Application of Two Spatial Verification Methods to Ensemble Forecasts of Low-Level Rotation

PATRICK S. SKINNER AND LOUIS J. WICKER

NOAA/National Severe Storms Laboratory, Norman, Oklahoma

DUSTAN M. WHEATLEY AND KENT H. KNOPFMEIER

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 24 September 2015, in final form 5 February 2016)

ABSTRACT

Two spatial verification methods are applied to ensemble forecasts of low-level rotation in supercells: a four-dimensional, object-based matching algorithm and the displacement and amplitude score (DAS) based on optical flow. Ensemble forecasts of low-level rotation produced using the National Severe Storms Laboratory (NSSL) Experimental Warn-on-Forecast System are verified against WSR-88D single-Doppler azimuthal wind shear values interpolated to the model grid. Verification techniques are demonstrated using four 60-min forecasts issued at 15-min intervals in the hour preceding development of the 20 May 2013 Moore, Oklahoma, tornado and compared to results from two additional forecasts of tornadic supercells occurring during the springs of 2013 and 2014.

The object-based verification technique and displacement component of DAS are found to reproduce subjectively determined forecast characteristics in successive forecasts for the 20 May 2013 event, as well as to discriminate in subjective forecast quality between different events. Ensemble-mean, object-based measures quantify spatial and temporal displacement, as well as storm motion biases in predicted low-level rotation in a manner consistent with subjective interpretation. Neither method produces useful measures of the intensity of low-level rotation, owing to deficiencies in the verification dataset and forecast resolution.

1. Introduction

An important component of the National Oceanic and Atmospheric Administration's (NOAA) Warn-on-Forecast project is the production of probabilistic, short-term guidance for tornado potential using an ensemble of convection-allowing numerical weather prediction models (Stensrud et al. 2009, 2013). Working toward this goal, many recent experiments have produced promising short-term (0–1 h) forecasts for low-level rotation in case studies of tornadic supercells (Dawson et al. 2012; Yussouf et al. 2013; Sobash and Wicker 2014; Jones et al. 2016; Wheatley et al. 2015; Yussouf et al. 2015) and quasi-linear convective systems (QLCSs; Snook et al. 2012; Putnam et al. 2014; Snook et al. 2015) initialized through the variational or ensemble Kalman filter (EnKF; e.g., Evensen 1994) based assimilation of Doppler radar data.[1] These forecasts have typically been visualized as a swath of probabilities that low-level vertical vorticity will exceed a certain threshold, which is then compared to observed tornado damage tracks or single-Doppler rotation tracks (Miller et al. 2013) to evaluate the forecast quality. These studies have regularly produced maximum probabilities of low-level rotation coincident with observed tornado tracks, qualitatively indicating a skillful forecast. However, automated, quantitative measures of ensemble forecast skill are needed for intercomparing large numbers of forecasts, as would be produced by an operational Warn-on-Forecast system. Development of consistent, objective measures of

Corresponding author address: Patrick Skinner, NOAA/National Severe Storms Laboratory, 120 David L. Boren Blvd., Norman, OK 73072.
E-mail: patrick.skinner@noaa.gov

DOI: 10.1175/WAF-D-15-0129.1

[1] Readers are referred to Meng and Zhang (2011) for a recent review of data assimilation in limited-area models.

forecast quality will assist in determining best practices for an operational Warn-on-Forecast system by quantifying differences in forecasts run with variable model parameters and allowing system performance in different meso- and synoptic-scale environments to be compared.

As an operational Warn-on-Forecast system will provide guidance on hazard potential in individual thunderstorms, feature-based verification of the forecasts will be required. Additionally, diagnostic information on specific storm features, such as mesocyclones, will be important for assessing the contributing error sources in individual forecasts. Spatial verification techniques [see Gilleland et al. (2009) and (2010) for recent reviews] are ideally suited for this problem, as they can avoid the double-penalty problem small position errors can induce in point-to-point verification (Wilks 2006); quantify differences between features in forecast and verification fields; and provide extensive diagnostic information on specific characteristics of features in the forecast and verification fields (e.g., Clark et al. 2014; Wolff et al. 2014; Cai and Dumais 2015; Pinto et al. 2015).

Spatial verification techniques have been widely employed to assess the skill of quantitative precipitation forecasts from convection-allowing numerical models (e.g., Davis et al. 2006b; Kain et al. 2008; Davis et al. 2009; Ebert 2009; Keil and Craig 2009; Schwartz et al. 2009; Marzban et al. 2009; Clark et al. 2010, 2011; Johnson et al. 2011; Johnson and Wang 2012; Johnson et al. 2013). Though many variations of spatial verification have been developed, those based on object identification and matching (Davis et al. 2006a,b; Ebert and Gallus 2009; Gallus 2010; Clark et al. 2012; Burghardt et al. 2014; Clark et al. 2014; Wolff et al. 2014; Pinto et al. 2015; Cai and Dumais 2015) or field deformation methods such as optical flow (Keil and Craig 2007, 2009; Marzban and Sandgathe 2010) are particularly appealing for verification of low-level rotation forecasts. This appeal is attributable to the ability of object-based and field deformation techniques to provide feature-based verification as well as to quantify components of forecast error. These two techniques have been recently used for applications similar to the objectives of the Warn-on-Forecast project. Clark et al. (2012, 2013) employed object-based techniques to compare the length of simulated updraft helicity tracks produced by an ensemble of convection-allowing models to observed tornado pathlengths and the optical-flow-based displacement and amplitude score (DAS; Keil and Craig 2009) has been used to quantify spatial errors for individual storms in a storm-scale ensemble forecast (Lange and Craig 2014).

This study applies an object-based verification technique based on the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2006a,b) and the displacement and amplitude score of Keil and Craig (2009) to ensemble forecasts of low-level rotation produced by the National Severe Storms Laboratory (NSSL) Experimental Warn-on-Forecast System (NEWS-e; Wheatley et al. 2015; Jones et al. 2016). The verification techniques are demonstrated using a series of four 1-h ensemble forecasts initialized at 15-min intervals by the NEWS-e for multiple tornadic supercells on 20 May 2013, including the parent storm of an [enhanced Fujita (EF) scale] EF5 tornado that struck Moore, Oklahoma (Atkins et al. 2014; Burgess et al. 2014). Forecasts are verified against single-Doppler azimuthal wind shear values calculated from Frederick (KFDR) and Oklahoma City (KTLX), Oklahoma, WSR-88D data interpolated to the model grid. Forecasts are first examined qualitatively to identify strengths and weaknesses (i.e., biased storm motion or overprediction of low-level rotation) expected to be captured by the objective verification scores. The two objective techniques are then applied to each forecast and evaluated on their ability to accurately and consistently match the subjective interpretation. The 20 May forecasts are then compared to other NEWS-e cases from the springs of 2013 and 2014 to assess the quality of the verification methods with varying storm mode and evolution.

Descriptions of the NEWS-e system and forecasts, development of an observational proxy for verification and of the two verification techniques are provided in section 2. An overview of the 20 May 2013 event and qualitative verification are presented in section 3, with objective verification and comparison to additional NEWS-e cases following in section 4. Conclusions and recommendations for future research are provided in section 5.

