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

BURKELY T. GALLO,[a,b] ADAM J. CLARK,[c,d] ISRAEL JIRAK,[b] DAVID IMY,[c] BRETT ROBERTS,[a,b,c] JACOB VANCIL,[a,b] KENT KNOPFMEIER,[a,c] AND PATRICK BURKE[c]

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

(Manuscript received 24 February 2023, in final form 2 January 2024, accepted 9 January 2024)

ABSTRACT: During the 2021 Spring Forecasting Experiment (SFE), the usefulness of the experimental Warn-on-Forecast System (WoFS) ensemble guidance was tested with the issuance of short-term probabilistic hazard forecasts. One group of participants used the WoFS guidance, while another group did not. Individual forecasts issued by two NWS participants in each group were evaluated alongside a consensus forecast from the remaining participants. Participant forecasts of tornadoes, hail, and wind at lead times of ~2–3 h and valid at 2200–2300, 2300–0000, and 0000–0100 UTC were evaluated subjectively during the SFE by participants the day after issuance, and objectively after the SFE concluded. These forecasts exist between the watch and the warning time frame, where WoFS is anticipated to be particularly impactful. The hourly probabilistic forecasts were skillful according to objective metrics like the fractions skill score. While the tornado forecasts were more reliable than the other hazards, there was no clear indication of any one hazard scoring highest across all metrics. WoFS availability improved the hourly probabilistic forecasts as measured by the subjective ratings and several objective metrics, including increased POD and decreased FAR at high probability thresholds. Generally, expert forecasts performed better than consensus forecasts, though expert forecasts overforecasted. Finally, this work explored the appropriate construction of practically perfect fields used during subjective verification, which participants frequently found to be too small and precise. Using a Gaussian smoother with $\sigma = 70$ km is recommended to create hourly practically perfect fields in future experiments.

SIGNIFICANCE STATEMENT: This work explores the impact of cutting-edge numerical weather prediction ensemble guidance (the Warn-on-Forecast System) on severe thunderstorm hazard outlooks at watch-to-warning time scales, typically between 1 and 6 h of lead time. Real-time forecast products in this time frame are currently provided on an as-needed basis, and the transition to continuous probabilistic forecast products across scales requires targeted research. Results showed that hourly probabilistic participant forecasts were skillful subjectively and statistically, and that the experimental guidance improved the forecasts. These results are promising for the implementation and value of the Warn-on-Forecast System to provide improved hazard timing and location guidance within severe weather watches. Suggestions are made to aid future subjective evaluations of watch-to-warning-scale probabilistic forecasts.

KEYWORDS: Ensembles; Forecast verification/skill; Forecasting; Mesoscale forecasting; Short-range prediction; Numerical weather prediction/forecasting

1. Introduction

Probabilistic forecasting is becoming increasingly common in the weather community as it shifts toward the Forecasting A Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) paradigm, in which a continuous flow of probabilistic information is envisioned. Forecasters heavily rely on radar and satellite observational trends for short-term probabilistic severe weather forecasts. Extrapolation studies stretch back to 1953 (Ligda 1953), with nowcasts first being issued by Bellon and Austin (1978). Throughout the history of radar data extrapolation techniques, prior work has shown that extrapolation of radar data is more skillful than numerical weather prediction (NWP) at forecasting precipitation intensity up to about 2 h in advance (Sun et al. 2014; WMO 2017; Wilson et al. 2020), after which NWP that assimilates radar data shows better skill. The skill of extrapolation versus NWP, however, can depend on factors such as scale, forcing, and topography (Wilson et al. 1998; Germann and Zawadzki 2002; Keil et al. 2014), with NWP handling synoptic forcing well, but struggling to handle smaller-scale features that radar can detect (e.g., gust fronts). The time frame in which the highest skill shifts from extrapolation to NWP lies between the watch and the warning scale, suggesting that this scale would benefit from probabilities to reflect the inherent uncertainty.

Currently, probabilistic forecasts are operationally issued by national centers within the National Weather Service (NWS)

Gallo's current affiliation: 16 WS, USAF, Offutt AFB, Nebraska.

Corresponding author: Burkely T. Gallo, burkely.gallo@us.af.mil

DOI: 10.1175/WAF-D-23-0033.1

such as the Storm Prediction Center (SPC) and the Weather Prediction Center (WPC) for severe convective storms and precipitation, respectively, and by local NWS Weather Forecast Offices (WFOs) for hazards such as snowfall and the probability of precipitation. Other products encompassing severe convective storms are mostly binary, such as severe thunderstorm and tornado watches and warnings. Between the watch and the warning time frames, SPC issues mesoscale convective discussions (MCDs) on an as-needed basis and local WFOs communicate with partners and provide Impact-based Decision Support Services (IDSS). However, the tools and framework of formal products to support the FACETs paradigm are not yet present in the watch-to-warning space, leaving uncertainty to be expressed only through words and forecaster intuition. This work explores the issuance of experimental hourly probabilistic forecasts at scales between the convective watch and warning, which is defined herein as 1–4 h of lead time, using guidance from an experimental ensemble of convection-allowing models (CAMs).

Much of the short-term NWP studied and available operationally focuses on the next-day time frame, with forecasts issued up to 36 or 48 h in the future. The High-Resolution Rapid Refresh, version 4 (HRRRv4; Dowell et al. 2022; James et al. 2022) hourly forecasts run to at least 18 h, including and extending beyond the watch-to-warning time frame. This guidance provides valuable insight and specificity for storm location, mode, and other convective attributes, and includes data assimilation. However, the Warn-on-Forecast System (WoFS; Stensrud et al. 2009, 2013) run at the National Severe Storms Laboratory, is an ensemble with 3-km horizontal grid spacing with *rapid* assimilation of multiple radar and satellite fields and conventional observations every 15 min. This rapid data assimilation enables forecasts of individual, ongoing convective storms and associated severe weather hazards. Forecasts run to 6 h at the top of each hour and to 3 h at the bottom of each hour. WoFS forecasts contain 18 ensemble members at the 3-km horizontal grid spacing common to many CAMs. A suite of probabilistic products is output from each run, including fields such as the probability of a specified event within neighborhoods of varying sizes around each grid point or percentile values of fields like hail size, 2–5-km updraft helicity (UH; Kain et al. 2008), and 10-m wind speed. This system is designed for a more skillful forecast of severe storm evolution with a few hours of lead time, particularly when the storm has existed for a few hours prior to the forecast initialization (Guerra et al. 2022).

WoFS has been evaluated and used in annual NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments (SFEs; Gallo et al. 2017; Clark et al. 2012; Kain et al. 2003a) since 2017. SFEs in the HWT bring together researchers, operational forecasters, and model developers to evaluate cutting-edge CAM guidance, postprocessing methods, and forecasting techniques. Introduction of the WoFS to the SFE has allowed for short-term experimental forecasting, which has largely focused on creating hourly probabilistic forecasts of varying types of severe convective weather. Initially, these short-term forecasts did not distinguish between hazard type, but since SFE 2020 participants have issued separate forecasts of tornadoes, hail, and wind. Time frames over which the forecasts are valid have varied but have remained within the watch-to-warning space. This work extends the usage of the WoFS in the HWT by directly examining the impact of WoFS on outlooks from forecasters issuing the same set of experimental forecasts, allowing us to determine the effect of WoFS at watch-to-warning scales.

A key component of the SFE involves next-day subjective evaluation exercises (Kain et al. 2003b). In these exercises, participants subjectively evaluate the performance of CAM guidance (including WoFS guidance) and their own experimental forecasts. Through both issuing forecasts using experimental guidance and completing evaluations, participants gain a better understanding of the strengths and weaknesses of different guidance. Insights gleaned from the subjective evaluation can illuminate aspects of the underlying CAM guidance, such as systematic biases in handling certain convective modes (e.g., cells growing upscale too quickly relative to observations), or can illuminate aspects of the forecast exercises themselves, such as workload concerns (e.g., if the generation of a new forecast type is likely too intensive to fit into the existing workflow of a WFO or SPC). Combining the objective and subjective verification techniques provides a holistic picture of the performance of both the guidance and the new forecasting techniques (Gallo et al. 2016, 2021; Miller et al. 2021), and paves the way for smoother operational implementation.

This work will examine the skill of hourly probabilistic forecasts issued by participants in the 2021 SFE, including determining how the skill differs between forecasts issued by participants with and without access to WoFS. A secondary focus of this work considers the skill of forecasts issued by individual operational forecasters relative to the skill of consensus forecasts from a group of meteorological professionals who were not necessarily operational forecasters by trade. By evaluating these experimental forecasts objectively and subjectively, we aim to answer the following questions: 1) How skillful are the hourly probabilistic forecasts? 2) How does WoFS impact the hourly probabilistic forecasts? 3) Are skill differences a function of the hazard being forecasted? 4) How does a consensus forecast compare to those of expert forecasters? Through the course of the 2021 SFE, during subjective evaluations, a fifth question arose: 5) What size of practically perfect forecast should be used for subjective evaluation?

Section 2 describes the methodology of this study, and section 3 covers the results from the subjective evaluations followed by the objective verification results. Exploration of how to verify the forecasts subjectively using practically perfect fields (Hitchens et al. 2013) composes the final part of the results, followed by the conclusions and directions for future work in section 4.

2. Methodology

a. *The Warn-on-Forecast System*

The 2021 SFE marks the fifth year of WoFS in the SFE, both in an evaluation context and to issue experimental

TABLE 1. Cases examined in the objective verification portion of this study. Bold numbers indicate that subjective verification also took place for these dates.

