Exploring the Watch-to-Warning Space: Experimental Outlook Performance during the 2019 Spring Forecasting Experiment in NOAA's Hazardous Weather Testbed

BURKELY T. GALLO,[a,b] KATIE A. WILSON,[a,c] JESSICA CHOATE,[d] KENT KNOPFMEIER,[a,c] PATRICK SKINNER,[a,c] BRETT ROBERTS,[a,b,c] PAMELA HEINSELMAN,[c,e] ISRAEL JIRAK,[b] AND ADAM J. CLARK[c,e]

[a] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma*
[c] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[d] *Department of Atmospheric Sciences, University of Illinois at Urbana–Champaign, Urbana, Illinois*
[e] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 20 October 2021, in final form 20 January 2022)

ABSTRACT: During the 2019 Spring Forecasting Experiment in NOAA's Hazardous Weather Testbed, two NWS forecasters issued experimental probabilistic forecasts of hail, tornadoes, and severe convective wind using NSSL's Warn-on-Forecast System (WoFS). The aim was to explore forecast skill in the time frame between severe convective watches and severe convective warnings during the peak of the spring convective season. Hourly forecasts issued during 2100–0000 UTC, valid from 0100 to 0200 UTC demonstrate how forecasts change with decreasing lead time. Across all 13 cases in this study, the descriptive outlook statistics (e.g., mean outlook area, number of contours) change slightly and the measures of outlook skill (e.g., fractions skill score, reliability) improve incrementally with decreasing lead time. WoFS updraft helicity (UH) probabilities also improve slightly and less consistently with decreasing lead time, though both the WoFS and the forecasters generated skillful forecasts throughout. Larger skill differences with lead time emerge on a case-by-case basis, illustrating cases where forecasters consistently improved upon WoFS guidance, cases where the guidance and the forecasters recognized small-scale features as lead time decreased, and cases where the forecasters issued small areas of high probabilities using guidance and observations. While forecasts generally "honed in" on the reports with slightly smaller contours and higher probabilities, increased confidence could include higher certainty that severe weather would not occur (e.g., lower probabilities). Long-range (1–5 h) WoFS UH probabilities were skillful, and where the guidance erred, forecasters could adjust for those errors and increase their forecasts' skill as lead time decreased.

SIGNIFICANCE STATEMENT: Forecasts are often assumed to improve as an event approaches and uncertainties resolve. This work examines the evolution of experimental forecasts valid over one hour with decreasing lead time issued using the Warn-on-Forecast System (WoFS). Because of its rapidly updating ensemble data assimilation, WoFS can help forecasters understand how thunderstorm hazards may evolve in the next 0–6 h. We found slight improvements in forecast and WoFS performance as a function of lead time over the full experiment; the first forecasts issued *and* the initial WoFS guidance performed well at long lead times, and good performance continued as the event approached. However, individual cases varied and forecasters frequently combined raw model output with observed mesoscale features to provide skillful small-scale forecasts.

KEYWORDS: Ensembles; Forecast verification/skill; Numerical weather prediction/forecasting; Severe storms; Thunderstorms; Probability forecasts/models/distribution

1. Introduction

Each year during the 5-week annual Spring Forecasting Experiment (SFE; Gallo et al. 2017; Clark et al. 2020), new forms of NWP, post-processing methods, and forecasting techniques are tested during the peak of the spring convective season. By demonstrating these new innovations to participants from the research, model development, and operational forecasting communities, researchers and forecasters can understand the unique challenges faced by each group in their daily routines. For example, researchers can better understand the time pressures constraining operational forecasters, while operational forecasters can learn the strengths and weaknesses of experimental guidance. This research-to-operations and operations-to-research feedback loop allows forecasters to experience and provide feedback on new tools, while enabling researchers to iteratively develop tools and beneficial guidance for forecasters. New versions of operational systems such as the HRRR (Benjamin et al. 2016; Alexander et al. 2020) and the High Resolution Ensemble Forecast system (HREF; Roberts et al. 2019) are frequently tested in SFEs, as are new forecasting products. For example, Day 2 individual hazard probabilities for tornadoes, severe hail [≥1 in. (2.54 cm)] and severe convective winds [≥58 mph (25.93 m s^{-1})] issued by Storm Prediction Center (SPC) forecasters were operationalized on 30 January 2020, after being tested for years in the SFE (Clark et al. 2020).

Much of the guidance examined during SFEs focuses on the 12–36-h time frame, and is initialized once or twice a day with forecasts extending to 36–60 h. A notable exception to

Corresponding author: Burkely T. Gallo, burkely.twiest@noaa.gov

DOI: 10.1175/WAF-D-21-0171.1

this is the HRRR, which provides hourly forecasts to 18 h and forecasts to 36 h every 6 h (Benjamin et al. 2016; Alexander et al. 2020). However, rapidly updating convection-allowing model (CAM) ensemble forecasts are more difficult to produce, in part because of the computational expense. The experimental Warn-on-Forecast System (WoFS; Wheatley et al. 2015; Heinselman et al. 2022, manuscript submitted to *Geophys. Monogr.*) is the most frequently updating system tested in the SFE, producing forecasts to 6 h (3 h) at the top (bottom) of every hour. The WoFS uses 15-min data assimilation to incorporate current radar and satellite observations (Jones et al. 2016), providing guidance from convective warning scales (~30 min) to watch or mesoscale convective discussion scales (~1–6 h). The SFE has examined WoFS since 2017 with activities ranging from surveying forecasters' interpretations of different ensemble guidance products (Wilson et al. 2019) to evaluating 1.5-km horizontal resolution versions of the WoFS (half the usual WoFS resolution; Clark et al. 2021a) to using WoFS for experimental forecasting activities (Clark et al. 2020, 2021a).

The WoFS is a unique CAM ensemble in many respects. Besides the data assimilation, the perspective of WoFS differs from traditional CAMs, as it was developed as an alternative to the "Warn-on-Detection" paradigm where forecasters issue warnings based on ongoing, observed storms (Stensrud et al. 2009, 2013). Instead, the "Warn-on-Forecast" concept visualizes a system where individual storms are sufficiently well-forecasted to allow forecasters to issue warnings based on the guidance, resulting in longer lead times for people in the storm's path. This vision extends beyond the warning to include longer lead time products such as SPC Mesoscale Convective Discussions and Weather Prediction Center Mesoscale Precipitation Discussions. After years of development and testing, the WoFS is the ensemble designed to realize the Warn-on-Forecast concept. However, as with any new tool, extensive testing and development work remains to successfully transition WoFS into operations, particularly with the shift to NOAA's Unified Forecast System (https://ufscommunity.org/), which has the goal of uniting NOAA's modeling efforts around a single dynamical core differing from the current WoFS configuration.

Verification methods for WoFS and other high-resolution CAMs were developed concurrently with the development of the WoFS modeling and data assimilation framework. Demuth et al.'s (2020) survey of forecaster needs regarding CAM ensembles shows a strong desire for useful verification metrics, particularly information about the reliability of the guidance and the probabilities generated by a given ensemble, and the performance of the ensemble in specific scenarios. Traditional, gridpoint-based verification for severe weather frequently suffers from the "double penalty" problem (Mass et al. 2002; Done et al. 2004; Gilleland et al. 2009), where a forecast displaced from the observation is penalized as both a "false alarm" and a "miss." However, such a forecast may still provide value to the forecaster or end user if it indicates that a hazard may be nearby during the forecast period (Kain et al. 2008). Several methods account for these displacement errors, including neighborhood approaches (see Schwartz and Sobash 2017 for an overview) and object-based verification (Wolff et al. 2014; Skinner et al. 2018). Object-based verification of the WoFS (e.g., Jones et al. 2018; Skinner et al. 2018; Flora et al. 2019, 2021) as well as other CAMs (e.g., Gallus 2010; Johnson et al. 2013, 2020; Potvin et al. 2019; Adams-Selin et al. 2019), have frequently examined shorter-term forecasts verified against radar or satellite proxies and few studies have applied object-based methodology to convective outlooks (e.g., Gallo et al. 2021).

SFE 2019's usage of the WoFS was the most extensive to date. Two forecasters were brought in specifically to forecast a series of outlooks through the evening using WoFS. Forecasters issued two rolling outlooks each hour and an outlook that was valid at a consistent time, resulting in shorter lead times as time progressed. Further details will be provided in the methodology section. Forecaster product usage data was also tracked and explored in Wilson et al. (2021). Wilson et al. (2021) examined product usage patterns, product-outlook performance correlations, and access pattern similarity across the entire forecasting task period in SFE 2019 (a subset of which is studied here; see section 2b for details on the outlook issuance task). They found that participants accessed the reflectivity, rotation, hail, and surface wind products most frequently, and accessed the rotation products after reflectivity products at least once daily. Rotation products included the 2–5-km hourly maximum updraft helicity (UH; Kain et al. 2010), which is frequently used as a skillful proxy for any severe weather hazard (Sobash et al. 2011).

While Wilson et al. (2021) focused on the product usage data, this work focuses on the experimental outlooks issued by the forecasters, and how those outlooks evolve in time. We explore whether forecasters were able to use WoFS guidance to hone in on the area of severe weather as the lead time decreased. We also examine how the descriptive and statistical characteristics of the outlooks change with decreasing lead time, answering questions such as the following: Did forecasts become more precise and confident (e.g., more extreme probability values) as the lead time decreased? Did the skill increase appreciably? Was the evolution of the forecasts similar from one hour to the next, or was there a point in the lead time where forecasts either improved or deteriorated by a large amount? We also consider the skill of the underlying WoFS UH probabilities, answering questions such as: How did the forecasters adjust forecasts compared to the raw model guidance? How skillful was the underlying WoFS guidance? With the increased usage of a Warn-on-Forecast System, these questions will become increasingly important in the watch-to-warning time scale.

2. Data and methodology

a. Model configuration and data

Participants in this experiment relied heavily on the WoFS, which used the WRF-ARW dynamical core (Skamarock et al. 2008). WoFS generated 18-member forecasts in the 0–6-h time frame on the hour, and in the 0–3-h time frame on the half hour. WoFS used a 36-member ensemble data assimilation process, where the full ensemble was updated every 15 min by Gridpoint Statistical Interpolation–ensemble Kalman filter data assimilation (Hu et al. 2017) of WSR-88D radar reflectivity and radial velocity

data (Wheatley et al. 2015), satellite cloud water path observations (Jones et al. 2016), and conventional observations when available. Data assimilation cycling for the first WoFS forecast of each day, 1900 UTC, began at 1700 UTC, so nine cycles of data assimilation occurred prior to the forecast being launched at 1900 UTC. The final forecast used by participants in this study was launched at 0000 UTC. Boundary conditions were taken from the 1200 UTC High-Resolution Rapid Refresh Ensemble (HRRRE; Kalina et al. 2021) forecast. WoFS had 3-km horizontal grid spacing and covered a 900 × 900 km^2 grid each day that focused on the area with the highest potential for severe weather in the CONUS, which was decided together with the SFE. A complete description of the WoFS configuration for spring 2019 is available in Jones et al. (2020).

Participants in this study accessed the WoFS guidance through a web page (currently found at https://wof.nssl.noaa.gov/realtime/). After SFE 2019, errors in soil moisture perturbations used to create the HRRRE initial conditions and in the assimilation of rotational velocity data were discovered and corrected, and WoFS cases were rerun and displayed on the web viewer. The soil moisture error artificially inflated the ensemble spread in environmental fields such as the temperature and dewpoint, leading to less accurate environmental fields in individual members and the potential for less realistic evolution of storm-scale forecasts. The skill of WoFS storm-scale guidance was negatively impacted by this soil moisture error, but impacts were expected to vary with the mesoscale environment and not significantly impair the ability of the system to provide short-term prediction of thunderstorms. The rotational velocity assimilation error involved the location of the observational data, and the most frequent effect was that the observations were erroneously discarded by quality control algorithms during the assimilation process. Despite the rerun files being displayed on the current WoFS web page, the original files that participants used were archived and forecasts from those files are verified herein to ensure that participant forecasts are being compared to model forecasts that were available to them at the time of the activity.

The ensemble probability of 2–5-km UH exceeding 60 m^2 s^{-2} (Skinner et al. 2018) at three different spatial neighborhoods is verified alongside participant forecasts to determine the skill of the underlying model guidance that participants used and whether participants' forecasts improved upon the guidance. Hourly 2–5-km UH probabilities serve as a proxy for the full suite of WoFS guidance, which includes many additional products. The representativeness of this product will vary from case-to-case (Wilson et al. 2021; e.g., lower skill anticipated for events that may not contain rotating storms), but 2–5-km UH has previously been shown to be a good proxy for severe weather (Sobash et al. 2011).

