Object-Based Verification of Short-Term, Storm-Scale Probabilistic Mesocyclone Guidance from an Experimental Warn-on-Forecast System

MONTGOMERY L. FLORA,[a,b,c] PATRICK S. SKINNER,[b,c,a] COREY K. POTVIN,[a,c] ANTHONY E. REINHART,[b,c] THOMAS A. JONES,[b,c,a] NUSRAT YUSSOUF,[b,c,a] AND KENT H. KNOPFMEIER[b,c]

[a] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[b] *Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma*
[c] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*

(Manuscript received 8 May 2019, in final form 25 August 2019)

ABSTRACT

An object-based verification method for short-term, storm-scale probabilistic forecasts was developed and applied to mesocyclone guidance produced by the experimental Warn-on-Forecast System (WoFS) in 63 cases from 2017 to 2018. The probabilistic mesocyclone guidance was generated by calculating gridscale ensemble probabilities from WoFS forecasts of updraft helicity (UH) in layers 2–5 km (midlevel) and 0–2 km (low-level) above ground level (AGL) aggregated over 60-min periods. The resulting ensemble probability swaths are associated with individual thunderstorms and treated as objects with a single, representative probability value prescribed. A mesocyclone probability object, conceptually, is a region bounded by the ensemble forecast envelope of a mesocyclone track for a given thunderstorm over 1 h. The mesocyclone probability objects were matched against rotation track objects in Multi-Radar Multi-Sensor data using the total interest score, but with the maximum displacement varied between 0, 9, 15, and 30 km. Forecast accuracy and reliability were assessed at four different forecast lead time periods: 0–60, 30–90, 60–120, and 90–150 min. In the 0–60-min forecast period, the low-level UH probabilistic forecasts had a POD, FAR, and CSI of 0.46, 0.45, and 0.31, respectively, with a probability threshold of 22.2% (the threshold of maximum CSI). In the 90–150-min forecast period, the POD and CSI dropped to 0.39 and 0.27 while FAR remained relatively unchanged. Forecast probabilities $> 60\%$ overpredicted the likelihood of observed mesocyclones in the 0–60-min period; however, reliability improved when allowing larger maximum displacements for object matching and at longer lead times.

1. Introduction

A fundamental aspect of NOAA's Warn-on-Forecast (WoF; Stensrud et al. 2009, 2013) project is producing rapidly updating, short-term (i.e., 0–6-h), storm-scale, probabilistic severe weather hazard guidance. Several case studies have demonstrated that experimental WoF systems can produce accurate short-term probabilistic guidance for hazards such as tornadoes (Snook et al. 2012; Yussouf et al. 2013a,b; Wheatley et al. 2015; Yussouf et al. 2015; Jones et al. 2016), hail (Snook et al. 2016; Labriola et al. 2017, 2019), and heavy rainfall (Yussouf et al. 2016; Lawson et al. 2018a). With continual development of the experimental WoF system, it is critical to objectively assess the impact of system configuration changes (e.g., improvements in data assimilation or increasing grid resolution) or inclusion of post-processing techniques (e.g., machine learning calibration) on probabilistic forecast performance. Recently, object-based frameworks have become increasingly common for the verification of convective-allowing model forecasts of various severe weather hazards (e.g., Gallus 2010; Johnson et al. 2013; Clark et al. 2014; Cai and Dumais 2015; Stratman and Brewster 2017; Skinner et al. 2018; Jones et al. 2018; Adams-Selin et al. 2019). Object-based verification can easily diagnose or intuitively account for displacement errors between a forecast and observations as well as provide object properties (e.g., orientation, aspect ratio, area) as additional forecast attributes for evaluation (Davis et al. 2006; Ahijevych et al. 2009). Skinner et al. (2018, hereafter S18) established a

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

Publisher's Note: This article was revised on 24 October 2019 to correctly identify the third author's affiliations, which were not complete when originally published.

DOI: 10.1175/WAF-D-19-0094.1

baseline for the performance of deterministic thunderstorm and mesocyclone predictions produced by the real-time prototype WoF system (WoFS), then known as the NSSL Experimental Warn-on-Forecast System for ensembles (Wheatley et al. 2015; Jones et al. 2016). Using an object-based framework, they determined that deterministic forecasts provided for both thunderstorms and mesocyclones across 32 spring cases were skillful overall. However, a limitation of the work was that no assessment of the accuracy and reliability of the WoFS probabilistic guidance was performed. As an extension of their work, we seek to develop a novel object-based verification method for storm-scale probabilistic guidance and apply it to WoFS mesocyclone guidance.

Objective verification of probabilistic mesocyclone forecasts from convective-allowing ensembles has thus far been performed in the next-day (6–36 h) paradigm using grid-based frameworks with neighborhood postprocessing (e.g., Gallo et al. 2016, 2018, 2019; Sobash et al. 2016a; Dawson et al. 2017). For next-day forecasts, there are multiple compelling reasons for utilizing neighborhood postprocessing. First, at these forecast lead times, intrinsic predictability limits restrict skillful forecasts to broader mesoscale regions rather than the scales representative of individual convective storms (Lorenz 1969). Second, a well-documented flaw of grid-based verification in high resolution forecasts is the infamous "double penalty", where a small spatial displacement between the forecast and an observation leads to both a missed observation and false alarm forecast (Ebert 2008). The result is an unduly negative evaluation of a forecast's predictive skill since, operationally, small spatial displacements are tolerable. Postprocessing techniques such as neighborhooding, filtering, or upscaling (i.e., coarsening the verification grid) applied to both forecasts and observations can relax the condition of an exact match and instead assess the scale at which forecasts have the best performance [for a comprehensive discussion on such techniques see Gilleland et al. (2009, 2010) and Schwartz and Sobash (2017)].

A difference between WoF and next-day ensemble forecasts is that WoF is intended to provide forecast guidance for individual thunderstorms (Stensrud et al. 2009, 2013). Grid-based verification of WoF guidance can quantify errors associated with the numerical model or data assimilation technique. However, the neighborhooding/filtering/upscaling techniques used by grid-based verification smooth spatial scales associated with convective storms. Therefore, a goal of this study is to develop a complementary verification technique for WoF guidance that retains storm-scale forecast information, but allows for operationally tolerable spatial displacements. A novel, object-based framework is developed to assess the accuracy and reliability of the WoFS probabilistic guidance. In this framework, forecast probability swaths associated with individual thunderstorms can be conceived as "probabilistic" forecast objects with a single, representative probability value. Conceptually, we assign a probability of event[1] occurrence within a storm-scale region bounded by the forecast envelope of the event location. The prescribed probability value predicts the likelihood of a given storm producing an event rather than the likelihood of an event impacting any particular point; this distinction and the advantages of event-based probabilistic forecasts are further discussed in section 3. Another advantage of an object-based framework is that it allows classification of each predicted and observed storm as a "hit," "miss," or "false alarm," permitting calculation of contingency table statistics, which can provide diagnostic information on specific forecast errors (e.g., Wolff et al. 2014) unlike more traditional methods (Brown et al. 2011). Additionally, object-based verification emulates initial forecaster interpretations of WoF guidance, where forecasters key in on coherent areas of interest in the WoFS model ouput rather than using the forecast information in a strictly point-by-point basis (Wilson et al. 2019).

Section 2 provides a description of the forecast and verification datasets and object identification methods for forecast and observed rotation tracks. The novel aspects of this study are discussed in section 3 including producing ensemble probabilities from forecast rotation tracks, the distinction between grid- and object-based verification of probabilities, and object matching and verification of probability swath objects. Verification results for WoFS probabilistic mesocyclone guidance over a dataset of 63 cases from 2017 and 2018 are provided in section 4. Finally, section 5 provides a summary and discusses limitations of the study and future work.

2. Forecast and verification data

a. *Description of the forecast dataset*

The WoFS is a rapidly updating ensemble data assimilation and prediction system. WoFS consists of a 36-member multiphysics ensemble (see S18, their Table 1) that uses the Advanced Research version of the Weather and Research Forecast Model (WRF-ARW; Skamarock et al. 2008) with 3-km horizontal grid spacing. WoFS is initialized with initial and lateral boundary conditions provided by the experimental 3-km High-Resolution Rapid Refresh Ensemble (HRRRE; Dowell

[1] The event considered in this study is a mesocyclone; however, the technique is applicable to any storm-generated hazard.

et al. 2016) on a 750 km × 750 km domain recentered daily over the region of greatest severe weather potential. Radar, satellite (i.e., GOES-16 cloud water path), and Oklahoma mesonet (when available) observations are assimilated every 15 min with conventional observations assimilated hourly using the ensemble adjustment Kalman filter (Anderson 2001) included in the Data Assimilation Research Testbed (DART) software. After five 15-min assimilation cycles (i.e., starting at 1900 UTC), 18-member forecasts (a subset of the 36 analysis members) are issued every 30 min and provide forecast output every 5 min for up to 6 h of lead time.

This study uses 63 cases generated during the 2017 and 2018 Hazardous Weather Testbed Spring Forecasting Experiments (HWT-SFE; Gallo et al. 2017) and 2018 Hydrometeorology Testbed Flash Flood and Intense Rainfall experiment (HMT-FFaIR; Barthold et al. 2015; Albright and Perfater 2018). The WoFS configuration described above was used during the 2017 and 2018 HWT-SFEs, but during the 2018 HMT-FFaIR the domain was enlarged to 900 km × 900 km, the Community Gridpoint Statistical Interpolation–based ensemble Kalman square root filter (GSI-EnKF; Hu et al. 2017) was used as the data assimilation scheme, and forecasts were initialized every hour between 1800 and 0400 UTC. The changes to the domain size and forecast length introduced during the 2018 HMT-FFaIR experiments were designed to focus on heavy rainfall forecasts at longer lead times. Overall model performance between both configurations was similar (not shown). Although forecast periods varied, to ensure that cases were weighted equally, only forecasts initialized at the top of the hour between 1900 and 0300 UTC were considered for our evaluations.