2. Methodology

a. *NEWS-e description*

The NEWS-e is a multiscale data assimilation system that uses a 36-member ensemble forecast from ARW, version 3.4.1 (Skamarock et al. 2008) as an initial state and assimilates conventional and Doppler radar observations using an EnKF technique provided by the Data Assimilation Research Testbed (DART; Anderson and Collins 2007; Anderson et al. 2009). For each NEWS-e case, a mesoscale parent grid with 15-km horizontal grid spacing is initialized at 0000 UTC the day of the target case from downscaled output from the Global Ensemble Forecast System. Conventional observations of

pressure, temperature, dewpoint temperature, and horizontal wind components provided by the NOAA Meteorological Assimilation Data Ingest System (Miller et al. 2007) are assimilated into the ensemble hourly to provide a representative mesoscale background for the inner domain. Following convection initiation, a finescale, one-way nested inner domain with 3-km horizontal grid spacing is initialized from the mesoscale analysis. Radar reflectivity and radial velocity data from three WSR-88D radars within the inner domain are objectively analyzed to a 6-km grid and assimilated using DART at 15-min intervals. When available, surface observations from the Oklahoma Mesonet are additionally assimilated into the inner domain. Storm-scale, 60-min forecasts are initialized following each cycle until genesis of the strongest tornado observed, according to *Storm Data*, for each case. Readers are referred to Wheatley et al. (2015) for a complete description of the NEWS-e methodology.

b. Observational dataset description

As observed vertical vorticity is not available without specialized observations, an imperfect proxy for low-level rotation must be developed to serve as a verification dataset. A natural choice to serve as this dataset is single-Doppler azimuthal wind shear, which is equal to half the vertical vorticity for solid-body rotation and is regularly used to create mesocyclone rotation tracks (e.g., Miller et al. 2013). For each case considered herein, azimuthal wind shear is calculated for radial velocity data from the nearest one or two WSR-88D radars to target storms following the process outlined by Newman et al. (2013). Radar data are first dealiased and nonmeteorological echoes are removed using the algorithm developed by Lakshmanan et al. (2014). Range-corrected azimuthal wind shear is then calculated using the linear least squares derivative (LLSD) method (Smith and Elmore 2004; Newman et al. 2013) for each available radar sweep. Finally, each sweep of azimuthal wind shear data is interpolated to the storm-scale NEWS-e domain using a Cressman scheme with a 3- (2-) km horizontal (vertical) radius of influence and multiplied by 2 to be equivalent to values of vertical vorticity using the assumption of solid-body rotation.

The objectively analyzed sweeps of azimuthal wind shear are aggregated to create a verification low-level rotation field with identical space and time dimensions as the forecast low-level rotation field by merging values from each sweep within a 5-min window centered on the forecast time. Advection correction is not applied to sweeps with a temporal offset; however, maximum errors resulting from storm motion are expected to be approximately equal to a single grid point (3 km) for cases considered herein.[2] Additionally, at least one azimuthal wind shear observation below 1500 m above ground level (AGL) is required at each point in the verification field in order to ensure that the low-level (defined herein as the lowest 2 km AGL) mesocyclone is being sampled. It is noted that azimuthal wind shear observations from at least one radar were available below 1500 m for the entirety of each rotation track considered herein.

c. Postprocessing of forecast and verification fields

A drawback of object- and optical-flow-based verification methods is that both feature a large number of tunable parameters. To mitigate the impact of subjectively determined parameters in verification, the forecast and observation fields are postprocessed to isolate the features of interest, as recommended by Wolff et al. (2014). As this study is concerned with forecasts of tornado potential, the presence of a low-level mesocyclone is used as an imperfect (Trapp et al. 2005), but best available, observational proxy for tornado occurrence. Therefore, NEWS-e vertical vorticity and WSR-88D azimuthal wind shear are processed in an attempt to isolate low-level mesocyclones.[3]

Low-level rotation is initially calculated as the average vertical vorticity (twice the azimuthal wind shear) in the 500–2000-m layer for each forecast (verification) field time step (Figs. 1a,e). Discrepancies between the initial layer-mean values in the verification and forecast fields are apparent, with the vertical vorticity field in NEWS-e members containing broader and a larger number of vorticity maxima than observations.[4] Variation in the spatial extent of rotational maxima is mitigated, as in Wheatley et al. (2015), by application of a 3×3 gridpoint maximum value filter (Figs. 1b,f). Maxima are then further broadened and

[2] For example, a $20\,\mathrm{m\,s^{-1}}$ storm motion and the maximum allowable temporal offset of 150 s would result in a 3-km spatial displacement at the analysis time.

[3] Low-level mesocyclones typically occur on a much smaller scale than the effective NEWS-e resolution at 3-km grid spacing [e.g., approximately 20 km or $7\Delta x$ Skamarock (2004)]. However, idealized simulations by Potvin and Flora (2015) have demonstrated that supercell processes can be adequately resolved at this grid spacing, leading to simulated low-level mesocyclones that behave similarly to those with finer grid spacing (Potvin and Flora 2015).

[4] One of the noisier ensemble members was selected for Fig. 1 in order to emphasize differences between the observations; however, each member exhibits a broader vertical vorticity field than the observations.

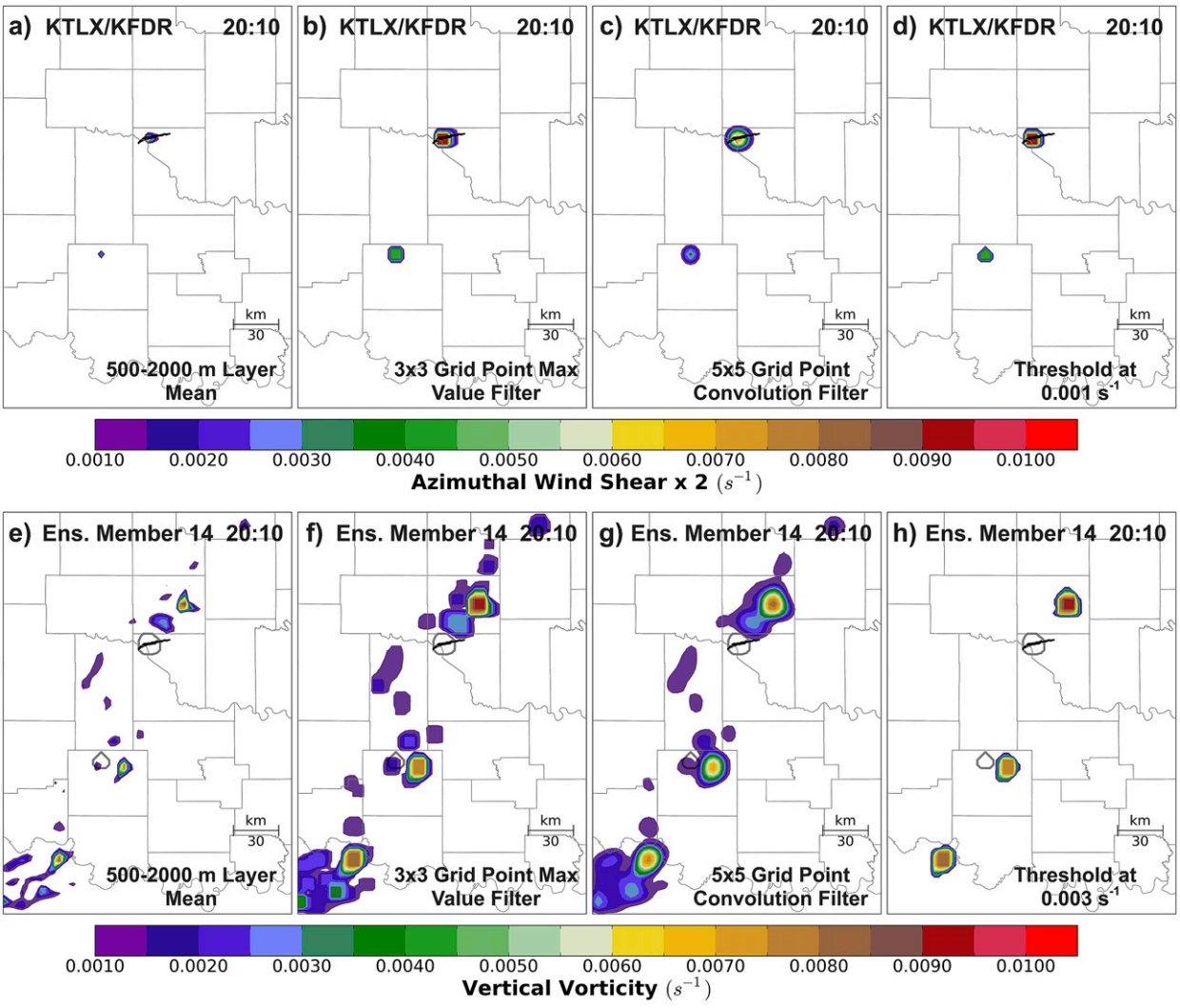

FIG. 1. Illustration of the postprocessing applied to the 20 May 2013 event for (top) twice the merged KTLX and KFDR azimuthal wind shear values and (bottom) the vertical vorticity output from a single NEWS-e member overlaid on an Oklahoma county map (thin gray lines). The (a),(e) 500–2000-m layer mean values are initially calculated for both fields. Afterward a (b),(f) 3×3 gridpoint maximum value and (c),(g) a 5×5 grid point convolution filter are applied to the mean values. The smoothed fields are then thresholded at (d) $\text{Ob}_{\text{thresh}} = 0.001\,\text{s}^{-1}$ or (h) $\text{Fcst}_{\text{thresh}} = 0.003\,\text{s}^{-1}$, and the values following the application of the maximum value filter are restored to grid points exceeding the threshold. All fields are valid at 2010 UTC, and the NEWS-e forecast was initialized at 1945 UTC. The $0.001\,\text{s}^{-1}$ contour of twice the azimuthal wind shear in (d) is plotted in dark gray in (e)–(h), and the damage path of the Moore tornado is marked in black.