Hourly ensemble neighborhood probabilities of UH > 60 m^2 s^{-2} were created as follows. First, a 2D field of the maximum UH from all 5-min instantaneous UH in a given member was extracted and combined to create an hourly maximum field for each member. A square neighborhood maximum filter with radii of either 9, 15, or 27 km (3, 5, and 9 grid points, respectively) was then applied, and the resultant field was smoothed using a 7.5-km exponential decay filter with weights of Eq. (1):

$$g = \exp\left[-\left(\frac{x^2}{r} + \frac{y^2}{r}\right)\right]. \quad (1)$$

These weights from Eq. (1) were then normalized by dividing each weight by the sum of all of the weights in the 7.5-km neighborhood. Finally, the probabilities of UH exceeding 60 m^2 s^{-2} were determined by how many members exceeded the threshold at each grid point.

b. *Outlook issuance activity*

Two different NWS forecasters participated each week during the 5 weeks of SFE 2019 (29 April–31 May 2019). Each hour starting from 2000 to 2100 UTC, forecasters issued three forecasts of severe weather, grouping together hail, severe convective wind, and tornado hazards into a single probability. Probabilities were defined in line with current SPC probabilistic definitions; namely, as the probability of severe weather within 25 miles of a point. These three forecasts consisted of 1) a 1-h rolling forecast valid starting at the end of the current hour, 2) a 4-h rolling forecast valid starting at the end of forecast 1), and 3) a targeted 1-h forecast that was always valid from 0100 to 0200 UTC (Fig. 1). The final forecasts were issued between 0000 and 0100 UTC, so forecasts 1 and 3 were the same at that hour. This study focuses on the targeted 1-h forecast (forecast 3), to look at the forecast evolution with decreasing lead time rather than the performance of forecasts at different times throughout the evening.

Participants began activities at 1700 UTC, with individual forecast generation beginning at 2100 UTC. Participants each created four targeted outlooks daily. Prior to 2100 UTC participants received a briefing from a retired SPC forecaster, met with WoFS program research scientists to build familiarity with WoFS, and/or completed an online training module focused on WoFS guidance concepts, depending on the day (Wilson et al. 2021). From 2000 to 2100 UTC, the two forecasters joined a larger group of SFE participants that issued the same set of three forecasts. Beginning with a group activity allowed the participants to get comfortable with the data, the drawing tool, and the ongoing weather via interactions with SFE participants and facilitators familiar with the weather and guidance.

Participants used a web-based drawing tool to create the probabilistic outlooks. Participants could draw contours at probability levels of 5%, 15%, 30%, 45%, and 60%. For the 1-h forecasts participants were advised to begin with the 15% contour, as the "practically perfect" forecasts (Hitchens et al. 2013) used in next-day subjective verification applied a small Gaussian smoother (σ = 40 km). As a result, a single Local Storm Report (LSR) used for next-day subjective verification resulted in a practically perfect forecast of 34%; thus a practically perfect forecast depicting a lone 5% or 15% contour without additional higher probability contours did not occur. This rule of thumb also lowered forecaster workload, as in cases with very active weather participants could focus on the guidance and make the best forecast possible, rather than worrying about drawing

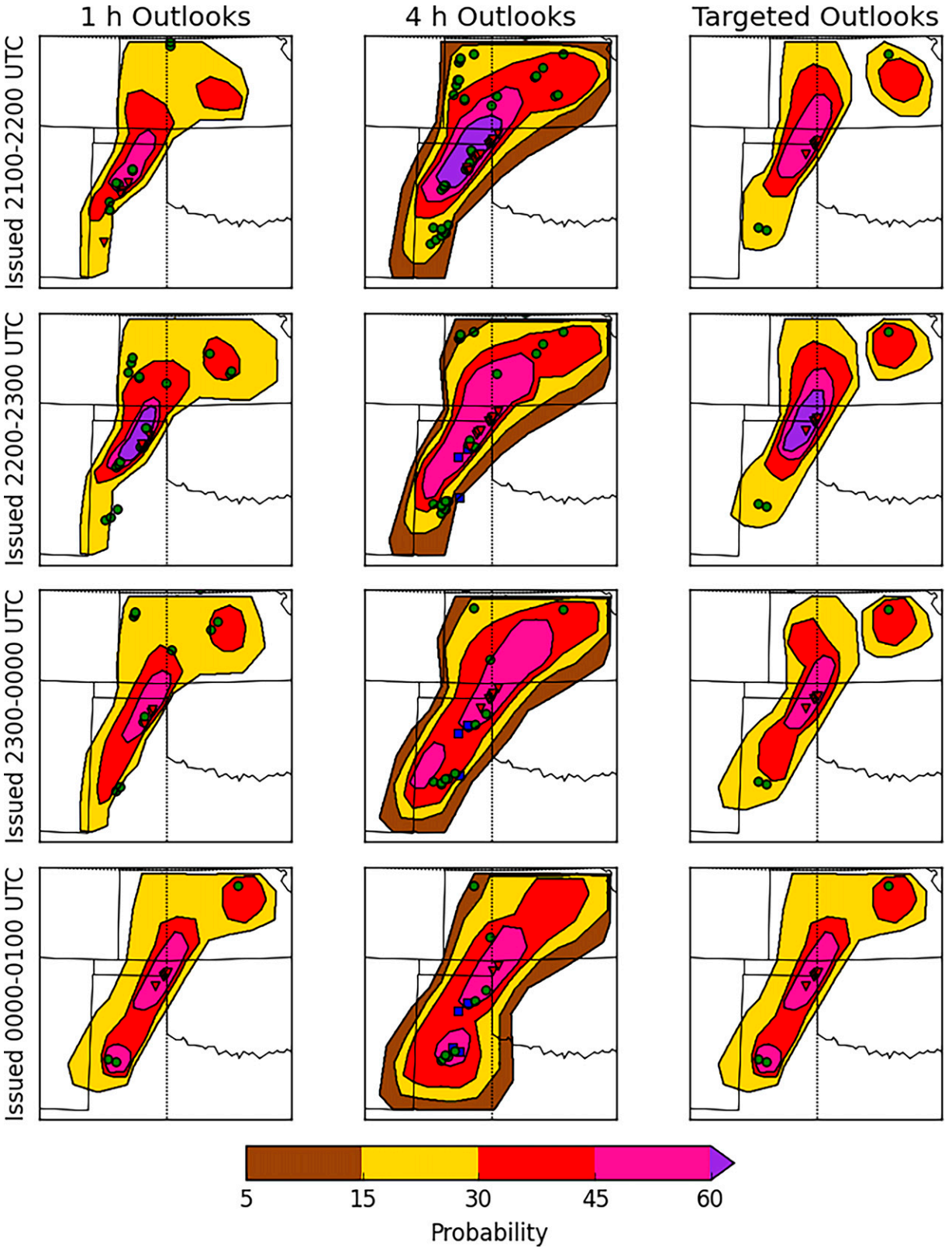

FIG. 1. Probabilistic outlooks issued by a single forecaster on 23 May 2019. Probabilities are color filled, and tornado (red inverted triangle), hail (green circle), and wind (blue square) reports are overlaid.

TABLE 1. A list of the SFE 2019 cases used in this study.

Week 1	Week 2	Week 3	Week 4	Week 5
30 Apr; 1, 2 May	7, 8 May	14, 15, 16 May	21, 22, 23 May	29, 30 May

many low-end probability lines. However, in cases of high uncertainty or low-end risk of severe weather, participants issued some 5% and 15% contours without any higher probabilities. After the initial hour, participants could load in and modify their prior outlooks, so they typically did not draw outlooks from scratch with each forecast issuance.

Cases selected herein are a subset of the full five weeks. While a research scientist was available throughout the activity every day, on the first day of the week participants received more guidance on WoFS and its performance, and were able to get a feel for issuing short-term probabilistic forecasts of severe weather; thus, that day is treated as a spinup case and is excluded from analysis. The evening activity was not conducted on Fridays, allowing participants to travel home. Finally, occasional WoFS availability issues occurred during the SFE, resulting in 13 cases examined herein (Table 1). These cases are examined both in aggregate across the full subset of cases and for individual cases, though differences in forecaster skill levels may occur between weeks as new forecasters participate in the experiment. However, all forecasters are issuing short-term probabilistic forecasts of severe weather, which are not part of a typical WFO process and may help decrease skill differences. Finally, having 10 forecasters enabled us to include forecasters with a variety of WFO experience and provide a broader sample than having 2 forecasters complete the exercise throughout the five weeks.

c. *Verification methods and metrics*

Experimental outlooks and WoFS UH probabilities were verified using entries from NCDC's *Storm Data* storm event database (https://www.ncdc.noaa.gov/stormevents/), to get the most accurate report data available. Reports of severe weather have noted shortcomings regarding areas of low population density, overestimation of wind speeds by some observers, and an underestimation of the spatial extent of some hazards (Witt et al. 1998; Doswell et al. 2005; Verbout et al. 2006; Trapp et al. 2006; Edwards et al. 2018). Thus, especially when looking at small time and space windows, verification results are sensitive to the presence or absence of reports. While next-day subjective verification used LSRs by necessity due to the low latency of these reports, objective verification herein can use the more exhaustive *Storm Data* entries. Since all forecasts are verified using the same data, relative performance should be unaffected by missed or erroneous reports. *Storm Data* entries were gridded to the WoFS domain and inflated to match the probabilistic definition of the forecasts as being for severe weather "within 25 mi of a point." Any point within 40 km (24.85 mi) of a report was considered a "hit" in standard 2×2 contingency table terminology. Prior to verification, participant outlooks were stored in Geo JavaScript Object Notation (GeoJSON) format before also being regridded to the 3-km WoFS grid. No interpolation took place between contour levels. WoFS UH probabilities were also verified at the probabilistic thresholds available for participants to draw (5%, 15%, 30%, 45%, and 60%) for fair comparison of the WoFS UH and the participant outlooks.

Forecast skill was assessed using traditional contingency-table-based metrics such as the area under the receiver operating curve (ROC area; Mason 1982), probability of detection (POD), false alarm ratio (FAR), success ratio (SR), and critical success index (CSI). ROC area values range from the worst possible score of 0 to the best possible score of 1, with 0.5 indicating the skill of a random forecast. ROC area values of ≥ 0.7 commonly indicate a skillful forecast (Buizza et al. 1999). POD ranges from a low value of 0, indicating that no observations were correctly forecast, to a high value of 1, indicating that all observations were correctly forecast. The SR equals 1 minus the FAR, so a score of 0 indicates that no forecasts successfully forecast an observation (e.g., all forecasts were false alarms), and a score of 1 indicates that no false alarms occurred. Reliability diagrams provide a complement to the ROC area which tends to reward POD at the expense of reliability for rare events such as tornado forecasting (Gallo et al. 2016).

A neighborhood maximum ensemble probability (Schwartz and Sobash 2017) approach to the fractions skill score (FSS; Roberts and Lean 2008) is applied herein. As the observations are binary (either severe weather happens within 40 km of a point or it does not) and during the FSS calculation the difference between the forecasted probabilities and the binary observations is used rather than the difference between the forecasted probabilities and a fractional set of observations. This formulation follows Roberts et al. (2020) and Wilson et al. (2021), who similarly applied the binary observation approach to severe convective storms, and is similar to the approach taken by Schwartz et al. (2010). Using binary observations of either 0 or 1 rather than fractional observations calculated by smoothing the reports avoids the issue of conflating the neighborhood and smoothing length scales noted by Schwartz and Sobash (2017). While the values of this score are lower than those calculated using the traditional FSS and are closely related to the Brier skill score (Brier 1950), avoiding the degrees of freedom associated with observational smoothing is important due to the small temporal and spatial scales examined herein. Given the watch-to-warning scales applied here, determining an optimal smoothing radius for observations at this scale is beyond the scope of the current study.

