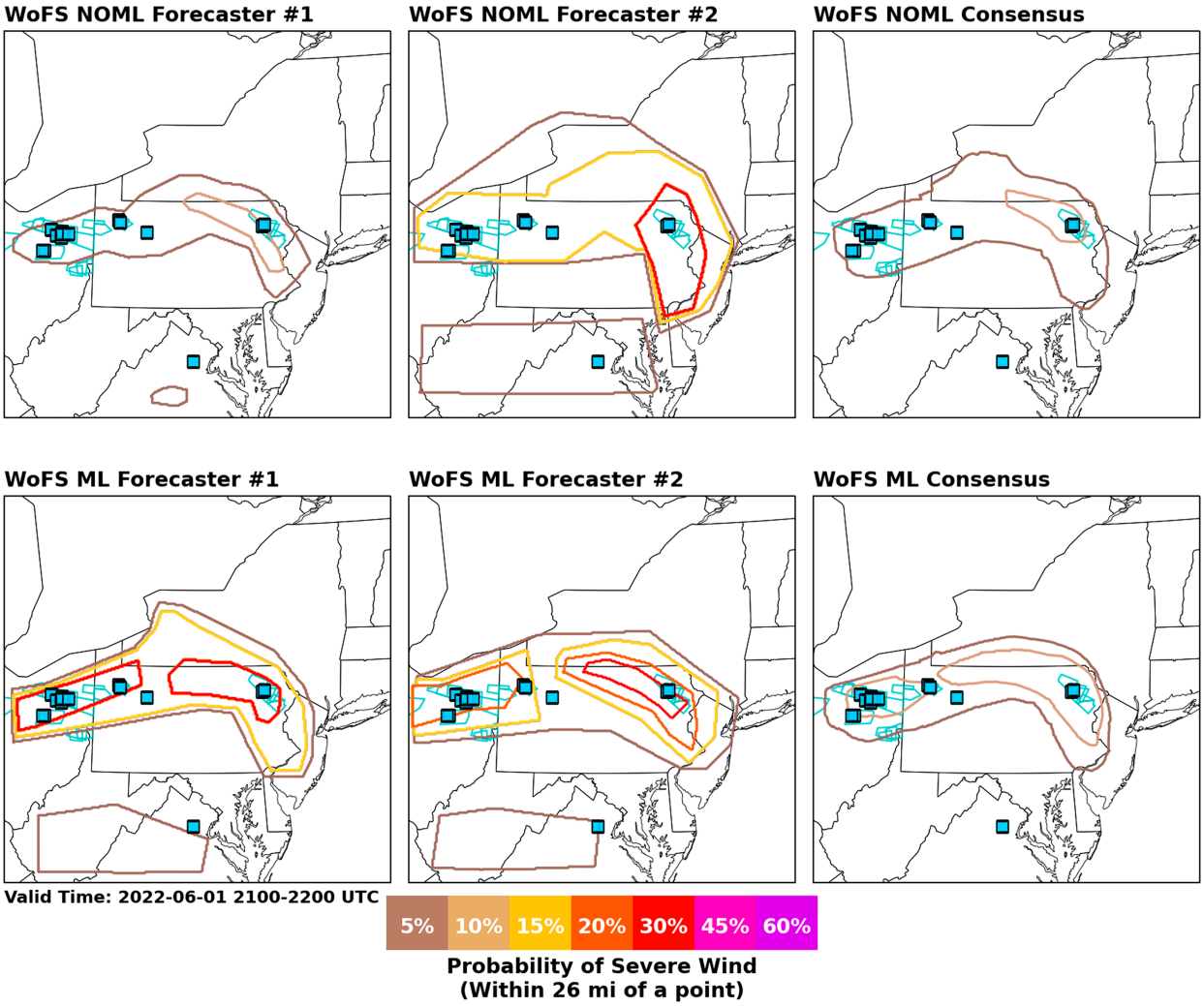

FIG. 3. Forecast outlooks generated as part of the afternoon forecasting activity highlighting the probability of severe wind gusts during 2100–2200 UTC 1 Jun 2022. The control group (NOML) outlooks are shown on the top row with the experimental group (ML) outlooks on the bottom row. The numbered forecasters are the NWS participants ("expert" forecasters). Observed wind reports are shown as blue boxes and severe weather warnings as blue polygons. These figures are modified versions of the official graphics available at https://hwt.nssl.noaa.gov/sfe/2022/.

diagram, the performance diagram (Roebber 2009), the reliability diagram (Hsu and Murphy 1986), and the metrics commonly associated with those diagrams. The ROC diagram measures the ability of the forecast probabilities to discriminate between events and nonevents. The probability of detection (POD) and the probability of false detection (POFD) are computed at a series of probability thresholds to calculate the ROC curve. We can summarize the ROC curve by measuring the area under the curve (AUC) where AUC = 0.5 indicates no skill, while AUC = 1.0 is a perfect discriminator. AUC, however, can be a misleading measure of discrimination since it values the correct prediction of events and nonevents equally.

Using the performance diagram, we can assess how our forecast probabilities discriminate between hits, misses, and false alarms. Instead of POD versus POFD, we can compare POD versus success ratio (SR) for a series of probability thresholds and similarly summarize the resulting curve by its area [known as the area under the performance diagram curve (AUPDC)]. Since the AUPDC is base rate sensitive (Boyd et al. 2013; Flora et al. 2021), it is a more useful metric for discrimination in the rare event case since it values correct prediction of events while ignoring correct predictions of nonevents.

Finally, we can assess how well the forecast probabilities match the observed conditional event frequencies using the reliability diagram (Hsu and Murphy 1986). For perfectly calibrated probabilities, the forecast probabilities should equal the event frequencies for different probability ranges (e.g., [0%–10%], [10%–20%], … [90%–100%]). The standard metric associated with the reliability diagram is the Brier skill score (BSS). Brier score is the mean squared error between the forecast probabilities and the binary outcomes. The Brier

TABLE 1. List of survey questions and their possible responses asked during the afternoon activity subjective evaluations associated with the WoFS ML activity in the HWT-SFE. The first two questions were asked of both the control and experimental group while the remaining questions were only asked of the experimental group.

Survey questions	Possible responses
How confident are you in your forecasts of the following hazards today (considering both the 2100–2200 UTC and the 2200–2300 UTC time periods) for each hazard?	Not at all, slightly, moderately, very, extremely
Approximately how many different WoFS products did you look at today when formulating your forecasts?	5 or less, 6–10, 11–15, 16+
How did the ML guidance affect your confidence in your forecasts today?	Open ended
How useful was the ML guidance when creating forecasts of the following hazards today (considering both the 2100–2200 UTC and the 2200–2300 UTC time periods) for each hazard?	Not at all useful, a little useful, somewhat useful, very useful
What did you like or dislike about the ML guidance?	Open ended
Where did the ML products fit in your workflow?	Open ended
(Optional) Please provide any additional comments that you have regarding the ML guidance and visualization	Open ended
How useful were the explainability graphics when creating forecasts of the following hazards today (considering both the 2100–2200 UTC and the 2200–2300 UTC time periods) for each hazard?	Not at all useful, a little useful, somewhat useful, very useful
How did the explainability graphics impact your confidence in your forecast, if at all?	Open ended

score of the forecast probabilities is compared against the performance of a constant climatological prediction (the sample climatology) to create a skill score. A positive BSS indicates that the prediction is better than always predicting the climatological frequency.

The forecast probabilities were verified with quality-controlled storm reports from the National Centers for Environmental Information (NCEI) Storm Data Publication. The reports for this study include thunderstorm wind gusts $\geq$ 50 kt, hail diameter $\geq$ 1 in. (1 in. = 2.54 cm), and any tornado. The observed storm reports were gridded in the WoFS domains and dilated to a radius of 39 km (corresponding to the SPC convective outlook risk definition of severe weather within 25 miles of a point). For brevity, the following verification metrics show each hazard's combined performance of the updated 2100–2200 UTC and 2200–2300 UTC outlooks. We should note, however, that the next-day subjective evaluations were performed with filtered, neighborhooded local storm reports available at the time of the evaluation and are subject to reporting latency.

e. *Qualitative analysis methods*

An inductive thematic analysis was used to identify and interpret patterns within the responses to the open-ended questions. An inductive approach is a method where themes or patterns are derived directly and organically from the data without any preexisting categories or frameworks (Braun and Clarke 2006). The thematic analysis approach was primarily based on the six-phase approach introduced in Braun and Clarke (2006). After data familiarization, i.e., reading through the responses, the individual responses are categorized in a "coding" phase, and larger descriptive themes can be derived from these codes. These "codes" are highly descriptive and require little analysis by the researcher (King 2004). To help with the initial codings, we used open-source Python packages such as the Natural Language Toolkit (NLTK; Bird et al. 2009) and a word cloud generator (Mueller 2023). After viewing those data, we created codes for each participant's response. From these initial codes, the authors derived themes and then vetted those themes against the individual responses

TABLE 2. As in Table 1, but for the next day activity.

Survey questions	Possible responses
After seeing the forecast verification, how confident would you be in using the WoFS while issuing a future forecast?	Not at all, slightly, moderately, extremely
After seeing the forecast verification, how confident would you be in using the WoFS ML guidance while issuing a future forecast?	Not at all, slightly, moderately, extremely
Please indicate the usefulness of WoFS for the following hazards today for each hazard.	Not at all, slightly, moderately, extremely
Please add any additional comments about the WoFS, the WoFS ML guidance, utility in forecast issuance of WoFS or the WoFS ML guidance, or yesterday's severe weather evolution that you feel would be helpful to the WoF team.	Open ended
Would you prefer consistent fields (explaining how the same set of predictors contributes to the prediction regardless of the storm) or storm-specific fields (using different predictors to contribute to each storm's prediction)?	Storm-specific fields, consistent fields, do not prefer one or the other.