Week 1	Week 2	Week 3	Week 4	Week 5
3, **4**, **5**, **6**, 7 May 2021	**10**, **12**, **13**, 14 May 2021	**17**, **18**, **19**, **20**, 21 May 2021	**24**, **25**, **26**, **27**, 28 May 2021	**1**, **2**, **3**, 4 Jun 2021

forecast types such as potential severe timing areas (Krocak 2020) and hourly probabilities of severe weather (Wilson et al. 2021, hereafter W21; Gallo et al. 2022, hereafter G22). WoFS forecasts launched at the top (bottom) of every hour generate forecasts to 6 (3) hours, with a latency of approximately 25 min to the arrival of the first data on the web viewer. Typically, all data are available by 45 min after the run initialization. Each day during SFE 2021 WoFS was first launched at 1700 UTC, and the final forecast was launched at 0300 UTC. See Jones et al. (2020) for a detailed description of the WoFS configurations during the 2021 SFE; configurations have been relatively consistent and object-based comparisons of year-to-year reflectivity forecasts have not revealed large changes in system performance (Guerra et al. 2022). Forecasts from the 2021 season can be found on the legacy WoFS viewer.[1]

b. *Experiment setup*

The experimental setup herein mirrors that of W21 and G22, with a few important differences. First, the experiment described herein took place during the last 2 h of the 2021 SFE (Gallo et al. 2017; Clark et al. 2022), from ~1415 to 1600 CDT, an abbreviated time relative to W21 and G22. Participants were also drawn from the researcher, modeler, academic, and operational forecaster participants of the 2021 SFE, and as such were not selected specifically for this experiment. While some SFE participants were not operational forecasters, all participants had the background of the full day's SFE activities prior to their issuance of these forecasts to help them acclimate to the forecast challenges of the day. Finally, in W21 and G22 the hourly probabilistic forecasts were issued for the tornado, hail, and convective wind threat jointly, whereas this work had participants issue separate tornado, wind, and hail probabilistic forecasts.

Throughout the five-week SFE, each participant issued their forecasts using web-based drawing tools. The SFE ran on weekdays from 3 May to 4 June 2021, excepting the Memorial Day holiday on 31 May 2021. WoFS data were available for all but one of those days, leading to 23 cases considered herein for objective verification (Table 1). For the forecasting activity, participants were divided into two groups. One group had access to and drew their probabilistic forecast contours over WoFS data (the WoFS group), and one group did not have access to WoFS data (the No WoFS group), instead drawing their forecasts in the SFE drawing tool. See Table 2 for contour levels available for participants to draw. The SFE drawing tool provided access to other data, such as hourly HRRRv4 (Dowell et al. 2022; James et al. 2022) forecasts, experimental CAM guidance initialized at 0000 and 1200 UTC, and radar data. Forecasters in the WoFS group could also access this model data and available diagnostic data. Two NWS forecasters in each of the WoFS and the No WoFS groups had their forecasts displayed individually the following day for subjective verification, and will be referred to hereafter as the "expert forecasts." Other participant forecasts were aggregated into a consensus forecast via the methodology described in the next section and will be referred to as the "consensus forecasts." Subjective verification took place for 18 of the 23 cases that were objectively verified (Table 1).

From 1415 to 1515 CDT (1915–2015 UTC), participants issued two initial forecasts of tornadoes, hail, and wind, valid 2200–2300 UTC and 2300–0000 UTC. Then, from 1515 to 1600 CDT (2015–2100 UTC), participants updated their forecasts valid 2200–2300 UTC and 2300–0000 UTC and issued individual hazard forecasts valid from 0000 to 0100 UTC. These valid times are 1–3 h from issuance time and were selected based on previous experience suggesting that the WoFS would add more value to forecasts with at least a lead time of 1 h, since participants would not be able to simply wait until near the end of the forecast issuance time and extrapolate the radar imagery to generate their forecasts. The later bounds of the forecasts were constrained by the availability of WoFS; participants typically used the 1900 and 2000 UTC runs of the WoFS to issue their forecasts (though they could also use data from 1700 UTC, 1800 UTC, and bottom-of-the-hour runs if they so desired). The 1900 UTC forecasts extend to 0100 UTC, therefore data from both the 1900 and 2000 UTC WoFS initializations could be used to issue all sets of forecasts. Participants could load their prior forecasts into the drawing tool as a starting point to update their forecasts (for the two recurring forecast periods), lessening the workload and allowing for three sets of individual hazard forecasts to be issued in the second part of the activity each afternoon. The three sets of forecasts issued from 1515 to 1600 CDT were analyzed using objective verification techniques described in the following section.

c. *Verification*

1) OBJECTIVE VERIFICATION

Participant forecasts were regridded from the native JSON files containing locations for each polygon vertex to the WoFS grid for verification. Gridded fields from the individual expert forecasters contained stepwise probabilities (e.g., each grid point within the 15% contour was assigned a value of 15% except for at grid points within that contour that were also within a higher probability contour such as 30%). Participant forecasts bound for the consensus were interpolated into a continuous field prior to averaging, but contour levels displayed after averaging were solely those available to draw. These consensus forecasts were then regridded to the WoFS grid, and were evaluated only at the probabilistic levels available for participants

[1] https://wof.nssl.noaa.gov/realtime/.

TABLE 2. Contour levels available for participants to draw for each hazard.

Hazard	Tornado	Wind	Hail
Contour levels	2%, 3%, 5%, 8%, 10%, 13%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%	5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%	5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%

when drawing their probabilistic contours. This approach ensured appropriate comparisons were made between the expert and consensus forecasts.

Tornado, wind, and hail reports from NCDC's *Storm Data* storm event database[2] were used to verify forecasts of their respective hazards. Reports were gridded to a 39-km neighborhood prior to verification, so each grid point within a 39-km radius of a report was considered a "hit." This definition approximates the current SPC convective outlook probabilistic definitions of severe weather within 40 km (~25 mi) of a point within the constraints of a 3-km grid. Only reports within the WoFS domain were considered. The number of *Storm Data* reports occurring within the WoFS domain across the 2021 SFE cases are detailed in Table 3.

Forecasts were evaluated utilizing areas under the receiver operating curve (AUC; Mason 1982), performance diagrams (Roebber 2009), fractions skill scores (FSS; Roberts and Lean 2008), and reliability diagrams. Initial FSSs were calculated following Roberts et al. (2020), W21, and G22, with a similar approach as Schwartz et al. (2010). This approach uses binary observations. Calculating the FSS in this manner bears some relation to the Brier score (Brier 1950), but prevents a muddying of results by eliminating the variable of the smoothing radius of observations in the practically perfect methodology (Hitchens et al. 2013). A secondary goal of this paper is to determine the optimal practically perfect neighborhood and smoothing radius for hourly probabilistic convective forecasts, which will be detailed in the following section.

2) PRACTICALLY PERFECT CALCULATIONS

Practically perfect contours (Hitchens et al. 2013) were created surrounding the available preliminary severe local storm reports (LSRs) during the 2021 SFE for real-time subjective verification purposes, using a Gaussian smoother with a kernel width of σ = 40 km. This level of smoothing was determined subjectively and used previously in SFEs by halving the radius used for experimental 4-h forecasts (σ = 80 km), which was in turn 40 km less than the radius used for 24-h forecasts (σ = 120 km). Since the definition of these probabilities was for a hazard within 25 mi (~40 km) of a point, the smoothing radius used decreased the σ by approximately one grid box per neighborhood as the time scale decreased.

During discussions in the SFE, it was frequently mentioned that the practically perfect contours did not seem to fit the original intention behind the practically perfect methodology—essentially, a forecaster would not be able to forecast reports with as high specificity as was displayed with the practically perfect probabilities. As such, postexperiment practically perfect fields were generated using neighborhood sizes varying from 0 to 39 km in 3-km increments, and smoothing radii varying from 10 to 120 km in 10-km increments. FSSs were then calculated using the expert forecasts from each hazard and the practically perfect fields with the goal of determining which neighborhood and σ combination most closely matched the expert forecasts generated. FSSs were calculated separately for WoFS expert forecasts and No WoFS expert forecasts in case access to the WoFS guidance impacted the optimal smoothing and neighborhood size of the hourly practically perfect fields.

3) SUBJECTIVE VERIFICATION

Next-day subjective verification took place following the forecasts issued Monday through Thursday. Participants used preliminary LSRs, Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) maximum estimated size of hail (MESH; Witt et al. 1998), and severe thunderstorm and tornado warnings to subjectively evaluate the prior day's forecasts. Forecasts issued by both groups were displayed simultaneously, with two expert forecasts and one consensus forecast from each group per row (Fig. 1). Participants were instructed to primarily use LSRs in their verification, with other datasets supplementing those impressions since it is recognized that LSRs occasionally take multiple days to be reported to the National Weather Service and transmitted to SPC, and therefore were not always available for subjective verification.

Practically perfect contours using a Gaussian smoother with width σ = 40 km were used for subjective verification (Fig. 2). At this level of smoothing, one report would create a practically perfect value of ~34%. As such, participants were instructed to draw at least a 30% contour in areas where they felt pretty sure that reports would occur, but to use the 5% and 15% contours as well for areas of lower confidence. Since the ideal practically perfect radius for this time and space scale was unknown during the 2021 SFE, the participants were asked to consider the raw LSR locations and density more so than trying to match specific contour values to the practically perfect contours when completing their subjective evaluations, but to consult them if they wanted to know what sorts of maximum probability values were "reasonable" according to the practically perfect methodology.

TABLE 3. Number of reports utilized for objective verification for each hazard and time frame examined herein.

	Tornado	Hail	Wind
2200–2300 UTC	13	80	52
2300–0000 UTC	16	116	47
0000–0100 UTC	7	98	49

[2] Available at https://www.ncdc.noaa.gov/stormevents/.