smoothed by applying a 5×5 gridpoint Gaussian convolution kernel (Figs. 1c,g). The smoothed rotation fields are then thresholded to isolate the strongest regions of rotation, and values following the application of the gridpoint maximum filter are restored to grid points exceeding the threshold (Figs. 1d,h), as was done in Davis et al. (2006a; their Fig. 2). Different thresholds are applied to twice the azimuthal wind shear ($\text{Ob}_{\text{thresh}} = 0.001\,\text{s}^{-1}$) and vertical vorticity ($\text{Fcst}_{\text{thresh}} = 0.003\,\text{s}^{-1}$) fields. The different thresholds are utilized to retain as much of the observed rotation field as possible without creating false positives and to remove as much of the weak, broad rotation as possible from the forecast field. The convolution and thresholding process is adapted from the MODE software (Davis et al. 2006a), and filter sizes and threshold values have been determined through trial and error.

d. *Object-based verification*

The object-based verification method used herein is based on the MODE software (Davis et al. 2006a,b) and has been developed using the scikit-image library for the Python programming language (Van der Walt

et al. 2014). Each contiguous area of low-level rotation at a specified time within the forecast and verification datasets is considered a single rotation object. The primary objective of a successful forecast is the prediction of a low-level rotation object in spatial and temporal proximity to an observed object. Therefore, the best match between forecast and observed objects for each ensemble member is found using a total interest score (Davis et al. 2009) weighted heavily on closeness in time and space. The total interest is calculated for every pair of forecast and observed rotation objects, including those offset in time, and is calculated as

$$I_{ij} = w_{dt}\left\{\left[\frac{(d_{\max} - d_{ij})}{d_{\max}}\right]\left[\frac{(t_{\max} - t_{ij})}{t_{\max}}\right]\right\} + w_a\left(\frac{a_{ij_{\min}}}{a_{ij_{\max}}}\right), \tag{1}$$

where I_{ij} represents the total interest between a given forecast and observed object, d_{ij} (t_{ij}) is the centroid distance (time offset) between the objects, $d_{\max}$ ($t_{\max}$) is the maximum allowable centroid distance (time offset) for matching objects, and a_{ij} is the area of an object, sorted such that the object with the larger area $a_{ij_{\max}}$ is the denominator. The spatiotemporal component of the total interest calculation is weighted far more heavily (w_{dt} = 0.9) than the areal component (w_a = 0.1), with the areal component intended to serve as a ''tiebreaker'' for objects with similar space and time offsets. Maximum allowable offsets in time $t_{\max}$ and space $d_{\max}$ are chosen according to the expected limits of usefulness for the forecast and set to 25 min and 30 km, respectively.

Matches between forecast and observed rotation objects are determined by the maximum available total interest value, with the condition that it exceeds 0.2 (verification score sensitivity to the total interest threshold is examined in the appendix). Each forecast object may only be matched to one rotation object. However, as a result of the ensemble and temporal aspects to object matching, a single observed object may be matched to many forecast objects within different ensemble members and at different times. For verification purposes, matched objects can be considered analogous to forecast ''hits,'' with unmatched observed objects ''misses'' and unmatched forecast objects ''false alarms.'' Classification in this manner allows quantities similar to the probability of detection (POD) and false alarm ratio (FAR) to be calculated using the standard contingency table formulation (Wilks 2006). Additionally, the object-based threat score (OTS) can be calculated according to the formula defined by Johnson et al. (2011) and Johnson and Wang (2013):

$$\mathrm{OTS} = \frac{1}{A_f + A_o}\left[\sum_{p=1}^{P} I^p(a_f^p + a_o^p)\right], \tag{2}$$

where A_f and A_o represent the total area of forecast and observed objects, and a_f and a_o the area of individual forecast and observed objects, respectively. The combined area of a pair of matched objects (denoted by superscript p) is weighted by total interest [I^p; which is equivalent to I_{ij} in Eq. (1)] and summed over each of P matched object pairs. In other words, the OTS represents the ratio of the matched object area, weighted by total interest, to the total object area within a given verification domain. A ''binary'' OTS (Johnson et al. 2011) can be calculated by setting the total interest to 1 for each matched object pair, so that the resulting score is the ratio of the matched object area to the total object area. A perfect forecast, where observed and forecast objects are identical in position, timing, and area, will result in weighted and binary OTS scores of 1.

In addition to the interest and OTS scores, individual properties of matched objects can be compared to quantify the performance of various aspects of the forecast. The ratio of maximum intensity of matched objects, the total centroid or time displacement, and the zonal and meridional components of matched-object centroid displacement are considered herein; however, many additional quantities may be considered according to case-specific verification priorities (e.g., Wolff et al. 2014; Cai and Dumais 2015; Pinto et al. 2015).

e. *Displacement and amplitude score*

The displacement and amplitude score developed by Keil and Craig (2007, 2009) is also used to verify the forecast and observed rotation fields. The DAS score is based on a pyramidal image matching algorithm (Keil and Craig 2007), which will calculate displacement vectors that morph a forecast field to a verification field, or vice versa, while minimizing amplitude-based error (Keil and Craig 2009). Displacement error is calculated as the magnitude of the displacement vectors and amplitude error is defined as the root-mean-square error between the verification and morphed fields, with both errors only calculated where the verification field is nonzero. Morphing is performed both from forecast to observation and observation to forecast fields, with total displacement (DIS) and amplitude (AMP) scores calculated as averages of the two morphs weighted by the number of nonzero grid points in the respective verification fields. Normalized DIS and AMP scores are calculated by dividing the total scores by a maximum

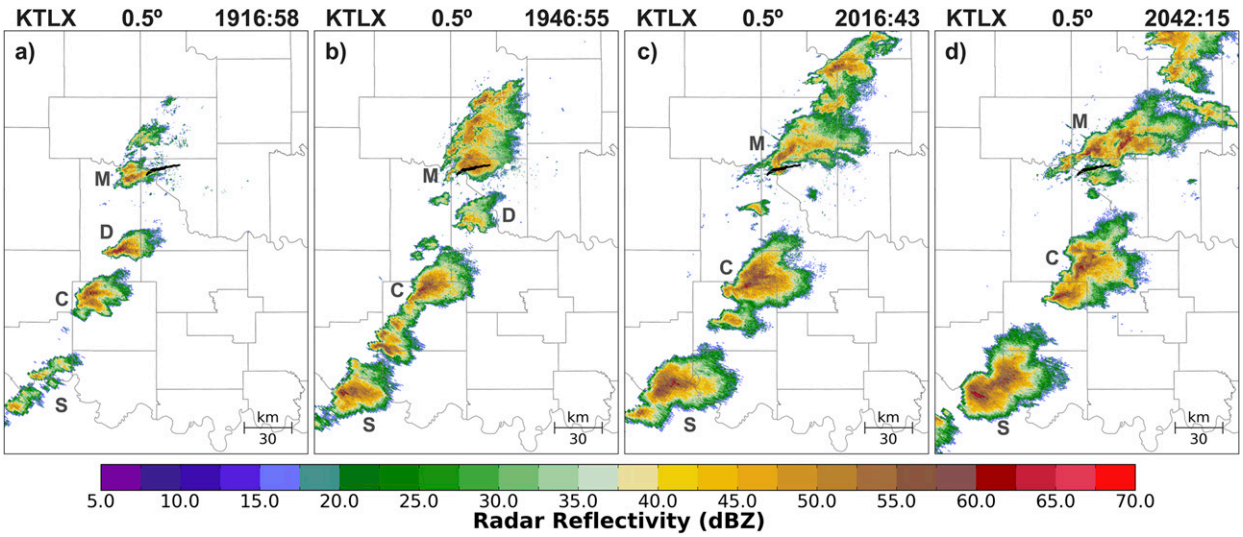

FIG. 2. KTLX 0.5° radar reflectivity at (a) 1916:58, (b) 1946:55, (c) 2016:43, and (d) 2042:15 UTC. The Moore supercell, central domain supercell, and southern supercell are annotated M, C, and S, respectively, and the dissipating storm in the center of the domain is annotated D. The damage track of the Moore tornado is marked in black.

allowable displacement and a characteristic intensity typical of the amplitude of the observed features. For the cases considered herein, the maximum allowable distance of features is set to 30 km, as in the object-based verification, and characteristic intensity is defined as the root-mean-square amplitude of the observed field, as done by Keil and Craig (2009). The combined DAS score is simply equal to the sum of the normalized DIS and AMP scores. Perfect forecasts will result in DIS and AMP scores of 0, and there is no upper limit to the scores. Readers are referred to Keil and Craig (2007, 2009) for a complete description of the displacement and amplitude score.