To evaluate the skill and reliability of WoFS probabilistic mesocyclone guidance, 60-min forecasts of updraft helicity (UH) in the 2–5- and 0–2-km layers above ground level (AGL) are examined in this study. Assessing UH in the two different layers can help determine if WoFS probabilistic mesocyclone guidance accurately distinguishes between supercells with and without low-level mesocyclones, which can be used as a proxy for tornado occurrence (Trapp et al. 2005). To examine the decrease in skill of the WoFS probabilistic model guidance with forecast lead time, the following four 60-min forecast periods were used: 0–60, 30–90, 60–120, and 90–150 min.

b. Description of the verification dataset

The verification dataset for the WoFS probabilistic mesocyclone guidance is developed using radar-derived rotation tracks rather than local storm reports, similar to several recent studies (e.g., Skinner et al. 2016; Dawson et al. 2017, S18). Although radar-derived rotation tracks are imperfect they avoid some limitations of using local storms reports, which suffer from poor estimates of intensity (Trapp et al. 2006; Verbout et al. 2006), nonmeteorological bias (Brooks et al. 2003; Doswell et al. 2005) and undersampling in rural areas (e.g., Potvin et al. 2019). Low- and midlevel (0–2 and 2–5 km AGL, respectively) radar-derived rotation tracks are generated from the maximum range-corrected NSSL Multi-Radar Multi-Sensor (MRMS) cyclonic azimuthal wind shear data (Smith and Elmore 2004; Miller et al. 2013; Smith et al. 2016; Mahalik et al. 2019) in each layer calculated every 5 min over the WoFS domain. Following quality control and interpolation onto the WoFS grid (fully described in S18), these azimuthal wind shear data are aggregated to produce 60-min rotation tracks. In S18, radar data in regions too close or too far (i.e., less than 5 km or greater than 150 km) from the nearest WSR-88D site were ignored to mitigate range-related impacts. However, in this study, radar data outside the 150-km radius or inside the 5-km radius are included in both the forecast and verification dataset. Recalculation of verification scores presented in S18 showed minimal sensitivity to including these data (not shown).

c. Object identification

1) Forecast and verification rotation tracks

The goal of the object identification is to isolate strong mid- and low-level rotation that may be associated with severe weather (e.g., winds $> 50\,\mathrm{m\,s^{-1}}$, hail > 1.0 in., or a tornado) in both the forecast and verification dataset. In S18, single thresholds based on the 99.95th percentile value in the forecast and verification dataset were used for object identification. However, there are known limitations to the single threshold method. Object identification in a single threshold method will be sensitive to small changes in the size and intensity of objects near the threshold. Additionally, without using an excessively high threshold, the single threshold method can perform poorly at separating distinct, overlapping features. A candidate object identification method well suited to mitigate these issues is the enhanced watershed algorithm, which identifies local maxima and then grows objects pixel by pixel from a quantized version of the original field until they reach a specified area or intensity criteria (Lakshmanan et al. 2009). Objects are restricted from growing into regions less than the given minimum threshold (e.g., midlevel UH $< 40\,\mathrm{m^2\,s^{-1}}$) and once an object is identified, a larger region surrounding the objects is demarcated as a no-grow region for additional objects ensuring separation (i.e., the *foothills* region in Lakshmanan et al. 2009).

TABLE 1. Parameters of the Hagelslag watershed algorithm for all identified objects. The minimum and maximum intensity thresholds (min_thresh and max_thresh, respectively) for the azimuthal wind shear reflect that of the rescaled values. A larger saliency criterion (size_threshold_pixels) than past studies (e.g., Sobash et al. 2016a) was required to prevent tracks from being broken into multiple objects. For more details on the parameters, the open-source Hagelslag Python package is available at https://github.com/djgagne/hagelslag.

	Azimuthal wind shear	Low-level UH	Mid-level UH	Ensemble probabilities
min_thresh	$0.003 \times 10^4\,s^{-1}$	$10\,m^2\,s^{-2}$	$40\,m^2\,s^{-2}$	0
max_thresh	$0.008 \times 10^4\,s^{-1}$	$50\,m^2\,s^{-2}$	$250\,m^2\,s^{-2}$	75
data_increment	2	5	5	10
size_threshold_pixels	200	200	200	200

For this study, we are using the enhanced watershed algorithm available in the open-source Hagelslag Python package (Gagne et al. 2016), which is a Python implementation of Lakshmanan et al. (2009). The parameters for the Hagelslag enhanced watershed algorithm (Table 1) were tuned to improve the identification of both MCS and supercell rotation tracks, but there are sensitivities to these parameters. Given that objects identified by the enhanced watershed algorithm are restricted from growing into regions less than the minimum threshold, a higher minimum threshold can shrink objects or potentially separate tracks where the intensity fluctuates below the minimum threshold (a limitation of the single threshold method as well). However, lowering the minimum threshold identifies weaker rotation tracks where the intensity inside the object is similar to the minimum threshold. To address this concern, we applied the image processing concept of hysteresis (Jain 1989; Lakshmanan et al. 2009) where objects are identified at a lower threshold, but must contain pixels above a second, higher threshold. Essentially, the lower minimum threshold is used to prevent shrinkage and/or separation of identified objects, but the additional threshold removes objects with weaker intensity. Rather than using the maximum intensity inside an object for the second threshold, which can be unrepresentative and isolated to a single point, we used the 75th percentile value; a value representative of quarter of the pixels within an object. The choice of a 75th percentile value threshold for mid- and low-level azimuthal wind shear was varied between 0.003 and $0.005\,s^{-1}$ with the identified objects matched against local storm reports to determine a representative value for "severe" rotation. Although increasing the intensity value improved matches against the local storm reports, there were diminishing returns in bulk verification metrics as increasing the threshold removed too many objects. We also did not strive for a perfect match owing to the underreporting bias noted above. A 75th percentile threshold of $0.0035\,s^{-1}$ was found to best balance these identification criteria for both mid- and low-level azimuthal shear.

Object identification thresholds for mid- and low-level UH swaths were determined by trying to produce a similar number of forecast objects as observed objects. This was motivated by Sobash et al. (2016b) and Sobash and Kain (2017), where forecast fraction skill score was maximized when the number of severe surrogate probabilistic forecasts was equivalent to the number of severe reports. The thresholds for low (mid)-level UH objects found to produce a forecast object count similar to the observed object count are $20\,m^2\,s^{-2}$ ($80\,m^2\,s^{-2}$). Although these values were not hypertuned, they still reflect the current WoFS dataset and may be defined suboptimally. Furthermore, we found that decreasing these values and thereby increasing the number of forecast objects improved the contingency table metrics [increased critical success index (CSI)], but degraded reliability. Similar to Sobash et al. (2016b) and Sobash and Kain (2017), we found that matching the forecast object count to the observed object count was an optimal trade-off between the contingency table metrics and reliability.

Another sensitivity to the watershed method is that a larger area threshold (or saliency criterion as denoted in Lakshmanan et al. 2009) is required to prevent separation and shrinkage. However, in the current implementation of Hagelslag, the separation of local maxima is a function of the area threshold. Thus, when using a larger area threshold it is possible that only a single rotation track is identified among a cluster of two or more tracks. To allow for identification of additional nearby tracks, we set the separation of local maxima at 30 km.

After identification, a series of quality control measures were applied. First, forecast and observed objects that did not meet a 90-km^2 minimum area threshold were removed. Next, forecast and observed objects with a minimum distance than 12 km were merged into a single object and objects with a duration less than 15 min were removed. Finally, the 75th percentile value threshold (i.e., the hystersis threshold) was applied to remove weaker rotation tracks identified by the watershed method.

2) FORECAST PROBABILITY SWATH OBJECTS

Forecast probability swaths associated with individual thunderstorms can be conceived as individual "probabilistic" forecast objects with a prescribed single, representative probability value. The parameters for the Hagelslag enhanced watershed algorithm for identifying

probability objects are provided in Table 1. The parameters were tuned for identifying probability objects for both MCS and supercell cases, but fail to distinguish between closely spaced rotation objects. The poorer performance in these cases is attributable to the sensitivity of the enhanced watershed algorithm to the scale of the phenomena to be identified (noted in Lakshmanan et al. 2009) and absence of universal parameters that cover all relevant spatial scales. After object identification of the probability swaths, the maximum gridpoint probability within an object is assigned to each gridpoint. Ideally, the likelihood of a mesocyclone occurring within a given storm is the total number of ensemble members producing a mesocyclone divided by the ensemble size, which is typically equal to the maximum probability within the object. However, in some cases, UH forecast objects among the ensemble members may not overlap at a single grid point (particularly at later lead times). In these cases, the maximum number of ensemble members forecasting a mesocyclone at a given point will be less than the total number of ensemble members forecasting a mesocyclone within a given storm. In these instances the maximum probability within the object will underestimate the ensemble probability of a mesocyclone occurring within a given storm.

3. Object-based verification of probabilistic guidance

a. *Generating the ensemble probability of mesocyclone occurrence*

In Schwartz and Sobash (2017), multiple methods for generating forecast probabilities from CAM ensembles were discussed. To generate gridscale ensemble probabilities, f_{ij} forecasts for $i = 1, \ldots, M$ grid points and $j = 1, \ldots, N$ ensemble members are converted to binary using an event threshold q (e.g., rainfall > 1 in.) to produce N binary probability fields (BP)

$$\mathrm{BP}(q)_{ij} = \begin{cases} 1, & \text{if} \quad f_{ij} \geq q, \\ 0, & \text{if} \quad f_{ij} < q, \end{cases} \tag{1}$$

where the binary probability fields are a function of the event threshold. The ensemble probability (EP) at the ith grid point is then calculated as an ensemble average of the binary probability fields

$$\mathrm{EP}(q)_i = \frac{1}{N}\sum_{j=1}^{N} \mathrm{BP}(q)_{ij}. \tag{2}$$

In this study, we adopt a similar definition, but the binary probability field of mesocyclone occurrence at the ith grid point for the jth member (BP_{ij}) is defined using the forecast mesocyclone objects

$$\mathrm{BP}_{ij} = \begin{cases} 1, & \text{if} \quad i \in S_j, \\ 0, & \text{if} \quad i \notin S_j. \end{cases} \tag{3}$$

where S_j is the set of grid points within the forecast mesocyclone objects for the jth ensemble member. Calculating the ensemble probability from the quality-controlled forecast mesocyclone objects, rather than using an event threshold on the raw time-aggregated UH forecasts, helps ensure that the probability swaths are associated with coherent forecast mesocyclone tracks. For this study, no additional alterations (e.g., upscaling, smoothing, filtering, neighborhooding) are made to the ensemble probabilities of mesocyclone occurrence.

b. *Grid-based verification of WoFS mesocyclone probabilistic guidance*

Forecast probability accuracy and reliability are traditionally evaluated in a grid-based framework where forecast probabilities and observations are verified on the native grid (e.g., 3-km grid for our study) or upscaled and evaluated on a coarser grid. The reliability of the 0–60-min low-level UH probabilistic guidance on the native 3-km grid is given in Fig. 1d with an example forecast shown in Fig. 1a. The gridscale forecast probabilities exhibit the sharpness and spatial scales of individual thunderstorms, but greatly overpredict the likelihood of a mesocyclone impacting a point (similarly for midlevel UH; not shown). The large overprediction bias of the WoFS probabilistic guidance on the native 3-km grid indicates considerable underdispersion. Quantifying and attributing the underdispersion in the WoFS is beyond the scope of this paper but warrants future research.