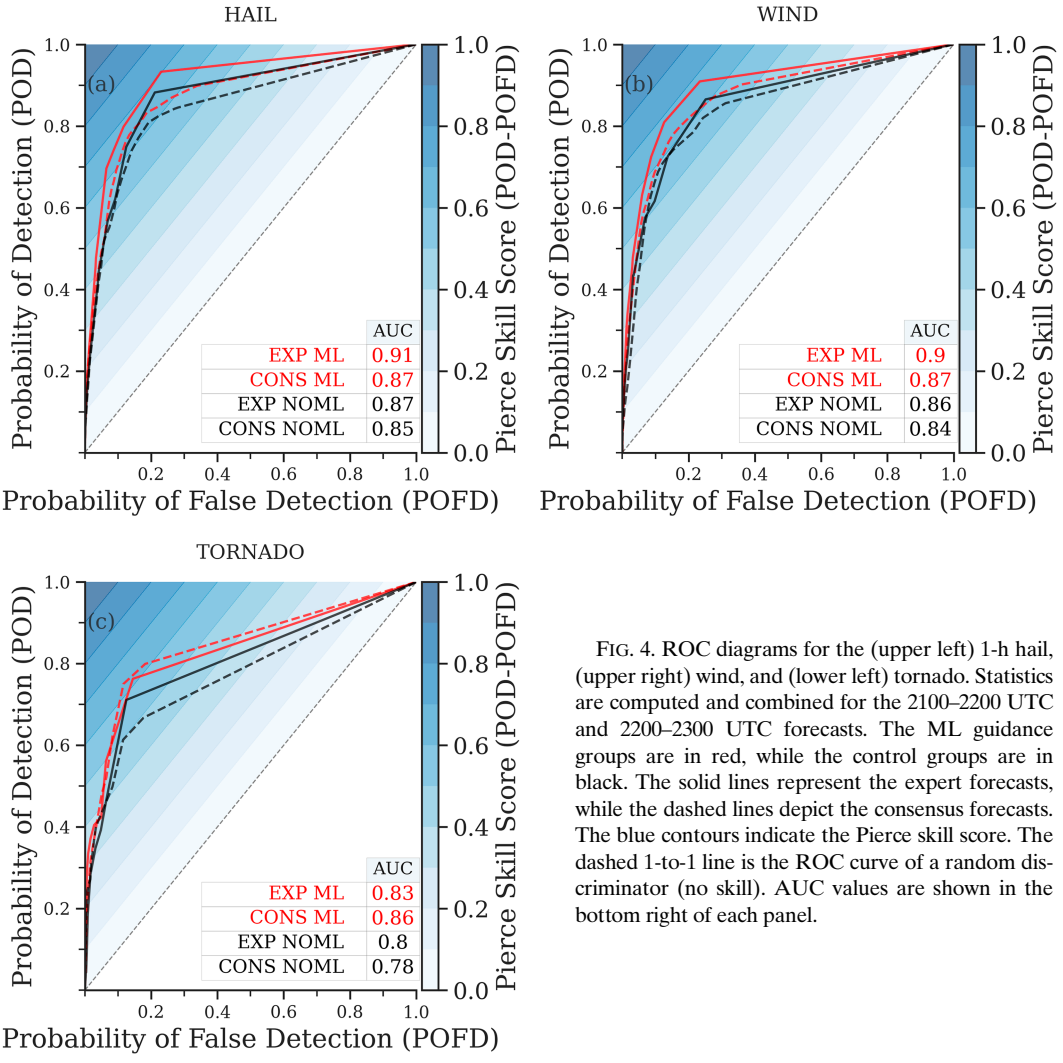

FIG. 4. ROC diagrams for the (upper left) 1-h hail, (upper right) wind, and (lower left) tornado. Statistics are computed and combined for the 2100–2200 UTC and 2200–2300 UTC forecasts. The ML guidance groups are in red, while the control groups are in black. The solid lines represent the expert forecasts, while the dashed lines depict the consensus forecasts. The blue contours indicate the Pierce skill score. The dashed 1-to-1 line is the ROC curve of a random discriminator (no skill). AUC values are shown in the bottom right of each panel.

to ensure the themes were consistent and representative. The authors then crafted the final themes and descriptions.

3. Results

a. Objective verification of the 1-h forecast outlooks

ROC and performance diagrams are shown in Figs. 4 and 5, respectively. This section will refer to the consensus and expert participants in the experimental group as CONS ML and EXP ML and the control group participants as CONS NOML and EXP NOML. The EXP ML group produced higher AUPDC and AUC for all three hazards than the EXP NOML group. The EXP ML group discriminated well for all three hazards, producing an AUC of 0.91, 0.9, and 0.83 for severe hail, severe wind, and tornado, respectively (Fig. 4). The EXP ML AUPDC increased by >50% against the EXP NOML for predicting tornadoes (Fig. 5c), but with the limited sample size, this should be interpreted cautiously. For both severe wind and hail, the EXP ML had similar percent increases (20%–25%) in AUPDC as compared to the EXP NOML (Figs. 5a,b). Among the three hazards, severe wind demonstrated the most pronounced discrepancy between EXP ML and CONS NOML, as depicted in Fig. 5b. This aligns with prior research indicating that ML guidance typically yields the most marked enhancement for severe wind relative to other hazards (Loken et al. 2020; Hill et al. 2020; Flora et al. 2021). Wind diagnostics in convection-allowing models frequently exhibit less proficiency than those for hail and tornadoes (Jirak et al. 2014; Hepper et al. 2016), potentially providing seasoned forecasters and ML with opportunities to add value. However, severe wind reports are inconsistent in quality (Trapp et al. 2006), particularly in densely forecasted regions where even subsevere winds can cause tree damage. We hypothesize that ML models might be more adept at detecting these "near" severe winds as we assume they are easier to predict. Consequently, ML-informed outlooks could better align with wind reports but may not equate to enhanced accuracy in predicting genuine severe wind events.

For severe wind and hail forecasts (Figs. 5a,b), the maximum critical success index (CSI) for the different groups

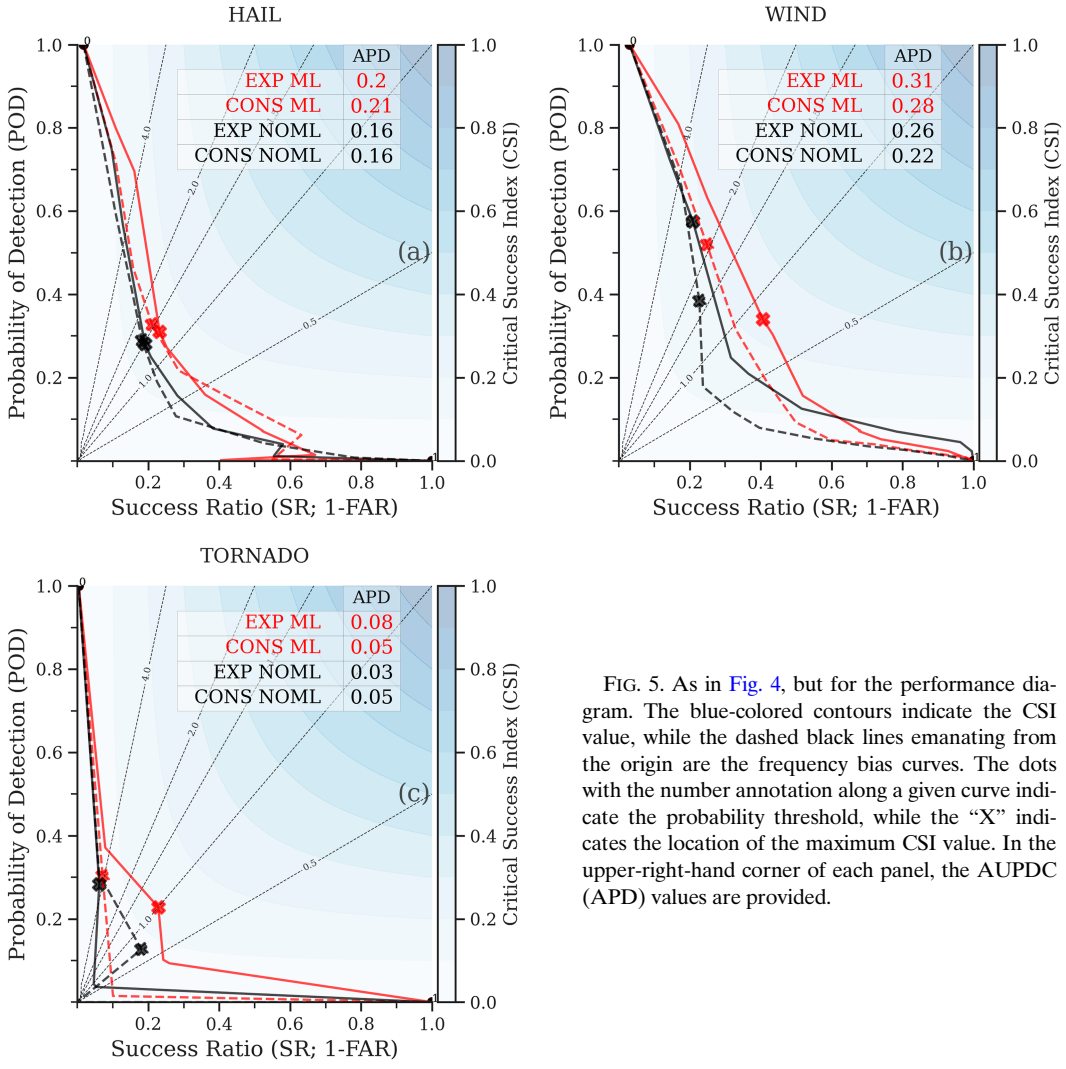

FIG. 5. As in Fig. 4, but for the performance diagram. The blue-colored contours indicate the CSI value, while the dashed black lines emanating from the origin are the frequency bias curves. The dots with the number annotation along a given curve indicate the probability threshold, while the "X" indicates the location of the maximum CSI value. In the upper-right-hand corner of each panel, the AUPDC (APD) values are provided.

occurred with a frequency bias > 1, which is typical for rare event problems (Baldwin and Kain 2006). Although EXP ML performed better than the other groups for tornado forecasts, the performance of all groups was much lower than for the other hazards. This is unsurprising given that few tornadoes occurred within the 2100–2300 UTC period during the SFE (Fig. 1). CONS ML performed similarly to or better than EXP NOML for all three hazards. We attribute the superior forecasts of the ML groups to their ability to leverage the explicit probabilities of severe weather hazards provided by the ML guidance. With little training/exposure (2–3 days at most), nonforecaster participants could leverage the novel guidance consistent with past HWT-SFE studies (Karstens et al. 2015; Gallo et al. 2024). If the EXP NOML group had more experience fully leveraging WoFS, they might have performed better than the CONS ML group.