3. Overview and subjective verification for 20 May 2013

a. Storm evolution

The period of interest for the 20 May 2013 case is from 1915 to 2100 UTC, during which 60-min NEWS-e forecasts were initialized at 1915, 1930, 1945, and 2000 UTC. This period covers the development of three supercells from the incipient through mature stages (Fig. 2), as well as the life cycle of the Moore tornado, which occurred from 1956 to 2035 UTC according to *Storm Data*. At 1915 UTC, developing convection is present across a line oriented from south-southwest to north-northeast within the verification domain (Fig. 2a). The eventual Moore supercell (denoted M in Fig. 2) is present as weak echoes southwest of the damage track, and the two strongest storms (denoted by D and C in Fig. 2) are located farther south in the center of the verification domain. The northern cell of these two storms (D) weakens and merges into the forward flank of the intensifying Moore supercell over the following hour while the southern cell (C) continues to intensify and produces two tornadoes east of Duncan, Oklahoma, between 1958 and 2022 UTC (Fig. 2). Initially multicell storms in the southern portions of the verification domain (S in Fig. 2) congeal into a third supercell by 2000 UTC (Figs. 2a–c), but no tornadoes associated with this supercell were documented during the period of interest.

The low-level mesocyclone path of each of the three supercells is apparent in rotation tracks created by merging maximum azimuthal wind shear values in the verification dataset for each time during the period of interest (Fig. 3). As would be expected, the azimuthal wind shear associated with the Moore supercell is far stronger and more expansive than is observed within the other two supercells. However, the strongest azimuthal wind shear values occur over the first half of the Moore damage track, with smaller maximum values, similar in magnitude to the other two tracks, over the second half of the track. Rotation tracks from the central and southern supercells within the domain are brief and intermittent, consistent with weaker low-level mesocyclones than the Moore supercell.

b. Subjective verification

Probabilistic rotation swaths are produced as the maximum probability of exceedance for mean 0–2 km

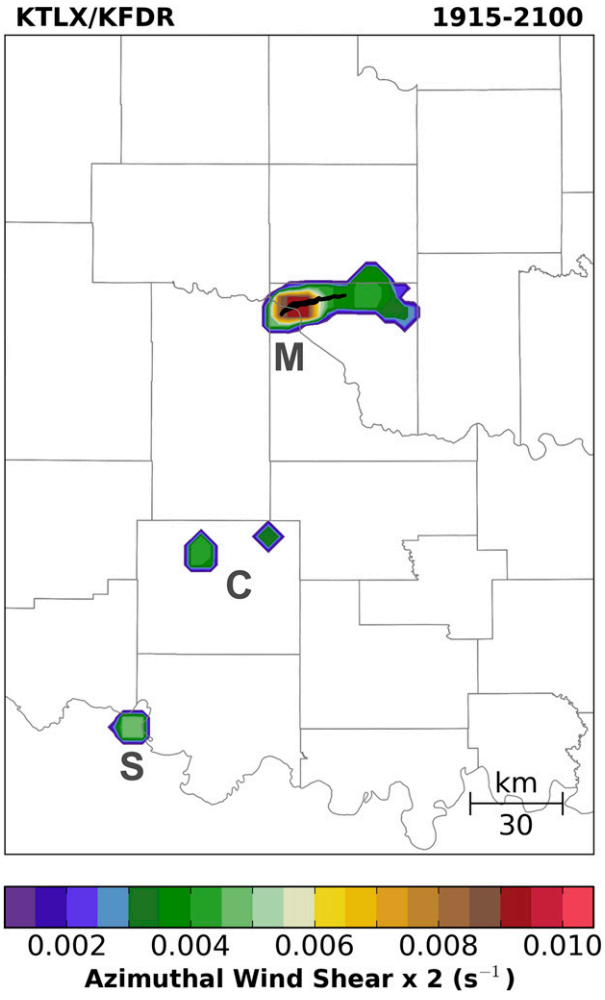

FIG. 3. Azimuthal wind shear swath (s^{-1}) from 1915 to 2100 UTC. Values represent twice the maximum mean 500–2000-m azimuthal wind shear during the time period at each grid point. The Moore supercell, central domain supercell, and southern supercell are annotated M,C, and S, as in Fig. 2.

AGL vertical vorticity greater than $Fcst_{thresh}$ at any time within the forecast period.[5] For the 1915 UTC forecast, probabilities of low-level rotation are generally less than 30% along the track of the Moore supercell and are displaced to the northeast of the observed track. Maximum probabilities of low-level rotation in the forecast occur along the track of the supercell in the center of the domain (Fig. 4a). Improvement in the forecast of the Moore supercell is apparent in each of the subsequent three forecasts, with maximum probabilities increasing to over 90% and shifting southwestward over the observed rotation and damage paths (Figs. 4b–d). Probabilities of low-level rotation along the track of the supercell in the center of the domain generally decrease following the 1930 UTC forecast and become displaced ahead of and to the south of the observed rotation track (Figs. 4b–d). Probabilities of low-level rotation associated with the track of the southern supercell remain below 40% for each forecast (Figs. 4a–d).

The intensity of the predicted low-level rotation is assessed by calculating the ensemble 90th percentile value of the mean 0–2-km vertical vorticity at each grid point, then taking the maximum gridpoint values for any forecast time to create a rotation swath (Figs. 4e–h). Similarly to the probabilistic swaths, these rotation tracks show intensifying rotation along the path of the Moore supercell in each successive forecast, with the highest values present along the first half of the observed Moore rotation track in the 1945 and 2000 UTC forecasts. Upward trends in both the spatial extent and intensity are additionally apparent in subsequent forecasts for the tracks of the two southern supercells, with 90th percentile values associated with the central domain supercell similar in magnitude to values along the track of the Moore supercell in the 1945 and 2000 UTC forecasts (Figs. 4g,h). This similarity in the 90th percentile values along the tracks of the Moore and central domain supercells contrasts the large differences in azimuthal wind shear between the two tracks (Fig. 3) and suggests a lack of intensity variation among ensemble members, likely attributable in part to forecasts poorly resolving the low-level mesocyclone with 3-km horizontal grid spacing (Potvin and Flora 2015).

Smaller temporal variability in forecast low-level mesocyclone intensity than observed azimuthal wind shear values is apparent when the ensemble mean and 90th percentile vertical vorticity values along the track of the Moore supercell are examined (Fig. 5). The ensemble mean magnitudes of forecast vertical vorticity generally increase with increasing time but exhibit relatively little variability across different forecasts (Fig. 5a). This lack of variability is confirmed in variance calculations of ensemble maximum vertical vorticity values, which remain below $1 \times 10^{-4}\,s^{-1}$ for the majority of each forecast (not shown). More variation is apparent when the 90th percentile values are considered, with the 1945 and 2000 UTC forecasts producing relative maxima at similar times to the observed maximum rotation within the Moore low-level mesocyclone. However, strong rotation is maintained after the dissipation of the Moore tornado in both the 1945 and 2000 UTC forecasts (Fig. 5b).

[5] Discrepancies between Fig. 4 and the probability swaths presented in Wheatley et al. (2015) are attributable to methodology differences in calculating low-level rotation fields. The same model output is used by both studies.