Finally, squared Pearson correlation coefficients (Wilks 2011) were used to look at the hour-to-hour variation between the outlooks, and the variation between the forecaster outlooks and the WoFS UH probabilities. Pearson correlation coefficients were chosen as they preserve the raw values of the probabilities during the calculation. Correlations were calculated between each pair of outlooks issued by an individual forecaster on a given day (e.g., between outlooks issued at 2100–2200 and 2200–2300 UTC, outlooks issued at 2100–2200 and 2300–0000

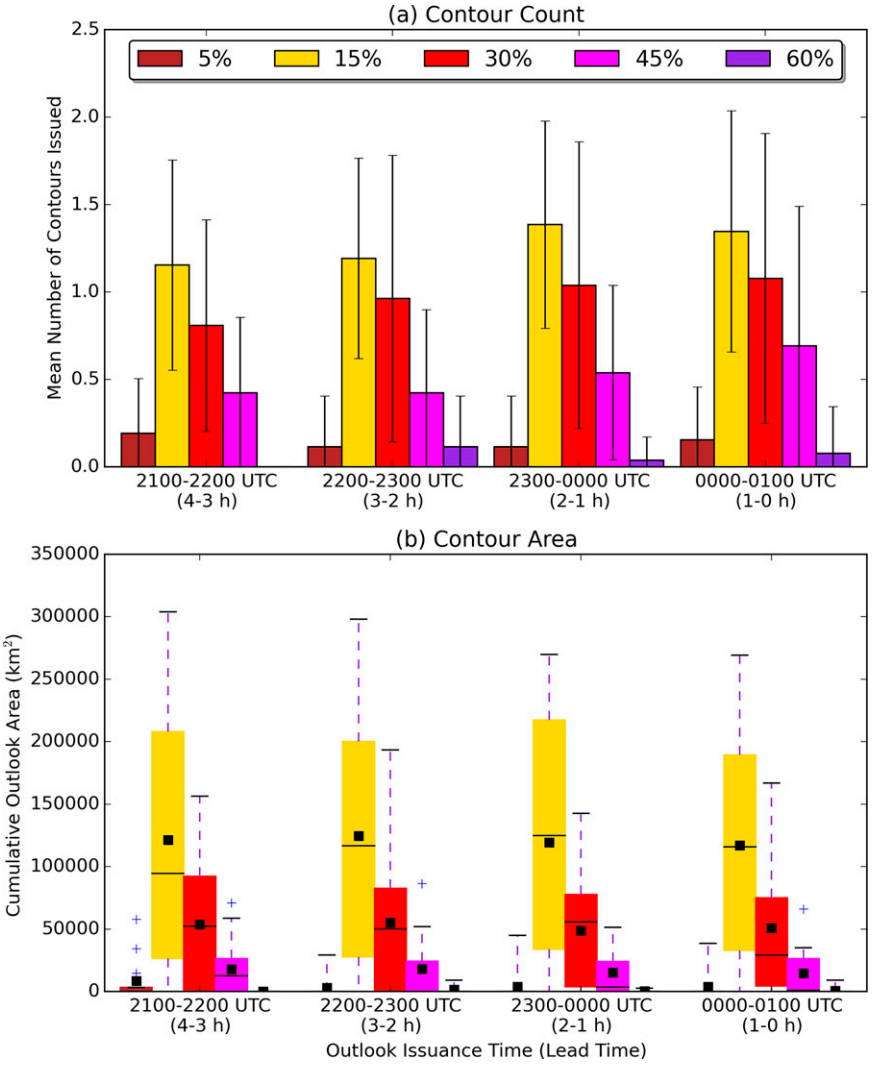

FIG. 2. Mean outlook (a) count and (b) area for targeted outlooks as a function of forecast issuance time. Error bars in (a) extend to ±1 standard deviation of the distribution. In (b), solid black lines represent the median, and black squares show the mean. Outliers are plotted as crosses.

UTC, etc. issued by forecaster 1 on 30 April 2019), and then the mean, minimum, and maximum correlations were determined for each pair across the 26 forecaster cases (13 days, with 2 forecasters working each day). Only outlooks that contained at least one contour were considered. Similarly, forecaster outlooks and WoFS UH probability correlations were computed for each forecaster case, but only correlations between outlooks and available guidance were analyzed (e.g., we excluded the 2100–2200 UTC forecaster outlook correlation with the 0000 UTC WoFS UH probabilities, as at 2100–2200 UTC the forecasters could not access the 0000 UTC guidance).

3. Results

a. Aggregate performance

We first characterize the experimental outlooks across the entire SFE, grouping together all 26 forecaster cases. An initial hypothesis was that the polygon sizes would decrease as lead time decreased and the area impacted by severe weather became clearer. Another initial hypothesis was that the probability values of the contours would trend toward extreme probability values (e.g., more low-end or high-end values) as lead time decreased, reflecting higher forecaster confidence in event occurrence or nonoccurrence as the valid time approached. Another mechanism by which more refined and confident forecasts might occur with decreasing lead time is that convection often initiates during the forecasting activities. Because WoFS rapidly assimilates satellite and radar data, its forecasts of ongoing storms are much more skillful than preconvective initiation, which should aid forecaster confidence.

Starting with the outlook characteristics, the mean number of 15% contours increased until ~2 h prior to the forecast valid time, before decreasing slightly in the final hour (Fig. 2a). However, the mean number of 30% probability contours

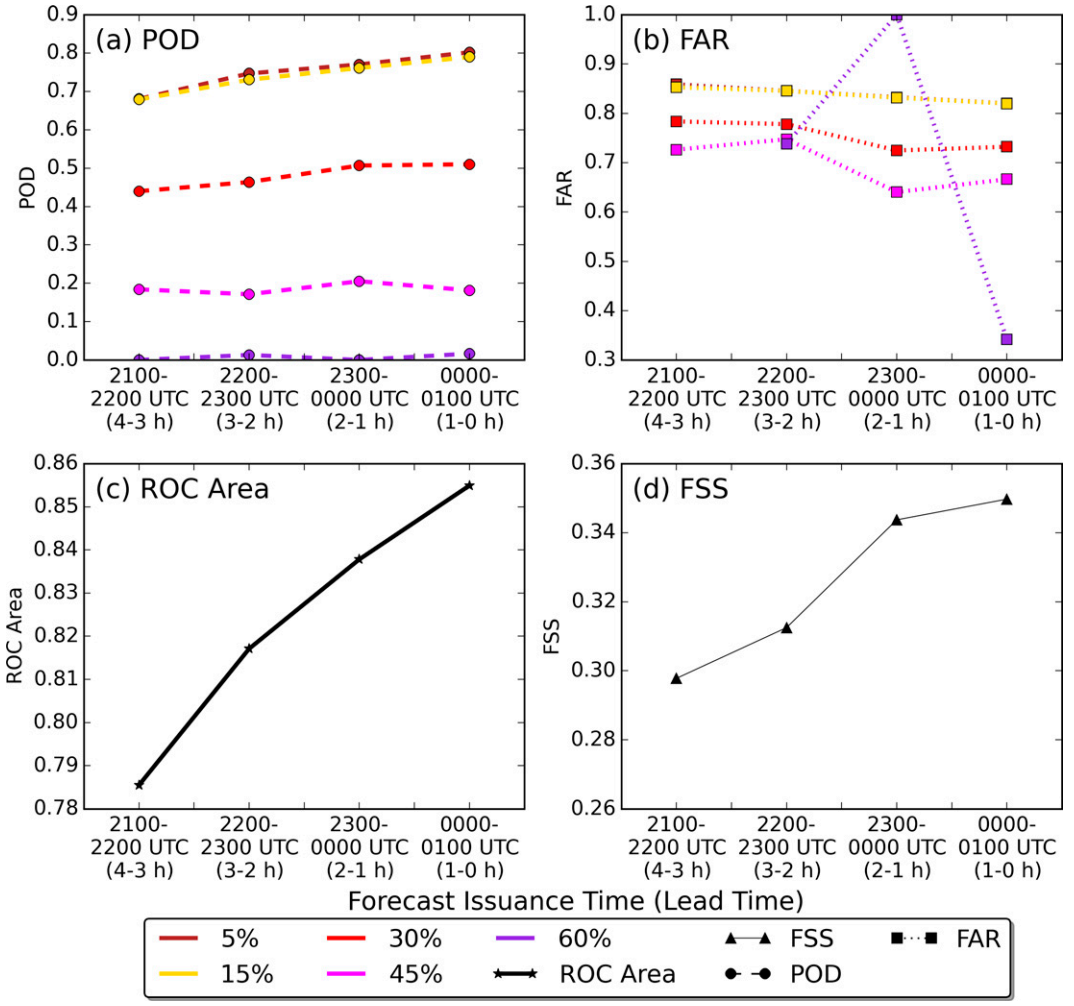

FIG. 3. (a) POD, (b) FAR, (c) ROC area, and (d) FSS for the targeted outlooks. Colored contours indicate the scores at different probabilistic thresholds, while black lines indicate metrics that encompass multiple probability levels.

increased steadily as the lead time decreased. This consistent increase in higher probability contours may indicate increasing forecaster confidence, suggesting that participants may have been pinpointing areas of interest. The highest contour possible for forecasters to draw, 60%, was drawn relatively infrequently (only 5 outlooks out of the 104 outlooks included in the analysis had a 60% contour). Thus, the small sample size prevents strong conclusions from being drawn about these highest contours; however, they are still included in analyses herein as a useful comparison to WoFS output, which frequently produced probabilities higher than 60%.

Ideally, a decrease in the area of the contours (especially lower probability contours) would accompany an increase in the number of higher probability contours, as forecasters honed in on areas of severe weather and potentially identified separate areas. A reduction in area could also decrease false alarm, which forecasters frequently consider when issuing forecast products (Brooks and Correia 2018).

Indeed, a slight decrease is seen in the 15% (30%) contour area means (medians), particularly at 2300–0000 and 0000–0100 UTC (Fig. 2b). The greatest decreases in mean area occurred between 2200–2300 and 2300–0000 UTC issuances for all contours. Thus, we also found slight support for the hypothesis that the forecast area decreased with decreasing lead time. The small magnitude of changes could be in part due to forecasters loading and adjusting prior forecasts allowing for large consistency hour-to-hour, as well as forecasters performing an unfamiliar task (probabilistic convective outlooks), which could cause fewer changes.

Moving from descriptive characteristics to verification statistics, we found minor improvements to overall skillful forecast performance across lead times (Fig. 3). ROC area and FSS both increased with decreasing lead time (Figs. 3c,d). Both an increase in POD and a decrease in FAR at the 5%, 15%, and 30% contours contributed to the increase in ROC area (Figs. 3a,b). The 45% contour showed a relatively large decrease in FAR

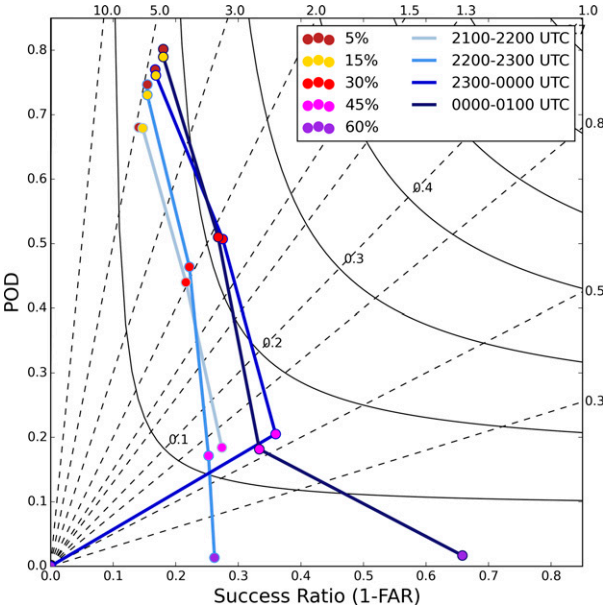

FIG. 4. Performance diagram for the forecaster outlooks across the entire sample period. Lighter line colors indicate earlier issuance times (i.e., longer lead time), and darker line colors indicate later issuance times (i.e., shorter lead time). Marker colors indicate the different probability thresholds.

between the 2200–2300 and 2300–0000 UTC issuance time, while the 45% contour POD remained consistent throughout. A performance diagram (Roebber 2009) shows the increase in CSI from the 2200–2300 to the 2300–0000 UTC issuance time occurring across largely the 30% and 45% thresholds (Fig. 4). CSI values increase from ~0.18 to ~0.24 for the 30% contour, an increase of about a third. Forecasters were able to improve both the POD and FAR with decreasing lead time at contours up to about 45%. The 2300–0000 UTC and 0000–0100 UTC forecast issuances are very similar, showing that the forecasts with ~2- and ~1-h lead time are performing consistently. Reliability diagrams concur with prior findings in that the statistics showed slight improvement with later lead times, and show general overforecasting (Fig. 5). The reliability of the 15% contour remained nearly identical through all of the forecasts, while the reliability of the 30% and 45% contours increased at later lead times, with the largest improvement occurring between the 2200–2300 UTC and 2300–0000 UTC forecast issuances.