The hail and wind reliability curves were close to the one-to-one line (Fig. 6). The wind probabilities, however, have an underforecast bias, especially for probabilities > 40%. The highest hail probabilities issued were around 50% (Fig. 6a) while the highest tornado probabilities issued were around 20% (Fig. 6c). The EXP ML group produced the most reliable forecasts and the highest BSS for all three hazards. Thus, the EXP ML group successfully translated the event-based ML probabilities into reliable spatial probabilities. We can also see that CONS ML and EXP ML issued slightly fewer probabilities in the bins between 20% and 50%, which improved the resolution of the forecasts and helped explain the somewhat higher BSS compared to the CONS NOML and EXP NOML groups.

b. *Comparing outlook areas between the ML and NOML groups*

To further distinguish the performance of the groups with access to the ML guidance from those without, we evaluated the comparative size of the outlook polygons produced by EXP ML and EXP NOML (Fig. 7). We restricted our analysis to severe wind and hail outlooks as the tornado outlooks

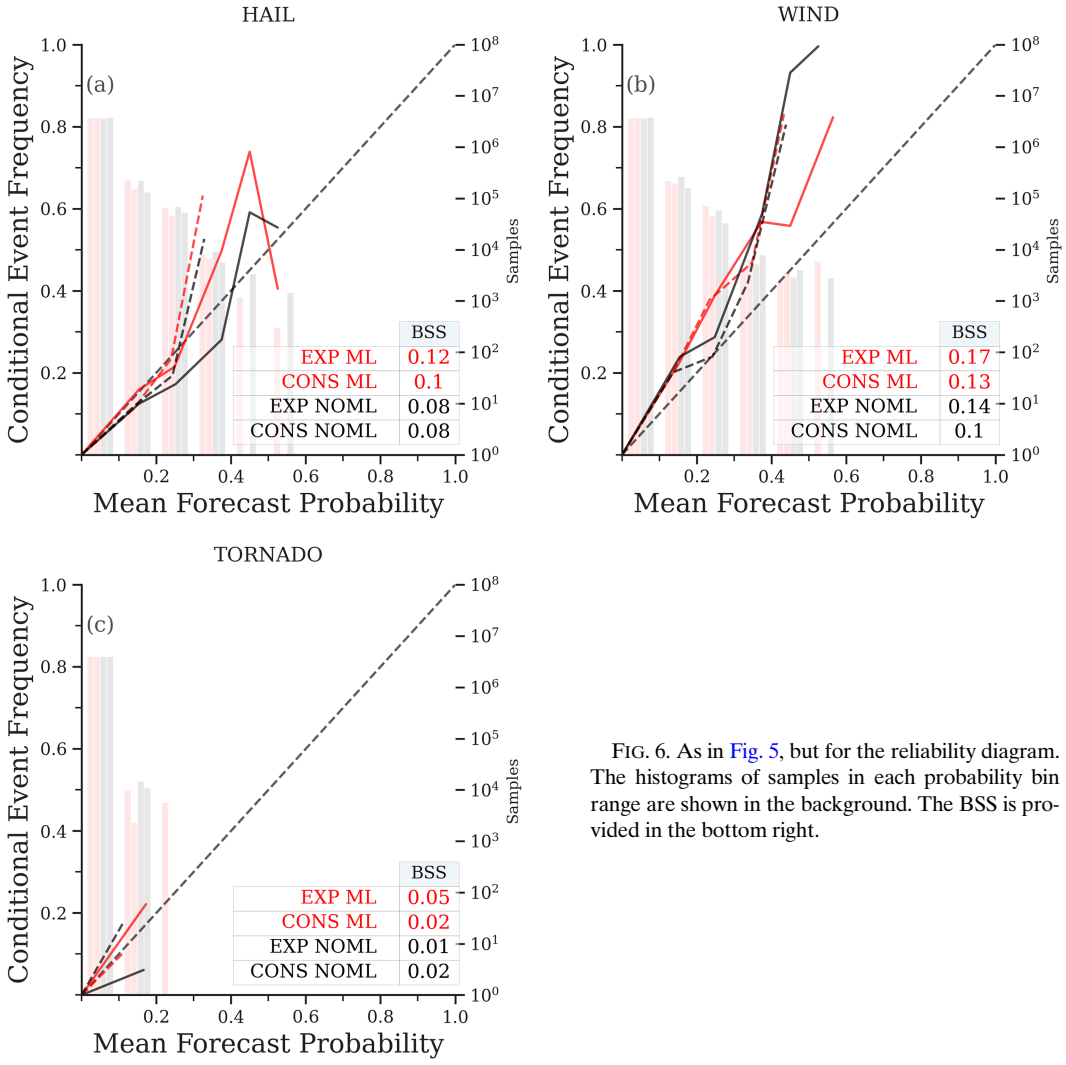

FIG. 6. As in Fig. 5, but for the reliability diagram. The histograms of samples in each probability bin range are shown in the background. The BSS is provided in the bottom right.

were too infrequent to create a meaningful sample size. Given that the hail and wind results are similar, we have omitted the hail results for brevity. Our goal is to assess whether the improvement in skill from using the ML guidance arises from the participants issuing smaller contours (i.e., honing in on the area of threat) or larger outlooks (i.e., reducing missed reports).

For wind probabilities $> 5\%$, the EXP ML did produce slightly larger area forecasts for many cases, but there is significant variance. For example, on 3 May 2022, the EXP ML area was nearly ten times larger than the EXP NOML, but the expert forecasters vastly underestimated the event compared to the ML groups (Fig. 8). The ML guidance highlighted not only the southern Ohio River valley but also northwest Ohio, which had a weak, nonsevere wind signal in the raw WoFS guidance (not shown). The northwest Ohio area, as a result, is included in the EXP ML outlooks, which does verify well (cf. Figs. 8 and 9). A similar situation occurred for the 16 May 2022 case, in which the EXP ML area better captured the event, and the NOML groups underestimated the event (not shown).

However, in some cases, the ML outlooks were smaller and more accurate. For example, on 26 May 2022, the EXP NOML forecast area was six times larger than the EXP ML area (Fig. 10). In this case, many participants had drawn relatively large 5% regions, but one of the EXP ML forecasters had drawn a highly focused threat area that compares well with the observed wind reports. In the 12 May 2022 Derecho case, though the WoFS generally captured the intense wind event, the CONS ML and EXP ML contours were better centered on the event, which was informed by the concise ML guidance locations (not shown). These results suggest that it is possible that with additional training, forecasters could take advantage of the WoFS and ML guidance to issue higher confidence forecasts over a smaller region. The EXP ML guidance did not always correctly capture the event, but when the EXP ML forecasts performed poorly, the EXP NOML also performed poorly. In general, when

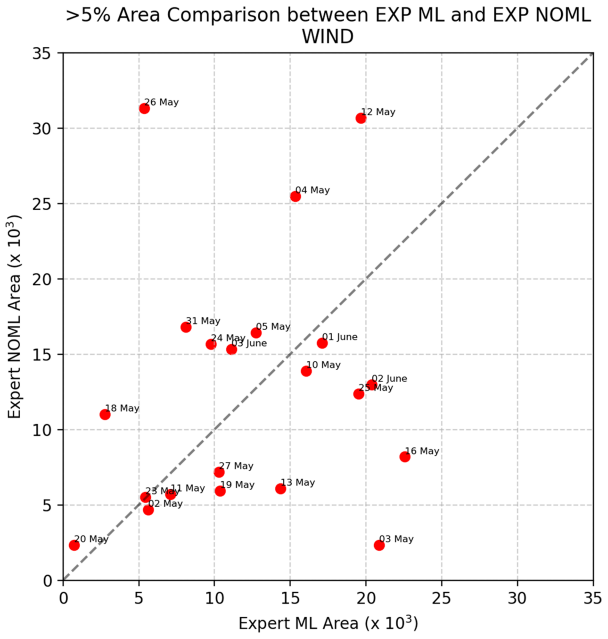

FIG. 7. Comparison of severe wind probabilities area > 5% between EXP ML and EXP NOML for the combined 2100–2200 UTC and 2200–2300 UTC periods. The area is in the number of 3 km × 3 km grid cells.

the EXP ML had a positive area bias compared to the EXP NOML, the EXP NOML underpredicted or failed to capture the event. Moreover, for cases where the EXP ML had smaller areas than the EXP NOML, the EXP ML forecasts better highlighted specific threat regions.

c. *Subjective verification of the 1-h forecast outlooks*

The initial and final forecasts generated by the participants were averaged to create the composite forecast ratings in Fig. 11. Plots showing the results separated by the initial and final outlooks and 2100–2200 UTC and 2200–2300 UTC periods are available in the 2022 HWT-SFE Preliminary Findings and Results Report (Clark et al. 2023).

Overall, the EXP ML forecasts were rated higher than the EXP NOML, with the most dramatic differences for wind, which matches the objective verification. The differences between EXP NOML and EXP ML were statistically significant ($p < 0.05$) for all three hazards. These results indicate that ML can provide significant value on top of the raw WoFS products. Furthermore, the average ratings for the CONS ML were higher than CONS NOML, though none of the differences were statistically significant.

d. *Analysis of the next day participant feedback*

The following section evaluates the survey responses to the afternoon activity questions in Table 2. Participants in the experimental (ML) and control (NOML) groups were asked about their confidence in using WoFS and ML guidance in the future and the usefulness of WoFS for each hazard after evaluating the forecast outlooks (Fig. 12). Overall, participants from both groups were moderately to very confident in using WoFS and ML guidance. Experimental group participants were not significantly more confident on average but had more reports of being "very" or "extremely" confident. Therefore, it is unclear whether seeing the forecast verification of the ML-influenced outlooks improved confidence. Confidence, however, is a varied and complex notion, and building confidence requires multiple exposures (Hoffman et al. 2017; Henderson et al. 2023; Cains et al. 2023). Though not interchangeable, trust and confidence are linked, and trust in guidance evolves with experience (Hoffman et al. 2013). Cains et al. (2023) emphasized that forecasters need personal, repeated experience with ML guidance to build trust. This allows forecasters to interrogate the guidance and develop mental bias corrections, which improves the forecaster's trust and confidence in the guidance.

As for usefulness, participants from both groups generally found WoFS "moderately useful" and "very useful," with more very useful responses for hail and wind (Figs. 13b,c). Though the participants using ML guidance reported WoFS as very useful or "extremely useful" more often for all three hazards, they found it significantly more useful for tornado guidance after seeing the forecast verification (Fig. 13a). Overall, the participants found the WoFS to be quite useful for all hazards, again demonstrating the usefulness of the WoFS for short-term hazard forecasting (Gallo et al. 2022, 2024). This is largely not only due to WoFS filling a crucial gap for watch-to-warning forecasting but also due to it fulfilling topics discussed in Demuth et al. (2020) such as displaying deterministic and probabilistic convection-allowing model output and guidance for IDSS.

e. *Analysis of the day-of participant feedback*

The following section evaluates the survey responses to the afternoon activity questions in Table 1. The experimental and control groups answered questions about their confidence in their outlooks and the number of products used, while the experimental group was asked additional questions regarding the ML guidance. The 2022 HWT-SFE Preliminary Findings and Results Report (Clark et al. 2023) contains a preliminary version of these figures and analysis. Note that the questions on participants' confidence or perceived usefulness of the ML guidance in Table 1 were asked before forecast verification of a given case.

Participants were asked how confident they were in their predictions for each hazard (Fig. 14). We hypothesize that the more precise ML guidance (compared to the raw WoFS output) will increase confidence when it aligns with the user's expectations. Still, at other times, when the ML guidance defies user expectations, it will likely decrease their overall confidence. Overall, both groups had similar confidence in their forecasts. While the differences were relatively small ($p > 0.05$), the experimental group responded more often as very or extremely confident about their tornado and hail forecasts. For wind forecasts, the experimental group was more likely to say they were "moderately" or very confident in their forecasts. On the contrary, the control group was more likely for all hazards to say that they were "slightly" confident in their forecasts.