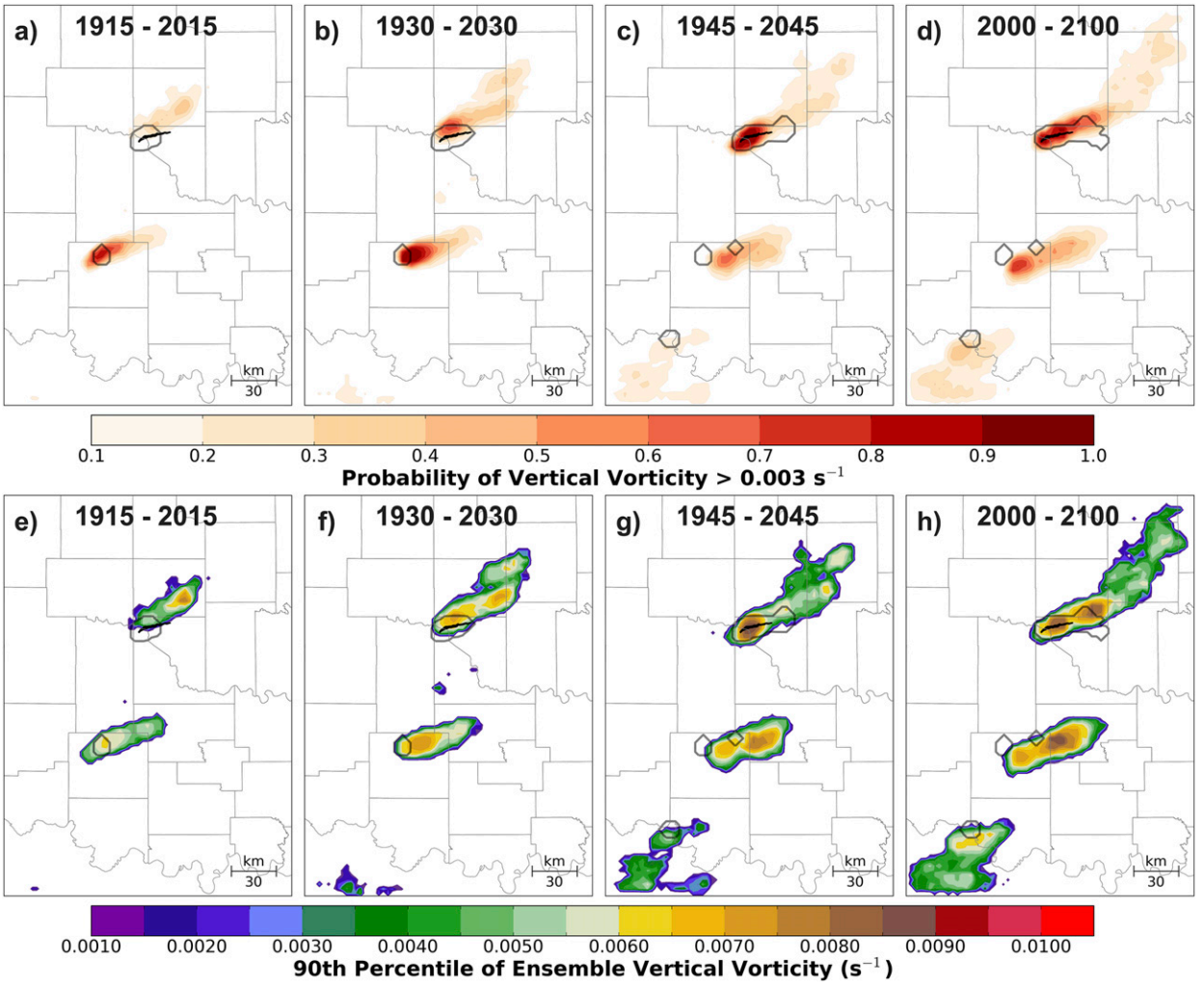

FIG. 4. Probabilities of 500–2000-m average vertical vorticity exceeding 0.003 s^{-1} ($Fcst_{thresh}$) at any time over 1-h forecasts initialized at (a) 1915, (b) 1930, (c) 1945, and (d) 2000 UTC. (e)–(h) The 90th percentile value of ensemble vertical vorticity along each swath. The 0.001 s^{-1} (Ob_{thresh}) contour of twice the observed azimuthal wind shear is plotted in dark gray, and the damage path of the Moore tornado is marked in black.

The evolution of the ensemble forecasts with time can be visualized by comparing the predicted locations of maximum vertical vorticity to the observed location within a 0.8° latitude × 1.0° longitude subset of the full verification domain containing the track of the Moore supercell (Fig. 6). Within this domain, each successive forecast provides an improvement in the predicted storm and low-level mesocyclone location valid for a specific time (cf. panels from left to right across each row in Fig. 6). Additionally, the positions of vertical vorticity maxima for a given time are more tightly clustered with each successive forecast, indicating smaller spread and higher confidence in the location of low-level rotation. While each successive forecast provides an improvement over the prior one for a specific time, degradation of each forecast with time is also apparent (cf. panels from top to bottom down each column in Fig. 6). Furthermore, an obvious high bias in storm speed, at times exceeding 5 m s^{-1}, is present, resulting in predicted vorticity maxima consistently located downstream of observations and becoming farther displaced with increasing forecast time. Similar high biases in storm speed have been regularly identified in recent probabilistic low-level rotation forecasts (e.g., Yussouf et al. 2013; Wheatley et al. 2015; Jones et al. 2016; Yussouf et al. 2015) and research is ongoing to identify the origins of the bias.

Subjectively identified characteristics of NEWS-e forecasts for 20 May 2013 may be summarized as follows:

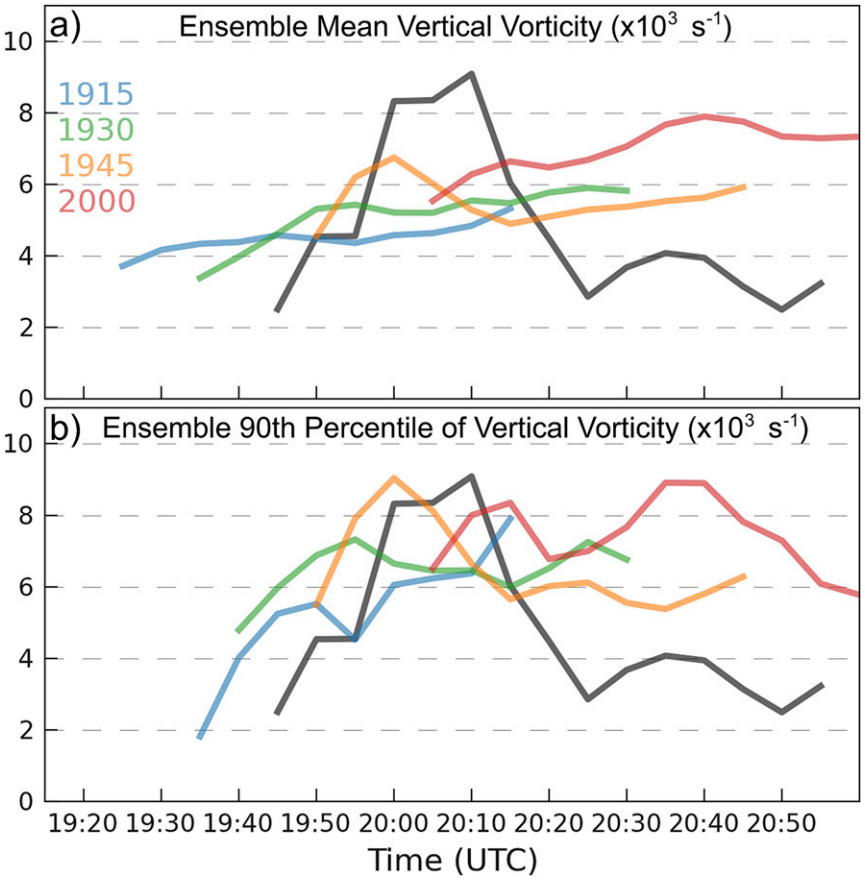

FIG. 5. Time series of the (a) ensemble mean and (b) 90th percentile values of vertical vorticity (s^{-1}) for forecasts initialized at 1915 (blue), 1930 (green), 1945 (orange), and 2000 (red) UTC. Maximum observed values of twice the azimuthal wind shear (s^{-1}) are plotted in black.

- Low-level rotation forecasts for each of the three supercells generally improve in accuracy with each successive forecast, with the largest improvements present along the track of the Moore supercell (Fig. 4).
- Forecasts of low-level rotation intensity are of lower quality than track forecasts and exhibit smaller spatio-temporal variation than observations (Figs. 4 and 5).
- All forecasts decay in quality with increasing forecast time, and a large high bias in storm speed is present (Fig. 6).