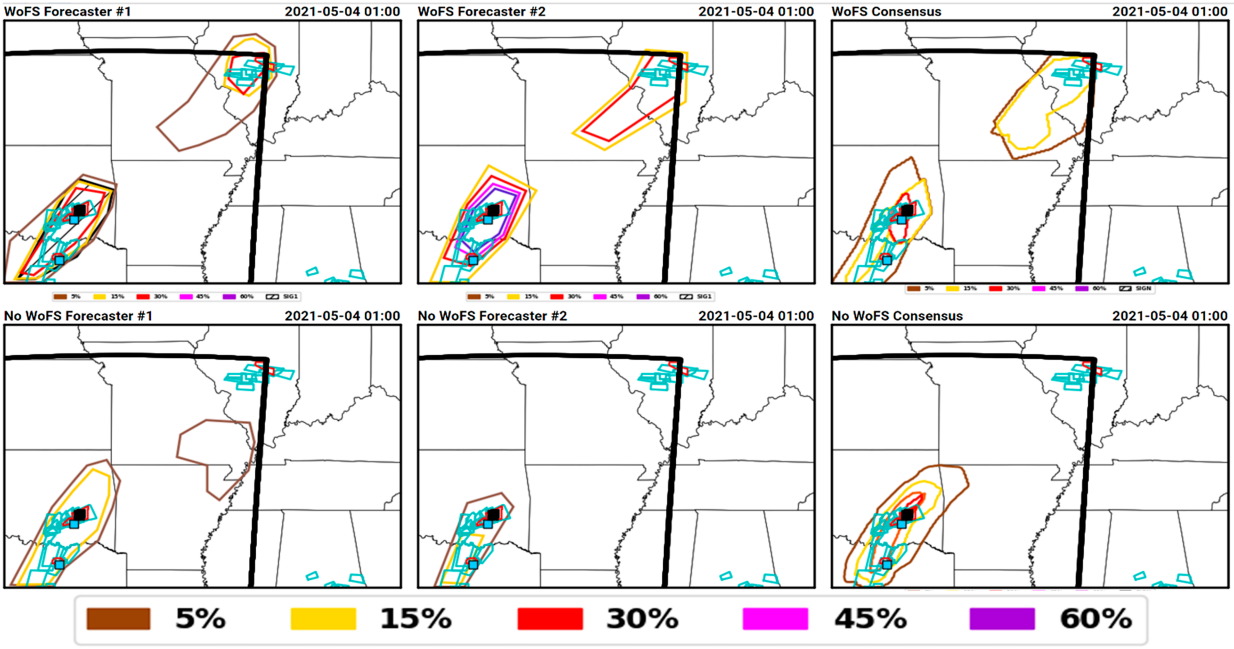

FIG. 1. Probabilistic wind forecast contours issued by (top) WoFS group participants and (bottom) No WoFS group participants, valid at 0000–0100 UTC 4 May 2021. WoFS domain boundaries are shown by the thick black lines. (left),(center) Individual expert forecasts and (right) consensus forecasts are shown. Blue squares indicate observed local storm reports of severe convective wind, and black squares indicate observed significant (≥74 mph) wind reports.

During the participant subjective evaluation, participants were asked to sort either the initial or the final forecasts into the following categories: excellent, above average, average, below average, poor, and forecast unavailable. Participants sorted separate forecasts for each tornado, hail, and wind forecast at each hour, resulting in each participant completing nine sorting tasks per day. Participant responses were converted into a 1–5 rating for analysis, with "1" representing "poor" and "5" representing "excellent." "Forecast unavailable" responses were excluded from analysis. All participants were asked to evaluate both the individual expert forecasts and the consensus forecasts.

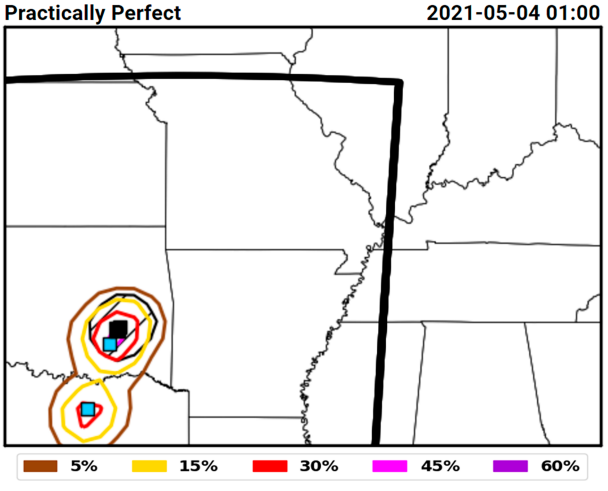

FIG. 2. An example of a practically perfect field for wind valid at 0000–0100 UTC 4 May 2021. Wind reports are overlaid as blue squares, significant wind reports are overlaid as black squares, and the colored contours show the practically perfect verification contours. The black hatched contour shows a practically perfect significant wind contour (not analyzed in this study). The thick black line shows the WoFS daily domain.

3. Results

a. *Subjective evaluations*

Participant ratings of the final forecasts issued over 2015–2100 UTC revealed differences between the expert WoFS forecasts and the expert No WoFS forecasts (Fig. 3). Consensus forecasts did not show large differences in mean ratings between the WoFS and the No WoFS groups for any hazard. The WoFS expert forecasts, however, achieved a higher mean rating for each hazard than the No WoFS expert forecasts. Wind forecasts saw the largest improvement in the mean rating, though the overall rating was lower than the hail and tornado WoFS group forecasts. Ratings were highest overall for the tornado forecasts, though tornadoes were the least frequent hazard observed during SFE 2021. Thus, many of these ratings may have rewarded low-probability tornado forecasts when tornadoes did not occur (e.g., "correct null" forecasts). A Welch's t test with a significance level of $\alpha = 0.05$ showed that for each hazard, the differences between the WoFS and No WoFS expert forecasts were significant. Differences

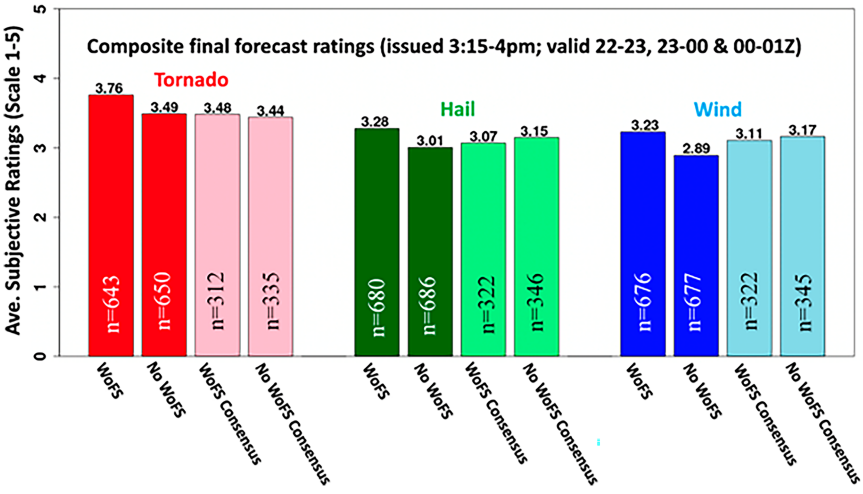

FIG. 3. Mean subjective participant ratings of the final forecasts aggregated over all hours for tornado, hail, and wind forecasts. The mean values are annotated at the top of the bar plots, and sample sizes are annotated on the individual bars.

between the consensus forecasts for both groups were not significant.

While participants found the WoFS guidance useful for all hazards, they found it to be most helpful for wind (Fig. 4). WoFS frequently provided useful guidance for hail as well, but was far less frequently indicated as useful for tornadoes. This low utility is likely due to the relatively few tornadoes occurring during the 2021 SFE, but may also be due to WoFS not highlighting the tornado threat well. An example of the utility of WoFS for wind forecasting can be shown by the case in Fig. 1, where the WoFS forecasters correctly had higher confidence in the southern area of storms, and highlighted a northern area of storms that were warned on by NWS forecasters. The No WoFS group missed the northern storms, and had lower probabilities on the southern storms (though the area was correct), leading to a better overall forecast from both the expert and the consensus WoFS forecasters.

When asked about specific products that they found most useful, participants indicated that WoFS ensemble products such as probabilities and percentiles were frequently useful (Fig. 5). Storm attribute fields such as reflectivity and 2–5-km UH, as well as explicit hazard guidance for wind speeds and hail, were the most useful fields. However, some participants did leverage products that blend the deterministic and ensemble frameworks, such as paintball plots (Roberts et al. 2019). Paintball plots enable users to visualize each ensemble member's forecast of a hazard simultaneously. While environmental fields were not commonly used, mixed-layer CAPE (MLCAPE) was said to be useful more frequently than the significant tornado parameter (STP; Thompson et al. 2003); MLCAPE may have been used to identify airmass boundaries. CAPE is utilized extensively in SFE evaluations and is required output for models to be evaluated during the SFEs (Clark et al. 2018). Hail size percentiles were also frequently used by participants, followed by wind and UH percentiles. The results found here, which show probabilities being more useful than percentiles, mirror those found in W21 from forecasters in the 2019 SFE. W21 also found high usage of the reflectivity paintballs. Together with W21, these results suggest that there is some consistency year-to-year in what products participants prefer when issuing severe convective storm forecasts using WoFS. Due to question framing, only two survey responses mentioned *why* specific products were useful, and further exploration of this question is important future work. From the limited responses here, participants liked the probability and percentile plots for summarizing ensemble information, and liked the reflectivity paintballs for visualizing convective evolution.