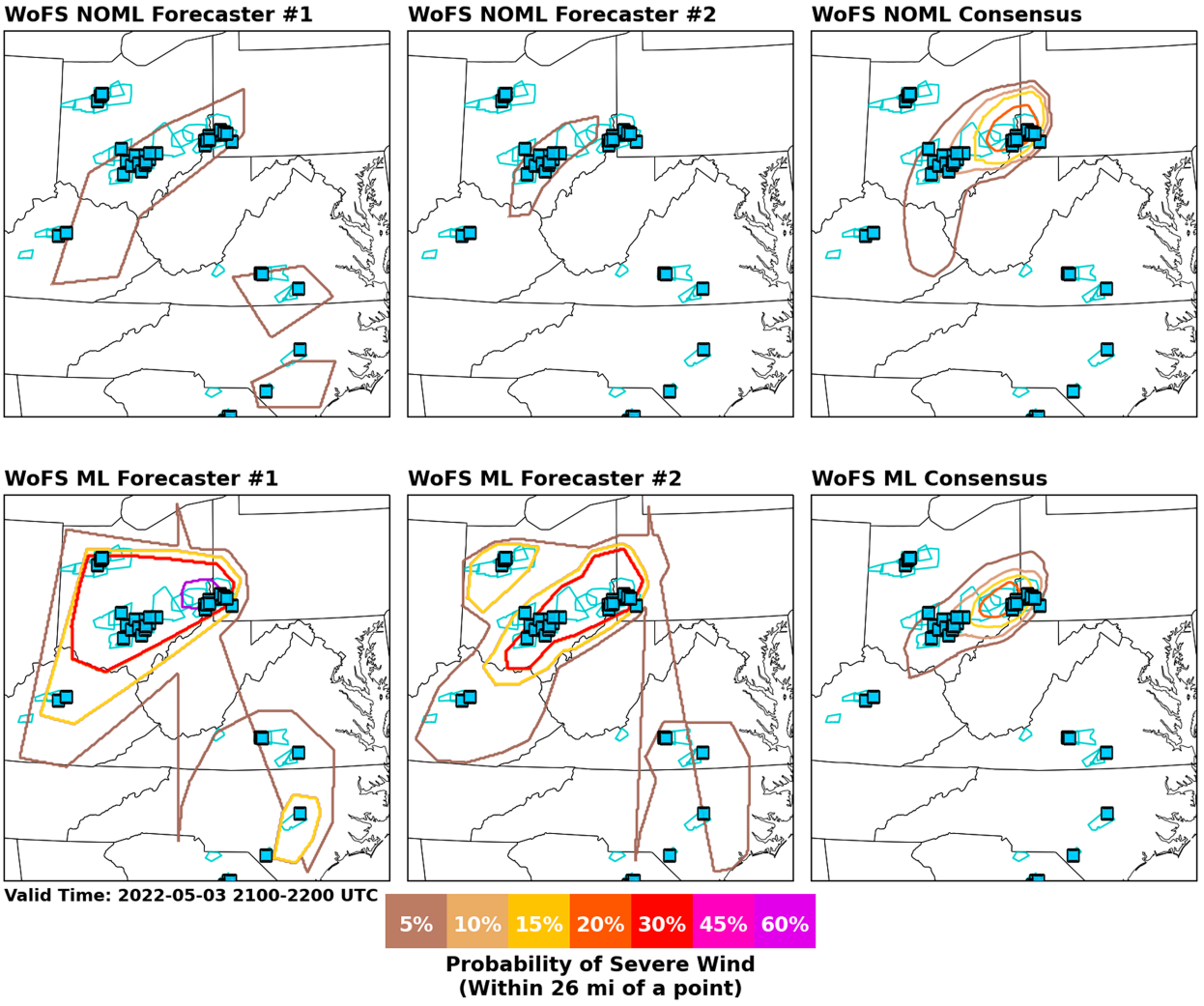

FIG. 8. As in Fig. 3, but for severe wind for 2100–2200 UTC 3 May 2022.

Though participants from both groups had similar confidence, we know from the objective and subjective verification above that the experimental group's outlooks outperformed the control group's outlooks. Most experimental group participants stated that the ML guidance positively affected their confidence. The largely positive sentiment is likely influenced by knowing that the ML guidance tends to verify well, which can positively impact trust (Nourani et al. 2020).

In terms of how the ML guidance impacted participant confidence, we identified the following themes in their responses:

- Refinement tool: Participants often used the ML guidance to refine and adjust their forecasts, i.e., adjusting contour locations and magnitudes. This was the most common theme among the participant responses.
- Issuing higher confidence guidance when consistent with other guidance: When WoFS and the ML guidance were consistent with their expectations, many participants felt confident to issue higher probabilities. However, when the guidance was inconsistent with their expectations, it decreased their confidence.
- Hazard-specific benefits: The ML guidance primarily boosted confidence for hail and wind probabilities with a more neutral impact on tornado probabilities. The neural effect of the ML guidance on the confidence of their tornado outlook was often associated with the limited predictability of tornadoes.
- Limited training/exposure: Some participants noted that their inexperience with the product impacted their confidence. For the few participants who viewed the ML guidance for more than one day, it increased their confidence.

With access to the explicit ML guidance, participants had a "ruler" against which to judge their outlooks. When the ML guidance was consistent with their expectations, they felt confident altering their outlooks and issuing higher probabilities, especially for severe wind and hail. Since the participants had limited training and exposure to ML guidance, they were reluctant to trust the ML guidance fully. These results are

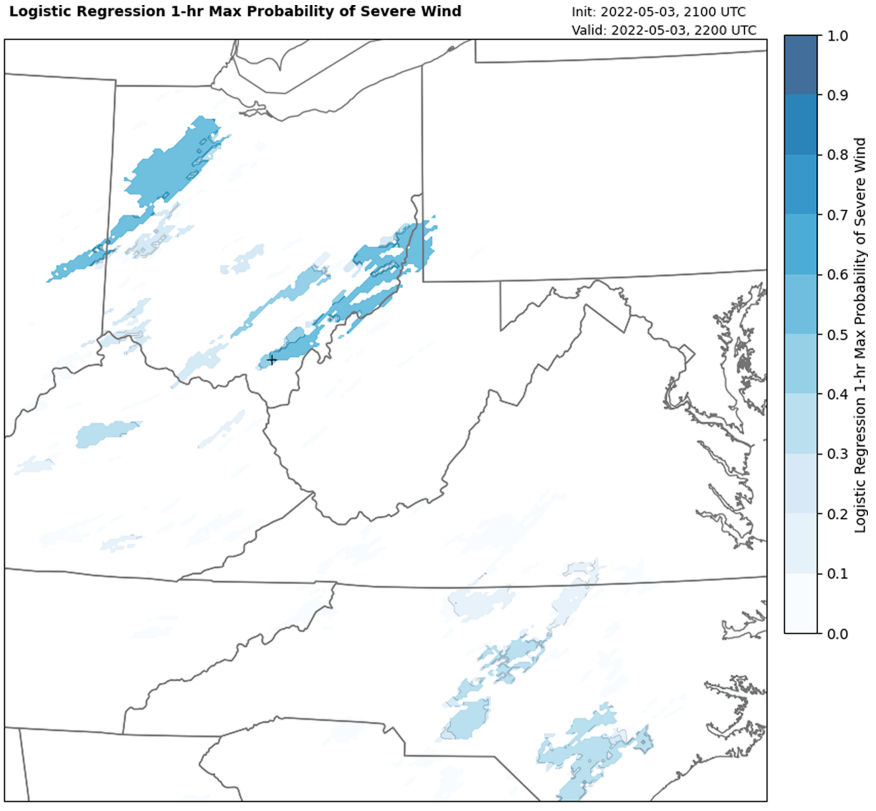

FIG. 9. WoFS-ML-Severe (logistic regression model) 1-h time-maximum composite of the probability of a severe wind being associated with an ensemble storm track. The forecast is valid for 2100–2200 UTC 3 May 2022.

consistent with Henderson et al. (2023) and Cains et al. (2023). With repeated exposure and training, forecasters tend to have increased trust in ML guidance (Cains et al. 2023).

The experimental group (ML) tended to self-report viewing more WoFS products than the control group (NOML; $p < 0.05$; Fig. 15). Specifically, while the control group frequently viewed fewer than 5 or between 6 and 10 different products, the experimental group often viewed between 11 and 15 or 16+ products on the WoFS web viewer. This observation is counter to our initial hypothesis. We had assumed that the WoFS ML guidance, which ingests multiple WoFS fields, might lead users to examine fewer products. However, to build confidence in the ML guidance, participants in the experimental group likely examined additional guidance. For example, if the ML products spotlighted an unexpected area (e.g., NW Ohio in Fig. 9), they were likely motivated to consult other WoFS products to improve their context. Among those who specified how ML products fit into their workflow ($N = 82$ of 152), most participants mentioned leveraging it at the start and the end of their forecast process, reinforcing the refinement theme.

Given that the experimental group participants tended to view more products, it may have contributed to the increased performance of the ML-guided outlooks versus the control group. However, viewing more guidance products does not always correlate with increased forecaster performance (Stewart et al. 1992). In a more thorough study of participants using WoFS, Wilson et al. (2021) did not find a significant relationship between the number of products viewed and forecast quality. We did not document the specific products viewed like Wilson et al. (2021); therefore, it is unclear how the additional products impacted either group. For example, without knowing the particular products viewed, participants may have self-reported products of a similar type (e.g., updraft helicity at different neighborhoods) as completely separate products.

Ultimately, as for viewing fewer products, Wilson et al. (2023) found that forecasters learned how to prioritize a subset of WoFS products over time, which helped reduce cognitive workload. Thus, with more exposure and training, we hypothesize that the ML guidance could alleviate cognitive load coincidence with other studies (Karstens et al. 2018; Calvo et al. 2022; Ehrmann et al. 2022) and increase forecaster confidence (Cains et al. 2023).

When the experimental group was asked about the usefulness of the ML guidance, the participants found the guidance to be "somewhat useful" or very useful most of the time (Fig. 16). The guidance was most helpful in assessing the wind threat, where most participants rated it as very useful. Very useful was also the most common answer for hail forecasts, whereas the tornado guidance was often rated somewhat useful.