The object-based and DAS verification methods are evaluated according to their ability to reproduce these subjectively assessed characteristics.

4. Spatial verification for 20 May 2013

a. Object-based verification

The quality of the object-based verification is initially assessed by considering gridpoint probabilities of matched and false alarm rotation objects (Fig. 7). As portions of the rotation objects may overlap spatially across different ensemble members, grid points may contain nonzero probabilities of being within both matched and false alarm objects (e.g., Figs. 7k,l). This overlap most often occurs when spatial and temporal displacement between forecast and observed objects approaches the maximum allowable offsets and the resulting total interest scores [Eq. (1)] are near the applied threshold of 0.2. Modification of the cutoff radii in space and time [d_{max} and t_{max} in Eq. (1)] or the total interest threshold can alter the distribution of false alarm and matched objects in these regions (see the appendix). However, altering this distribution primarily impacts the binary OTS, POD, and FAR, and will have a smaller impact on the weighted OTS owing to low values of total interest near the maximum time and space radii. The location of the highest probabilities of matched objects mirrors the location of the maximum low-level vertical vorticity in each forecast (Figs. 4 and 6). This similarity provides confidence that most observed rotation

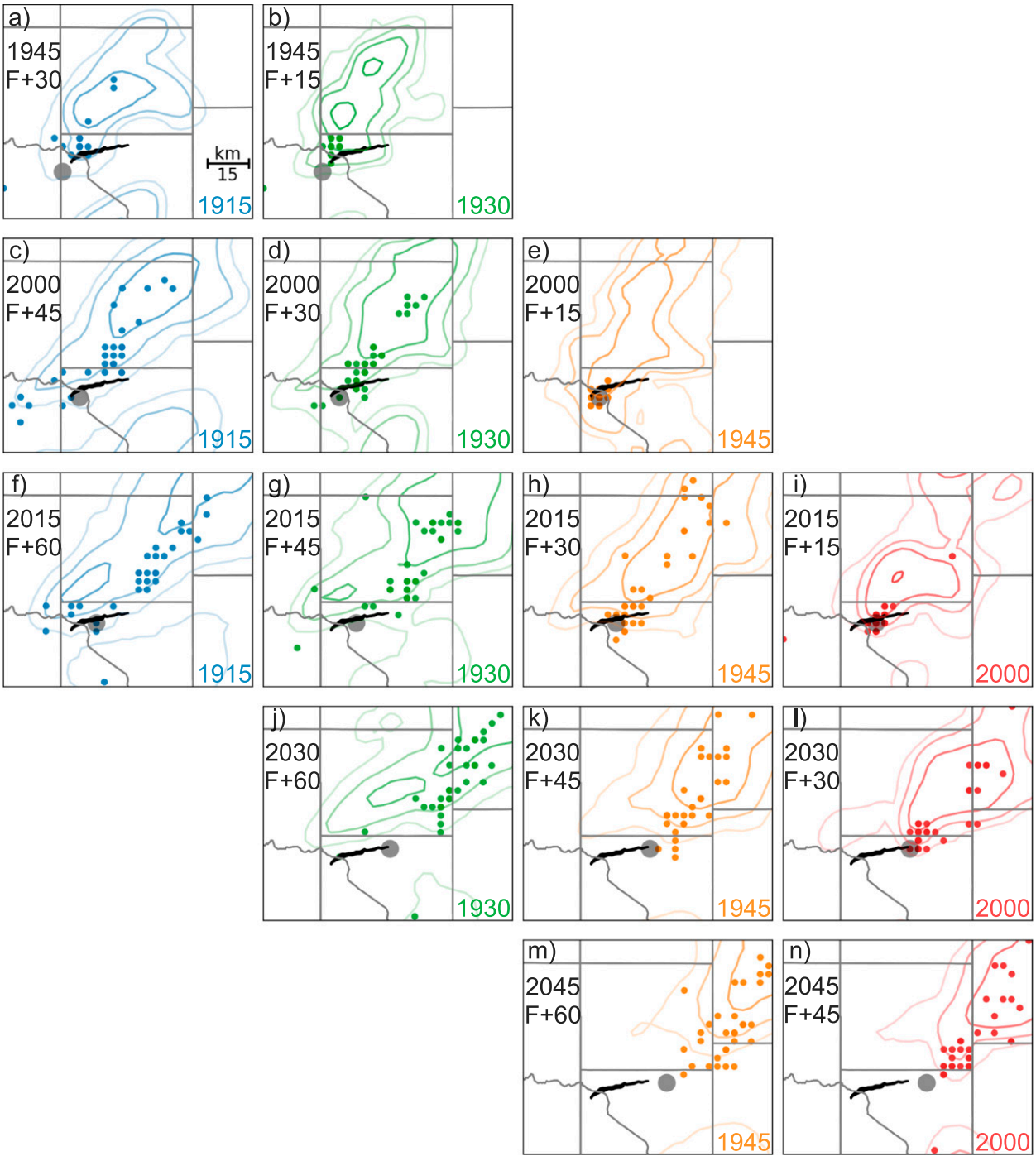

FIG. 6. Locations of ensemble member vertical vorticity maxima along the track of the Moore supercell at (a),(b) 1945, (c)–(e) 2000, (f)–(i) 2015, (j)–(l) 2030, and (m),(n) 2045 UTC. Each maximum is color coded according to the forecast initialization time, which is provided at the bottom right of each panel with the time and forecast minute at the top left. Multiple maxima may be present at a given grid point. Ensemble-mean simulated reflectivity at the lowest model level is contoured at 20, 30, 40, and 50 dBZ, with increasing opacity indicating higher values. The location of the maximum observed azimuthal wind shear is plotted as a gray dot, and the damage path of the Moore tornado is marked in black.