Given the slight increase in skill of the forecast contours, it is reasonable to question how much forecasters adjusted their forecasts hour-to-hour. Forecast continuity and consistent messaging (Mileti and Sorensen 1990) are important considerations for forecasters and researchers alike (Williams and Eosco 2021). Some research suggests that users may lose trust in forecasts if they "flip-flop" between different forecast outcomes (Weyrich et al. 2019), though not all emerging research of message consistency within a weather context indicates that the public is averse to changing guidance (Burgeno and Joslyn 2020). While the forecasts issued herein were not meant for end-users, this perception of message consistency means that forecasters likely still tried to avoid drastic changes to the outlooks unless they were certain of the change. Forecasters also inherited their prior outlooks, and in most cases a dramatic shift in forecast thinking is unlikely. Therefore, the resultant forecasts often did not change much, and small incremental improvements occurred with subsequent outlooks. Forecast correlations show the similarity or difference between forecasts issued at different hours (Fig. 6). Mean correlation values between forecasts of adjacent hours were typically around 0.7. The smallest mean correlations between single hour forecasts occurred between forecasts issued at 2200 UTC and 2300 UTC, corresponding to when the largest changes in area and verification scores occurred. Maximum correlations exceeded 0.87 for all pairs of forecasts, showing that in some cases forecasters had only minor changes between the initial and the final outlook issuance time, despite decreasing lead time. Minimum correlations also showed that large changes to the forecasts sometimes occurred, such as on 14 May 2019 (see section 3c).

b. *WoFS performance*

Given the high hour-to-hour correlation between forecasts issued during the experimental forecast activity and the small improvements in skill, we also investigate the underlying WoFS guidance that forecasters were using through the 2–5-km UH probabilities. If the WoFS UH probabilities remain relatively unchanged from run to run, it is reasonable to assume that forecasters would also keep their outlooks consistent, particularly prior to convective initiation and maturation. Besides updating observational data, participants relied heavily on WoFS guidance, as it provided new ensemble information every half-hour. An example of the 2–5-km UH probabilities at the three available neighborhoods for the 23 May 2019 0000 UTC WoFS initialization is shown in Fig. 7.

WoFS 2–5-km UH probabilities generated skillful forecasts according to the ROC area at the 15- and 27-km neighborhoods, although the 9-km neighborhood ROC area values hovered around 0.7 at all initializations except for 2100 UTC (Fig. 8c). The WoFS POD of the 15% contours remained steady or increased slightly at shorter lead times for all neighborhoods, with the largest increases at the 27-km neighborhood (Fig. 8a). The POD of higher probability contours improved more as lead time decreased, perhaps due to the repeated assimilation of radar and satellite data. The larger the neighborhood size, the larger the POD at all probabilistic thresholds. FAR also increased slightly with increasing neighborhood size (Fig. 8b), though not as drastically as the POD did. Most neighborhoods and probability thresholds show a decrease in POD and an increase in FAR with the 2200 UTC initialization relative to the 2100 UTC initialization. This time (2200 UTC) corresponds to 1700 CDT/1800 EDT, when convective initiation often is occurring and severe weather is increasing (Krocak and Brooks 2020). This change across the expected time of convective initiation may pose a challenge for WoFS, which performs better when storms have been established long enough for data assimilation to produce accurate storm-scale analyses in model initial conditions. However, difficulty with convective initiation does not explain why the 2100 UTC initialization performs so well in terms of ROC area, POD,

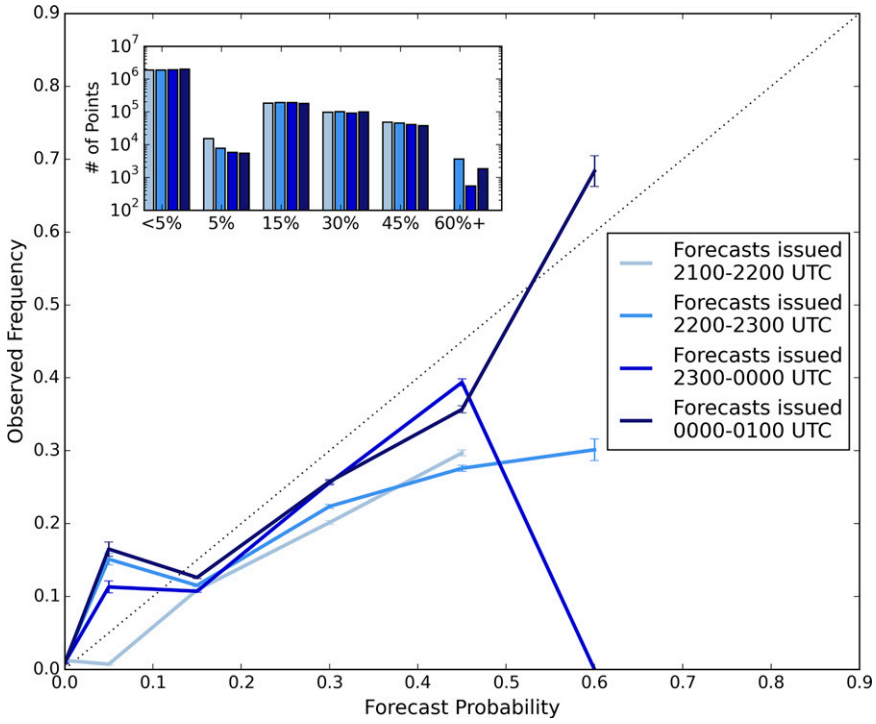

FIG. 5. Reliability diagram for all experimental forecaster outlooks issued across SFE 2019. The dashed line indicates perfect reliability. Inset bar chart shows the number of points in each bin. Error bars encompass the 95% confidence interval.

and FAR at all neighborhood sizes. ROC area at 2100 UTC was higher than subsequent lead times at all neighborhoods (Fig. 8c), perhaps due to increased ensemble spread at longer lead times decreasing the impact of ensemble underdispersion (Flora et al. 2019). Forecasters improved upon the ROC area of all initializations except for at 2100 UTC, and the FSS at 2200 UTC and 2300 UTC (Table 2). Unlike the model performance, forecaster ROC area and FSS increased consistently with decreasing lead time, showing that forecasters could correct for model deficiencies at shorter lead times.

Comparing forecaster performance to WoFS UH probabilities using a 27-km neighborhood shows how forecasters diverge from the guidance (Fig. 9). WoFS UH probabilities, like the human forecasters, perform best at 2300 UTC and 0000 UTC and show incremental improvement across thresholds. Unlike the human forecasters, the 2200 UTC initialization performs particularly poorly. Forecasters generally have a higher POD than WoFS at the 15% and 30% thresholds, but lower PODs at the 45% contours. FARs of forecasters at most thresholds are also larger than the UH probabilities. WoFS UH probabilities have higher CSIs than forecasters at the 15%, 45%, and 60% thresholds, but similar CSI scores at the 30% threshold. WoFS has far more instances of 60% probability than the forecasters which may be tied to 60% being the highest possible probability that forecasters could issue, leading them to reserve it for high-confidence scenarios with multiple reports likely. Smaller neighborhood probabilities generated by WoFS have lower skill scores (not shown).

Reliability diagrams of the WoFS UH probabilities show underforecasting at the lower probability thresholds of 5% and 15%, and overforecasting at probability thresholds of 45% and 60% (Figs. 10a–c). The 30% threshold shows mixed results, with the 15-km neighborhood probabilities (Fig. 10b) having nearly equal observed and predicted frequencies. Thus, the forecasters could improve upon the guidance at higher probability thresholds by improving the FAR that resulted from the guidance overforecasting. Initializations show no clear trend in reliability. Larger neighborhoods are the most reliable at 5% and 15% thresholds, while the smaller neighborhoods underforecast at those thresholds. All neighborhoods show a general increase in the observed frequency as the forecast probability increases. At the 45% and 60% probability thresholds, the smallest neighborhood is the most reliable, despite still overforecasting. The inset relative frequency plots show WoFS's ability to forecast higher probabilities than the forecasters, as the bin with the 60%+ probabilities has more points than the next-smallest bin. As expected, the number of points in that largest bin increases as the initialization time decreases, as WoFS hones in on and becomes more confident in the area that will see severe weather.

From the prior analyses, it becomes clear that there was more difference in the performance of the WoFS forecast and the forecaster outlook than there was between two sets of forecaster outlooks (e.g., WoFS guidance initialized at 2300 UTC and forecaster outlooks issued at 2300 UTC differed more than forecaster outlooks issued at 2200 UTC and 2300 UTC). The mean and maximum correlations between the

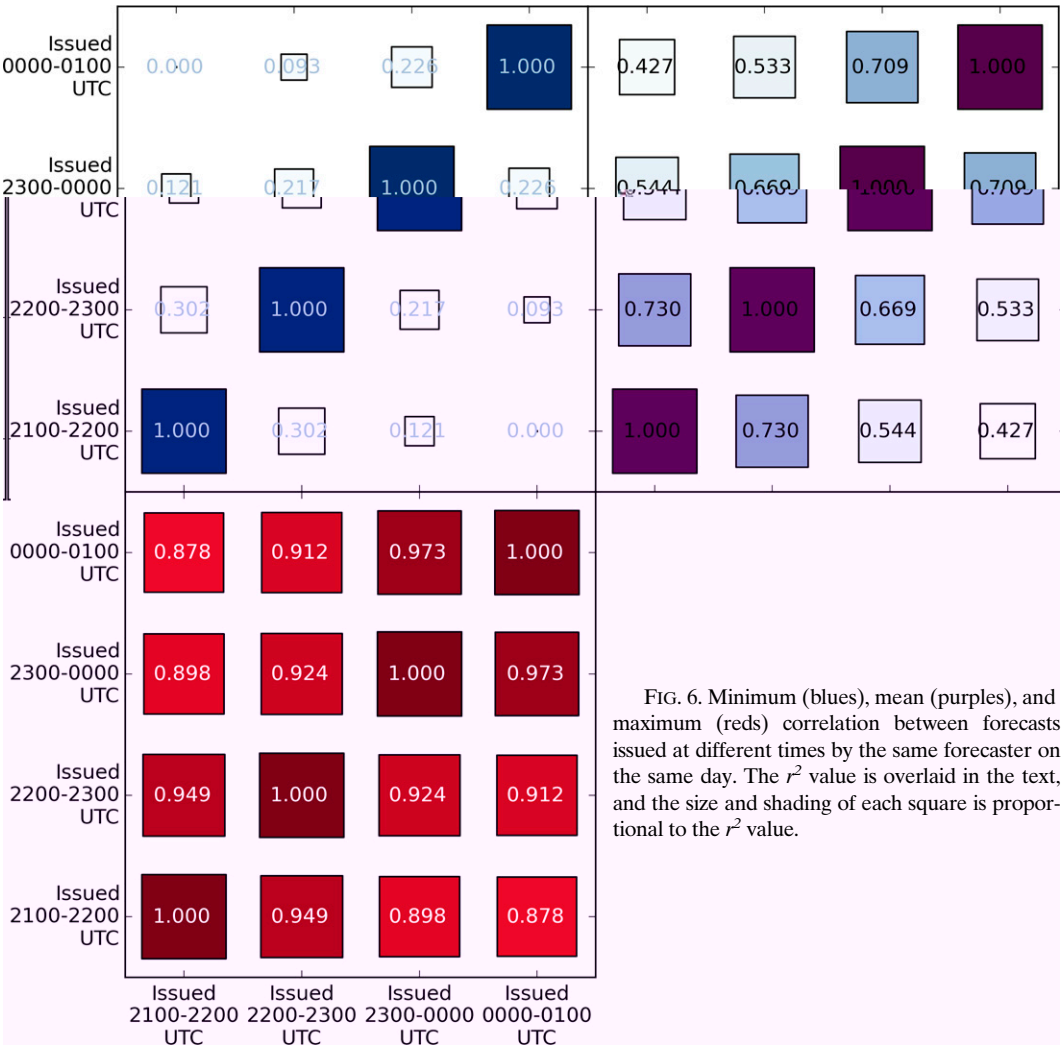

FIG. 6. Minimum (blues), mean (purples), and maximum (reds) correlation between forecasts issued at different times by the same forecaster on the same day. The r^2 value is overlaid in the text, and the size and shading of each square is proportional to the r^2 value.

guidance and the forecaster outlooks at each time (Fig. 11) further demonstrate the lower correlations relative to hour-to-hour forecaster outlooks (Fig. 6). We also surmise that a higher correlation between WoFS guidance and forecaster outlooks corresponds to the forecaster using that product more heavily.

Based on the expectation that guidance with a smaller neighborhood is more appropriate in the shorter-term and guidance with a larger neighborhood is more appropriate in the longer term, we hypothesized that higher correlations with the 9-km WoFS neighborhood probabilities would occur with later forecaster outlooks, while the higher correlations with the 27-km WoFS probabilities would occur with earlier forecaster outlooks. Instead, we see that the mean correlations between the forecaster outlooks and guidance at the same time decreases at all neighborhoods as the lead time decreases.