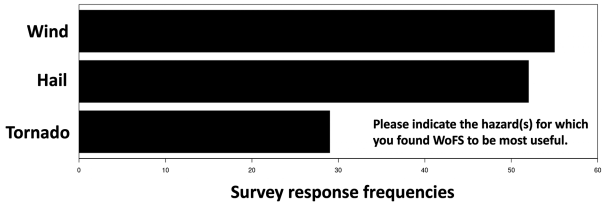

FIG. 4. WoFS group participant responses indicating for which hazard(s) they found WoFS to be most useful during the previous day. Participants could select multiple hazards.

b. *Forecast characteristics*

A brief look at the number of contours issued by each forecaster reveals that WoFS expert forecasters issued, on average, more contours than the No WoFS expert forecasters for all hazards (Fig. 6). Analysis of consensus forecast results in this manner is less useful due to the consensus algorithm formulation, and results were mixed (not shown). For tornadoes, WoFS expert forecasters issued more contours on average at all probability thresholds, while for wind and hail the WoFS forecasters issued more contours at probability thresholds of

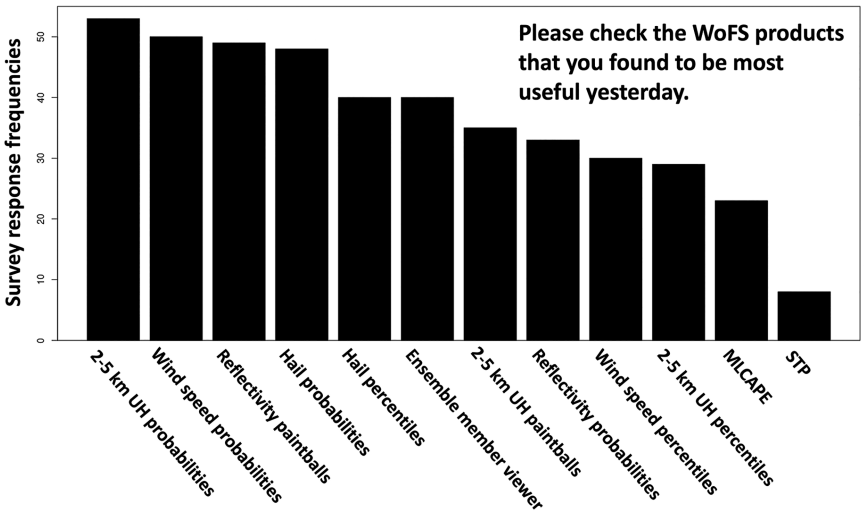

FIG. 5. WoFS group participant indications of which WoFS products were most useful when generating their forecasts the previous day. Participants could select multiple products in their responses.

15% or higher. From looking at the individual cases, there are two main mechanisms by which the WoFS experts issued more contours. Both mechanisms are demonstrated in Fig. 1. In this case, the WoFS expert forecasts had higher probabilities than the No WoFS expert forecasts. Also, the WoFS forecasters were able to pick up on multiple areas of convection. In Fig. 1, convection occurred in the northeastern corner of the domain as well as the main area of focus. Although there were no reports associated with the storms during subjective evaluation, the storms were intense enough that severe thunderstorm warnings were issued by the local NWS WFO. In other cases, multiple areas of relatively higher probability are generated by the WoFS expert forecasters within a broader-scale area of threat, while the No WoFS expert forecasts maintain a continuous area of probability. This confidence in multiple precise areas is not always warranted, as demonstrated by Fig. 7. During this 27 May event, the broad high-confidence forecast issued by one WoFS expert and the multiple areas issued by the other WoFS expert are not necessarily better than the focused, singular areas issued by the No WoFS experts. The skill of these higher-probability forecasts will be explored in the next section via objective verification.

c. *Objective verification*

Areas under the ROC curve (AUC) showed all forecasts as being skillful, with forecasts from all groups valid for all times and all hazards exceeding a score of 0.7 across the full SFE (Fig. 8). Typically, the expert forecasts performed better than the consensus forecasts, though this was not universally true for all hazards and forecast times. The tornado forecasts valid 2200–2300 UTC had the highest skill, followed by the wind forecasts valid 2300–0000 UTC. For the tornado forecasts, the WoFS expert forecasts scored the highest in AUC for all hours. The WoFS expert forecasts also scored highest for the wind and hail forecasts valid 2200–2300 UTC. These results suggest that WoFS is providing value to short-term forecasts when compared to forecasts issued without WoFS guidance. WoFS group forecasts most clearly improved for hail and tornado forecasts, particularly at longer lead times. This result differs somewhat from the subjective evaluations showing the largest improvement in the wind forecasts, and may be due to aggregating all subjective evaluation results across times for the hazards. While some improvement is also seen in wind forecasts, it is less than what is seen for the later tornado or hail forecasts. Most of the improved skill is derived from the improved POD at low probability thresholds (e.g., 2%, 5% for tornado forecasts, and 5%, 15% for wind and hail forecasts). The probability of false detection (POFD) of some of the WoFS group forecasts at the low probability threshold increases relative to the No WoFS forecasts, indicating some increased false alarm from the WoFS group forecasts. For a few hazards and hours, the consensus forecasts have higher AUCs than the expert forecasts. At these longer lead times, when uncertainty is inherently larger, the consensus forecasts may achieve higher scores by virtue of their construction.

Unlike the AUC, reliability showed better performance for the consensus forecasts relative to the expert forecasts (Fig. 9). By combining multiple individual forecasts into the consensus forecasts, extremes in individual forecasts were smoothed out, leading to overall lower probabilities across a more widespread area. Forecasts were generally overforecasts, although the tornado forecasts were more reliable than the other hazards and at times demonstrated nearly perfect reliability. Hail consensus forecasts at 2300–0000 UTC and 0000–0100 UTC were quite reliable to 15% and 30%, respectively, while the wind consensus forecasts overforecast for probabilities less than 45%. Expert forecasts all showed strong overforecasting for the hail and wind hazards, with similar performance between the WoFS and No WoFS groups. The consensus forecasts also showed relatively little difference between the

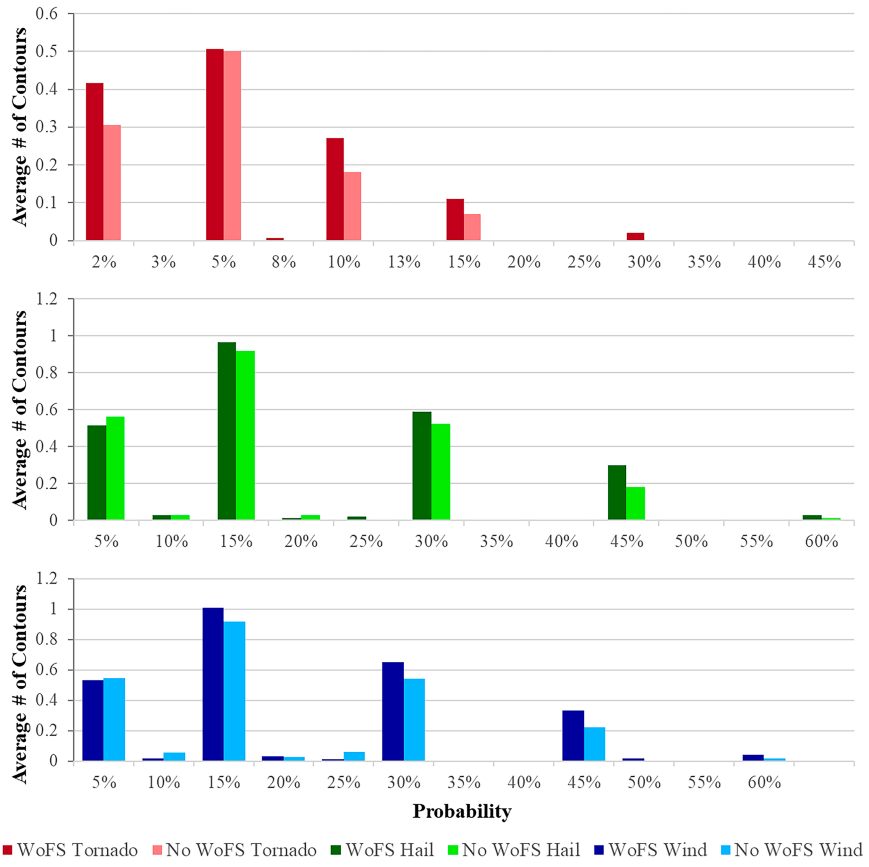

FIG. 6. The average number of probability contours in any expert forecast across SFE 2021. Averages are computed across all valid times and between the two expert forecasts issued on any given day. Note the different ranges of probabilistic contours available for tornadoes vs wind and hail.

WoFS and No WoFS groups. Looking at the count of high probability values and comparing the WoFS and No WoFS expert forecasts (Fig. 9, top row, inset solid bars) shows that the WoFS expert forecasts had more points in the high-probability tornado bins. This pattern of the WoFS expert forecasts having more high-probability points was consistent for all hours of the tornado forecasts, but present only for the 2200–2300 UTC and 2300–0000 UTC forecasts for wind and the 2200–2300 UTC hail forecasts.