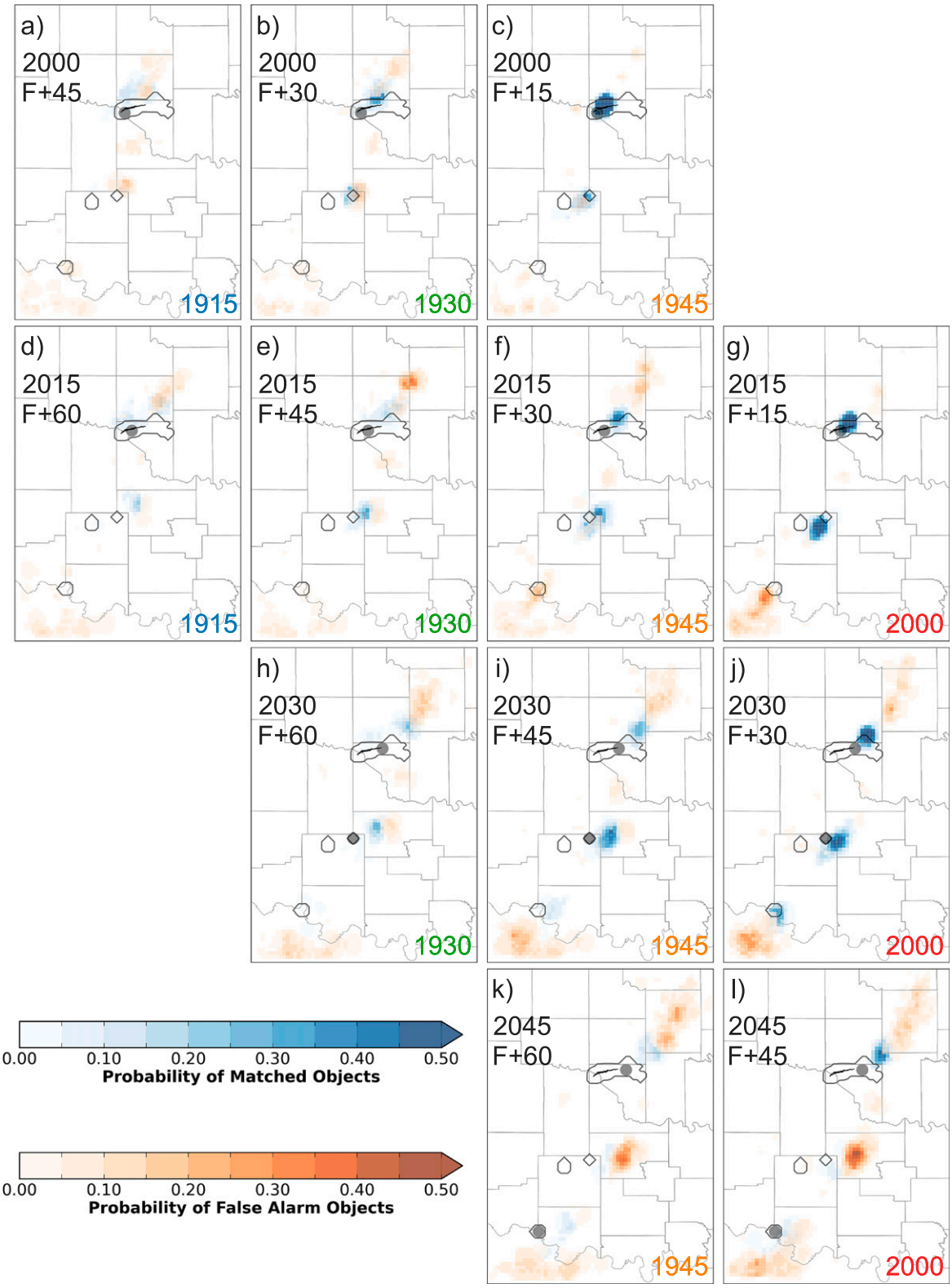

FIG. 7. Ensemble probabilities of matched (blue) and false alarm (orange) rotation objects at (a)–(c) 2000, (d)–(g) 2015, (h)–(j) 2030, and (k),(l) 2045 UTC for each grid point. Individual objects are not plotted, and each grid point may be within multiple matched and false alarm objects across different ensemble members. The forecast initialization time is provided at the bottom right of each panel, and the centroid of observed rotation objects is plotted as a gray dot when present. Observed azimuthal wind shear contour, Moore tornado damage track, and distance scale are as in Fig. 3.

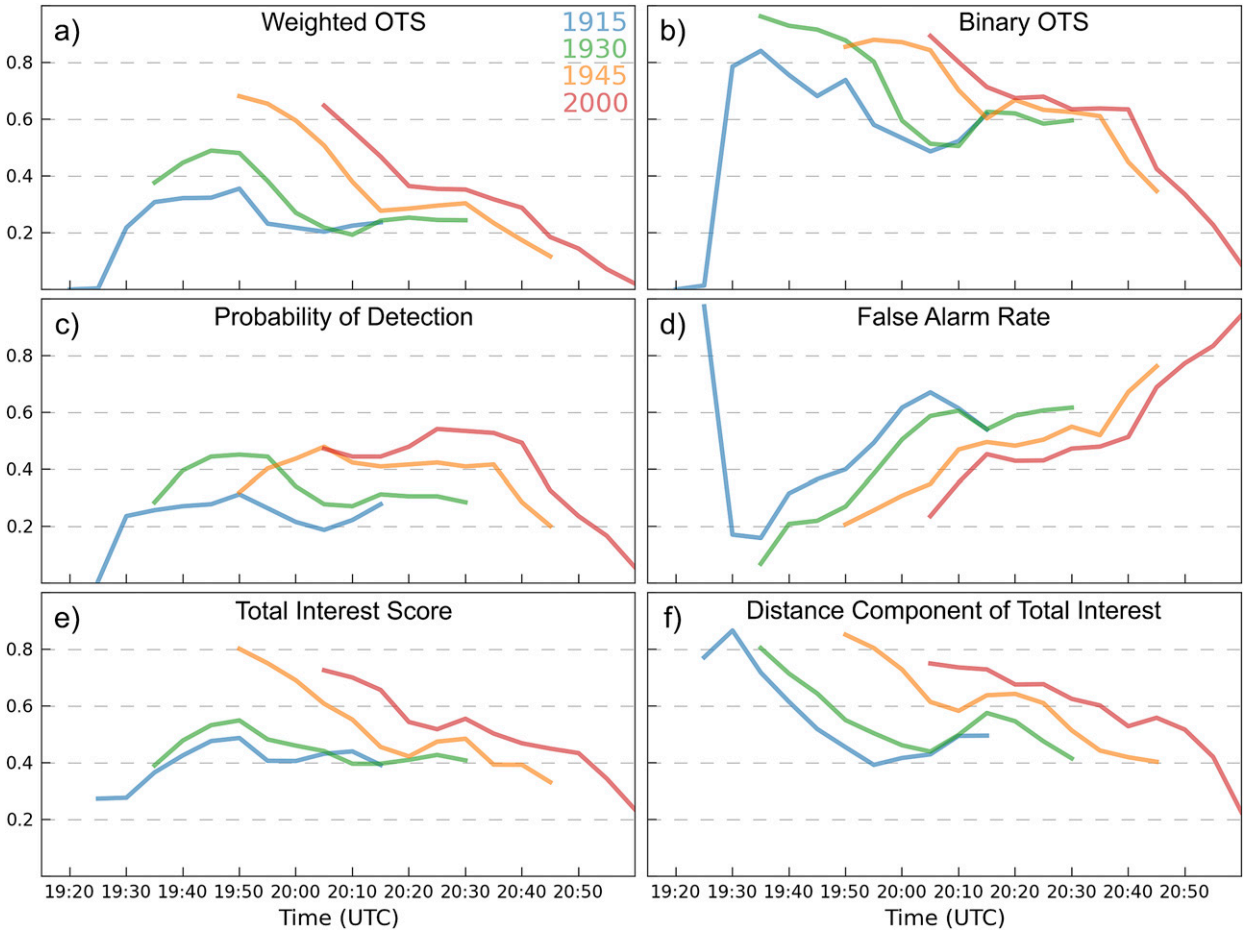

FIG. 8. Ensemble-mean time series of (a) weighted object-based threat score, (b) binary object-based threat score, (c) probability of matching observed objects (probability of detection), (d) percentage of unmatched forecast rotation objects (false alarm rate), (e) total interest score of matched objects, and (f) the distance component of the total interest score for rotation objects within the verification domain. Each forecast is color coded as in Fig. 5.

objects are being correctly matched to corresponding forecast rotation objects.

The evolution of matched rotation objects in successive forecasts is similar to the subjective evaluation. A dramatic increase in the probability of matched objects is present between the forecasts initialized at 1915 and 2000 UTC, particularly along the track of the Moore supercell (Fig. 7). Additionally, each subsequent forecast from 1915 UTC results in higher concentrations of matched objects at a given time and closer proximity between the matched objects, suggesting improvements in both accuracy and forecast confidence with increasing time (each row in Fig. 7). In contrast to the improvement in forecast quality with each successive forecast, the forecast objects become less concentrated and farther displaced from observed objects with increasing forecast time in each forecast (columns in Fig. 7.) This degradation of forecast quality with increasing forecast time is also apparent in the number of false alarm objects. Low probabilities of false alarm objects are present over a relatively large region of the verification domain during the latter portion of each forecast (Figs. 7a,d,e,h,i,k,l). These objects are associated with the development of spurious rotation objects within individual ensemble members during the forecast. Additional regions of more highly concentrated false alarm objects are apparent along the track of the two southern supercells. These objects occur as the forecast objects become too far displaced in space (Figs. 7k,l) or time (Figs. 7f,g) to exceed the total interest score threshold.

Ensemble-mean[6] properties of rotation objects within the verification domain can be used as summary measures of forecast quality (Figs. 8–10). Weighted object-based

[6] Ensemble-mean properties for object- and optical flow-based verification are produced by averaging verification statistics calculated individually for each member.