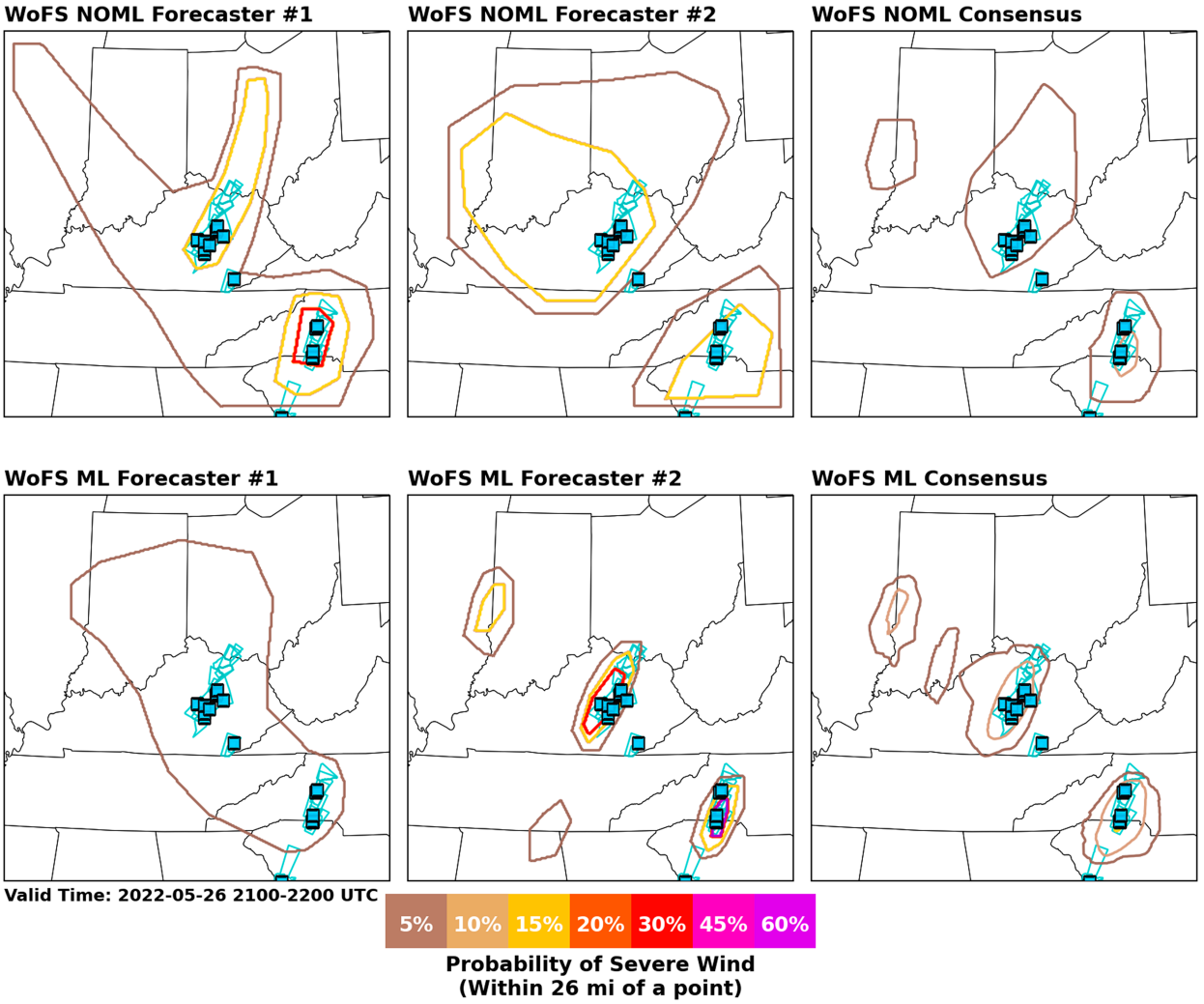

FIG. 10. As in Fig. 3, but for severe wind for 2100–2200 UTC 26 May 2022.

The experimental group had access to explainability products during the forecasting exercise. These are graphics attached to each ensemble track in the domain, and they contain contextual information about a consistent set of hazard-dependent top five predictors, i.e., those predictors that are the most influential in the training dataset. However, for one of the survey questions, participants were presented with a static image (Fig. 17) of global (same predictors for all storms) and local (storm-specific) explainability products and asked which of them they preferred. Storm-specific fields were most commonly chosen, followed by "Do not prefer one or the other" (Fig. 17). The participants also had the option to write a suggestion in the "other" response. Some participants used this response to indicate that they wanted more training with the product, to see the product demonstrated during an event, or to see the product shown for an event with more storms before making their selections. Others indicated a nuanced take, where different preferences would match different scenarios. One participant said, "I would prefer the global attributes if I were forecasting near the maximum in severe wind. If I were forecasting in an area that does not see stronger winds often, I may prefer the local variables." Another participant suggested incorporating 2–3 parameters from each option into a single plot. This feedback is being used to improve the explainability products and expand the utility of the ML guidance.

The participants found slightly less utility in the supplementary explainability products than in the ML guidance (Fig. 18). However, the explainability products were still valuable, and most participants indicated they were at least somewhat useful for all hazards. Participants may have required substantial time to explore the explainability products since this was their first use in the SFE. Some participants' comments reflected the need to understand the explainability products better, but many participants commented that they liked the ideas behind these products.

Participants were asked open-ended questions about whether the explainability products affected their confidence (if at all).

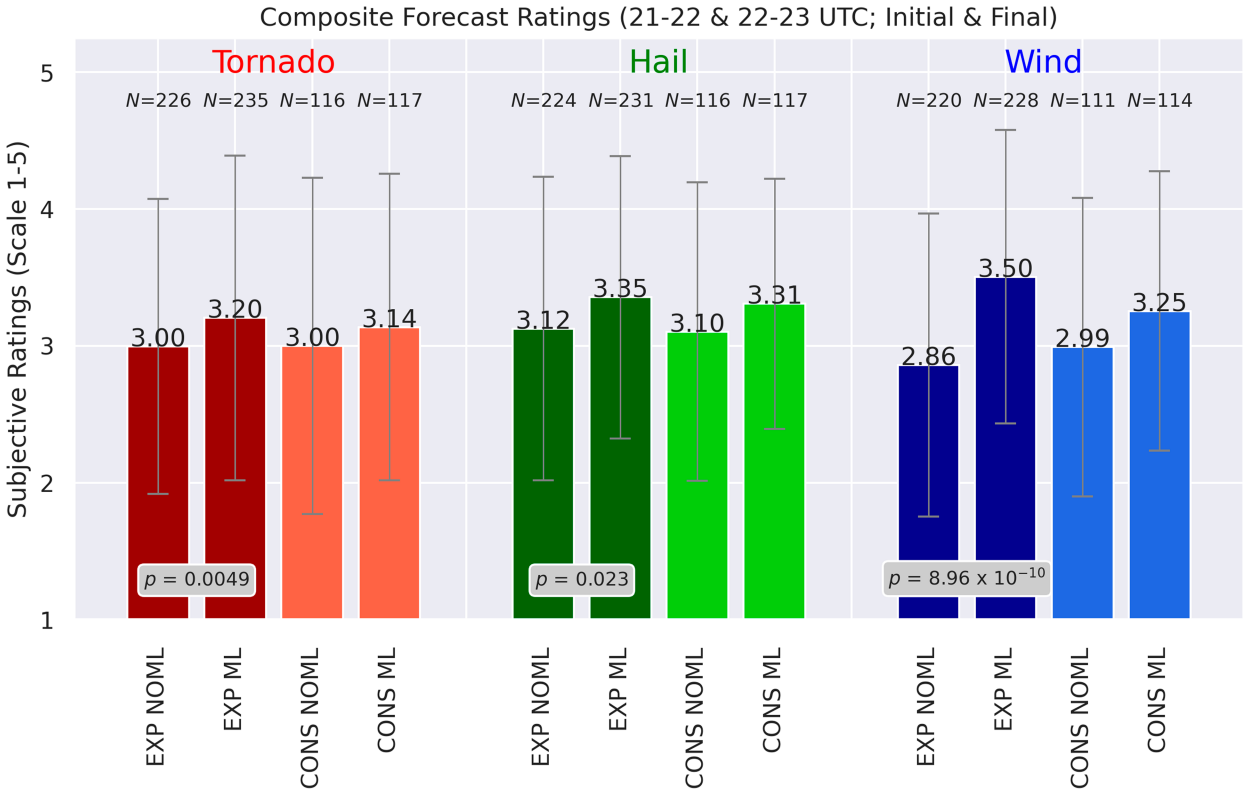

FIG. 11. Average subjective ratings for EXP ML, EXP NOML, CONS ML, and CONS NOML for all three hazards averaged for the 2100–2200 UTC and 2200–2300 UTC periods, combining the initial and final outlooks. The standard deviation is shown with a gray error bar. The p values from Welch's t test comparing the EXP ML and EXP NOML outlooks are overlaid on the histogram bars for each hazard. The sample size is provided for each rating.

The sentiment was mixed for those who reported using the explainability products (N = 104 of 122). Many felt neutral toward the products and its impact on their confidence. We identified the following themes in their responses:

- Lack of use: Many participants reported barely using the explainability products (or not at all). The lack of use was attributed to unfamiliarity with the products and uncertainty about leveraging the information to alter their outlooks.
- Peering into the black box: For those participants who felt explainability products positively affected the confidence in their forecasts, they often attributed this sentiment to a better understanding of the ML inputs.
- Information overload: Some participants found viewing the explainability products often contributed to information overload.

Information overload is a theme that has been identified in other explainable AI (XAI) user studies (Gunning et al. 2021; Cains et al. 2023). Though not in the context of weather forecasting, Gunning et al. (2021) found examining explanations can hinder user performance due to the increased cognitive load. Gunning et al. (2021) also found that explanations were not always helpful except when the AI output was incorrect or in edge cases. Nevertheless, it is worth investigating whether providing additional context with explainability products could reduce the forecaster's need to search for context in other products, thereby reducing workload.

Participants' responses were mixed when asked what they liked or disliked about the ML guidance. We identified the following themes in their responses:

- Spatial precision: Participants found the ML guidance useful for identifying smaller, targeted threat areas with higher probabilities.
- Explainability graphics: Many participants liked the accompanying explainability graphics and their visualization.
- Concerns and limitations: Participants were concerned about image segmentation, the use of monochromatic colorbars, and disagreement of ML guidance with other WoFS products like HAILCAST and 80-m wind speeds.
- Learning curve: Many participants reported liking the visualization and product design but recognized that additional training and exposure were required to leverage the guidance fully.

Participants found the ML guidance helpful in increasing their confidence in wind and hail forecasts (consistent with the hazard-specific benefits theme). Some users liked the explainability products and seeing guidance in an event-based framework (e.g., being able to pinpoint areas of focus). However,

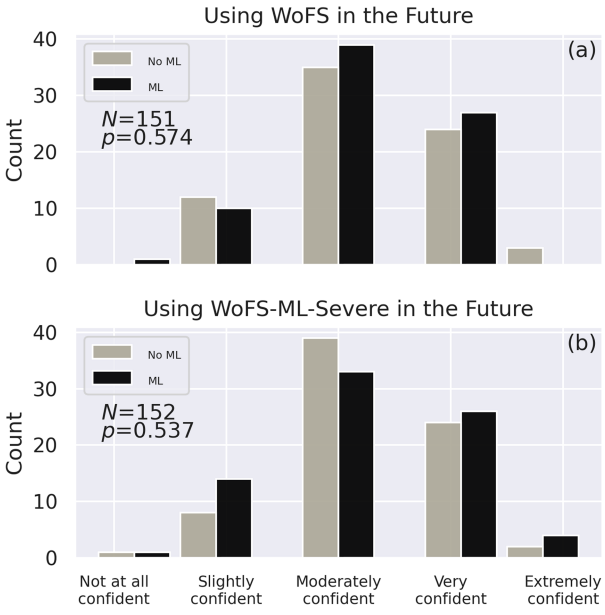

FIG. 12. Participant responses to the questions, (a) "After seeing the forecast verification, how confident would you be in using the WoFS while issuing a future forecast?" and (b) "After seeing the forecast verification, how confident would you be in using the WoFS ML guidance while issuing a future forecast?". WoFS NOML is the control group, while WoFS ML is the experimental group.

some users found the storm objects occasionally too large or were confused about interpreting probabilities in an event-based framework. For example, some users had questions/concerns about storm object size compared to other ensemble probabilities available on the WoFS web viewer. Participants were concerned that the larger object size would be prohibitive for warning-based guidance. These mixed results highlight a crucial issue in developing new forecast technology: different end users want different things, and creating a generalized system to meet disparate goals is difficult (Nourani et al. 2020). For example, there is an issue developing user-based guidance not only for different forecasting centers (e.g., WFOs versus SPC) but also for different users based on their expertise. Explainability is more advantageous for the more advanced user (Bayer et al. 2022) while novices are more likely to overtrust ML guidance (Nourani et al. 2020; Bayer et al. 2022). The end user may also need to become reasonably familiar with a product before its usefulness or benefit can be determined. Therefore, we continue collaborating with NWS forecasters to incorporate their feedback into improving training materials and ML products.