Utilizing performance diagrams (Roebber 2009) to investigate the tornado forecasts further, the WoFS expert forecasts improved CSI at higher probabilistic forecast thresholds (e.g., 10%, 15%), by increasing POD and decreasing FAR simultaneously (Fig. 10). Better performance from expert WoFS forecasts relative to expert No WoFS forecasts was consistent for most probabilistic thresholds and forecast hours, particularly after the 2200–2300 UTC forecasts. A bias closest to 1 occurred with the 30% threshold, and high biases were seen in the lower forecast probabilities. These results demonstrate how high AUCs as seen in Fig. 8 can be achieved by overforecasting at low probabilities. In the case of severe convective storms, the asymmetric penalty function often applies, where the perceived cost of a false alarm is much lower than the cost of a missed event and incentivizes overforecasting (Jolliffe and Stephenson 2012). Conversely, low biases at high forecast thresholds may be due to many factors, including insufficient confidence in the forecast scenario or background knowledge that a 45% tornado probability correlates to an SPC High Risk in a convective outlook. While participants were told that their forecasts were for smaller temporal scales than a convective outlook and that they should feel free to use higher probability contours, we cannot rule out that the SPC categories and probabilities influenced participants as they attempted this new forecasting task. Overall, these results show how WoFS guidance improves short-term forecasts, supporting earlier work showing WoFS's utility at short time frames (Jones et al. 2018; Skinner et al. 2018; Flora et al. 2019, 2021; Guerra et al. 2022; G22). Hail and wind results (not shown) have similar trends for POD, but the CSI remains relatively consistent between forecasts suggesting that skill persists in all groups at longer lead times.

The final objective metric evaluated herein is the FSS. FSSs show increased performance for hail and wind forecasts relative to tornado forecasts for all time periods (Fig. 11). Although there are mixed results as to which forecasts perform best depending on hazard and valid time, WoFS expert forecasts

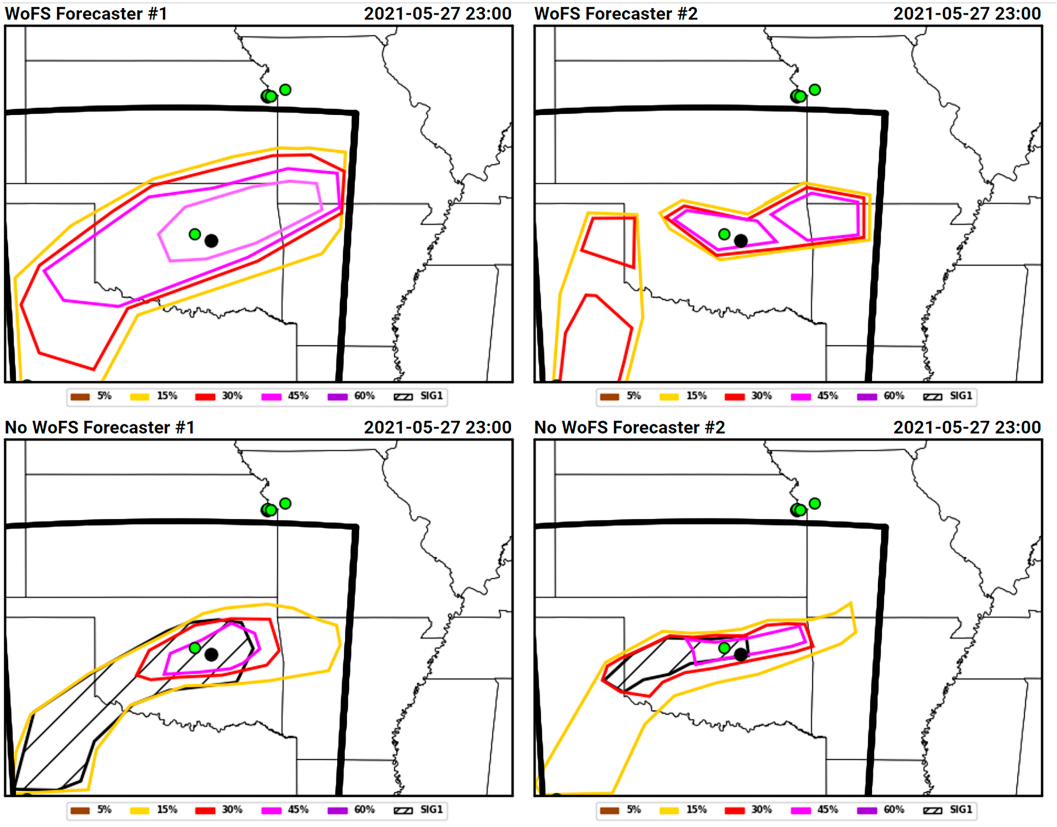

FIG. 7. As in Fig. 1, but only expert hail forecasts from 2200 to 2300 UTC 27 May 2021. Note that hail reports in Kansas and Missouri are outside of the WoFS domain bounds for this day. Green circles indicate hail LSRs, and the black circle indicates a significant (≥2-in) hail LSR.

score higher than No WoFS expert forecasts for most hazards at most hours. Exceptions to this include the 2200–2300 UTC tornado and hail forecasts, and the 0000–0100 UTC wind forecasts. For tornadoes, the consensus forecasts perform similarly to the expert forecasts, but are typically worse than the WoFS expert forecasts and better than the No WoFS expert forecasts. The expert forecasts are typically better than the consensus forecasts for hail, but the consensus forecasts were frequently better than the expert forecasts for the wind. WoFS provided the most benefit to the consensus forecasts for hail, while results from tornadoes and wind showed that the WoFS consensus forecasts performed worse than the No WoFS consensus forecasts. For the tornado forecasts, this is likely an issue of small sample size, while for the wind forecast the reliability diagrams show generally more overforecasting from the WoFS group relative to the No WoFS group. When looking at which group had the highest probability contour for a given hour and case, groups were relatively even. WoFS consensus forecasts had a higher maximum tornado (wind) probability for 26/51 (27/51) forecasts, and No WoFS consensus had a higher maximum tornado (wind) probability for 25/51 (24/51) forecasts. The WoFS may have also provided participants with a focus for high-confidence contours, leading to smaller and more precise consensus forecasts for the WoFS group relative to the No WoFS group. While precision and confidence are ideal with a perfectly accurate forecast, precise and inaccurate areas would decrease the FSS. Trends in the tornado expert forecasts showed decreasing skill with longer lead time for the No WoFS group forecasts, while the WoFS group forecasts showed similar performance between the 2200–2300 UTC forecasts and the 0000–0100 UTC forecasts, with decreased skill at 2300–0000 UTC.

d. Determining an appropriate practically perfect probability for 1-h forecasts

While practically perfect forecasts were provided to participants (e.g., Fig. 2), participants frequently indicated that the areas were likely too small and precise for a forecaster to realistically replicate. To figure out what combination of smoothing and neighborhoods should be used to verify the 1-h forecasts, several different practically perfect forecasts were compared to the expert forecasts using FSS. Essentially, this analysis swapped the forecasts and observations, assuming that the expert forecasts were the correct size and treating them as the "observations," while the varying practically perfect fields were the forecasts. This process was repeated for all three hazards in the WoFS and No WoFS group separately. Practically perfect forecasts for all hazards and both groups scored best for the