Decreased correlations may be partially due to the increasing importance of observations as the event approaches. Though WoFS assimilates observations, displacement errors still occur in the guidance, particularly for newly initiated storms, and these errors will become more apparent to forecasters as lead time to the outlook period decreases. Accounting for those errors would drop the mean and maximum correlations at the final hour (Fig. 11). Mean correlations were largest for the 27-km neighborhood and smallest for the 9-km neighborhood, perhaps indicating that forecasters were drawing areas that were closer in size to the larger neighborhood probabilities. Given that the probability definition for the outlooks was the probability of severe weather within 24.85 miles (40 km) of a point, the size of the forecaster outlooks should more closely match the larger neighborhood probabilities generated by WoFS. The maximum correlation for any forecaster-case shows the highest correlations between outlooks and the 9-km and the 27-km WoFS probabilities, suggesting that forecasters relied on those two products more than the 15-km guidance. The lower correlation with the 15-km guidance may be due to its radius being between two extremes; forecasters may have felt that the 9- and 27-km guidance provided a sufficient envelope for their

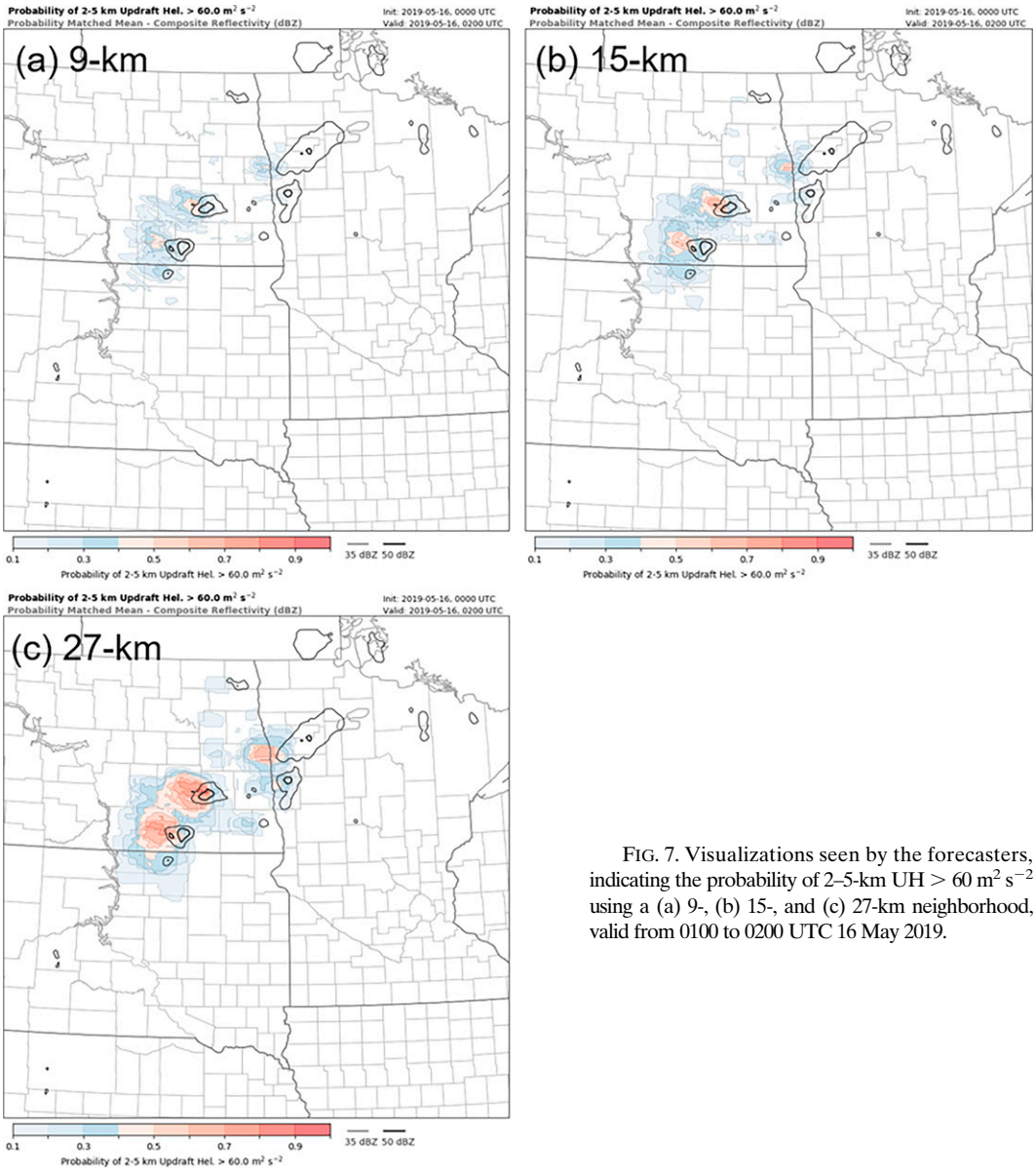

FIG. 7. Visualizations seen by the forecasters, indicating the probability of 2–5-km UH > 60 m^2 s^{-2} using a (a) 9-, (b) 15-, and (c) 27-km neighborhood, valid from 0100 to 0200 UTC 16 May 2019.

analyses, or may have prioritized the largest and smallest radius when needing to issue forecasts in a relatively short amount of time.

Mean correlation values with prior forecasts also tended to increase in the later forecasts (e.g., the 2100 UTC 9-km WoFS guidance was on average more correlated with the 2200 UTC forecaster outlook than with the 2100 UTC forecaster outlook). This pattern of increased correlation with guidance from prior WoFS initializations may be due to many factors. Forecasters might be building upon prior WoFS runs as they gain information about feature consistency and certainty. If a feature is present in roughly the same location for multiple runs and a forecaster highlights it in their outlook, and excludes other, less consistent features, the correlation with prior outlooks depicting the consistent feature could increase. Persistent run-to-run features in WoFS guidance are likely to originate from features consistently assimilated into the initial conditions, rather than simply initiated by the ensemble. Spurious features in individual WoFS members are also likely mitigated by analyzing the ensemble probabilities. WoFS forecasts also have some latency, so forecasters may build outlooks largely off of the prior hour's guidance and then make small adjustments based on the current run when it becomes available.

c. *Case performance*

Given the small variations in performance when collectively examining the SFE cases, we next examine individual cases,

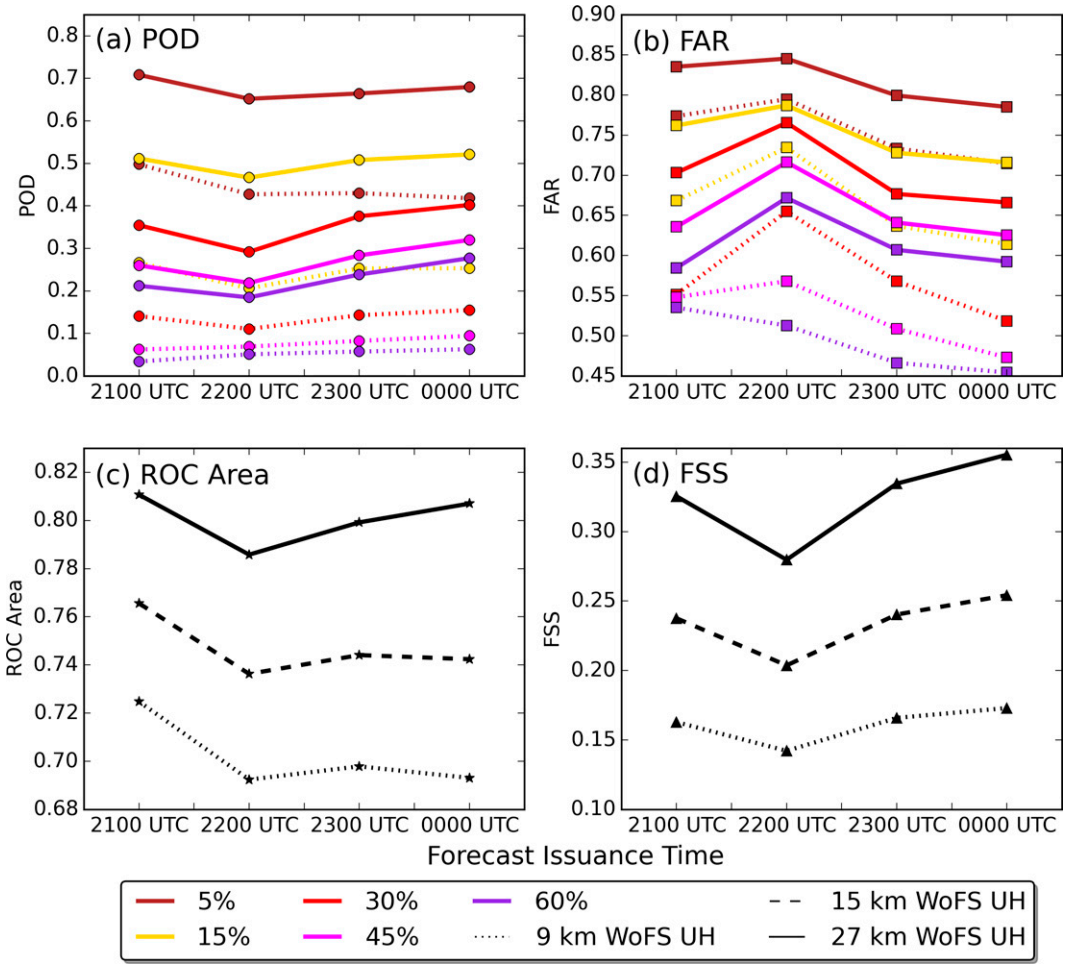

FIG. 8. As in Fig. 3, but for WoFS UH probabilities with neighborhoods of 9 km (dotted lines), 15 km (dashed lines), and 27 km (solid lines). In (a) and (b) the 15-km neighborhood was excluded for ease of readability; values fell between the 9- and 27-km neighborhoods at all probability thresholds.

to determine whether the aggregation of cases masked differences in performance under certain circumstances. For these analyses, the two forecasters for each case are grouped together.

The ROC areas (Fig. 12a) vary between the WoFS UH probabilities or the forecaster outlooks performing best, although whether forecasters perform better or worse than the guidance is typically consistent on any given day. In some cases, like 30 April, 21 May, and 23 May, forecasters improve on the WoFS guidance at all neighborhoods. In other cases, such as on 1 May, 14 May, and 16 May, the forecaster outlooks and the guidance perform similarly. In two cases, 7 May and 8 May, the forecaster outlooks have lower ROC areas than all guidance for the first two forecasts, but improve for the 2300–0000 UTC and 0000–0100 UTC issuances and perform similarly to or slightly worse than the WoFS UH probabilities. In many of the cases, the ROC area of forecaster outlooks peaks at the 2300 UTC issuance before declining. In a few cases (16 May and 30 May), ROC areas improve drastically at the final forecast issuance. Both of these effects are likely due to forecasters extrapolating existing storms, honing in on the area likely to be impacted, and becoming more

TABLE 2. ROC area and FSS for the WoFS initializations and forecaster outlooks over the course of all SFE 2019.

ROC area	2100 UTC	2200 UTC	2300 UTC	0000 UTC	FSS	2100 UTC	2200 UTC	2300 UTC	0000 UTC
9-km WoFS	0.7248	0.6924	0.6979	0.6931	9-km WoFS	0.1629	0.1422	0.1659	0.1730
15-km WoFS	0.7656	0.7363	0.7441	0.7424	15-km WoFS	0.2376	0.2037	0.2403	0.2542
27-km WoFS	0.8107	0.7858	0.7992	0.8070	27-km WoFS	0.3254	0.2797	0.3345	0.3553
Forecasters	0.7855	0.8171	0.8378	0.8549	Forecasters	0.2978	0.3125	0.3437	0.3497

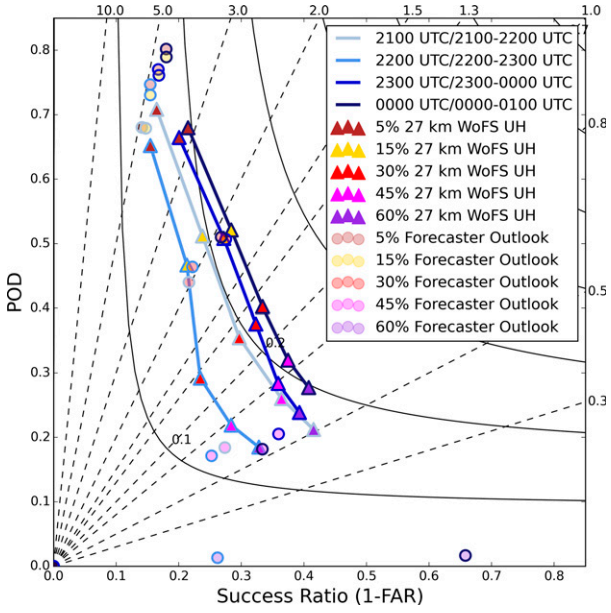

FIG. 9. A performance diagram showing the WoFS UH probability performance using the 27-km neighborhood (triangles; lines), with forecaster performance from Fig. 4 plotted for reference.

confident in the occurrence of severe weather. WoFS guidance consistently had higher ROC areas at larger neighborhoods, but did not always have increasing ROC areas as lead time decreased (Fig. 12a, grayscale markers). In five cases, 2100 UTC WoFS guidance ROC area was higher than 0000 UTC (e.g., 30 April, 15 May). Some of these cases saw WoFS UH probabilities increase for storms that did not produce severe reports, while others had a displacement in the simulated storms relative to observed storms. In some weakly forced cases, such as on 2 May or 30 May, the later iterations of the WoFS perform better once they are able to assimilate ongoing convection.