Traditionally, neighborhood maxing and spatial smoothing is applied to forecast probabilities to correct for underdispersion, which can improve reliability. To improve the reliability of the forecast probabilities on the native 3-km grid without altering the observations requires substantial spatial smoothing ($\sigma = 300$ km), which is unsurprising as a given point in the WoFS domain had 0.02% chance of being within observed low-level rotation over the 63 cases. For a rare event, reliable forecast probabilities (especially on high resolution grids) will tend to be low, near the climatological frequency, especially as predictability decreases (Murphy 1991). This smoothing can limit the usefulness of WoFS probabilistic guidance to human forecasters for hazards associated with individual thunderstorms between the watch and warning time scales. This is because the smoothed probabilities can be misinterpreted as each thunderstorm having a low likelihood of producing an

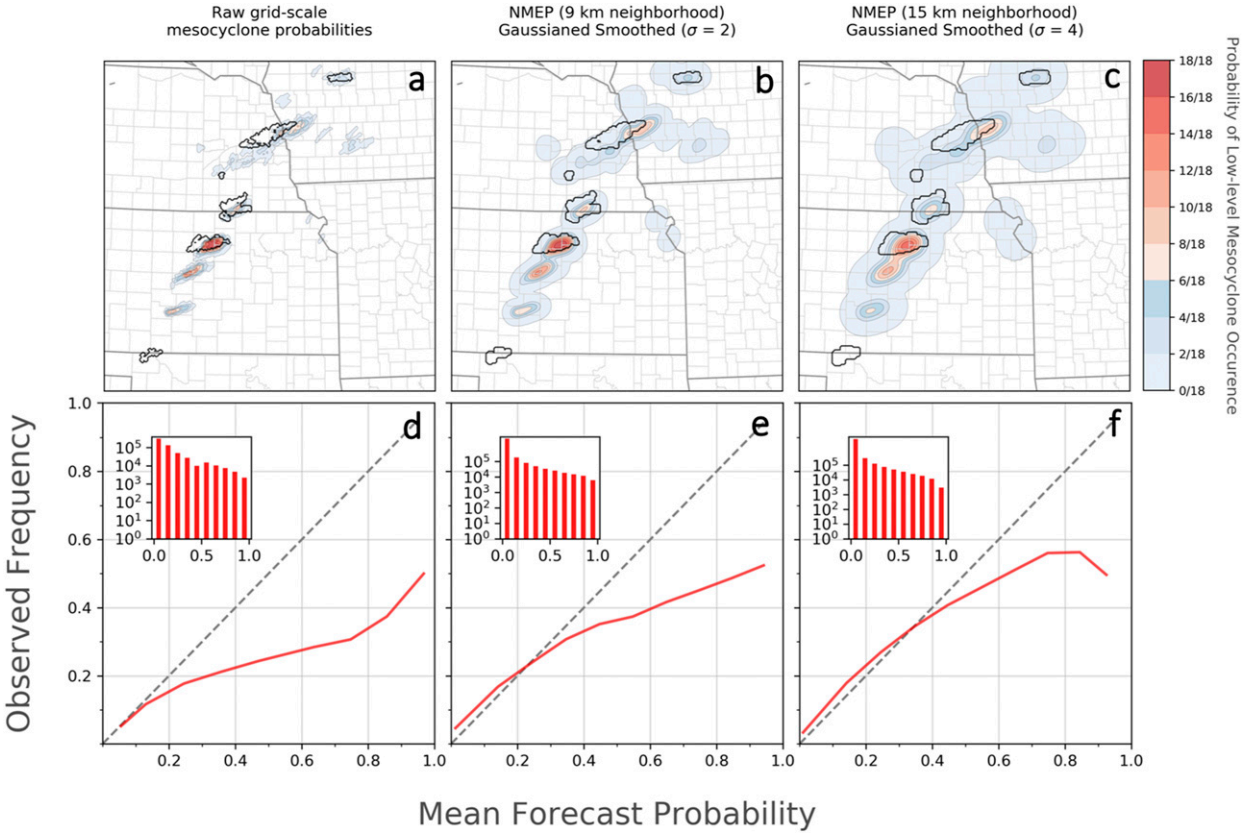

FIG. 1. (top) The 0–60-min probabilistic forecast of low-level mesocyclone occurrence initialized at 2300 UTC 1 May 2018 with (a) forecast probabilities and observations on the native 3-km grid and no postprocessing, (b) NMEP in 3×3 gridpoint neighborhood with Gaussian smoothing ($\sigma = 2$) and 3×3 gridpoint maximum value filter applied to the observations, and (c) NMEP in 5×5 gridpoint neighborhood with Gaussian smoothing ($\sigma = 4$) and 5×5 gridpoint maximum value filter applied to the observations. Observed hour-long low-level rotation tracks are outlined with black contours. (bottom) Reliability diagrams for the 0–60-min WoFS low-level updraft helicity probabilities calculated for all 63 cases and evaluated in a grid-based framework; (d)–(f) correspond to probabilities calculated in the manner described for (a)–(c).

event rather than an event impacting *any particular point* having a low likelihood; this ambiguity was pointed out in Ebert et al. (2011) for heavy rainfall forecasting.

It is possible to retain higher probabilities (e.g., >50%) using neighborhood maxing in combination with smoothing, but again at the cost of spatial resolution, as shown in Figs. 1b, 1c, 1e, and 1f. In Figs. 1b and 1e (Figs. 1c,f), the neighborhood maximum ensemble probability (NMEP; Schwartz and Sobash 2017) is calculated within a 3×3 (5×5) gridpoint neighborhood and smoothed with a 6-km (12-km) Gaussian filter a while 3×3 (5×5) gridpoint maximum filter was applied to the observations. It is worth noting that these neighborhoods are much smaller than those used for next-day convective-allowing ensembles (e.g., 40-km smoothing and maximum value radii are typical for next-day verification). Although improved reliability and higher probabilities are present in both cases (more so in Fig. 1f), much of the thunderstorm-scale forecast information has been filtered out. For example, the high probabilities associated with four distinct supercells in Kansas are strongly damped or aggregated into broad, coarser regions of forecast probabilities (cf. Fig. 1a with Fig. 1b or Fig. 1c). Ultimately, the forecast probabilities are unreliable on the native 3-km grid owing to underdispersion and improving reliability through postprocessing techniques obscures storm-scale information.

c. *Distinction between grid- and object-based verification of probabilities*

Figure 1a suggests WoFS, which uses rapidly cycled data assimilation to produce accurate storm-scale initial conditions, is capable of producing highly confident short-term forecasts of a rare event. To retain unsmoothed, high forecast probabilities valid at finer spatial scales we are drawing a distinction between *spatial* probabilities and *event* probabilities, which is illustrated in Fig. 2. Event probabilities predict the likelihood of a given storm producing an event *within a neighborhood*

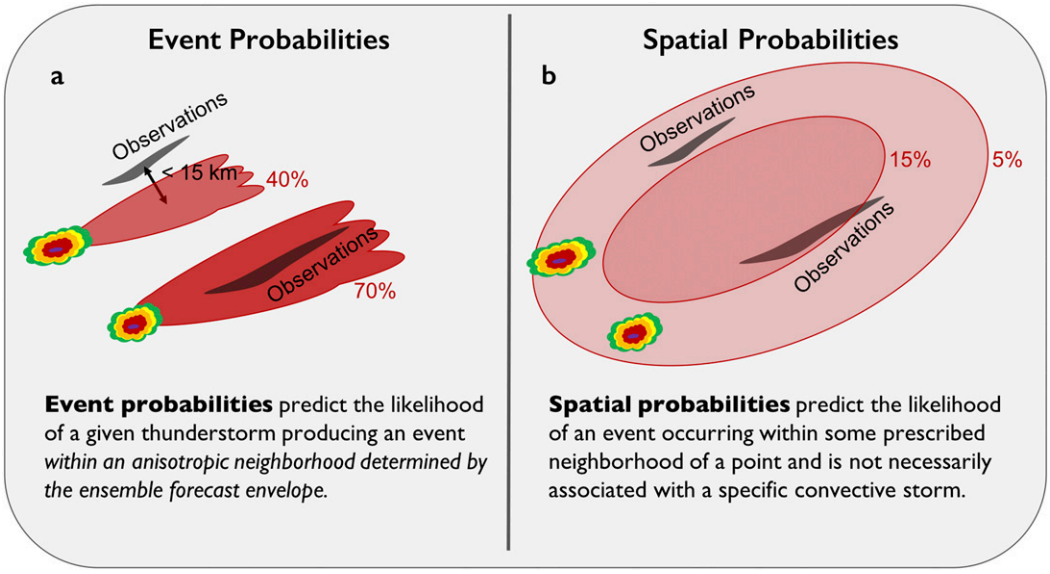

FIG. 2. Illustration of distinction between spatial and event reliability of probabilistic forecasts. (a) Event reliability measures the consistency of probabilistic forecasts associated with an individual thunderstorm within an anisotropic neighborhood determined by the forecast ensemble envelope [forecast probabilities (shown in red) are the likelihood of the event occurring]. (b) Spatial reliability measures the consistency of probabilistic forecasts of an event occurring within some prescribed neighborhood of a point and are not associated with a specific convective storm [forecast probabilities (shown in red) are the likelihood of the event impacting a particular point].

determined by the ensemble forecast envelope while spatial probabilities predict the likelihood of an event occurring *within some prescribed neighborhood* of a point and are not necessarily associated with a specific convective storm. Therefore, one can measure the consistency of probabilistic forecasts in complementary event- or spatial-based frameworks (e.g., the consistency of the spatial probabilities was assessed in section 3b). The event probability framework is tolerant of small spatial displacements between ensemble member forecasts of a mesocyclone, but is conditional on the predicted mesocyclones developing within the same parent thunderstorm. This effectively changes the interpretation of the forecast probabilities from the likelihood of an event occurring within a prescribed radius of a point to the likelihood a particular storm will produce an event. The ensemble-determined footprint is flow dependent and can grow in time as forecast uncertainty increases. The use of a static neighborhood in traditional methods, on the other hand, measures forecast quality at the same spatial scales for each available lead time. Event-based verification permits the consistency of WoFS's probabilistic guidance for rare events to be assessed for predictions of individual convective storms.

d. *Verification of probability swaths in an object-based framework*

We focus on two questions for evaluating WoFS probabilistic guidance:

1) Are probabilistic mesocyclone forecasts for individual thunderstorms skillful?
2) Are probabilistic mesocyclone forecasts for individual thunderstorms reliable?

To answer the first question, we apply object matching between the probability and observed rotation tracks objects. Object matching allows for calculation of verification metrics based on traditional contingency table statistics (i.e., hits, misses, and false alarms), which are intuitive and easily interpreted. Traditionally, matched forecast objects are classified as "hits," unmatched forecast objects as "false alarms," and unmatched verification objects as "misses." However, probability forecast objects generated from multiple predicted UH swaths (e.g., broad MCS probability objects) may overlap with several observed mesocyclones, especially at later lead times. In these situations, the number of "hits" in a single forecast will vary depending on whether matched forecast or observed objects are counted. Based on the contingency table, the total number of possible "hits" is the number of observed objects. Thus, when "hits" were classified as matched forecast objects, the number of hits was reduced within the contingency table, resulting in lower probabilistic forecast skill (roughly a 0.1 drop in CSI; not shown). Furthermore, to remain consistent in the contingency table, if "hits" are classified as matched forecast objects, then in situations with multiple

observed objects overlapping a single forecast object, all but one observed object would be considered a "miss." As this situation arises within probability swath objects associated with MCSs or nearby cellular convection, we classify "hits" as the number of observed rotation track objects that are matched to forecast probability objects.

The verification metrics in this study are limited to those that consider only hits, misses, and false alarms, which can be visualized using a performance diagram (Roebber 2009). Traditionally, probabilistic forecasts exceeding a probability threshold are considered a "yes" forecast and probabilistic forecast less than the probability threshold are "no" forecasts. These classifications allow for the contingency-table-based probability of detection (POD), false alarm ratio (FAR), success ratio (SR; 1 − FAR), frequency bias (or simply bias), and CSI to be used to quantify the skill of WoFS probabilistic mesocyclone forecasts. These metrics do not address the impact of correct negatives, which is a known limitation of object-based methods (Davis et al. 2009). Probability forecast objects can be labeled as "no" forecasts through a probability threshold, but they remain a poor sample of the "true" number of correct negatives for rare-event forecasting, given the majority of the forecast domain is not within any object. The necessity of ignoring correct negatives precludes the use of traditional probabilistic forecast verification metrics such as Brier skill score (BSS), the receiver operating curve (ROC), and area under the ROC (AUC). In this study, the probability thresholds used for defining probability swath objects and calculating contingency table metrics are the discrete ensemble probabilities [(1/18, 2/18, . . . , 18/18)].