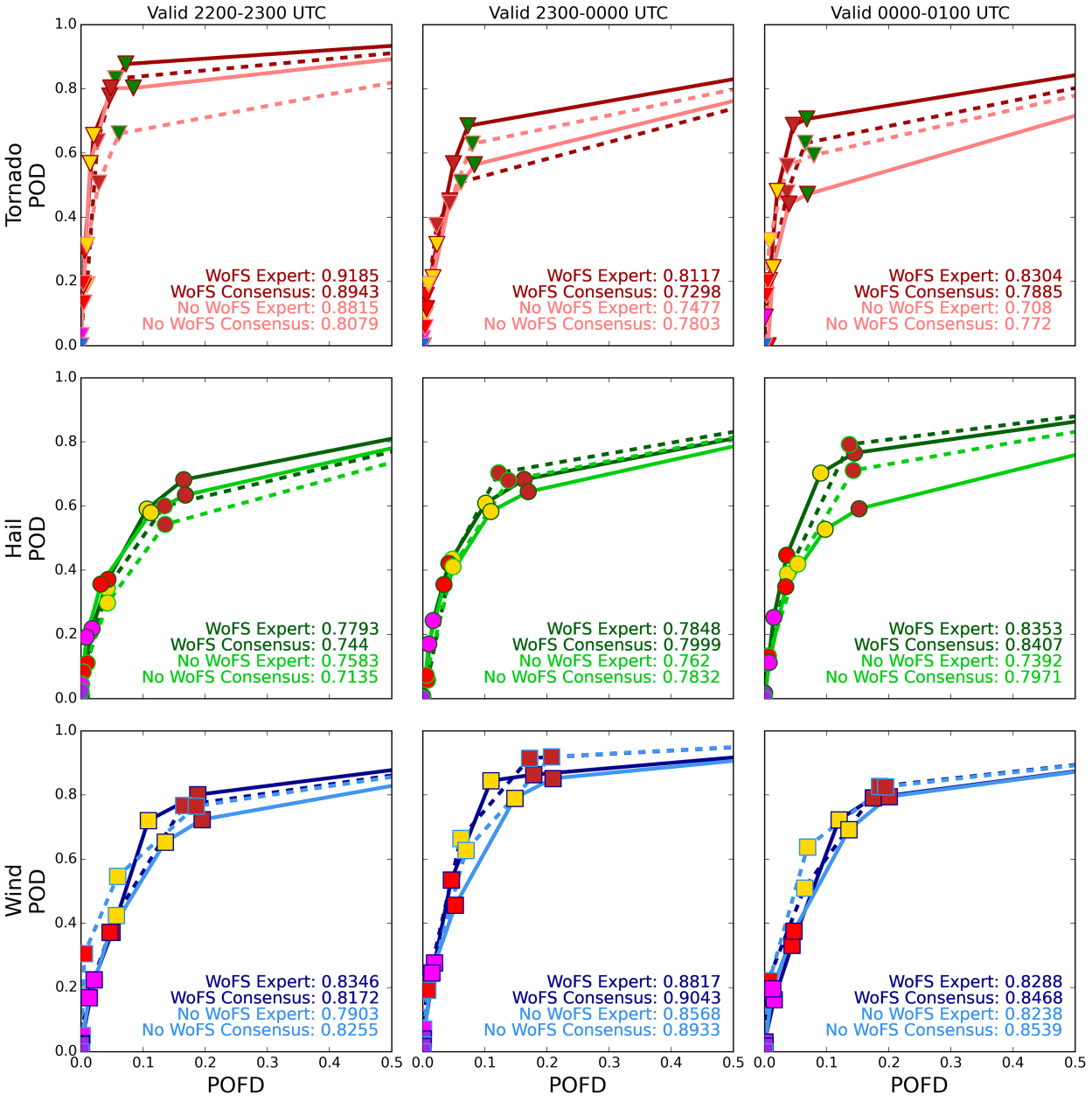

FIG. 8. ROC curves for expert (solid lines) and consensus (dashed lines) forecasts valid at (left) 2200–2300, (center) 2300–0000, and (right) 0000–0100 UTC. The WoFS group is indicated by darker colors, while the No WoFS group is indicated by lighter colors. Probabilistic tornado forecast thresholds of 2%, 5%, 10%, 15%, 30%, 45%, and 60% are indicated by the green, brown, yellow, red, pink, purple, and blue inverted triangles, respectively. Probabilistic hail and wind forecast thresholds of 5%, 15%, 30%, 45%, and 60% are indicated by the brown, yellow, red, pink, and purple squares (wind) and circles (hail), respectively. AUCs are annotated on each plot in the bottom right. Note the abbreviated *x* axis.

largest neighborhood size (Fig. 12), though the optimal smoothing radius varied.

While larger neighborhoods are typically going to be more skillful than smaller neighborhoods, this work finds that the skill should eventually asymptote or decrease if the base rate of the phenomena is much lower than the forecasted probabilities of the phenomena. In this case, expert tornado forecasts were 5% or less most of the time, but practically perfect probabilities of 50%+ could occur at small smoothing radii. If the forecast probabilities and observations in the FSS calculation differ sufficiently, eventually the FSS will decrease due to the probabilities saturating higher neighborhoods with values far above the base rate. When the base rate becomes sufficiently low, the numerator in the FSS equation becomes dominated by the low observational base rate, and the denominator becomes dominated by the high forecast probabilities. This

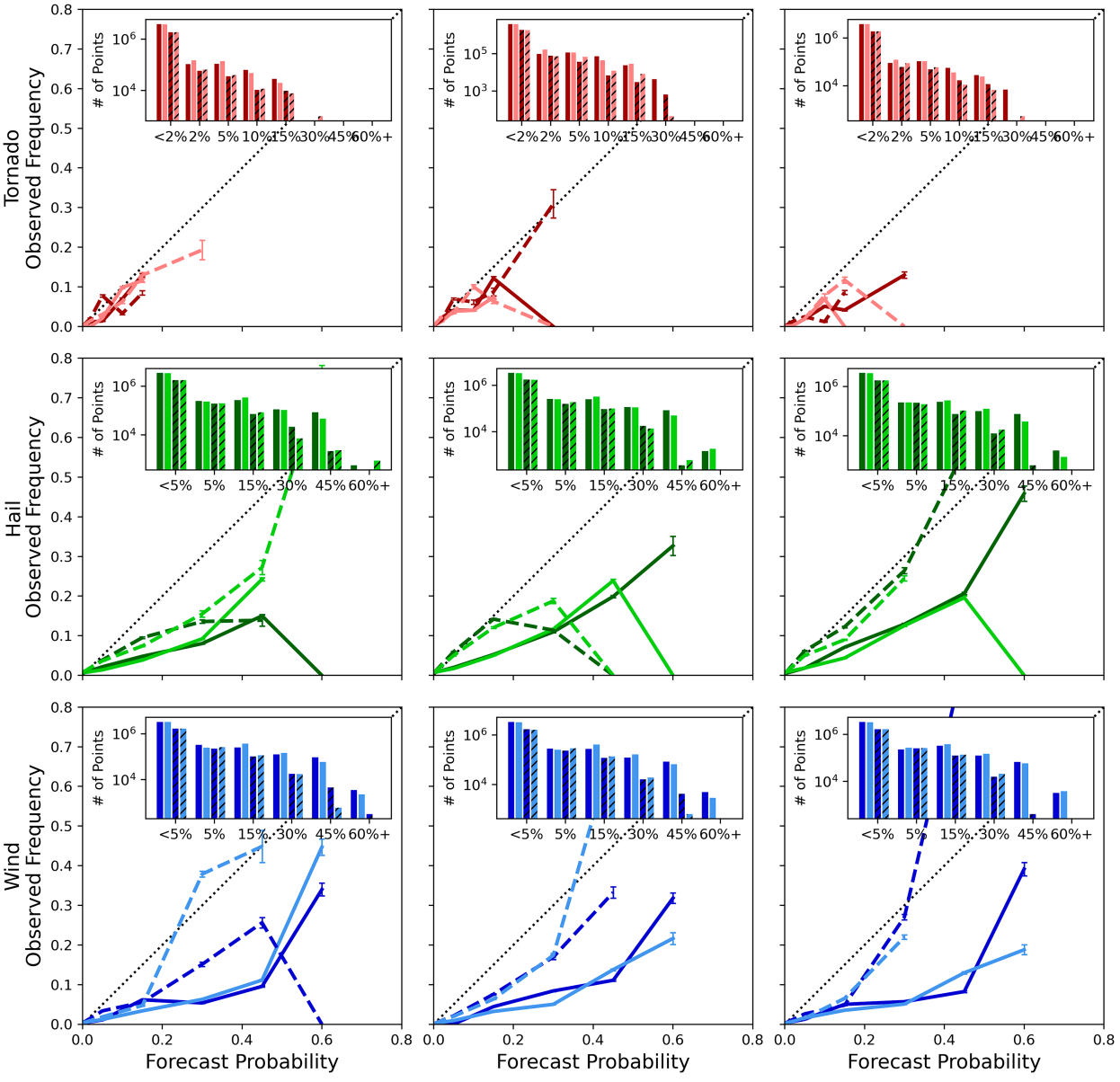

FIG. 9. Reliability diagrams curves for expert (solid lines) and consensus (dashed lines) forecasts valid at (left) 2200–2300, (center) 2300–0000, and (right) 0000–0100 UTC. The WoFS group is indicated by darker colors, while the No WoFS group is indicated by lighter colors. (top) Tornado (red), (middle) hail (green), and (bottom) wind (blue) forecasts are shown separately. Confidence intervals are displayed around each point. The dotted black line indicates perfect reliability. The inset bar plots show the number of points in each bin, with solid (dotted) bars showing the expert (consensus) forecasts. The WoFS group is indicated by darker colors, while the No WoFS group is indicated by lighter colors.

mismatch leads to decreasing skill when the base rate of the observations is much different than the maximum forecast probability issued, like in our prior example of the practically perfect forecasts exceeding 50% while the expert tornado forecasts were frequently an order of magnitude less. This skill decrease can be visualized in the tornado subplots along vertical lines corresponding to some of the smaller smoothing radii (roughly 10–60 km). The probabilities of tornadoes were frequently much lower than the hail and wind forecast probabilities during the course of this study, resulting in the decreasing skill with larger neighborhood being more easily visualized with the tornado forecasts relative to the other hazards. Similar patterns were seen in tornado forecasts at all hours, and an idealized model with a fixed base rate and varying probabilistic forecasts showed similar behavior (not shown).

Tornadoes require the most smoothing of any of the three hazard types to maximize FSS. Wind forecasts maximize the FSS at the lowest smoothing level of all hazards for all