FSS performance (Fig. 12b) frequently mirrors the ROC area, though in some cases the FSS is better able to weight increasing forecaster confidence and increased probabilities. This is illustrated on 14 May, when forecasters correctly adjusted their probabilities northward and increased the probabilities as lead time decreased [see section 3c(2)]. While the ROC area for that case decreased slightly from 2200 UTC onward, the FSS shows steady improvement. For WoFS UH probabilities, particularly using the largest neighborhood, FSS scores often increase as lead time decreases, though some cases show mixed results between initializations. FSS shifts are likely due to increasing probabilities from the ensemble as lead time decreases and the ensemble members coalesce around the paths and severity of individual storms.

Case performance showed some link to the number of reports, as four of the five days with ten or more reports had high ROC areas compared to the other cases (Fig. 12a; 30 April, 7 May, 15 May, and 22 May). Cases with fewer reports generally had lower ROC areas, but some of these cases had large improvement in ROC area as the event approached, particularly for the forecasters (e.g., 23 May,

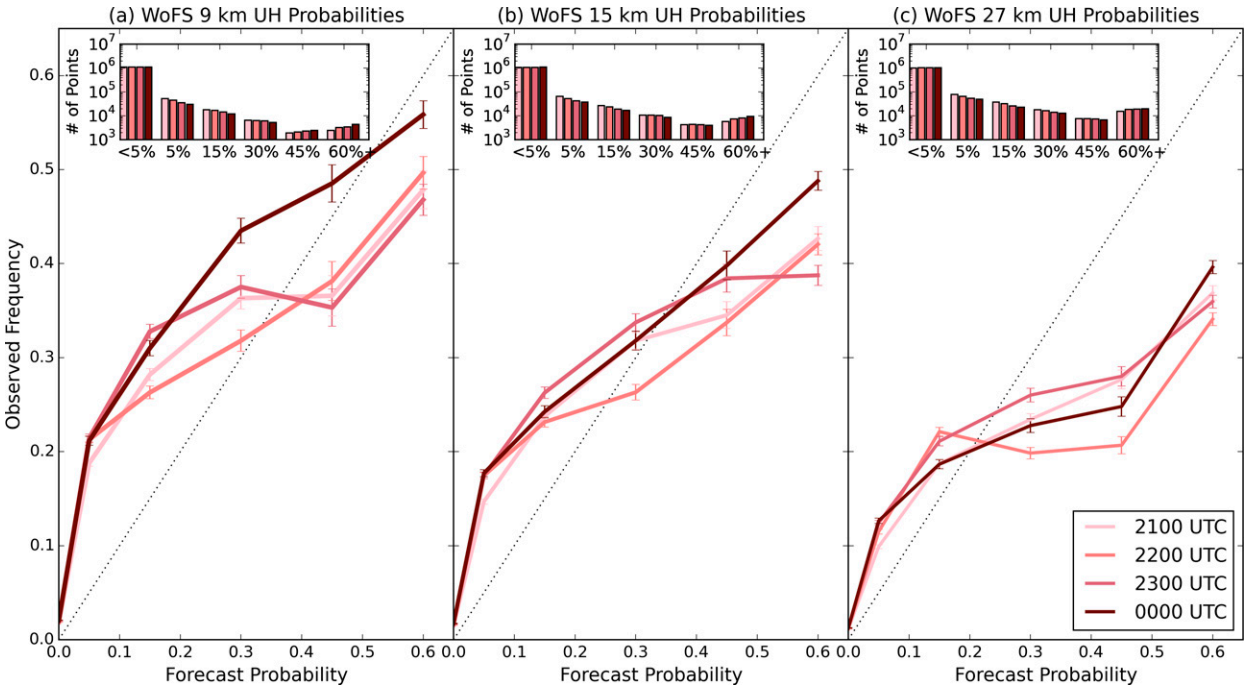

FIG. 10. Reliability of WoFS probabilities of UH > 60 $m^2 s^{-2}$ at different neighborhoods. Each point corresponds with a probabilistic threshold that forecasters could draw (5%, 15%, 30%, 45%, and 60%), and error bars encompass the 95% confidence interval. Times are WoFS initialization times.

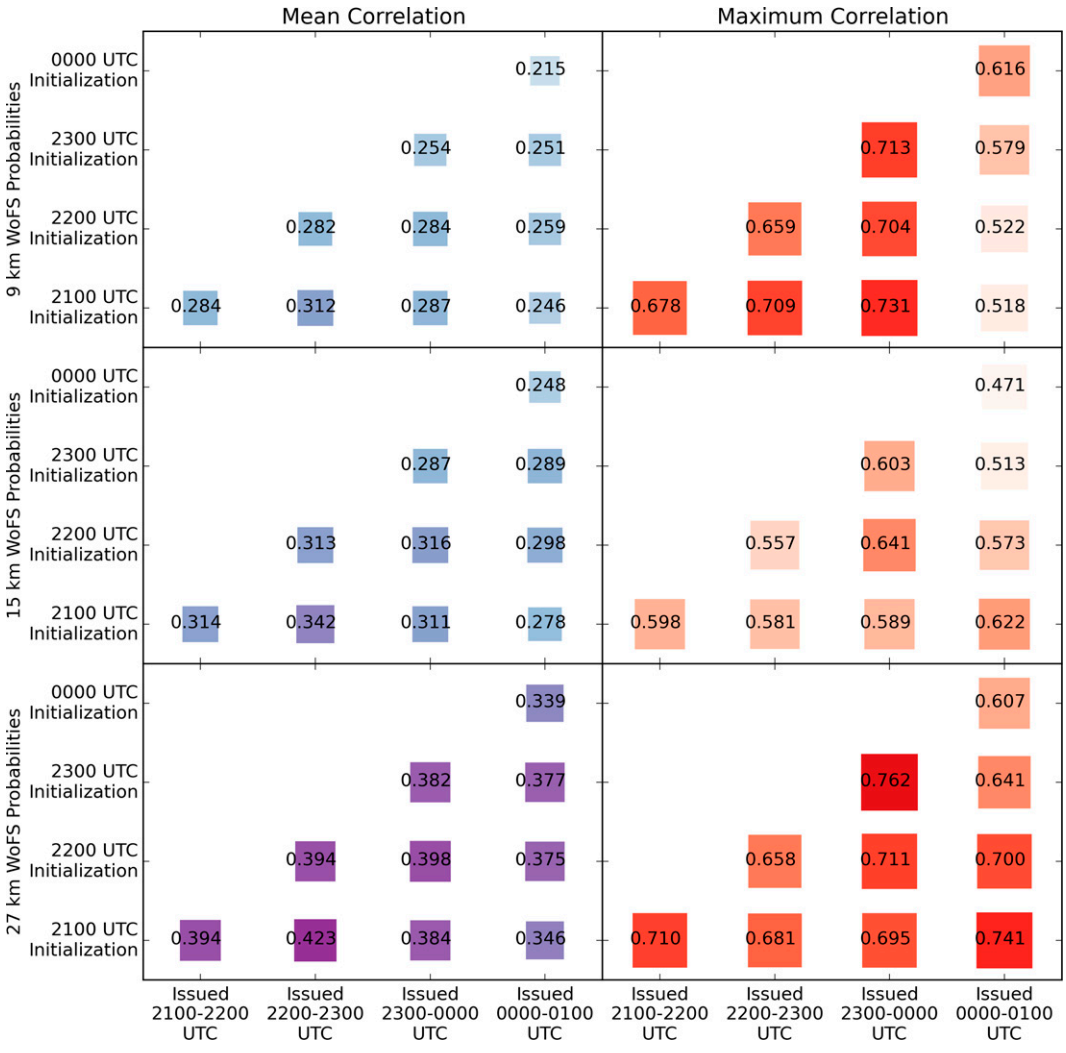

FIG. 11. Mean (purples) and maximum (reds) correlations between all of the experimental forecasts issued by forecasters and raw WoFS guidance for all days and issuance times. Correlations are only shown for forecasts that would have been available to forecasters at the given issuance time.

21 May, 30 May). Whether the guidance or the forecaster outlooks performed better in terms of ROC area was mixed, with experimental outlooks scoring consistently better on three of the five days with 10+ reports and mixed performance from the forecaster outlooks on 2 of the 5 days with 10+ reports. FSS showed a general decrease as the number of reports decreased (Fig. 12b), although 16 May was one of the worst-performing cases in terms of FSS despite having the second-highest number of reports. Generally, the WoFS UH probabilities at the highest neighborhood and the forecaster outlooks had similar FSS scores for cases with more than 10 reports, although both the WoFS UH probabilities and the forecaster outlooks had one case out of these five that performed best throughout the case.

Performance diagrams for 21 May (Fig. 13) show the emphasis forecasters place on capturing severe weather in at least the 15% contour, which maintains high POD throughout the case. Forecaster performance increases in the 2300–0000 UTC (Fig. 13c) and 0000–0100 UTC (Fig. 13d) outlooks relative to the 2100–2200 UTC (Fig. 13a) and 2200–2300 UTC (Fig. 13b) outlooks, with improved POD and SR. WoFS UH probabilities show no clear trend as the initializations progress, as the main area of cells was depicted too far west, and probabilities over another report site decreased as time went on. At 0000 UTC, the location error had decreased, though the probability magnitude error had not, leading to small improvements relative to 2200 UTC and 2300 UTC. However, the forecasters maintained high probabilities where WoFS decreased UH probabilities, and introduced higher probabilities closer to the reports, though they also suffered from some location error. Better performance from the forecasters for this case show that the aggregated statistics may mask days with marked improvement in the forecaster outlooks as lead time decreases.

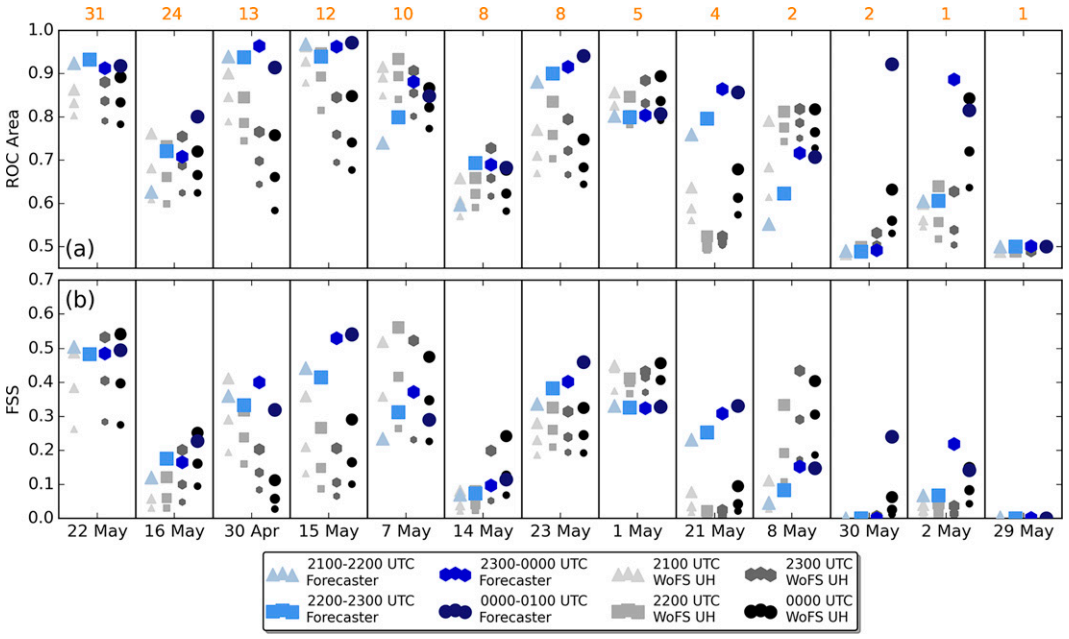

FIG. 12. (a) ROC area and (b) FSS for the forecasters (blues) and WoFS UH guidance (grays) for each case in the study. Earlier initializations or issuance times in each case are lighter colors, while later initializations or issuance times are in darker colors. The size of the markers indicating the WoFS guidance corresponds to the neighborhood of the probability, with the smallest symbols indicating the 9-km neighborhood probabilities and the largest symbols indicating the 27-km neighborhood probabilities. The orange text at the top of the graphic indicates the number of reports within the domain for each day.