To address the second question on assessing the reliability of the probabilistic mesocyclone forecast, we can use the event reliability definition from Fig. 2. Similar to grid-based reliability, probability objects can be binned based on their representative probability and compared against the observed frequency. For our study, the observed frequency is defined as the number of matched probability objects divided by the total (matched and unmatched) number of probability objects in a given probability bin. Unlike the contingency table metrics, probability objects are binned on every other discrete ensemble probability [(1/9, 2/9, . . . , 9/9)] as large variations in number of samples exist when binning on each discrete probability. Furthermore, using the method from Bröcker and Smith (2007), through bootstrap resampling we can generate a set of reliable forecasts from our dataset and compute variations in the observed frequencies. From the observed frequency variations, we can display the 5th and 95th percentile for each probability bin (known as consistency bars), which allows for immediate interpretation of the *reliability* of the reliability diagram. The extent to which the probabilistic forecasts are reliable is reflected by whether the observed frequencies fall within the consistency bars rather than the "distance from the diagonal."

The object matching in S18 used a simplified version of the total interest score [Davis et al. 2006; see Eq. (1) in S18] that included only the minimum spatial displacement and centroid and timing displacements. The timing displacement factor is not considered for the 60-min forecast periods used in this study. A match must exceed a minimum total interest score of 0.2, which effectively reduces the matching distance. To explore the sensitivity of forecast skill and reliability to matching distance, the maximum distance for both centroid and minimum displacement used in the total interest score is varied between 0, 9, 15, and 30 km and is hereafter referred to as the matching neighborhood.

The method for generating gridscale probabilities and identifying probability swaths as objects is summarized in Fig. 3. First, forecast rotation track objects are identified and quality controlled from the raw UH field for all ensemble members [Fig. 3a; section 2c(1)]. The gridscale ensemble probability of mesocyclone occurrence is then calculated from the forecast rotation track objects (Fig. 3b; section 3a), and probability swath objects are identified using the enhanced watershed algorithm with the maximum probability value assigned to the swath object [Fig. 3c; section 2c(2)].

It is worth noting that the probability object identification for the broad regions of probability in Nebraska (near the "A") is a limitation of the enhanced watershed algorithm. The enhanced watershed algorithm will stop growing objects once they satisfy the area criterion and an appropriate minimum area is based on the scale of the phenomena. Therefore, in cases of larger probability swaths, shrinkage of the identified probability swath may produce larger centroid and boundary displacement from observed object than if the object was identified to its full extent.

4. Results

a. Contingency table metrics

The performance of the probabilistic low- and mid-level UH forecasts for different matching neighborhoods and forecast lead times are shown in Figs. 4 and 5, respectively. The location of perfect performance, indicated by a CSI of 1, is in the upper-right corner, but for a probabilistic forecast with nonzero spread a perfect CSI is not possible (Hitchens et al. 2013). Additionally, the maximum CSI should correspond with POD comparable to SR (i.e., bias $\approx$ 1) to discourage forecast

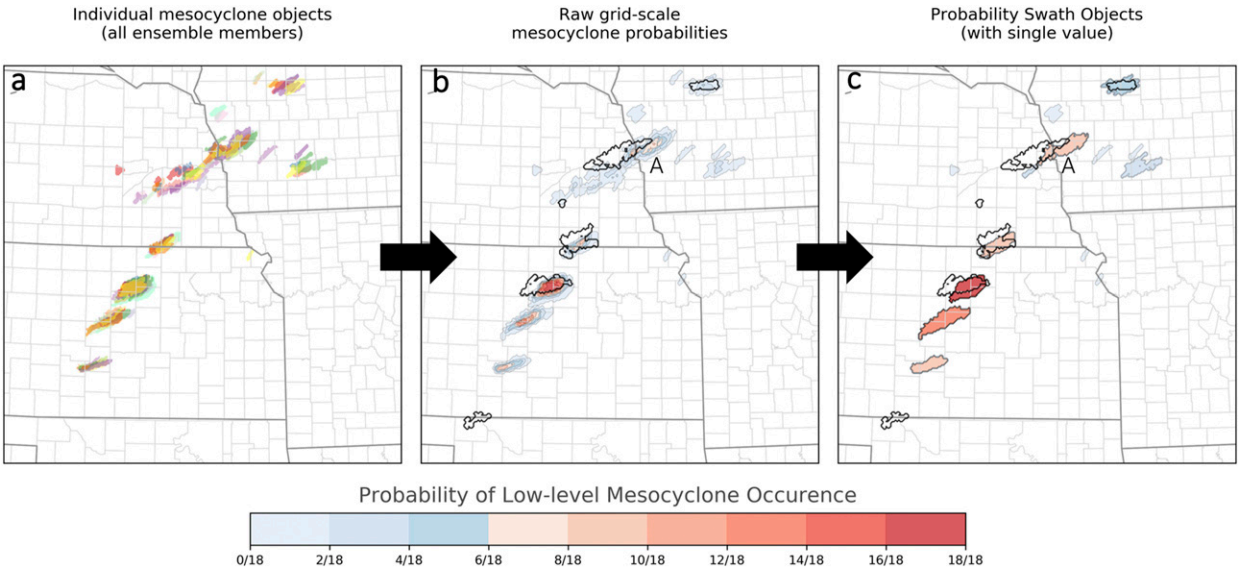

FIG. 3. Illustration of transforming individual ensemble member mesocyclone objects into probabilistic mesocyclone objects with a single, representative probability value. (a) Paintball plot of forecast mesocyclone objects identified from raw updraft helicity aggregated over 60 min, then quality controlled as described in section 2c(1). (b) Gridscale probabilities calculated from the mesocyclone objects as described in section 3a. (c) Probability objects are identified using the enhanced watershed algorithm and assigned the maximum probability occurring in the object (shown as the filled color). The technique is demonstrated using a 0–60-min probabilistic forecast of low-level mesocyclone occurrence initialized at 2300 UTC 1 May 2018. Observed hour-long low-level rotation tracks are outlined with black contours. The large probability swath near point A denotes a potential limitation of the watershed algorithm where objects can be shrunk compared to the raw probability field.

"hedging" (e.g., overforecasting to correctly predict observations).

The maximum CSI for low-level UH probability swaths tends to correspond with a probability threshold of 22.2% (4/18), independent of the lead time or matching neighborhood. The maximum CSI value ranges from 0.26 to 0.31 (based on the matching neighborhood) in the 0–60-min period (Fig. 4a) and drops to 0.21–0.27 in the 90–150-min period (Fig. 4d). Focusing on the probability threshold = 22.2% (4/18), the POD and SR for low-level UH in the 0–60-min period at the 30-km matching neighborhood is 0.46 and 0.47 leading to a bias close to 1 (0.97; Fig. 4a). These POD and SR values correspond to correct predictions of ≈50% of the observed low-level rotation tracks (with a similar success rate) out to 60 min of lead time. Even with a 0-km matching neighborhood (indicating overlapping forecast and observed objects), the WoFS low-level probabilistic guidance correctly predicted 40% of observed low-level rotation tracks. Looking at the different lead times for the 22.2% (4/18) probability threshold, the POD drops to 0.39 (30-km matching neighborhood) for the 90–150-min lead time (Fig. 4d). However, the SR remains relatively unchanged as the lead time increases. One explanation for the consistent SR values with increasing lead time may be that convection initiation at later lead times is poorly forecasted, resulting in an increasing number of misses without a corresponding increase in false alarms. The trend in POD with lead time results in a steady drop in bias to 0.85 (30-km matching neighborhood) at the 90–150-min lead time (Fig. 4d).

In general, as the probability threshold increases beyond 11.1% (2/18), there is a shift toward bias below 1, which is largely attributable to storm-scale predictability limits. Storm decay at later lead times in some ensemble members coupled with greater ensemble spread in mesocyclone location (increasing the likelihood of nonoverlaping UH tracks in members) will cause forecast probabilities associated with an individual thunderstorm to decay with lead time (Cintineo and Stensrud 2013; Flora et al. 2018). Therefore, the maximum probability for all probability forecast objects will decrease with increasing lead time. Thus, the number of probability forecast objects at lower (higher) probability thresholds will grow (drop) with increasing lead time effectively lower the bias at higher probability thresholds. This increasing number of probability objects at lower probability thresholds also explains why the contingency table metrics for probability thresholds ≤ 11.1% (2/18) appear insensitive to forecast lead time. Overall, the contingency table metrics and trends with increasing

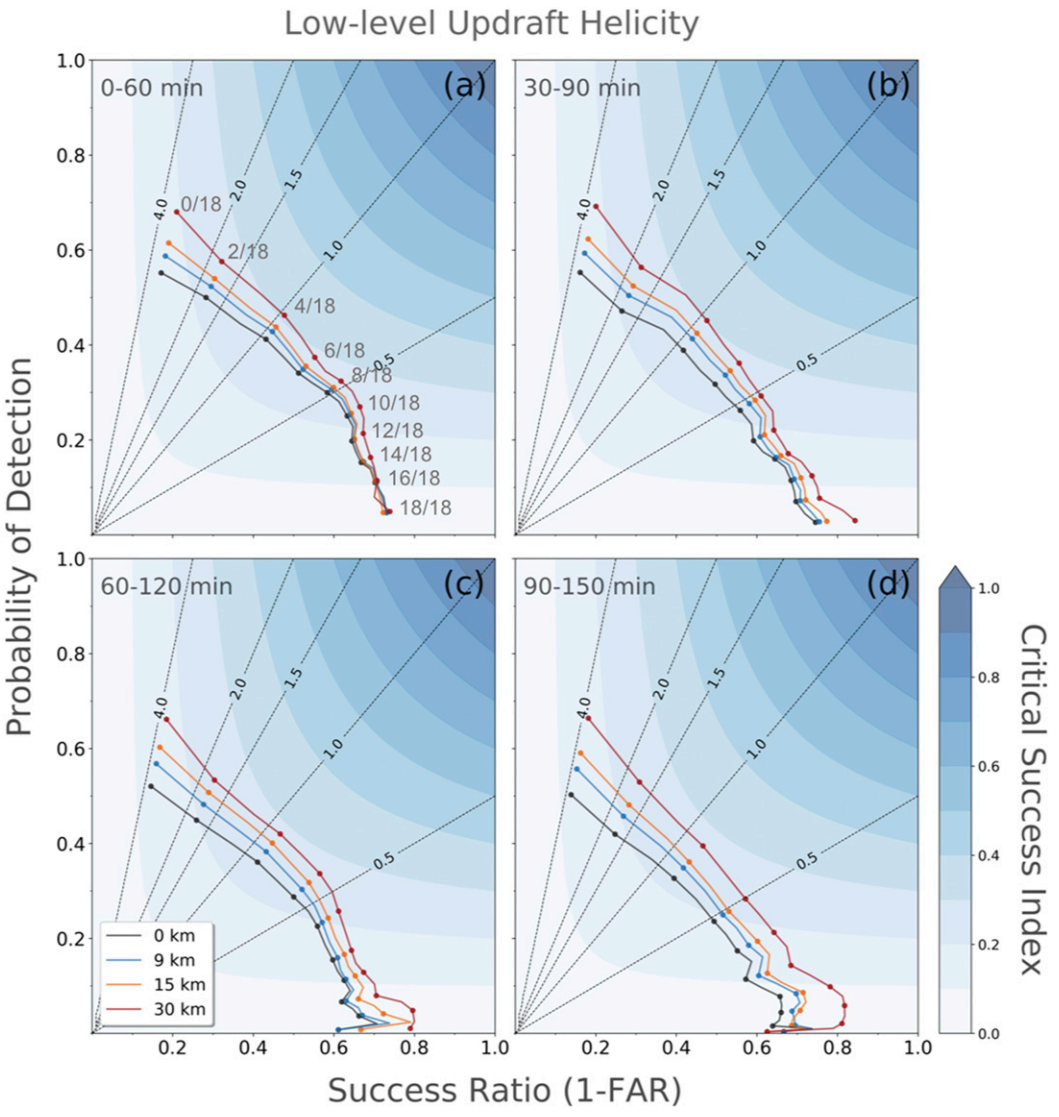

FIG. 4. Performance diagrams for WoFS low-level (0–2 km AGL) mesocyclone probability swath objects using 0-, 9-, 15-, and 30-km matching neighborhoods (gray, blue, orange, and red, respectively) and valid at (a) 0–60, (b) 30–90, (c) 60–120, and (d) 90–150 min. The dots represent the different probability thresholds [plotted every 11.1% (2/18)].