4. Summary

AI is revolutionizing severe weather forecasting for lead times < 1 h out to multiple days. A novel AI system was recently developed using machine learning (ML) within the Warn-on-Forecast System (WoFS), a state-of-the-art convection-allowing ensemble forecast system. Termed the WoFS-ML-Severe system, it was evaluated as part of the virtual 2022 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (HWT-SFE; Clark et al. 2023). We performed a two-group experiment where the control group had access to the full WoFS suite while an experimental group had access to both the WoFS and the ML guidance. Participants from both groups generated hourly outlooks for tornadoes, hail, and wind. The outlooks were verified objectively against observed reports and subjectively against all available verification

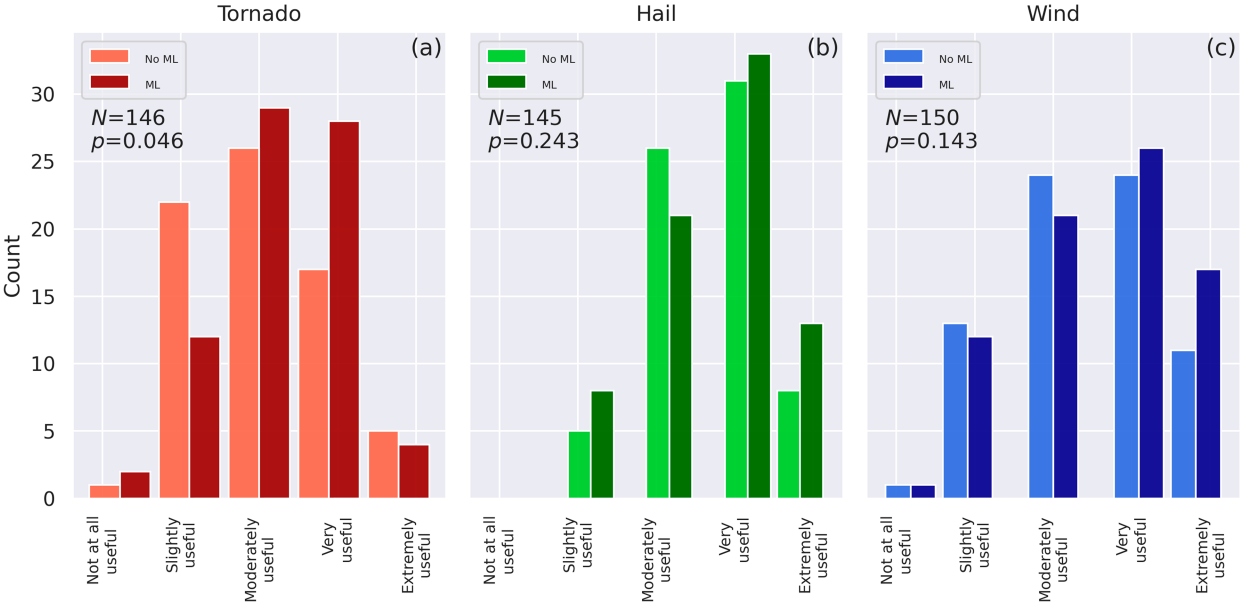

FIG. 13. Participant responses to the question, "Please indicate the usefulness of WoFS for the following hazards today." Dotted bars are from the group without access to ML guidance, while solid bars are from the group that had access to the ML guidance.

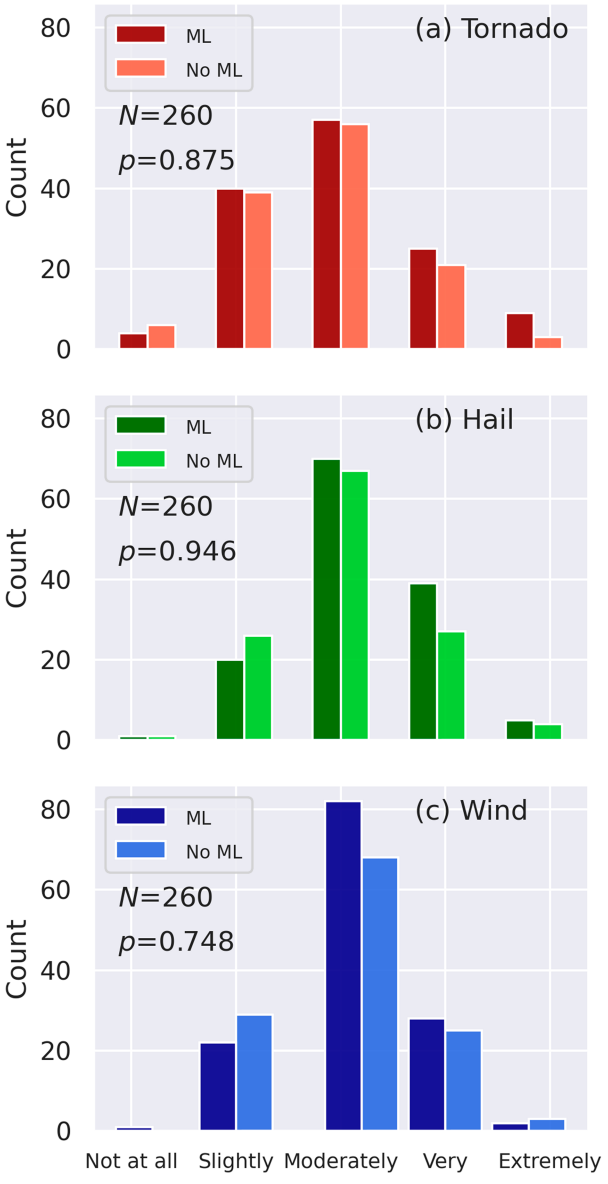

FIG. 14. Participant responses to the question "How confident are you in your forecasts of the following hazards today (considering both the 2100–2200 UTC and the 2200–2300 UTC periods)?" Participants responded separately for the (a) tornado, (b) hail, and (c) wind hazards.

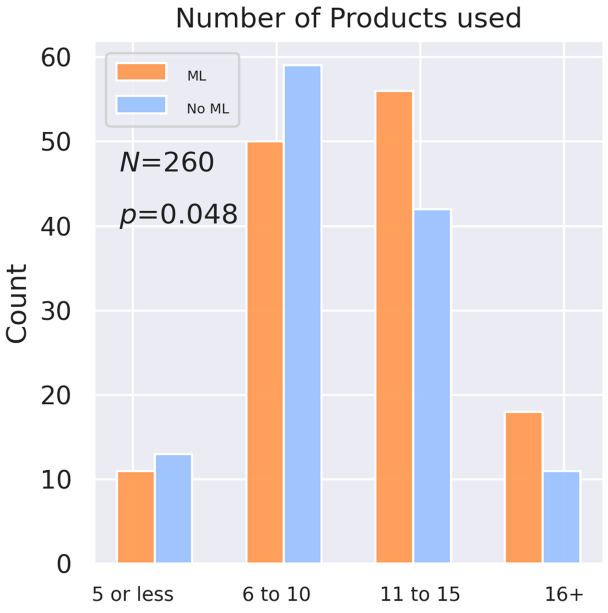

FIG. 15. Participant responses to the question "Approximately how many different WoFS products did you look at today when formulating your forecasts?" Participants were given the response options shown here, i.e., the number of products was prebinned in the responses.

datasets. Participants were asked questions after issuing their outlooks and the following day after the subjective evaluations.

Based on our analysis, we conclude the following:

- Participants, especially NWS forecasters, with access to the ML guidance performed better than participants without access to the ML guidance. ML-guided outlooks were the most reliable and had the best discrimination. The most significant improvement margin was for severe wind and the least for tornadoes.
- The ML-guided outlooks often better captured the spatial coverage of storm reports than the non-ML outlooks. This includes being more spatially focused with higher confidence for some cases or capturing being more diffuse in other cases.
- The experimental group outlooks received significantly higher subjective ratings than the control group outlooks. The wind outlooks received the highest ratings of the three hazards.
- The experimental and control groups reported being moderately confident in their outlooks (before verification). Though the experimental group had more reports of being very or extremely confident, the overall difference between the groups was not significant.
- Using ML guidance did not reduce the forecaster workload in this experiment. ML users tended to look at more WoFS products on average than non-ML users in addition to the explainability products. They were generally as confident as users without access to the ML guidance. This is consistent with Cains et al. (2023) as participants require personal, repeated experience with guidance before they can trust it and gain confidence.
- Although the responses were generally positive for the explainability products, some participants found them confusing and occasionally unnecessary. More work must be done to explore when operational explainability products are necessary and what information should be shown. This is especially true since recent research has suggested that increased cognitive load used to interpret explanations can hinder user performance (Gunning et al. 2021).

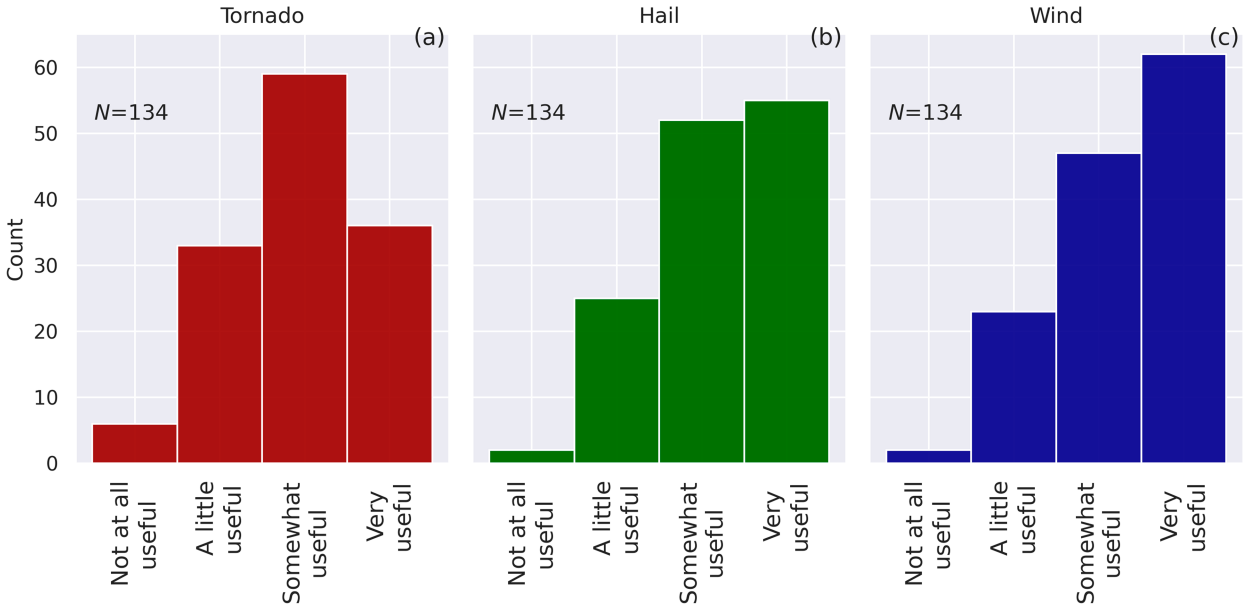

FIG. 16. Participant responses to the question "How useful was the ML guidance when creating forecasts of the following hazards today (considering both the 2100–2200 UTC and the 2200–2300 UTC periods)?"