Finally, we will examine three example cases (Fig. 14) selected based on ROC area and FSS: 1) 15 May, the best-performing case in terms of forecaster skill, 2) 14 May, one of the consistently worst-performing cases,[1] and 3) 30 May, a case with large improvement in the final forecast issuance.

1) 15 MAY 2019

On 15 May 2019, the focus for convection was a cold front moving from North Dakota into Minnesota, accompanied by a subtle upper-level shortwave trough. Early light precipitation across northern North Dakota made the northern extent of severe convection uncertain. Convection initiated on the front by 2100 UTC, becoming organized and cellular by 2200 UTC. Severe hail initially developed in the southern storm by 2300 UTC, and hail reports occurred throughout the line by 0000 UTC. The northern part of the line, closer to the stronger dynamical forcing, grew upscale and produced both hail and wind reports between 0100 and 0200 UTC, including an estimated significant wind gust [≥75 mph (33.5 m s^{-1})]. The southern storms, however, began to weaken and dissipate by the period of interest. Since storms had been producing severe reports prior to 0100–0200 UTC, this case had relatively high predictability. Forecasters highlighted a narrow corridor (Fig. 14a), introducing higher probabilities with decreasing lead time. Initial focus was on the southern storms, perhaps due to limited instability in northern North Dakota. However, as the northern storms began to grow upscale, forecasters increased their northern probabilities, with one forecaster introducing a 60% contour in the 0000–0100 UTC outlook that almost perfectly encompassed the severe reports. WoFS indicated some high UH probabilities in eastern North Dakota at the 2100 UTC initialization, but with the 2200 UTC initialization probabilities decreased and were widespread along the front. The 2300 UTC and 0000 UTC guidance emphasized the southern storms, depicting UH probabilities exceeding 70%. The consistently high probabilities issued by forecasters and the correct shift to focus on the northern storms decreased the correlation with WoFS guidance over time. In this case, the lower correlation at later initialization times showed the forecaster correcting for WoFS underdoing the severity of the northern storms as shown by the probability of UH > 60 m^2 s^{-2}.

2) 14 MAY 2019

Two areas of interest occurred within the WoFS domain on 14 May 2019 (Fig. 14b). Ongoing convection across Iowa at 1300 UTC created forecast questions about air mass recovery and remnant boundary placement, which could provide a focus for later severe weather. While most of Iowa remained clear of additional convection, storms initiated between 2300 and 0000 UTC over northern Missouri and quickly produced severe hail from 0000 to 0400 UTC, including several reports between 0100 and 0200 UTC. A second area of convection

[1] 29 May 2019 also performed very poorly, but was a low-end case with only one contour issued at one issuance time by a single forecaster. One wind report occurred. Thus, we focus on a poorly performing case with more reports and probability contours.

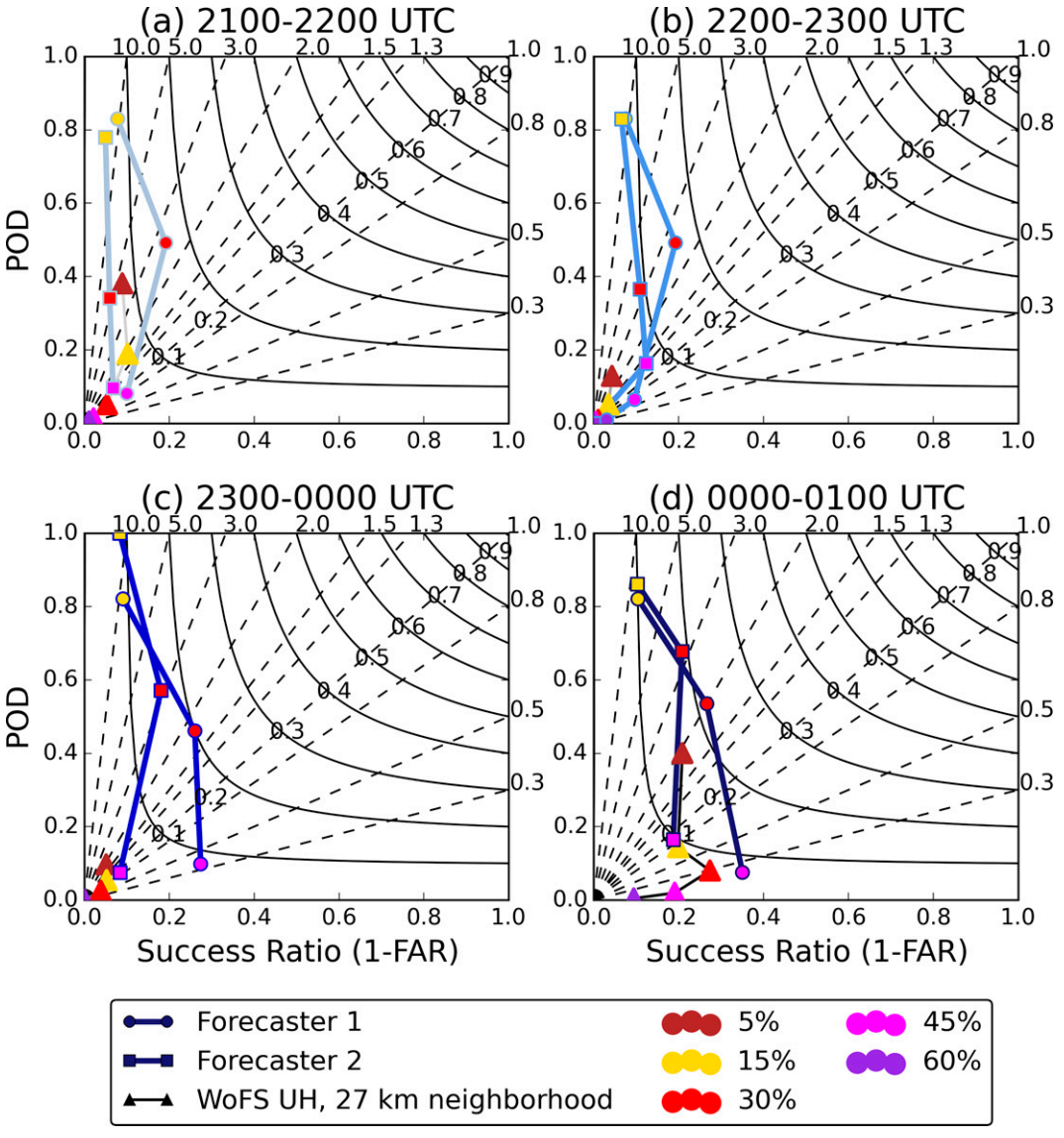

FIG. 13. Performance diagram for forecaster outlooks and WoFS guidance on 21 May 2021, issued or initialized at (a) 2100, (b) 2200, (c) 2300, and (d) 0000 UTC. The connected square and circle markers represent the two forecasters working this case, and the triangles indicate WoFS UH probabilities.

initiated in southwestern Minnesota and propagated southward into Iowa, producing two hail reports between 0000 and 0100 UTC before dissipating. WoFS UH probabilities highlighted both of these areas starting with the 2200 UTC run. The 2100 UTC WoFS initialization had low UH probabilities (<40%) across Missouri and no probabilities in Iowa. In the 2200 UTC initialization, probabilities increased across Missouri, but south of the eventual hail reports. Low probabilities first appeared in Iowa in this initialization. The 2300 UTC and 0000 UTC WoFS had higher UH probabilities, and shifted the Missouri probabilities northward, closer to the observed hail reports. Outlook correlations with WoFS guidance were lowest at 2100 UTC, increased at 2200 and 2300 UTC, and declined at 0000 UTC, with a larger decline at smaller neighborhoods. Overall, correlations between forecaster outlooks and WoFS guidance were lower than 15 May 2019, perhaps reflecting that this event was difficult to forecast for humans and model guidance alike. It may also be more difficult to improve upon model guidance when predictability is lower. A higher correlation was not necessarily optimal in this case, as WoFS introduced probabilities of severe weather in Iowa, where the convection produced no severe weather during the time of interest.

3) 30 MAY 2019

The 30 May 2019 event provided three areas of focus (Fig. 14c). The first was a cold front moving across the eastern CONUS, with the main forecast challenge for participants being

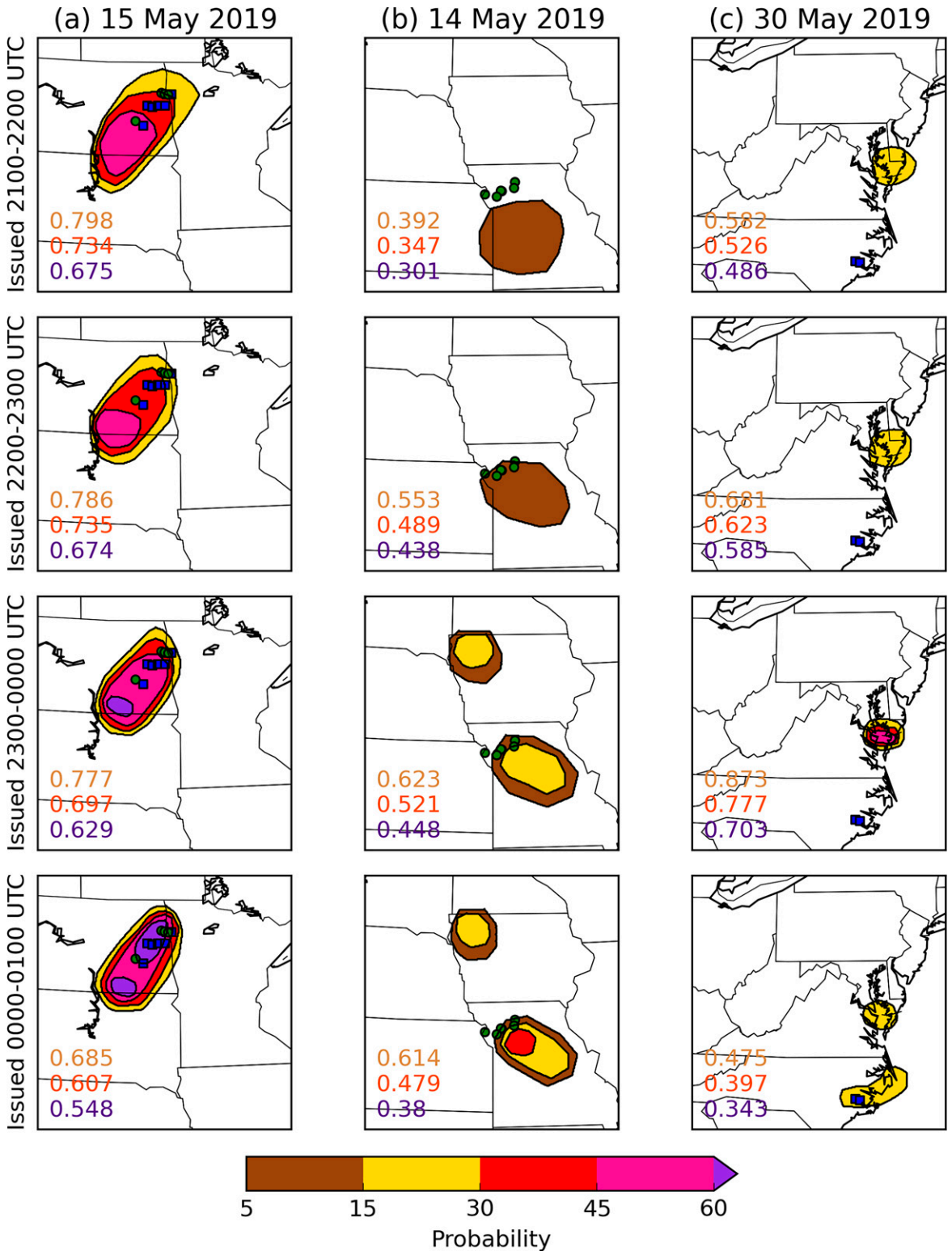

FIG. 14. Forecast evolution of one forecaster's outlooks valid 0100–0200 UTC for (a) 15 May 2019, (b) 14 May 2019, and (c) 30 May 2019. Overlaid text is the correlation of the forecaster's outlook with WoFS guidance generated using a 9-km neighborhood (purple), 15-km neighborhood (orange), and 27-km neighborhood (brown). Tornado (red inverted triangle), hail (green circle), and wind (blue square) reports are overlaid.

when the convection would move offshore. Two convective lines developed; one across Maryland, Pennsylvania, and northern Virginia that was largely synoptically driven, and one across North Carolina that developed off of the Appalachian Mountains. The northern band moved offshore at 2300 UTC, concurrent with the southern band beginning to produce severe wind reports. Participants correctly diagnosed the northern band as being primarily offshore from 0100 to 0200 UTC, drawing low probabilities across Maryland and Delaware. The southern band was thought to be displaced sufficiently from the better dynamics to remain subsevere; however, it did produce two wind reports between 0100 and 0200 UTC. Forecasters and WoFS UH probabilities did not anticipate this threat until late, with forecasters issuing a 15% contour in their 0000–0100 UTC outlooks encompassing the wind report and ~50% probabilities occurring in the 0000 UTC WoFS. Capturing these reports resulted in a high ROC area and increased FSS for the final initialization in Fig. 12, as the POD increased.