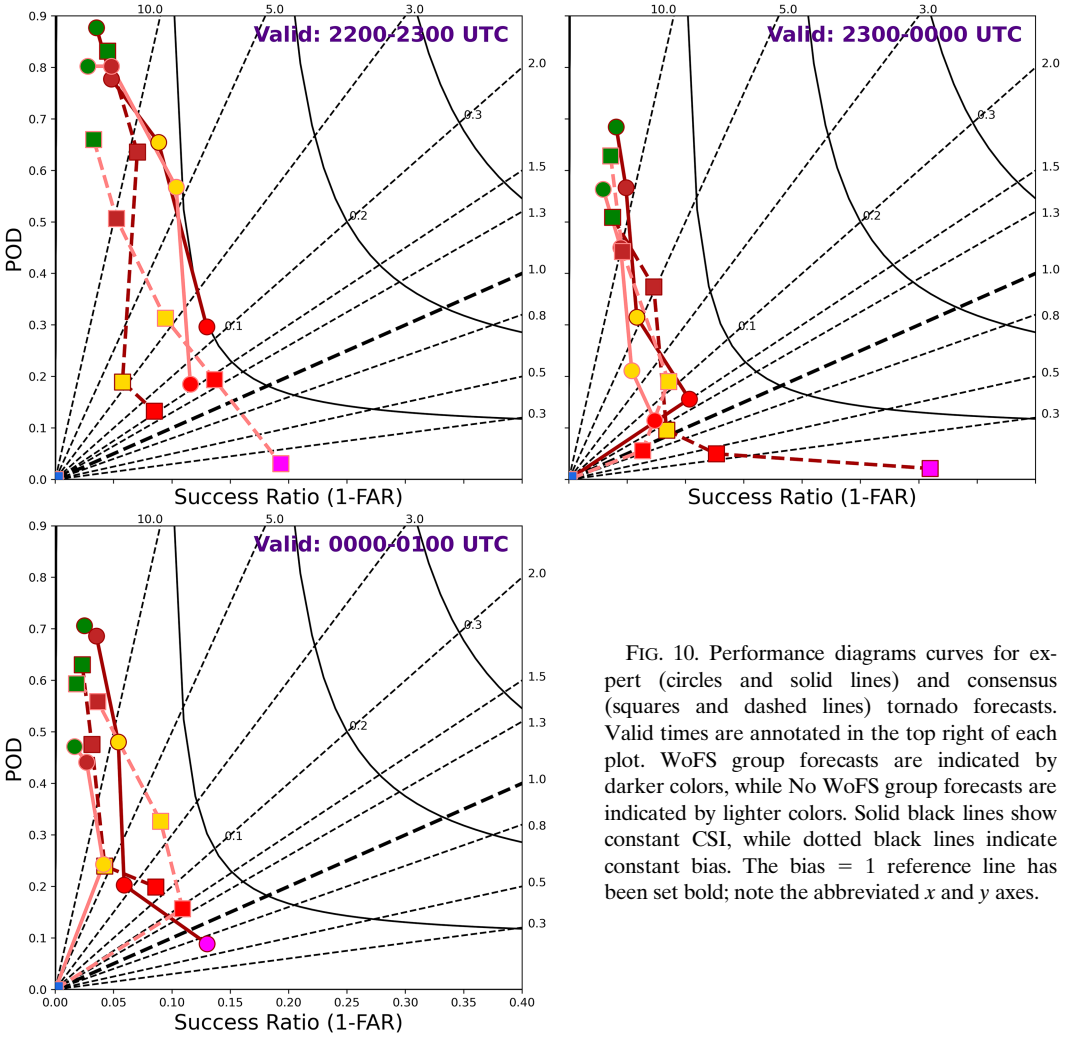

FIG. 10. Performance diagrams curves for expert (circles and solid lines) and consensus (squares and dashed lines) tornado forecasts. Valid times are annotated in the top right of each plot. WoFS group forecasts are indicated by darker colors, while No WoFS group forecasts are indicated by lighter colors. Solid black lines show constant CSI, while dotted black lines indicate constant bias. The bias = 1 reference line has been set bold; note the abbreviated x and y axes.

forecast hours. This may be in part due to forecasters drawing larger areas for wind hazards, which can be created by multiple convective modes. Averaging σ values of each of the hazards and forecast hours, it is recommended to use $\sigma = 70$ km for the practically perfect hourly forecasts in future SFEs. However, these results also show that the WoFS expert forecasts maximized FSS at a *lower* σ than the No WoFS expert forecasts for most hazards at most hours, by 10–20 km. This signal is consistent enough that a lower σ may need to be used if WoFS is used for all hourly forecasts, and future work should evaluate experimental data from SFE 2022 (Clark et al. 2023) using this framework to determine if using WoFS merits a reduction in the amount of smoothing done in the practically perfect fields used for subjective evaluation.

4. Conclusions and future work

This work initially sought to answer four questions about short-term, hourly probabilistic forecasts generated during the 2021 Spring Forecasting Experiment (SFE). Those questions were: 1) How skillful are the hourly probabilistic forecasts? 2) How does WoFS impact the hourly probabilistic forecasts? 3) Are skill differences a function of the hazard being forecasted? and 4) How does a consensus forecast compare to those of expert forecasters?

Expert NWS forecasters and other SFE participants generated hourly probabilistic forecasts of tornadoes, wind, and hail valid at 2200–2300, 2300–0000, and 0000–0100 UTC. The expert forecasts were subjectively evaluated individually the following day, while the other participants' forecasts were aggregated into a consensus. Some participants had access to WoFS, while others did not, setting up comparisons between those two groups, as well as comparisons between expert and consensus forecasts.

To the initial question of forecast skill, this work found the hourly probabilistic forecasts to be quite skillful, having high areas under the ROC curve and FSSs on par with other work computing FSS with a binary observation field. The areas under the ROC curve show that the participants were able to

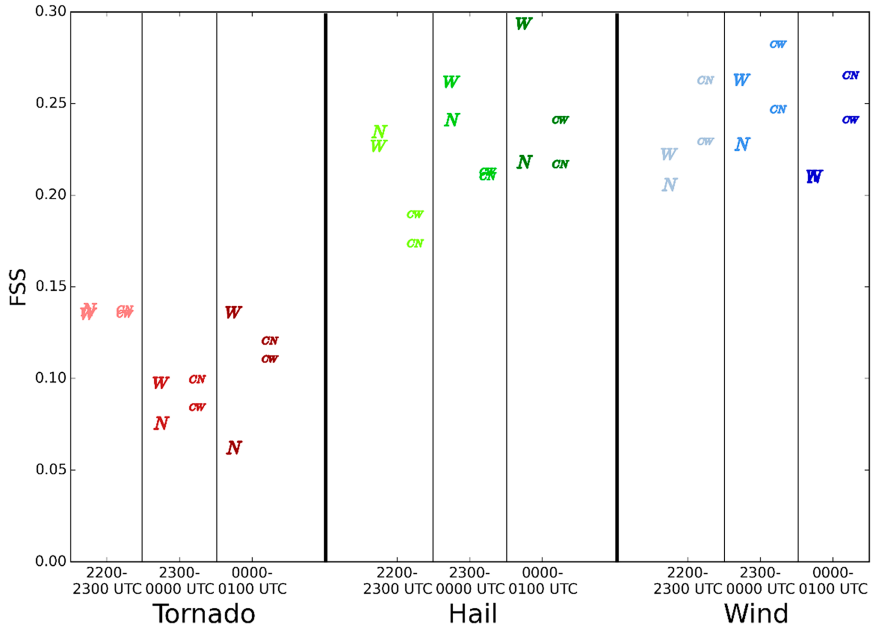

FIG. 11. FSS values for the experimental forecasts. WoFS expert ("W" markers), No WoFS expert ("N" markers), WoFS consensus ("CW" markers), and No WoFS consensus ("CN" markers) forecasts are shown. Darker colors indicate later times, while the hazards are differentiated by color. Times are further separated by thin black lines, while the hazards are separated by thick black lines.

separate events from nonevents within the WoFS domain, though reliability diagrams show that there is overforecasting occurring. There were some differences between the hazards, speaking to question 3, in that the tornado forecasts were the most reliable of any of the hazards while maintaining high AUCs. However, the tornado forecasts also had lower FSSs than hail or wind. Therefore, we find differences between the forecast skill of different hazards, but no clear "winner" in terms of which hazard was forecasted the best by our participants. The small sample size of the tornadoes over the hours considered in this study may limit the generalization of these results, and we recommend further testing over a larger sample of cases with tornado reports to examine whether the results would change for this hazard.

Access to WoFS appears to improve the hourly probabilistic forecasts of all hazards, which is most easily summarized by the subjective ratings of the expert forecasts. For all hazards, the WoFS expert forecasts were subjectively rated higher than the No WoFS expert forecasts. For objective metrics, the AUC and the FSS both show WoFS group forecasts as more skillful forecasts than the No WoFS forecasts. From the performance diagrams, the WoFS expert forecasts frequently increased POD and decreased FAR at high probabilistic thresholds, leading to an improved CSI relative to No WoFS expert forecasts. Therefore, the evidence for question 2 strongly supports the positive impact of WoFS on probabilistic forecasts in the watch-to-warning time frame.

The final research question, regarding a consensus versus expert forecasts showed mixed results. The consensus forecasts were generally more statistically reliable overall than the expert forecasts. Otherwise, the expert forecasts generally outperformed the consensus forecasts objectively in AUC, FSS, and on performance diagrams. Subjectively, the consensus forecasts were frequently rated more highly than the No WoFS expert forecasts, but below the expert WoFS forecasts.

As part of this work, the question arose as to what an optimal practically perfect forecast should look like at hourly scales. Based on FSS results from the experimental forecasts and practically perfect forecasts, a 39-km neighborhood and a σ of 70 km should be used to generate the practically perfect contours used in subjective verification of hourly forecasts during future SFEs.