In summary, the experimental group outlooks tended to outperform the control group outlooks, but it may not be entirely attributed to the ML guidance. For example, those accessing the ML guidance tended to self-report viewing more products, which may partially explain their improved outlooks. On the other hand, having more forecast information available does not necessarily result in forecast improvement (Stewart et al. 1992; Wilson et al. 2021). Also, given that the SFE participants are eager to adopt novel guidance and have a wide range of forecasting expertise, those in the experimental group may have overtrusted the ML guidance. Nourani et al. (2020) found that novice users are likelier to adopt novel guidance, especially if they are unaware of the errors and biases. Our results are promising, but how they generalize to all NWS forecasters is unclear.

Another limitation is that only the honor system was in place for control group participants not to use the ML products. However, facilitators frequently reminded control group participants to refrain from viewing the ML guidance, and given the virtual attendance, the two groups were always separated from each other.

More work is also required to explore the role of explainability and the ability of ML to reduce forecaster workload. Our main goal was to capture first impressions from the participants and use their feedback to guide ML development in the future. Given the novelty of the guidance and the participants' short "training" period, it is unsurprising that the ML guidance did not necessarily make the forecasting process easier or faster in this experiment (Cains et al. 2023). We also did not set out to evaluate the forecaster workload properly in this study. Though participants could access data outside the WoFS webviewer, many of them were likely heavily relying on WoFS due to the nature of the SFE. In an operational setting, NWS forecasters have other duties and thus may delegate less time to the WoFS and the ML guidance. It is unclear what impacts these could have on the usefulness of the ML guidance. Multiple studies have highlighted the need for more forecaster education and training, especially with novel forms of probabilistic guidance (Doswell 2004; Wilson et al. 2019a; Roebber and Smith 2023).

Another major takeaway from the HWT-SFE activity is that participants successfully took a product meant for one task (i.e., event-based prediction) and translated it into another (traditional neighborhood-based prediction). The event-based framework was developed because it reflected how NWS forecasters interpreted WoFS forecasts. It is designed to be closer to the warning side of watch-to-warning (Flora et al. 2019; Wilson et al. 2019a). SPC forecasters, however, operate on larger spatiotemporal scales closer to the watch side of watch-to-warning. It is encouraging that a product meant for one end user appears useful to another with different needs.

When product developers gain users' feedback on their products, they gain important insights into improving the guidance for the end user (Demuth et al. 2020). Keeping the end user in mind is crucial as care must be taken to introduce new tools into the forecast process with minimal negative impact (Stuart et al. 2006). We have identified several ways to improve upon the WoFS-ML-Severe guidance:

- Improving the ensemble storm-track identification: Some users have concerns about the ensemble storm-track area. The current tracks show the full ensemble spread in storm location. Detailing the full spread is not always useful; instead, we can tailor the spread to display a complementary "focused" area.
- Training on verification data other than human reports: Human reports are biased in numerous ways (McGovern et al. 2022), and instead, using remotely sensed data like

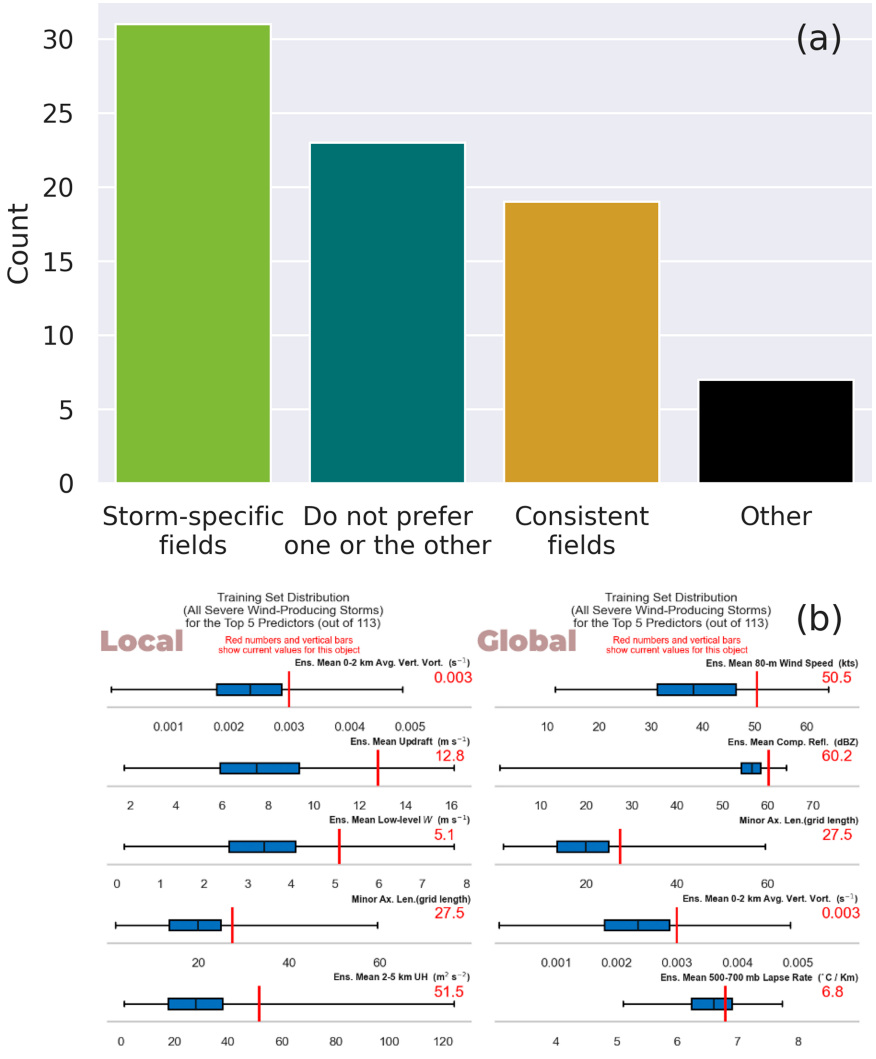

FIG. 17. (a) Participant responses to the question "Would you prefer consistent fields (explaining how the same set of predictors contribute to the prediction regardless of the storm) or storm-specific fields (using a different set of predictors contributing to each storm's prediction)?" An other response with a write-in was also available. Responses in this category are discussed in the text. (b) An example of (left) local and (right) global sets of predictors for the explainability graphics. Local predictor fields would change depending on the storm object, while global predictor fields would remain the same between objects. Participants were shown this image before answering which set of predictors they preferred.

radar-derived hail size or even NWS warning polygons alleviate many of those biases. However, these other data sources also introduce biases of their own.

- Providing conditional intensity: Users have asked for a complementary prediction of severe weather magnitude, i.e., maximum expected wind speed or hail size.
- Highlighting well-assimilated storms: Guerra et al. (2022) found that WoFS accuracy improves with more data assimilation cycles. Identifying ensemble storm tracks associated with well-assimilated storms could improve forecasters' trust in the guidance.

One limitation of the traditional HWT-SFE framework is that outlooks are generated early in the day (e.g., valid between 2000 and 2200 UTC), while the most hazardous convection occurs later (e.g., 2300–0300 UTC). The WoFS system has the advantage of becoming more accurate after several data assimilation cycles following the initiation of storms (Guerra et al. 2022; Gallo et al. 2022). Ongoing real-time collaboration with NWS forecasters helps us to evaluate the usefulness of the ML guidance at these later times.

One potential avenue for future research involves investigating the use of spatial probabilities compared to storm-based probabilities for forecasters throughout the watch-to-warning decision-making process. The current approach in the Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) program uses neighborhood-based

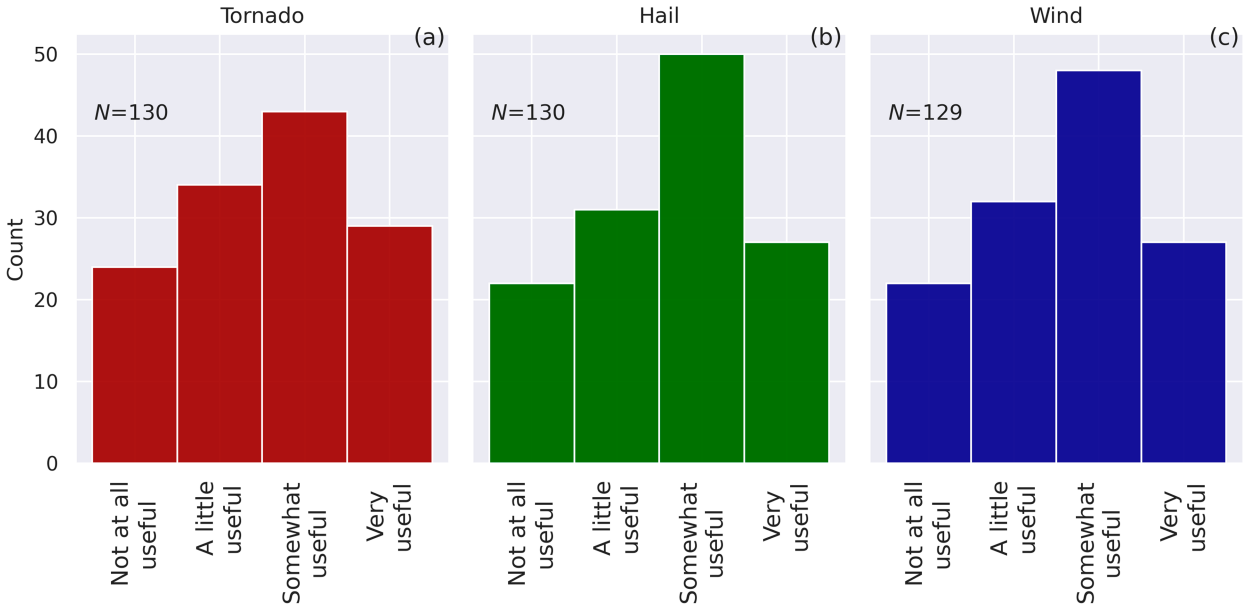

FIG. 18. Participant responses to the question "How useful were the explainability graphic when creating forecasts of the following hazards today (considering both the 2100–2200 UTC and the 2200–2300 UTC periods)?"

probabilities at all lead times. However, at shorter lead times, forecasters are more accustomed to relying on storm-specific guidance, such as the ProbSevere system (Cintineo et al. 2020). Especially as the WoFS transitions to operations, there is an increasing need to couple complementary event-based and grid-based guidance in the watch-to-warning space.