The final convection of interest was a lone supercell in northern Virginia, which tracked across the state and produced reports from ~2130 to 0010 UTC. This storm was well organized, and determining if it would remain severe until it moved offshore was a large forecast challenge. WoFS guidance from the 2100–2300 UTC initializations showed high UH probabilities continuing until the storm moved offshore, with decreased probabilities in the 0000 UTC run. However, the storm rapidly began to weaken after 0010 UTC, and it produced no reports during the time of interest. As such, forecasters' high probabilities issued at 2300–0000 UTC were a false alarm. Correlations with WoFS UH probabilities were highest at 2300 UTC, as forecasters agreed with the high probabilities in WoFS guidance at this time.

4. Conclusions

During the 2019 SFE, two forecasters issued experimental forecasts valid from 0100 to 0200 UTC using guidance from the experimental Warn-on-Forecast System (WoFS). The aim was to explore forecast skill in the time frame between severe convective watches and severe convective warnings during the peak of the spring convective season. Trends in forecaster contour count, area, and skill showed small changes when looking at the full duration of the SFE (Figs. 2, 3), with ROC area and FSS increasing incrementally in forecaster outlooks as lead time decreased, driven by a higher POD and a lower FAR. The largest changes in outlook skill were between the 2200–2300 and the 2300–0000 UTC issuances, with two to three hours of lead time (Figs. 4, 5). However, forecasters did incrementally adjust and improve their outlooks, as seen through correlations between forecaster outlooks (Fig. 6).

The relatively small changes in skill of the forecaster outlooks may be linked to the performance of the underlying WoFS 2–5-km UH probabilities, which also showed small improvements in performance as lead time decreased (Fig. 8). For ROC area and FSS, the skill of WoFS UH probabilities at 2100 and at 0000 UTC were very similar. This result is somewhat surprising given WoFS's advanced data assimilation, although all forecasts (even at long lead times) were skillful and forecasters were frequently unable to improve upon WoFS forecasts, particularly at high probabilities (Fig. 9). Forecasters did especially well at increasing the POD of low probability contour levels relative to WoFS guidance.

Individual cases vary, encompassing both cases where forecasters performed better than WoFS guidance and where forecasters performed similarly to WoFS guidance. Generally, forecasters improved as lead time decreased more consistently than the WoFS UH probabilities did on a case-by-case basis (Fig. 12, 13), showing that forecasters can compensate for potentially erroneous aspects of WoFS guidance. Three case studies illustrate different ways the forecasts evolved, showing the variation that participants encountered during SFE 2019 (Fig. 14).

To answer the questions raised in the introduction, forecasts seemed to "hone in" slightly, with higher probability contours and smaller areas as lead time decreased. However, forecasts in some cases (e.g., 30 May 2019), honed in by decreasing the probabilities due to increased certainty that severe weather would not occur. Forecaster outlook skill increased slightly and consistently as lead time decreased, while the skill of the WoFS UH probabilities varied with time. The largest changes in skill seemed to occur between the 2200 UTC guidance/forecasts and the 2300 UTC guidance/forecasts, perhaps due to timing relative to the most widespread convective initiation, which most frequently occurred between 2000 and 2200 UTC in our set of cases. Finally, forecasters frequently blended the full suite of WoFS guidance with observations of the mesoscale environment and ongoing convection. These results show that WoFS can help forecasters to issue skillful short-term forecasts, particularly given that forecasters were focusing heavily on WoFS guidance during this experimental task.

Results from this work suggest that WoFS will be a valuable tool for operational forecasters issuing short-term forecasts of severe convective hazards, particularly since its "long-range" forecasts were quite skillful. While many prior studies of WoFS have showed skill in the 0–3-h time range, this work shows skill extending beyond those first three hours. This work also shows how forecasters can improve upon the guidance in many cases. Since the Forecasting a Continuum of Environmental Threats paradigm (FACETs; Rothfusz et al. 2018) emphasizes probabilistic guidance, demonstrating skillful watch-to-warning scale probabilistic forecasts shows that forecasters have the ability to issue skillful and reliable probabilities at these temporal and spatial scales.

Additional work in subsequent SFEs has looked at separating these "all-hazard" forecasts into individual hazard forecasts of tornado, hail, and wind based on WoFS guidance. While experimental forecasts in SFE 2020 and SFE 2021 have taken place over one or two hours and not over the course of an entire evening, they may still provide insight into which hazards WoFS is particularly beneficial for. Additional work in SFE 2021 had forecasters issue short-term hourly probabilistic hazard forecasts with and without WoFS. While subjective ratings of those forecasts showed that forecasts using WoFS performed better than those not using WoFS for each hazard (Clark et al. 2021b), objective verification work remains. Finally, the full scope of this activity included the verification work discussed

here and an examination of forecaster product usage in Wilson et al. (2021). Future work of this type could include semistructured interviews or focus groups asking forecasters about their forecast process. Activities taking place in NOAA's Hazardous Weather Testbed frequently have the unique potential to answer questions about model guidance, forecast products, and forecaster usage of products. A multidisciplinary approach creates a more thorough understanding of experimental products than a single approach, and we recommend further research efforts that concurrently examine the applications, forecaster processes, and outcomes resulting from forecasters' use of newly developed products, tools, and guidance.

Acknowledgments. We thank the participants of this experiment for their efforts, as well as all those who contributed to SFE 2019 and the evening activity. This includes evening facilitators: Drs. Corey Potvin, Kimberly Hoogewind, Nusrat Yussof, and Derek Stratman. 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 (BTG, KAW, JC, KK, PS, BR). Authors PH and AJC completed this work as part of regular duties at the federally funded NOAA National Severe Storms Laboratory. Author ILJ completed this work as part of regular duties at the federally funded NOAA Storm Prediction Center. Finally, we thank four anonymous reviewers, whose comments helped to improve the clarity of the manuscript and the visualization of the figures.

Data availability statement. De-identified 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 two 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

Adams-Selin, R. D., A. J. Clark, C. J. Melick, S. R. Dembek, I. L. Jirak, and C. L. Ziegler, 2019: Evolution of WRF-HAILCAST during the 2014–16 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, **34**, 61–79, https://doi.org/10.1175/WAF-D-18-0024.1.

Alexander, C., and Coauthors, 2020: Rapid Refresh (RAP) and High Resolution Rapid Refresh (HRRR) model development. *30th Conf. on Weather Analysis and Forecasting (WAF)/26th Conf. on Numerical Weather Prediction (NWP)*, Boston, MA, Amer. Meteor. Soc., 8A.1, https://rapidrefresh.noaa.gov/pdf/Alexander_AMS_NWP_2020.pdf.

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

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.

Brooks, H. E., and J. Correia Jr., 2018: Long-term performance metrics for National Weather Service tornado warnings. *Wea. Forecasting*, **33**, 1501–1511, https://doi.org/10.1175/WAF-D-18-0120.1.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Wea. Forecasting*, **14**, 168–189, https://doi.org/10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2.

Burgeno, J. N., and S. L. Joslyn, 2020: The impact of weather forecast inconsistency on user trust. *Wea. Climate Soc.*, **12**, 679–694, https://doi.org/10.1175/WCAS-D-19-0074.1.

Clark, A. J., and Coauthors, 2020: A real-time, simulated forecasting experiment for advancing the prediction of hazardous convective weather. *Bull. Amer. Meteor. Soc.*, **101**, E2022–E2024, https://doi.org/10.1175/BAMS-D-19-0298.1.

——, and Coauthors, 2021a: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bull. Amer. Meteor. Soc.*, **102**, E814–E816, https://doi.org/10.1175/BAMS-D-20-0268.1.

——, and Coauthors, 2021b: Spring forecasting experiment 2021 preliminary findings and results. Experimental Forecast Program, NOAA Hazardous Weather Testbed, 86 pp., https://hwt.nssl.noaa.gov/sfe/2021/docs/HWT_SFE_2021_Prelim_Findings_FINAL.pdf.

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.

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

Doswell, C. A., III, H. E. Brooks, and M. P. Kay, 2005: Climatological estimates of daily local nontornadic severe thunderstorm probability for the United States. *Wea. Forecasting*, **20**, 577–595, https://doi.org/10.1175/WAF866.1.

Edwards, R., J. T. Allen, and G. W. Carbin, 2018: Reliability and climatological impacts of convective wind estimations. *J. Appl. Meteor. Climatol.*, **57**, 1825–1845, https://doi.org/10.1175/JAMC-D-17-0306.1.

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.

Gallus, W. A., 2010: Application of object-based verification techniques to ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 144–158, https://doi.org/10.1175/2009WAF2222274.1.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

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.

Hu, M., G. Ge, H. Shao, D. Stark, K. Newman, C. Zhou, J. Beck, and X. Zhang, 2017: Gridpoint statistical interpolation user's guide version 3.6. Developmental Testbed Center, 158 pp., https://dtcenter.org/com-GSI/users/docs/.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2013: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts. *Mon. Wea. Rev.*, **141**, 3413–3425, https://doi.org/10.1175/MWR-D-13-00027.1.

——, ——, Y. Wang, A. Reinhart, A. J. Clark, and I. L. Jirak, 2020: Neighborhood- and object-based probabilistic verification of the OU MAP ensemble forecasts during 2017 and 2018 Hazardous Weather Testbeds. *Wea. Forecasting*, **35**, 169–191, https://doi.org/10.1175/WAF-D-19-0060.1.

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 I: Combined radar and satellite assimilation. *Wea. Forecasting*, **31**, 297–327, https://doi.org/10.1175/WAF-D-15-0107.1.

——, 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., 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.

——, S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542, https://doi.org/10.1175/2010WAF2222430.1.

Kalina, E. A., I. Jankov, T. Alcott, J. Olson, J. Beck, J. Berner, D. Dowell, and C. Alexander, 2021: A progress report on the development of the High-Resolution Rapid Refresh ensemble. *Wea. Forecasting*, **36**, 791–804, https://doi.org/10.1175/WAF-D-20-0098.1.

Krocak, M. J., and H. E. Brooks, 2020: An analysis of subdaily severe thunderstorm probabilities for the United States. *Wea. Forecasting*, **35**, 107–112, https://doi.org/10.1175/WAF-D-19-0145.1.

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

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Mileti, D. S., and J. H. Sorensen, 1990: Communication of emergency public warnings: A social science perspective and state-of-the-art assessment. Tech. Rep. ORNL-6609, Oak Ridge National Laboratory, 159 pp., https://doi.org/10.2172/6137387.

Potvin, C. K., and Coauthors, 2019: Systematic comparison of convection-allowing models during the 2017 NOAA HWT Spring Forecasting Experiment. *Wea. Forecasting*, **34**, 1395–1416, https://doi.org/10.1175/WAF-D-19-0056.1.

Roberts, B., I. Jirak, A. Clark, S. Weiss, and J. 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 R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

——, 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.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

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.

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

Stensrud, D. J., and Coauthors, 2009: Convective-scale Warn-on-Forecast System: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, 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.

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.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the U.S. tornado database: 1954–2003. *Wea. Forecasting*, **21**, 86–93, https://doi.org/10.1175/WAF910.1.

Weyrich, P., A. Scolobig, and A. Patt, 2019: Dealing with inconsistent weather warnings: Effects on warning quality and intended actions. *Meteor. Appl.*, **26**, 569–583, https://doi.org/10.1002/met.1785.

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

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Williams, C. A., and G. M. Eosco, 2021: Is a consistent message achievable?: Defining "message consistency" for weather enterprise researchers and practitioners. *Bull. Amer. Meteor. Soc.*, **102**, E279–E295, https://doi.org/10.1175/BAMS-D-18-0250.1.

Wilson, K. A., P. L. Heinselman, P. S. Skinner, J. J. Choate, and K. E. Klockow-McClain, 2019: 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.

——, 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, E. D. Mitchell, J. T. Johnson, and K. W. Thomas, 1998: Evaluating the performance of WSR-88D severe storm detection algorithms. *Wea. Forecasting*, **13**, 513–518, https://doi.org/10.1175/1520-0434(1998)013<0513:ETPOWS>2.0.CO;2.

Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Wea. Forecasting*, **29**, 1451–1472, https://doi.org/10.1175/WAF-D-13-00135.1.