Future work can extend this verification framework to the forecasts generated during the 2022 SFE. The 2022 SFE used a very similar setup to this work, but with two groups of participants who either did or did not have access to machine learning products within WoFS (Flora et al. 2021). The time frames were also slightly different than those examined herein. However, preliminary subjective results show a subjective rating difference similar in size to the difference shown here between WoFS and No WoFS group forecasts (Clark et al. 2023). If the objective metrics showed a similar sizable increase in performance for the participants using the machine learning guidance, it would suggest that this particular hazard guidance is extremely useful to participants in generating hazard-specific forecasts. Future work can also examine the ability of hourly forecasts to be decomposed from larger-scale outlooks, since it is likely that operational forecasters would not have time to draw hourly probabilistic forecasts for every hour. An approach to creating timing information on

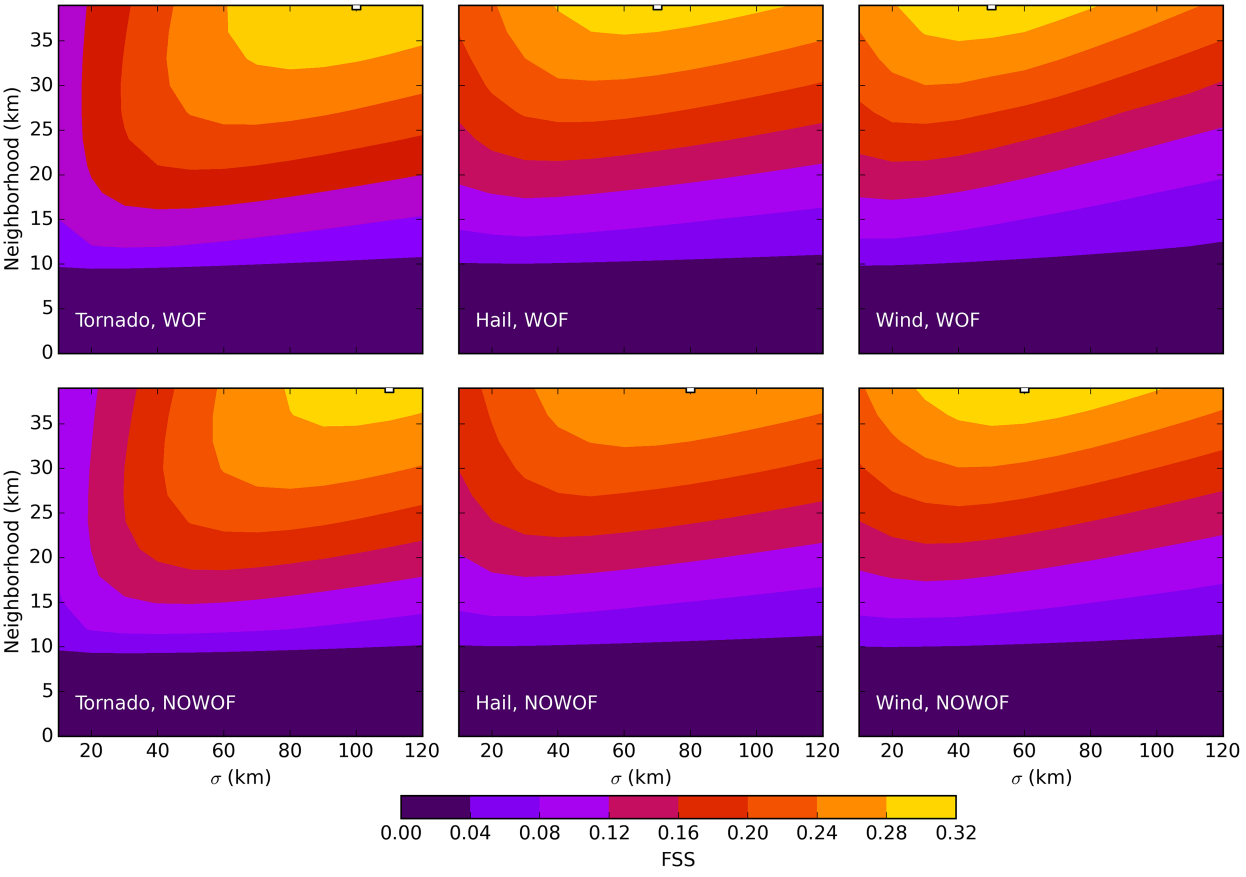

FIG. 12. FSS values for the WoFS and No WoFS expert forecasts for each hazard valid at 2200–2300 UTC, verified against practically perfect forecasts generated using different neighborhoods and smoothing levels. The white squares indicate the location of the highest FSS.

hazards using the HREF and the current convective outlook is currently being generated internally by the SPC for the NWS (Jirak et al. 2020), and perhaps a similar approach could be taken using WoFS for short-term, hourly updating probabilistic information at smaller time scales. However, continued work remains to understand the impact of WoFS on forecaster decision-making prior to operational implementation. Future SFEs and other testbed experiments remain crucial to ensuring the research-to-operations, operations-to-research loop can result in useful guidance to the forecaster.

Acknowledgments. The authors thank all of the participants and facilitators of SFE 2021, which took place virtually. Despite the virtual nature, participants were enthusiastic and engaged, and made the collection of these data very insightful with their feedback as they were conducting the activity. Thanks also go to Dr. Harold Brooks for discussions surrounding the base-rate paradigm, which aided in interpretation of the hourly practically perfect field FSS results, and Dr. Tim Supinie for assistance with refining Figs. 1 and 2. Funding for this work was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA16OAR4320115, U.S. Department of Commerce (B. T. G., B. R., J. V., K. K.). Authors D. I., P. B., and A. J. C. completed this work as part of regular duties at the federally funded NOAA National Severe Storms Laboratory. Author I. L. J. completed this work as part of regular duties at the federally funded NOAA Storm Prediction Center. The authors would also like to thank three anonymous reviewers, whose helpful comments improved the clarity of the manuscript.

Data availability statement. Deidentified datasets (e.g., experimental outlook forecasts) stored internally at NSSL may be shared upon request and free of charge following a reasonable period of time for data analysis and publishing (approximately 2 years). Warn-on-Forecast System (WoFS) model output is also stored internally at NSSL and may be shared upon request. Reports of severe weather used for verification were obtained from the *Storm Data* public page: https://www.ncdc.noaa.gov/stormevents/.

REFERENCES

Bellon, A., and G. L. Austin, 1978: The evaluation of two years of real time operation of a short-term precipitation forecasting procedure (SHARP). *J. Appl. Meteor.*, **17**, 1778–1787, https://

doi.org/10.1175/1520-0450(1978)017<1778:TEOTYO>2.0.CO;2.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

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

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

——, and Coauthors, 2022: The second real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bull. Amer. Meteor. Soc.*, **103**, E1114–E1116, https://doi.org/10.1175/BAMS-D-21-0239.1.

——, 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.

Dowell, D. C., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Flora, M. L., 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.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, https://doi.org/10.1175/WAF-D-15-0134.1.

——, 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, 2021: Exploring convection-allowing model evaluation strategies for severe local storms using the Finite-Volume Cubed-Sphere (FV3) model core. *Wea. Forecasting*, **36**, 3–19, https://doi.org/10.1175/WAF-D-20-0090.1.

——, 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.

Germann, U., and I. Zawadzki, 2002: Scale-dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology. *Mon. Wea. Rev.*, **130**, 2859–2873, https://doi.org/10.1175/1520-0493(2002)130<2859:SDOTPO>2.0.CO;2.

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.

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.

James, E. P., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part II: Forecast performance. *Wea. Forecasting*, **37**, 1397–1417, https://doi.org/10.1175/WAF-D-21-0130.1.

Jirak, I. L., M. S. Elliott, C. D. Karstens, R. S. Schneider, P. T. Marsh, and W. F. Bunting, 2020: Generating probabilistic severe timing information from SPC Outlooks using the HREF. *Severe Local Storms Symp.*, Boston, MA, Amer. Meteor. Soc., 3.1, https://ams.confex.com/ams/2020Annual/webprogram/Paper367695.html.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. John Wiley & Sons, 292 pp.

Jones, T. A., P. Skinner, K. Knopfmeier, E. Mansell, P. Minnis, R. Palikonda, and W. Smith Jr., 2018: Comparison of cloud microphysics schemes in a Warn-on-Forecast System using synthetic satellite objects. *Wea. Forecasting*, **33**, 1681–1708, https://doi.org/10.1175/WAF-D-18-0112.1.

——, 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., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003a: Collaboration between forecasters and research scientists at the NSSL and SPC: The Spring Program. *Bull. Amer. Meteor. Soc.*, **84**, 1797–1806, https://doi.org/10.1175/BAMS-84-12-1797.

——, M. E. Baldwin, P. R. Janish, S. J. Weiss, M. P. Kay, and G. W. Carbin, 2003b: 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.

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

Keil, C., F. Heinlein, and G. C. Craig, 2014: The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quart. J. Roy. Meteor. Soc.*, **140**, 480–490, https://doi.org/10.1002/qj.2143.

Krocak, M. J., 2020: If we forecast it, they may (or may not) use it: Sub-daily severe weather timing information and its utility for forecasters, stakeholders, and end users. Ph.D. dissertation, University of Oklahoma, 144 pp., https://shareok.org/handle/11244/324932.

Ligda, M. G., 1953: The horizontal motion of small precipitation areas as observed by radar. M.I.T. Tech. Rep. 21, 60 pp.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Miller, W. J. S., and Coauthors, 2021: Exploring the usefulness of downscaling free forecasts from the Warn-on-Forecast System. *Wea. Forecasting*, **37**, 181–203, https://doi.org/10.1175/WAF-D-21-0079.1.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, https://doi.org/10.1175/BAMS-D-18-0041.1.

——, B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Wea.*

Forecasting, **35**, 2293–2316, https://doi.org/10.1175/WAF-D-20-0069.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

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

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.

Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

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

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

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

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

Sun, J., and Coauthors, 2014: Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bull. Amer. Meteor. Soc.*, **95**, 409–426, https://doi.org/10.1175/BAMS-D-11-00263.1.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261, https://doi.org/10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2.

Wilson, J., N. A. Crook, C. K. Mueller, J. Sun, and M. Dixon, 1998: Nowcasting thunderstorms: A status report. *Bull. Amer. Meteor. Soc.*, **79**, 2079–2099, https://doi.org/10.1175/1520-0477(1998)079<2079:NTASR>2.0.CO;2.

——, D. Megenhardt, and J. Pinto, 2020: NWP and radar extrapolation: Comparisons and explanation of errors. *Mon. Wea. Rev.*, **148**, 4783–4798, https://doi.org/10.1175/MWR-D-20-0221.1.

Wilson, K. A., B. T. Gallo, P. S. Skinner, A. J. Clark, P. L. 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.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. 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.

WMO, 2017: Guidelines for nowcasting techniques. WMO Doc. WMO-1198, 82 pp., https://library.wmo.int/doc_num.php?explnum_id=3795.