In ongoing work, we want to expand the types of WoFS-ML-Severe and explainability products. The explainability products will be modified to include additional context information, and we plan to explore additional ways to provide global and local explanations. Any-severe (matched to any severe weather report) and any-significant-severe weather [matched to any significant severe weather report (EF2+ tornadoes, ≥2-in. hail, ≥60-kt wind)] products are also being developed to complement the existing WoFS-ML-Severe suite. We will build on the findings of this experiment to improve training materials and increase forecaster exposure to leverage the full benefits of WoFS and WoFS-ML-Severe.

Acknowledgments. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under the NOAA-University of Oklahoma Cooperative Agreement NA2OAR4320204, U.S. Department of Commerce. This material is based upon work supported by the Joint Technology Transfer Initiative Program within the NOAA/OAR Weather Program Office under Award NA22OAR4590171. We thank the participants of this experiment for their efforts and all those who contributed to SFE 2022. The authors thank Patrick Burke for informally reviewing an early version of the manuscript. The authors also acknowledge the team members responsible for generating the experimental WoFS output, which includes Joshua Martin, Kent Knopfmeier, Brian Matilla, Thomas Jones, Patrick Skinner, Brett Roberts, Nusrat Yussouf, and David Dowell. We also thank Mariana Cains for helping with the literature review on user-centered product design.

Data availability statement. The nonidentifiable data collected in this study (i.e., the experimental outlook forecasts and participant responses) will be made available upon request and free of charge following a reasonable period for data analysis and publishing (approximately 2 years).

REFERENCES

Abras, C., D. Maloney-Krichmar, and J. Preece, 2004: User-centered design. *Encyclopedia of Human-Computer Interaction*, W. Bainbridge, Ed., Sage Publications, 445–456.

Argyle, E. M., J. J. Gourley, Z. L. Flamig, T. Hansen, and K. Manross, 2017: Towards a user-centered design of a weather forecasting decision support tool. *Bull. Amer. Meteor. Soc.*, **98**, 373–382, https://doi.org/10.1175/BAMS-D-16-0031.1.

Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, https://doi.org/10.1175/WAF933.1.

Bayer, S., H. Gimpel, and M. Markgraf, 2022: The role of domain expertise in trusting and following explainable AI decision support systems. *J. Decis. Syst.*, **32**, 110–138, https://doi.org/10.1080/12460125.2021.1958505.

Bird, S., E. Klein, and E. Loper, 2009: *Natural Language Processing with Python*. O'Reilly Media Inc., 504 pp.

Boyd, K., K. H. Eng, and C. D. Page, 2013: Area under the precision-recall curve: Point estimates and confidence intervals. *Machine Learning and Knowledge Discovery in Databases*, H. Blockeel et al., Eds., Springer, 451–466.

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

Cains, M. G., and Coauthors, 2023: Exploring what AI/ML guidance features NWS forecasters deem trustworthy. *22nd Conf. on Artificial Intelligence for Environmental Science*, Denver, CO, Amer. Meteor. Soc., 8A.2, https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/419371.

Calvo, L., I. Christel, M. Terrado, F. Cucchietti, and M. Pérez-Montoro, 2022: Users' cognitive load: A key aspect to successfully communicate visual climate information. *Bull. Amer. Meteor. Soc.*, **103**, E1–E16, https://doi.org/10.1175/BAMS-D-20-0166.1.

Chen, T., and C. Guestrin, 2016: XGBoost: A scalable tree boosting system. arXiv, 1603.02754v3, https://doi.org/10.48550/arXiv.1603.02754.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020: NOAA ProbSevere v2.0—ProbHail, ProbWind, and ProbTor. *Wea. Forecasting*, **35**, 1523–1543, https://doi.org/10.1175/WAF-D-19-0242.1.

Clark, A. J., and Coauthors, 2023: The third real-time, virtual spring forecasting experiment to advance severe weather prediction capabilities. *Bull. Amer. Meteor. Soc.*, **104**, E456–E458, https://doi.org/10.1175/BAMS-D-22-0213.1.

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.

Ehrmann, D. E., S. N. Gallant, S. Nagaraj, S. D. Goodfellow, D. Eytan, A. Goldenberg, and M. L. Mazwi, 2022: Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. *Nat. Med.*, **28**, 1331–1333, https://doi.org/10.1038/s41591-022-01833-z.

Flora, M. L., C. K. Potvin, A. McGovern, and S. Handler, 2024: A machine learning explainability tutorial for atmospheric sciences. *Artif. Intell. Earth Syst.*, **3**, e230018, https://doi.org/10.1175/AIES-D-23-0018.1.

——, P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental warn-on-forecast system. *Wea. Forecasting*, **34**, 1721–1739, https://doi.org/10.1175/WAF-D-19-0094.1.

——, 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.

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

——, and Coauthors, 2022: Exploring the watch-to-warning space: Experimental outlook performance during the 2019 spring forecasting experiment in NOAA's hazardous weather testbed. *Wea. Forecasting*, **37**, 617–637, https://doi.org/10.1175/WAF-D-21-0171.1.

——, A. J. Clark, I. Jirak, D. Imy, B. Roberts, J. Vancil, K. Knopfmeier, and P. Burke, 2024: WoFS and the wisdom of the crowd: The impact of the warn-on-forecast system on hourly forecasts during the 2021 NOAA hazardous weather testbed spring forecasting experiment. *Wea. Forecasting*, **39**, 485–500, https://doi.org/10.1175/WAF-D-23-0033.1.

Guerra, J. E., P. S. Skinner, A. Clark, M. Flora, B. Matilla, K. Knopfmeier, and A. E. Reinhart, 2022: Quantification of NSSL warn-on-forecast system accuracy by storm age using object-based verification. *Wea. Forecasting*, **37**, 1973–1983, https://doi.org/10.1175/WAF-D-22-0043.1.

Gunning, D., E. Vorm, J. Y. Wang, and M. Turek, 2021: DARPA's explainable AI (XAI) program: A retrospective. *Appl. AI Lett.*, **2**, e61, https://doi.org/10.1002/ail2.61.

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.

Hepper, R. M., I. L. Jirak, and J. M. Milne, 2016: Assessing the skill of convection-allowing ensemble forecasts of severe MCS winds from the SSEO. Preprints, *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 16B.2, https://ams.confex.com/ams/28SLS/webprogram/Paper300134.html.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **148**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

——, 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.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, https://doi.org/10.1175/WAF-D-12-00113.1.

Hoffman, R. R., M. Johnson, J. M. Bradshaw, and A. Underbrink, 2013: Trust in automation. *IEEE Intell. Syst.*, **28**, 84–88, https://doi.org/10.1109/MIS.2013.24.

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

Hsu, W.-r., and A. H. Murphy, 1986: The attributes diagram a geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecast.*, **2**, 285–293, https://doi.org/10.1016/0169-2070(86)90048-8.

Jirak, I. L., C. J. Melick, and S. J. Weiss, 2014: Combining probabilistic ensemble information from the environment with simulated storm attributes to generate calibrated probabilities of severe weather hazards. Preprints. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 2.5, https://ams.confex.com/ams/27SLS/webprogram/Paper254649.html.

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

——, and Coauthors, 2020: Assimilation of *GOES-16* radiances and retrievals into the warn-on-forecast system. *Mon. Wea. Rev.*, **148**, 1829–1859, https://doi.org/10.1175/MWR-D-19-0379.1.

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

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

——, and Coauthors, 2018: Development of a human–machine mix for forecasting severe convective events. *Wea. Forecasting*, **33**, 715–737, https://doi.org/10.1175/WAF-D-17-0188.1.

——, R. Clark III, I. L. Jirak, P. T. Marsh, R. Schneider, and S. J. Weiss, 2019: Enhancements to storm prediction center convective outlooks. *Ninth Conf. on Transition of Research to Operations*, Phoenix, AZ, Amer. Meteor. Soc., J7.3, https://ams.confex.com/ams/2019Annual/webprogram/Paper355037.html.

King, N., 2004: Using templates in the thematic analysis of text. *Essential Guide to Qualitative Methods in Organizational Research*, C. Cassell and G. Symon, Eds., Sage Publications, 257–270.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Wea. Forecasting*, **32**, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1.

——, ——, 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.

Likert, R., 1932: A technique for the measurement of attitudes. *Arch. Psychol.*, **22**, 55.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Wea. Forecasting*, **35**, 1605–1631, https://doi.org/10.1175/WAF-D-19-0258.1.

McGovern, A., R. Lagerquist, D. J. Gagne II, G. E. 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.

——, I. Ebert-Uphoff, D. J. Gagne II, and A. Bostrom, 2022: Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environ. Data Sci.*, **1**, e6, https://doi.org/10.1017/eds.2022.5.

——, R. J. Chase, M. Flora, D. J. Gagne II, R. Lagerquist, C. K. Potvin, N. Snook, and E. Loken, 2023: A review of machine learning for convective weather. *Artif. Intell. Earth Syst.*, **2**, e220077, https://doi.org/10.1175/AIES-D-22-0077.1.

Metz, C. E., 1978: Basic principles of ROC analysis. *Semin. Nucl. Med.*, **8**, 283–298, https://doi.org/10.1016/S0001-2998(78)80014-2.

Mueller, A., 2023: word_cloud: A little word cloud generator in python. GitHub, https://github.com/amueller/word_cloud.

Nourani, M., J. T. King, and E. D. Ragan, 2020: The role of domain expertise in user trust and the impact of first impressions with intelligent systems. arXiv, 2008.09100v1, https://doi.org/10.48550/arXiv.2008.09100.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

——, 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.

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

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-4751STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Sobash, R. A., G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Wea. Forecasting*, **35**, 1981–2000, https://doi.org/10.1175/WAF-D-20-0036.1.

Stewart, T. R., W. R. Moninger, K. F. Heideman, and P. Reagan-Cirincione, 1992: Effect of improved information on the components of skill in weather forecasting. *Organ. Behav. Hum. Decis. Processes*, **53**, 107–134, https://doi.org/10.1016/0749-5978(92)90058-F.

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.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting*, **21**, 408–415, https://doi.org/10.1175/WAF925.1.

Wilson, K. A., P. L. Heinselman, P. S. Skinner, J. J. Choate, and K. E. Klockow-McClain, 2019a: Meteorologists' interpretations of storm-scale ensemble-based forecast guidance. *Wea. Climate Soc.*, **11**, 337–354, https://doi.org/10.1175/WCAS-D-18-0084.1.

——, and Coauthors, 2019b: Exploring applications of storm-scale probabilistic warn-on-forecast guidance in weather forecasting. *Virtual, Augmented and Mixed Reality. Applications and Case Studies*, J. Chen and G. Fragomeni, Eds., Lecture Notes Computer Science, Vol. 11575, Springer, 577–572.

——, B. T. Gallo, P. Skinner, A. Clark, P. Heinselman, and J. J. Choate, 2021: Analysis of end user access of warn-on-forecast guidance products during an experimental forecasting task. *Wea. Climate Soc.*, **13**, 859–874, https://doi.org/10.1175/WCAS-D-20-0175.1.

——, and Coauthors, 2023: The NOAA weather prediction center's use and evaluation of experimental warn-on-forecast system guidance. *J. Oper. Meteor.*, **11**, 82–94, https://doi.org/10.15191/nwajom.2023.1107.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. W. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 286–303, https://doi.org/10.1175/1520-0434(1998)013<0286:AEHDAF>2.0.CO;2.