lead time for probabilistic forecasts of low-level UH are similar to those for midlevel UH (Fig. 5). The probability threshold corresponding with the maximum CSI in the midlevel UH objects varies between 22.2% (4/18) and 33.3% (6/18), dependent on forecast lead time. Using the probability threshold = 33.3% (6/18), the SR is greater than the POD in the 0–60-min period, unlike the low-level UH. At later lead times, however, the maximum CSI of the midlevel UH forecasts generally have a bias of 1 (Figs. 5c,d). The CSI for the midlevel UH forecasts tend be slightly less than corresponding thresholds in the low-level UH forecasts (cf. Fig. 5 and Fig. 4). This is in contrast to the results of S18 that found midlevel UH forecasts had slightly higher CSI than low-level UH in the deterministic verification. A possible explanation is that the current study includes more summertime events where the WoFS may be overpredicting midlevel rotation. There was also a similar drop in POD in the midlevel UH forecasts as compared to the low-level UH, but nearly constant SR at the later lead times leading to the bias dropping below 1. Ultimately, the differences between UH in the two layers are very small and may not be significant.

Last, some additional characteristics of low- and midlevel UH probability swath object accuracy in the performance diagrams are noted. First, separation between

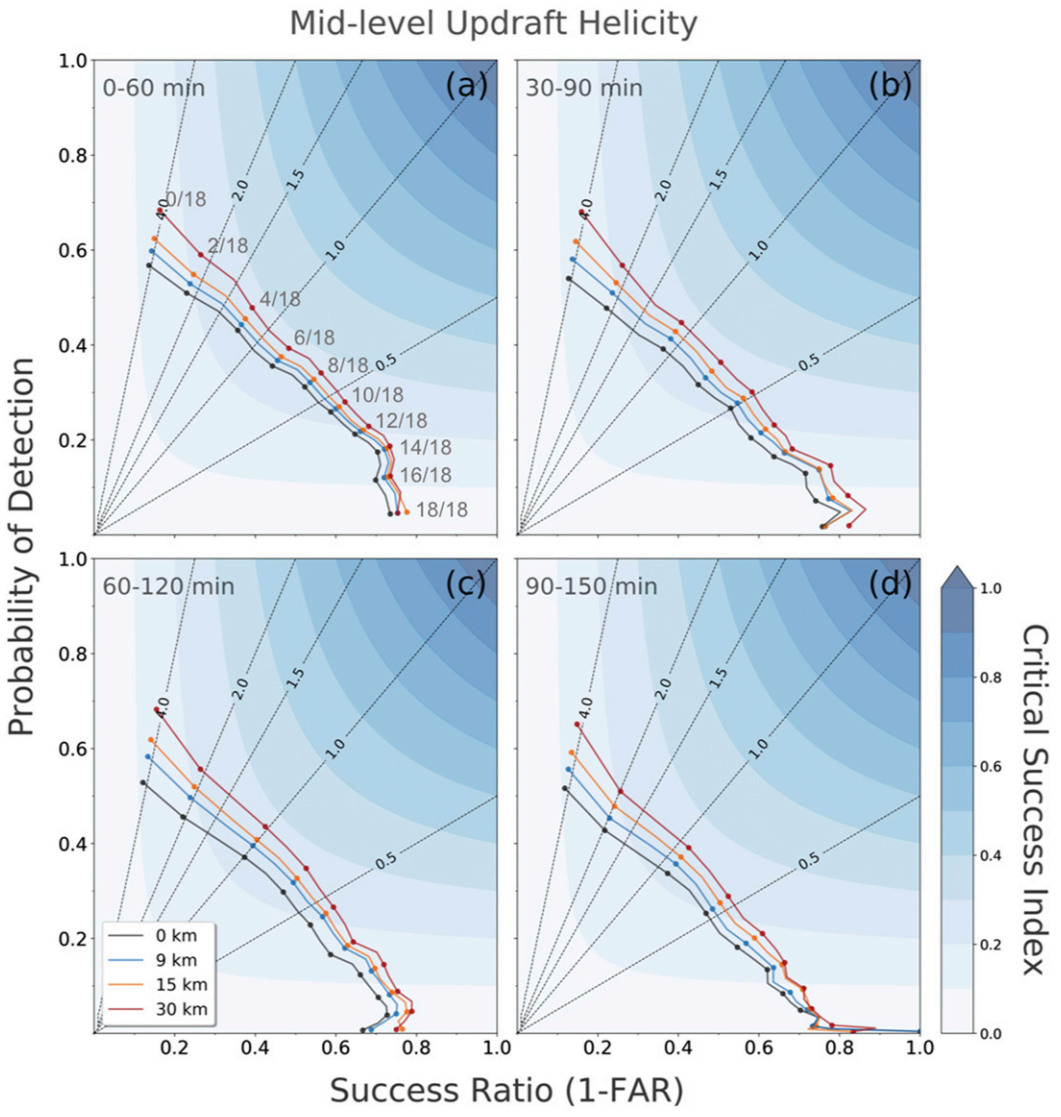

FIG. 5. As in Fig. 4, but for midlevel (2–5 km AGL) updraft helicity probability swath objects.

the performance curves at different matching neighborhoods decreases as the probability threshold increases. This is unsurprising as increasing the probability threshold progressively reduces the number of "yes" forecasts, resulting in lower number of possible hits and a low POD regardless of the matching neighborhood. Second, separation between the performance curves at different matching neighborhood does not change markedly with forecast lead time. As will be shown in section 4c, the centroid displacement between forecast and observed objects grows markedly with lead time. Therefore, the lack of lead time sensitivity to neighborhood in the contingency table score is likely attributable to the minimum spatial displacement in the total interest score used for object matching (i.e., objects may overlap but have a larger centroid displacement at longer lead times).

b. *Reliability diagrams*

Figures 6 and 7 show the reliability of the low- and midlevel UH probabilistic forecasts for the different matching neighborhoods and forecast lead times, respectively. Traditionally, for optimal reliability, the curves ought to lie along the diagonal from left to right with curves falling to bottom right (upper left) having an over (under) forecasting bias. Using the method of Bröcker and Smith (2007), we can compute consistency bars for the observed frequencies in each probability bin. Thus, we can assess how "reliable" the reliability estimates are. Additionally, the inset histograms are the

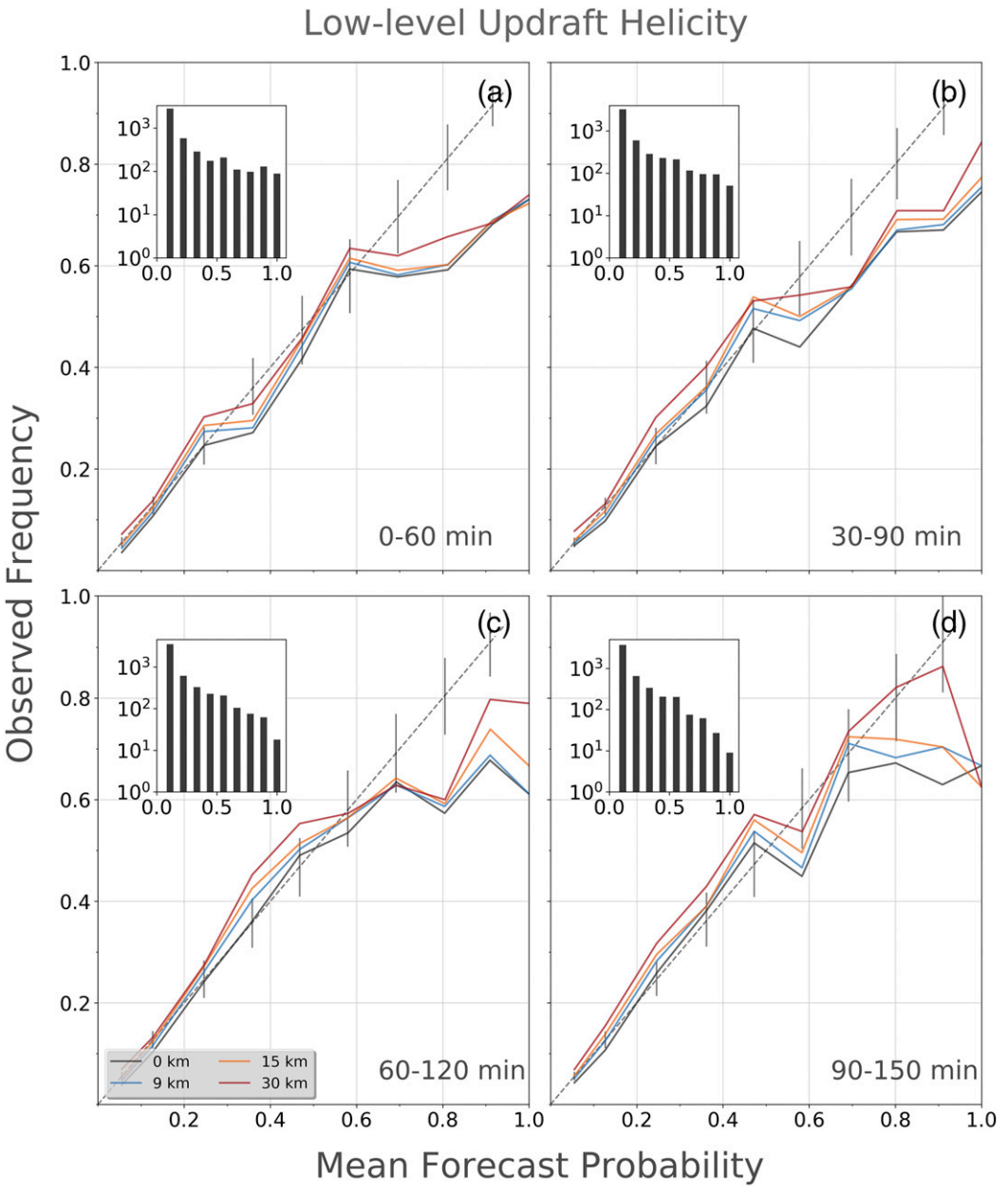

FIG. 6. Reliability diagrams for WoFS low-level mesocyclone probability swath objects using 0-, 9-, 15-, and 30-km matching neighborhoods (gray, blue, orange, and red, respectively) and valid at (a) 0–60, (b) 30–90, (c) 60–120, and (d) 90–150 min. The bin increment of forecast probabilities is 11.1% (1/9). The inset (gray bar graph) is the forecast histogram for the 0-km matching neighborhood. The dashed line represents perfect reliability. The vertical line along the diagonal was the error bars for the observed frequency in each bin based on the method in Bröcker and Smith (2007).

number of probability objects in each probability bin [in increments of 11.1% (1/9)] for the 0-km matching neighborhood.

Low-level UH forecast probability objects < 60% (Fig. 6a) have a near-perfect reliability in the 0–60-min period with increasing reliability at greater matching neighborhoods, but an overprediction of mesocyclone likelihood is present for probability values greater than 60%. Overprediction of forecast probabilities greater than 60% in the 0–60-min time period are attributable to underdispersion in WoFS forecasts (Fig. 1). In the inset histograms for both mid- and low-level UH, the forecast sharpness decays with increasing lead times as the number of probability objects at probabilities greater

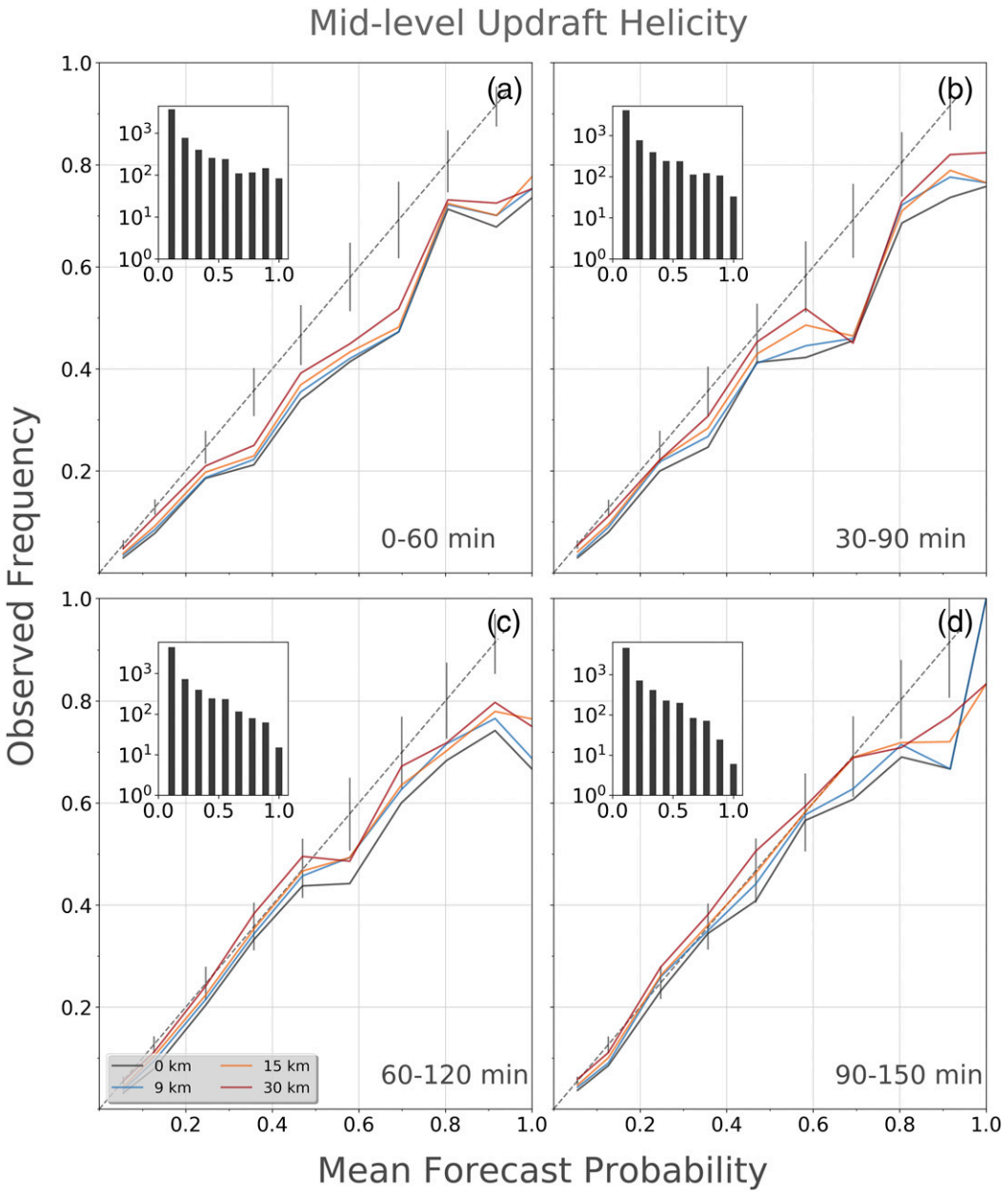

FIG. 7. As in Fig. 6, but for midlevel updraft helicity probability swath objects.

than 77.7% (7/9) greatly drops off. As explained above, the decay in probabilities with increasing lead time is attributable to the storm-scale predictability.

Sensitivity of the reliability for mid- and low-level UH probabilistic forecasts was generally lead-time and bin dependent. Increasing the matching neighborhood does increase the number of observed objects in a given bin, but does not necessarily improve the reliability. The greatest sensitivity to the matching neighborhood was evident for probabilities greater than >60%, especially as lead time increases. However, the probability swath values for low-level UH matched to observations using a 30-km matching neighborhood in the 60–120- and 90–150-min periods generally deviates from the observed frequency by less than 10% (Figs. 6c,d).

Mid-level UH forecast probabilities < 30% are also reliable in the 0–60-min period, but the forecast probabilities >40% have a larger overprediction bias than low-level UH (Fig. 7a). For example, in the 0–60-min period, probability swath objects near the 60% bin for midlevel UH only overlap with observed rotation 40% of time. However,

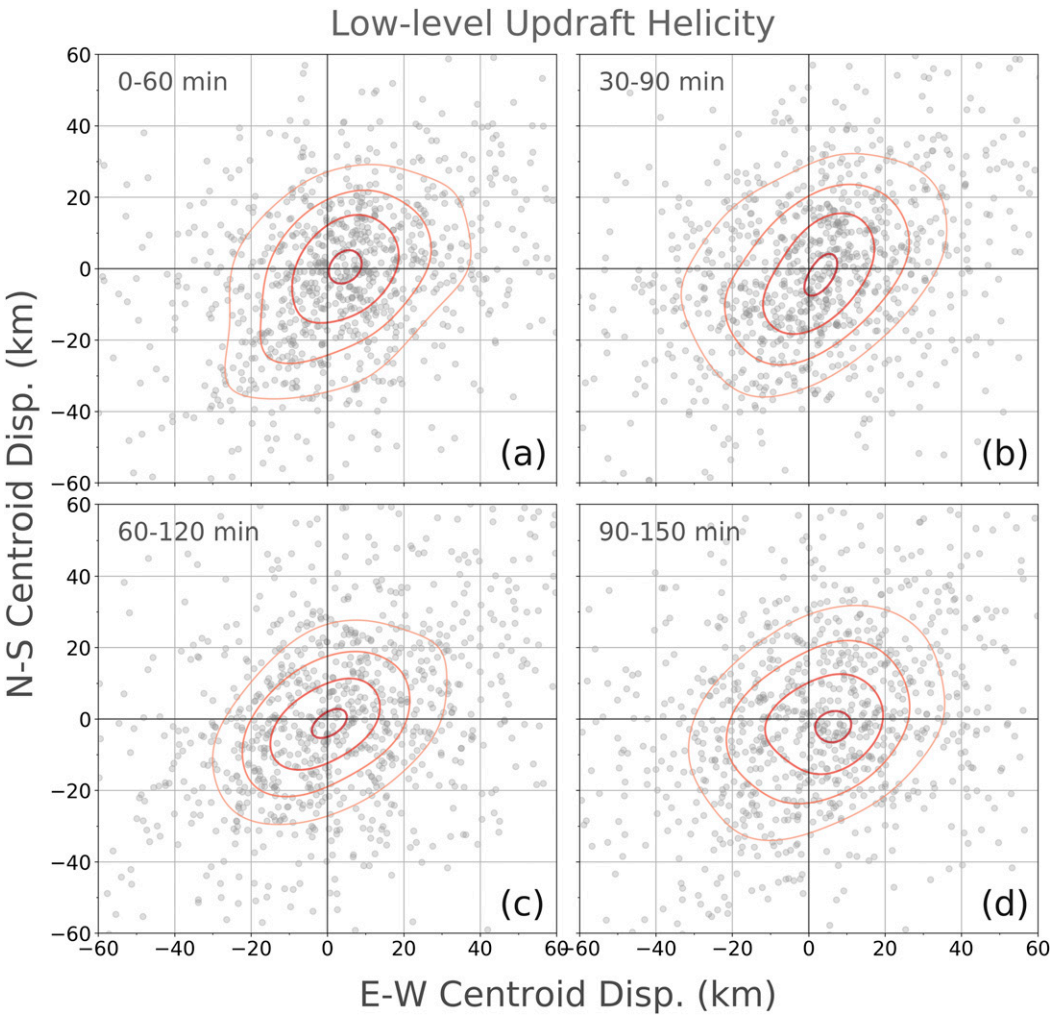

FIG. 8. Scatterplots of the east–west and north–south centroid displacements (km) of matched objects for hour-long low-level updraft helicity probability objects valid at (a) 0–60, (b) 30–90, (c) 60–120, and (d) 90–150 min. KDE contours of the 95, 97.5, 99, and 99.9 percentile values of each distribution are overlain to illustrate the evolution of centroid displacement with lead time.

at later lead times, midlevel UH forecast probabilities $>70\%$ are generally more reliable than the low-level UH forecast probabilities (cf. Figs. 6c,d and Figs. 7c,d).

c. *Centroid displacement*

Finally, centroid displacement between matched objects is examined to identify potential storm motion biases, which have been noted in subjective evaluations of WoFS probabilistic guidance (Yussouf et al. 2013b; Wheatley et al. 2015; Yussouf et al. 2015) as well as in objectively evaluated deterministic products (Skinner et al. 2016). Figures 8 and 9 show the centroid displacement between the matched observed and forecast objects with kernel density estimate (KDE) contours overlaid for low- and midlevel UH, respectively. The KDE technique implemented here applies a Gaussian kernel with a smoothing bandwidth determined from a general optimization algorithm to each point within the parameter space (Scott 1992). Kernels for each point are summed to provide a measure of the density of points and quantify biases in the displacement between the forecast and observed objects. As discussed in section 2c(2), since the enhanced watershed algorithm uses minimum area as a stopping criteria, probability swath objects in some cases will be shrunk, potentially changing their centroid and boundary displacement from observed objects. However, the impact of the enhanced watershed algorithm is primarily on the highest KDE contour when compared with probability objects identified using a single threshold method (not shown). The highest concentration of centroid displacements for both mid- and low-level UH (Figs. 8 and 9) are within

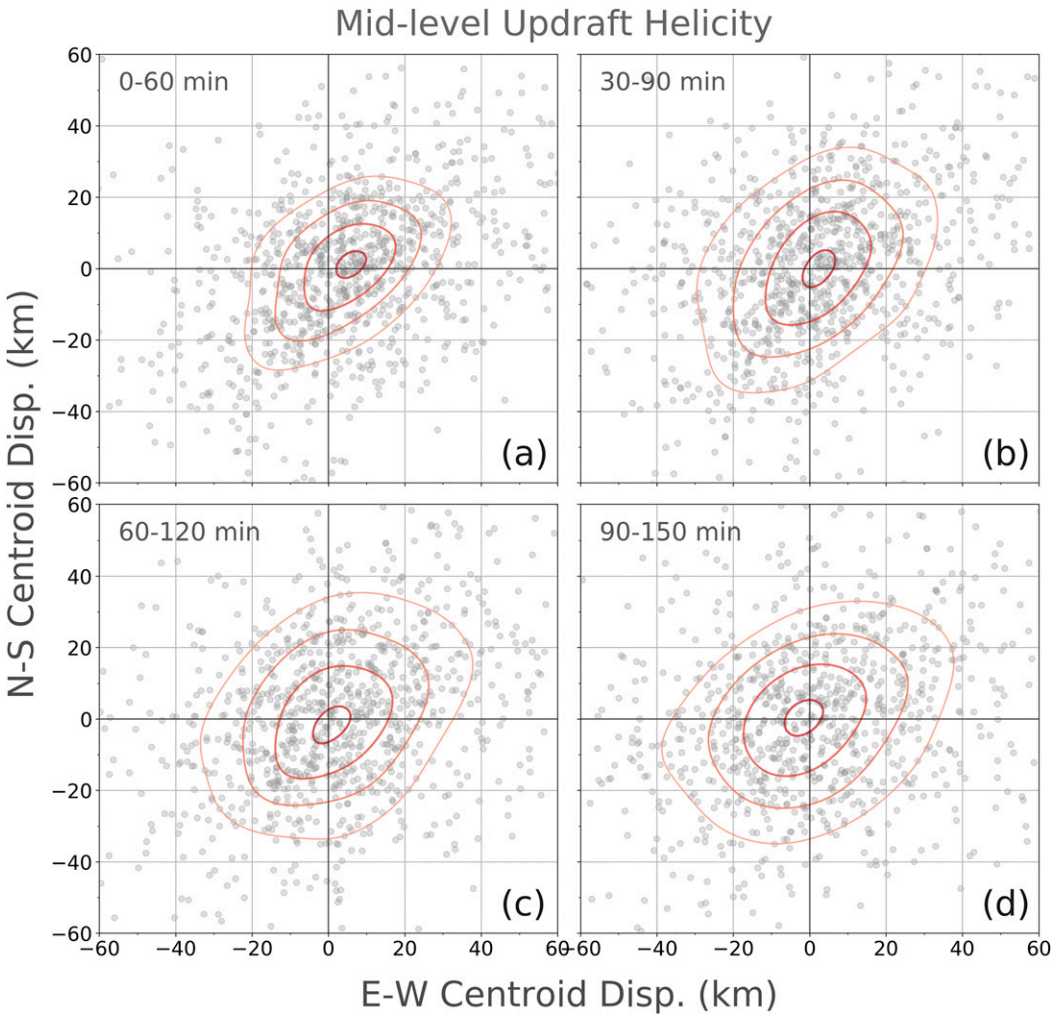

FIG. 9. As in Fig. 8, but for midlevel updraft helicity probability swath objects.

30 km, consistent with S18. Deviations larger than the matching neighborhoods tested in this study are a by-product of forecast probability objects in MCSs being much larger than observed rotation tracks. Often, the large probability objects associated with MCSs can have overlapping observed objects, but the centroids are displaced up to 60–90 km.

Centroid displacement for both low- and midlevel UH, based on the 99.9th percentile contour (innermost), has an inconsistent bias with forecast lead time with a slight eastward displacement (≈5 km) in the 0–60-min forecast period (Fig. 8a and Fig. 9a, respectively) shifting to minimal bias in the 60–120-min forecast period (Fig. 8d and Fig. 9d, respectively). In the 90–150-min forecast period, there remains minimal bias in the midlevel UH forecast (Fig. 9d), but the eastward bias returns for the low-level UH forecasts (Fig. 8d). We suspect the bias is an artifact of different track lengths between the UH and azimuthal wind shear tracks and in addition to the object identification and matching methods. Differences between UH and azimuthal shear track lengths can be related to variation in storm motion, but also to variation in storm intensity or duration, which would also result in centroid displacement between matched object pairs. Thus, attributing centroid displacement biases solely to differences in storm motion is difficult since biases in predicted intensity or longevity could produce similar centroid displacements. At all forecast lead times, the 95th and 97.5th percentile contours (two outermost) are similar between the low- and midlevel UH and roughly centered on the origin (Fig. 8 and Fig. 9). The area of the 95th percentile contours are similar for low- and midlevel UH except in the 90–150-min forecast period where low-level UH is a bit broader compared to the midlevel UH indicating a larger variance in centroid displacement between

matched objects (cf. Fig. 8d and Fig. 9d). In general, the outermost KDE contour (95th percentile) expands with increasing lead time, especially for low-level UH. As noted in section 4a, the centroid displacement between forecast and observed objects grows markedly, but there was a lack of lead time sensitivity to matching neighborhood in the contingency table score. Therefore, the minimum displacement is likely dampening the effects of the larger centroid displacements for the contingency table metrics (i.e., forecast and observed objects overlap, but have larger centroid displacement). Ultimately, the orientation of the contours are along the expected climatological storm track, suggesting the centroid displacements most likely represent differences in track length (and relative centroid position) that are not necessarily a result of a biased storm motion. Additionally, artifacts in MRMS rotation tracks are more common in the 0–2-km layer than for 2–5 km (owing to more ground clutter), so the bias may be influenced by limitations of the verification dataset as well as differences in the forecast.

5. Conclusions

A fundamental goal of the WoF project is to provide probabilistic guidance of severe weather hazards associated with individual thunderstorms. As a first effort, S18 established a baseline for WoFS deterministic thunderstorm and mesocyclone forecast products. The current study extends S18 by verifying the accuracy and reliability of WoFS hour-long probabilistic mesocyclone track forecasts. As grid-based verification showed, the WoFS probabilistic mesocyclone guidance on the native 3-km grid greatly overpredicts the likelihood of a mesocyclone impacting a particular point. This overprediction bias is an indication of considerable underdispersion in the WoFS. It is possible to improve the grid-based reliability by upscaling the forecasts and observations, but doing so obscures probabilities associated with individual storms. Despite the overprediction bias, WoFS probabilistic guidance on the native 3-km grid has been found to be useful in operational settings (Wilson et al. 2019). For example, Choate et al. (2018) found that paintball plots, which show the separate rotation tracks for all ensemble members on a single figure, were by far the most commonly used products in the SFE. These differences between grid-based verification metrics and forecaster usage have motivated the development of a novel, complementary verification method for evaluating short-term, storm-scale probabilistic guidance. This method uses an object-based framework where probability swaths associated with individual storms are treated as forecast objects and prescribed a single, representative probability. This approach tolerates spatial differences between forecasts and observations by defining a user-specified matching distance. Importantly, unlike in the grid-based framework, the forecast probabilities are not smoothed or upscaled, which preserves forecast likelihood of mesocyclones occurring within individual thunderstorms. Last, this verification method was designed with the human forecast decision model for WoFS probabilistic guidance in mind and is intended to match the expected forecaster usage of probability swaths (e.g., Wilson et al. 2019). The primary findings from applying the object-based verification technique to WoFS probabilistic mesocyclone guidance forecasts for 63 cases during 2017 and 2018 are as follows:

- The highest skill, in terms of CSI, of the WoFS mesocyclone probabilistic guidance was approximately associated with a probability threshold of 22.2% (4/18).
- The highest skill in the 0–60-min forecast period for low-level UH probabilistic forecasts had a POD, SR, and CSI of 0.47, 0.46, and 0.31, respectively. In the 90–150-min forecast period, the POD and CSI dropped to 0.39 and 0.27 while SR remained relatively unchanged.
- WoFS probabilistic low-level mesocyclone guidance is reliable for forecast probabilities $< 60\%$ at all forecast lead times using a 0-km matching neighborhood size, but an overprediction of mesocyclone likelihood is present at probability values $> 60\%$.
- Mid- and low-level probabilistic mesocyclone forecasts had similar contingency table metrics, reliability, and centroid displacement of matched pairs.
- The highest concentrations of centroid displacements (as indicated by KDE contours greater than the 99.9th percentile) in matched objects remained under 30 km (which is the approximate size of the NWS warning polygon) up to lead times of 90–150 min.

The object-based framework developed herein can be adapted to evaluate the performance and reliability of other severe weather hazards (e.g., hail, heavy rainfall, and severe winds) as well as changes in performance across different WoFS system configurations. In future work, it will be important to distinguish between the skill and reliability of probabilistic rotation forecasts in MCSs versus supercells. Our expectation is that mesocyclone forecasts will be more skillful for discrete supercells in a favorable environment than for rotation associated with MCSs (e.g., S18). It is also important to explore the impact of timing errors on the performance and reliability of WoFS mesocyclone guidance. In future work, 15- or 30-min probability swath objects could be used to explore the impact of timing errors.

Other techniques beyond simple object-based verification should be explored in future work. No single verification method adequately describes the different attributes of forecast performance and it is crucial

to develop complementary verification measures. For example, in a WoF framework, Skinner et al. (2016) explored multiple verification techniques of deterministic forecasts of low-level mesocyclones. Although the object-based methods were favored in that study, more work exploring different spatial verification methods is warranted. There are also promising, new techniques such as ensemble structure–amplitude–location (eSAL; Radanovics et al. 2018) or verification that leverages information theory (Lawson et al. 2018b) which could be suited for short-term, storm-scale probabilistic guidance.

There are limitations of the current method that will need to be improved upon in future iterations. First, we are using imperfect observation data coupled with an imperfect object identification method. Though extensive efforts were made to tune the object identification algorithms used in this study, the number of objects identified is sensitive to the scale of the phenomena to be identified. Observed rotation tracks and probability swaths, especially when considering different storm modes, can span a wide spectrum of spatial scales. Thus, it is difficult to find universal parameter settings for any object identification algorithm that covers all relevant scales in this problem. This limitation, however, could potentially be mitigated by improving observations of mesocyclones and accurately categorizing storm mode in simulated and observed reflectivity. We also plan to explore the use of the multiple hypothesis tracking (MHT; Lakshmanan et al. 2013) method in the Warning Decision Support System–Integrated Information (WDSS-II) software to improve identification of rotation tracks in the time-aggregated azimuthal wind shear data. It will also be possible to mitigate limitations in the object identification method at higher resolution where discriminating between intense and weak rotation is improved.

A goal of this study was not only to assess current WoFS probabilistic guidance, but also provide a framework to objectively assess the impacts of potential postprocessing techniques (e.g., machine learning calibration). Applications of artificial intelligence methods are becoming more common in the meteorological community with methods spanning from traditional machine learning algorithms to sophisticated deep learning methods (McGovern et al. 2017). Postprocessing techniques utilizing machine learning can potentially improve the skill and reliability of WoFS probability swath objects by correcting model biases. Developing verification techniques suited to short-term, storm-scale probabilistic guidance is a necessary first step to evaluating machine learning and other promising postprocessing methods.

Acknowledgments. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. We thank Adam Clark for informally reviewing an early version of the manuscript. Valuable local computing assistance was provided by Gerry Creager, Jesse Butler, Jeff Horn, Karen Copper, and Carrie Langston. We also acknowledge Harold Brooks for useful comments and discussion and the three anonymous reviewers whose comments made for an improved manuscript.

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.

Ahijevych, D., E. Gilleland, B. G. Brown, and E. E. Ebert, 2009: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts. *Wea. Forecasting*, **24**, 1485–1497, https://doi.org/10.1175/2009WAF2222298.1.

Albright, B., and S. Perfater, 2018: *2018 Flash Flood and Intense Rainfall Experiment.* Weather Prediction Center, 97 pp.

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney, 2015: Improving flash flood forecasts: The HMT-WPC flash flood and intense rainfall experiment. *Bull. Amer. Meteor. Soc.*, **96**, 1859–1866, https://doi.org/10.1175/BAMS-D-14-00201.1.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, https://doi.org/10.1175/WAF993.1.

Brooks, H. E., C. A. Doswell, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the United States. *Wea. Forecasting*, **18**, 626–640, https://doi.org/10.1175/1520-0434(2003)018<0626:CEOLDT>2.0.CO;2.

Brown, B., E. Gilleland, and E. E. Ebert, 2011: Forecasts of spatial fields. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed. I. T. Jolliffe and D. B. Stephenson, Eds., Vol. 1, Wiley, 95–117.

Cai, H., and R. E. Dumais, 2015: Object-based evaluation of a numerical weather prediction model's performance through forecast storm characteristic analysis. *Wea. Forecasting*, **30**, 1451–1468, https://doi.org/10.1175/WAF-D-15-0008.1.

Choate, J. J., A. Clark, B. T. Gallo, E. Grimes, P. L. Heinselman, P. S. Skinner, and K. A. Wilson, 2018: Examining the use of the NSSL experimental warn-on-forecast system for ensembles for the prediction of severe storms through short-term forecast outlooks during the 2018 spring forecasting experiment. *29th Conf. on Severe Local Storms*, Stowe, VT, Amer. Meteor. Soc., 3A.5, https://ams.confex.com/ams/29SLS/webprogram/Paper348346.html.

Cintineo, R. M., and D. J. Stensrud, 2013: On the predictability of supercell thunderstorm evolution. *J. Atmos. Sci.*, **70**, 1993–2011, https://doi.org/10.1175/JAS-D-12-0166.1.

Clark, A. J., R. G. Bullock, T. L. Jensen, M. Xue, and F. Kong, 2014: Application of object-based time-domain diagnostics for tracking precipitation systems in convection-allowing models.

Wea. Forecasting, **29**, 517–542, https://doi.org/10.1175/WAF-D-13-00098.1.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

——, ——, ——, and J. Halley-Gotway, 2009: The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1.

Dawson, L. C., G. S. Romine, R. J. Trapp, and M. E. Baldwin, 2017: Verifying supercellular rotation in a convection-permitting ensemble forecasting system with radar-derived rotation track data. *Wea. Forecasting*, **32**, 781–795, https://doi.org/10.1175/WAF-D-16-0121.1.

Doswell, C. A., 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.

Dowell, D., and Coauthors, 2016: Development of a high-resolution rapid refresh ensemble (HRRRE) for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 8B.2, https://ams.confex.com/ams/28SLS/webprogram/Paper301555.html.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, https://doi.org/10.1002/met.25.

——, M. Turk, S. J. Kusselson, J. Yang, M. Seybold, P. R. Keehn, and R. J. Kuligowski, 2011: Ensemble tropical rainfall potential (eTRaP) forecasts. *Wea. Forecasting*, **26**, 213–224, https://doi.org/10.1175/2010WAF2222443.1.

Flora, M. L., C. K. Potvin, and L. J. Wicker, 2018: Practical predictability of supercells: Exploring ensemble forecast sensitivity to initial condition spread. *Mon. Wea. Rev.*, **146**, 2361–2379, https://doi.org/10.1175/MWR-D-17-0374.1.

Gagne, D. J., II, A. McGovern, N. Snook, R. Sobash, J. Labriola, J. K. Williams, S. E. Haupt, and M. Xue, 2016: Hagelslag: Scalable object-based severe weather analysis and forecasting. *Sixth Symp. on Advances in Modeling and Analysis Using Python*, New Orleans, LA, Amer. Meteor. Soc., 447, https://ams.confex.com/ams/96Annual/webprogram/Paper280723.html.

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.

——, A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended probabilistic tornado forecasts: Combining climatological frequencies with NSSL WRF ensemble forecasts. *Wea. Forecasting*, **33**, 443–460, https://doi.org/10.1175/WAF-D-17-0132.1.

——, ——, ——, ——, ——, and ——, 2019: Incorporating UH occurrence time to ensemble-derived tornado probabilities. *Wea. Forecasting*, **34**, 151–164, https://doi.org/10.1175/WAF-D-18-0108.1.

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.

——, D. A. Ahijevych, B. G. Brown, and E. E. Ebert, 2010: Verifying forecasts spatially. *Bull. Amer. Meteor. Soc.*, **91**, 1365–1376, https://doi.org/10.1175/2010BAMS2819.1.

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/.

Jain, A., 1989: *Fundamentals of Digital Image Processing*. Prentice Hall, 569 pp.

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.

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

——, P. Skinner, K. Knopfmeier, E. Mansell, P. Minnis, R. Palikonda, and W. Smith, 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.

Labriola, J., N. Snook, Y. Jung, B. Putnam, and M. Xue, 2017: Ensemble hail prediction for the storms of 10 May 2010 in south-central Oklahoma using single- and double-moment microphysical schemes. *Mon. Wea. Rev.*, **145**, 4911–4936, https://doi.org/10.1175/MWR-D-17-0039.1.

——, ——, ——, and M. Xue, 2019: Explicit ensemble prediction of hail in 19 May 2013 Oklahoma City thunderstorms and analysis of hail growth processes with several multimoment microphysics schemes. *Mon. Wea. Rev.*, **147**, 1193–1213, https://doi.org/10.1175/MWR-D-18-0266.1.

Lakshmanan, V., K. Hondl, and R. Rabin, 2009: An efficient, general-purpose technique for identifying storm cells in geospatial images. *J. Atmos. Oceanic Technol.*, **26**, 523–537, https://doi.org/10.1175/2008JTECHA1153.1.

——, M. Miller, and T. Smith, 2013: Quality control of accumulated fields by applying spatial and temporal constraints. *J. Atmos. Oceanic Technol.*, **30**, 745–758, https://doi.org/10.1175/JTECH-D-12-00128.1.

Lawson, J. R., J. S. Kain, N. Yussouf, D. C. Dowell, D. M. Wheatley, K. H. Knopfmeier, and T. A. Jones, 2018a: Advancing from convection-allowing NWP to warn-on-forecast: Evidence of progress. *Wea. Forecasting*, **33**, 599–607, https://doi.org/10.1175/WAF-D-17-0145.1.

——, C. Potvin, and M. Flora, 2018b: Information, predictability, and verification at the thunderstorm scale. *29th Conf. on Severe Local Storms*, Stowe, VT, Amer. Meteor. Soc., 124, https://ams.confex.com/ams/29SLS/webprogram/Paper348764.html.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

Mahalik, M. C., B. R. Smith, K. L. Elmore, D. M. Kingfield, K. L. Ortega, and T. M. Smith, 2019: Estimates of gradients in radar moments using a linear least squares derivative technique.

Wea. Forecasting, **34**, 415–434, https://doi.org/10.1175/WAF-D-18-0095.1.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

Miller, M. L., V. Lakshmanan, and T. M. Smith, 2013: An automated method for depicting mesocyclone paths and intensities. *Wea. Forecasting*, **28**, 570–585, https://doi.org/10.1175/WAF-D-12-00065.1.

Murphy, A. H., 1991: Probabilities, odds, and forecasts of rare events. *Wea. Forecasting*, **6**, 302–307, https://doi.org/10.1175/1520-0434(1991)006<0302:POAFOR>2.0.CO;2.

Potvin, C. K., C. Broyles, P. S. Skinner, H. E. Brooks, and E. Rasmussen, 2019: A Bayesian hierarchical modeling framework for correcting reporting bias in the U.S. tornado database. *Wea. Forecasting*, **34**, 15–30, https://doi.org/10.1175/WAF-D-18-0137.1.

Radanovics, S., J.-P. Vidal, and E. Sauquet, 2018: Spatial verification of ensemble precipitation: An ensemble version of SAL. *Wea. Forecasting*, **33**, 1001–1020, https://doi.org/10.1175/WAF-D-17-0162.1.

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

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.

Scott, D. W., 1992: *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley and Sons, 360 pp.

Skamarock, W., 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.

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

Smith, T. M., and K. L. Elmore, 2004: The use of radial velocity derivative to diagnose rotation and divergence. *11th Conf. on Aviation, Range, and Aerospace*, Hyannis, MA, Amer. Meteor. Soc., P5.6, https://ams.confex.com/ams/pdfpapers/81827.pdf.

——, 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.

Snook, N., M. Xue, and Y. Jung, 2012: Ensemble probabilistic forecasts of a tornadic mesoscale convective system from ensemble Kalman filter analyses using WSR-88D and CASA radar data. *Mon. Wea. Rev.*, **140**, 2126–2146, https://doi.org/10.1175/MWR-D-11-00117.1.

——, Y. Jung, J. Brotzge, B. Putnam, and M. Xue, 2016: Prediction and ensemble forecast verification of hail in the supercell storms of 20 May 2013. *Wea. Forecasting*, **31**, 811–825, https://doi.org/10.1175/WAF-D-15-0152.1.

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

——, G. S. Romine, C. S. Schwartz, D. J. Gagne, and M. L. Weisman, 2016a: Explicit forecasts of low-level rotation from convection-allowing models for next-day tornado prediction. *Wea. Forecasting*, **31**, 1591–1614, https://doi.org/10.1175/WAF-D-16-0073.1.

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016b: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

Stensrud, D. J., and Coauthors, 2009: Convective-scale Warn-on-Forecast system. *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.

Stratman, D. R., and K. A. Brewster, 2017: Sensitivities of 1-km forecasts of 24 May 2011 tornadic supercells to microphysics parameterizations. *Mon. Wea. Rev.*, **145**, 2697–2721, https://doi.org/10.1175/MWR-D-16-0282.1.

Trapp, R. J., G. J. Stumpf, and K. L. Manross, 2005: A reassessment of the percentage of tornadic mesocyclones. *Wea. Forecasting*, **20**, 680–687, https://doi.org/10.1175/WAF864.1.

——, 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.

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.

Wilson, K. A., and Coauthors, 2019: Exploring applications of storm-scale probabilistic warn-on-forecast guidance in weather forecasting. *International Conference on Human–Computer Interaction (HCII 2019): Virtual, Augmented and Mixed-Reality, Applications and Case Studies*, J. Chen and G. Fragomeni, Eds., 577–572, https://doi.org/10.1007/978-3-030-21565-1_39.

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.

Yussouf, N., J. Gao, D. J. Stensrud, and G. Ge, 2013a: The impact of mesoscale environmental uncertainty on the prediction of a tornadic supercell storm using ensemble data assimilation approach. *Adv. Meteor.*, **2013**, 1–15, https://doi.org/10.1155/2013/731647.

——, E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013b: The ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma City tornadic supercell storm using single- and double-moment microphysics schemes. *Mon. Wea. Rev.*, **141**, 3388–3412, https://doi.org/10.1175/MWR-D-12-00237.1

——, D. C. Dowell, L. J. Wicker, K. H. Knopfmeier, and D. M. Wheatley, 2015: Storm-scale data assimilation and ensemble forecasts for the 27 April 2011 severe weather outbreak in Alabama. *Mon. Wea. Rev.*, **143**, 3044–3066, https://doi.org/10.1175/MWR-D-14-00268.1.

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