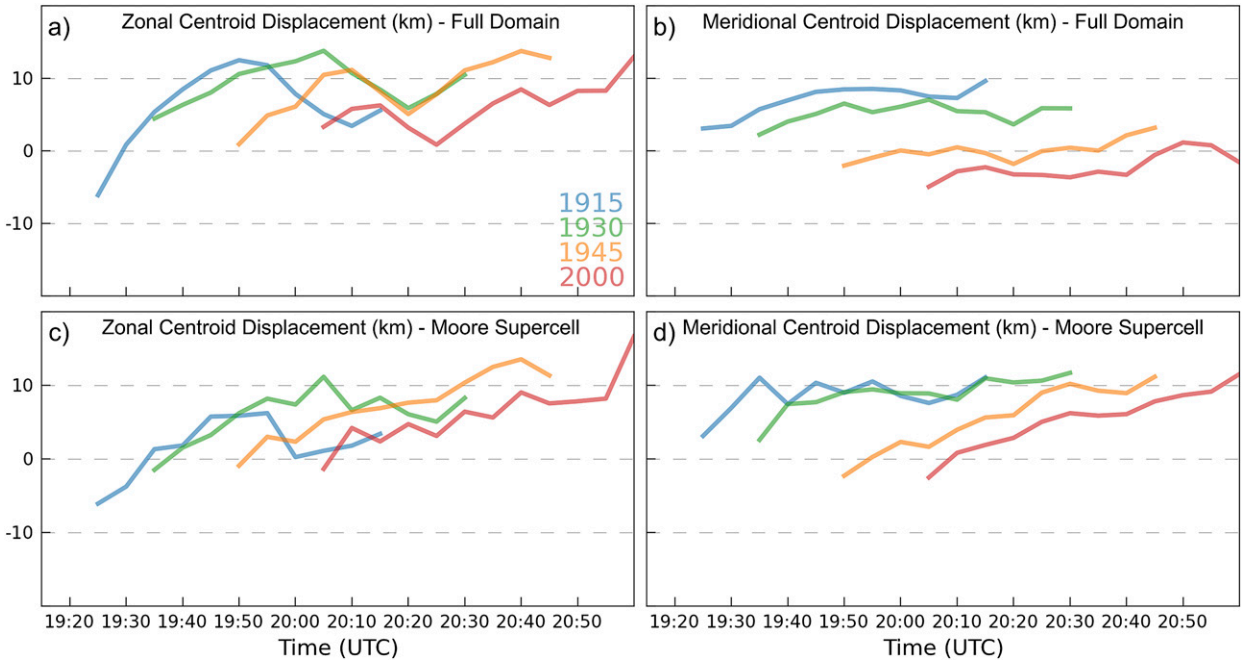

FIG. 9. As in Fig. 8, but for the (left) zonal and (right) meridional components of ensemble-mean object centroid displacement for matched objects in the (a),(b) full verification and (c),(d) Moore supercell domain. Positive values of zonal (meridional) displacement indicate an eastward (northward) offset in forecast rotation objects.

threat scores calculated for each forecast match the expected evolution from subjective verification (Fig. 8a). An increase in OTS is apparent with each successive forecast, with a large improvement in skill noted between the first half of the 1945 and 2000 UTC forecasts compared to the 1915 and 1930 UTC forecasts (Fig. 8a). This increase is a reflection of the two early forecasts predicting strong rotation along the track of the Moore and central domain supercells prior to the observed development of a low-level mesocyclone (Fig. 5). The temporal offset in these forecasts results in relatively low total interest scores (Fig. 8e) despite a distance component indicating little spatial error (Fig. 8f). Timing errors are also responsible for the dramatic variation in binary OTS, probability of detection, and false alarm rate over the first 15 min of the 1915 UTC forecast (Figs. 8b–d), where the majority of predicted rotation objects are present more than 20 min (t_{max}) prior to the observed rotation objects, resulting in a high false alarm rate (Fig. 8d). The variation over the initial portions of the 1915 UTC forecast illustrates the advantages of summary verification measures weighted by total interest.

Incremental improvements with each forecast are additionally apparent in the binary OTS (Fig. 8b), but less prominent than those in the weighted OTS, consistent with improvements in both the ensemble probability of producing matching rotation objects to observations (binary OTS) and the spatiotemporal accuracy of predicted rotation objects (weighted OTS). The relatively low probability of detection values (Fig. 8c), which remain below 0.6, are a result of a low concentration of matched objects in the two southern supercells prior to the 2000 UTC forecast (Fig. 7) and are consistent with lower probabilities of low-level rotation along the tracks of the southern two supercells (Figs. 4a–d). A dramatic increase in POD, and related binary OTS, occurs if the subset of the domain containing the track of the Moore supercell is considered (not shown), as would be expected given the higher probabilities of forecast low-level rotation within that storm (Fig. 4).

The degradation of forecasts with increasing forecast time is also present in the ensemble-mean object-based skill scores. Each forecast exhibits a decrease in the distance component of the total interest score (Fig. 8f) coupled with an increase in the false alarm rate (Fig. 8d) with increasing forecast time. The increase in false alarm rate is attributable to both the development of spurious rotation objects as well as rotation objects too far displaced in space and time to be matched (Fig. 7). The positive bias to storm motion is the primary contributor to the reduction in the distance component, and by extension reductions in interest score and weighted OTS, over the course of a given forecast.

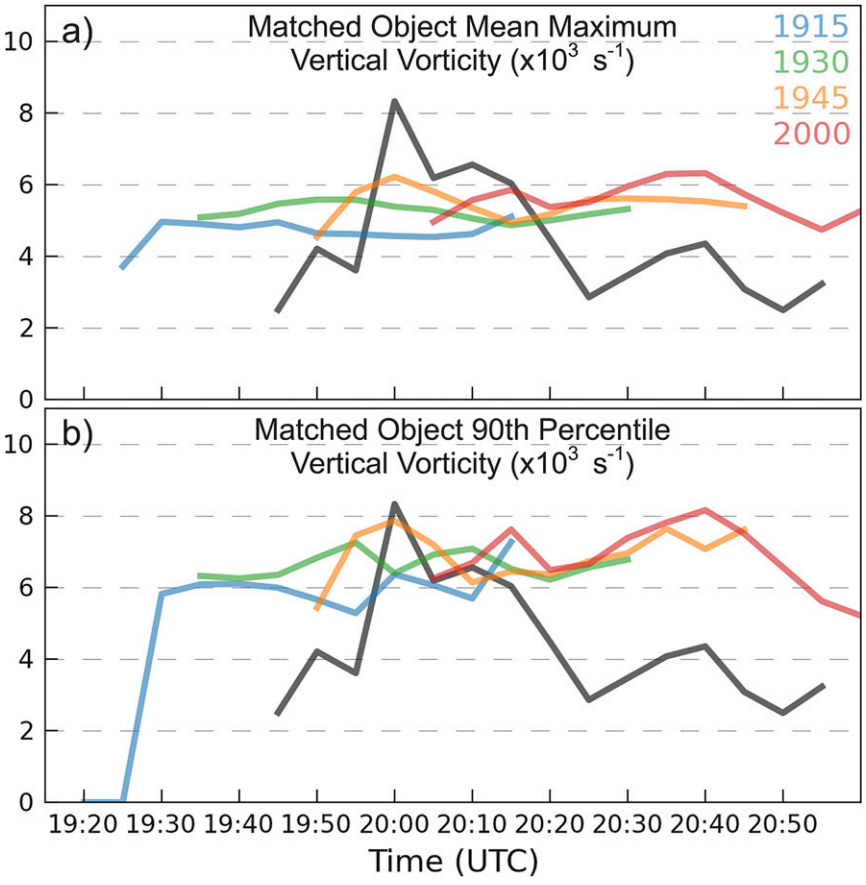

FIG. 10. As in Fig. 5, but for the (a) ensemble mean and (b) 90th percentile values of maximum vertical vorticity for matched rotation objects.

Storm motion biases may be directly quantified by considering the ensemble-mean zonal and meridional centroid displacement for matched objects (Fig. 9). Positive zonal displacement values, indicating an eastward bias in storm motion, are present and become larger with increasing forecast time in nearly every forecast (Figs. 9a,c). Small reductions in zonal centroid displacement during the mid- to latter portions of each forecast, apparent as periods of negative slope in Figs. 9a and 9c, are attributable to spurious rotation objects trailing the primary supercells being matched to the observed rotation objects (not shown). These spurious matches most often occur with rotation objects associated with the southern two supercells; if the subset of the verification domain containing only the track of the Moore supercell is considered, a more consistent increase in centroid distance with time is apparent (Figs. 9c,d). An incremental, southward displacement in matched-object centroid position is additionally apparent (Figs. 9b,d), which is consistent with the observed southward shifts in rotation tracks (Fig. 4) and vorticity maxima (Fig. 6) with successive forecasts.

Both the ensemble mean and 90th percentile values of maximum vertical vorticity in matched rotation objects exhibit little variation among different forecasts and during the duration of each forecast (Fig. 10). The general similarity between the object-based and subjective (Fig. 5) comparisons of maximum vertical vorticity is expected, provided the maximum vertical vorticity values within the domain occur in matched forecast objects. The similarity of Figs. 5 and 10 provides additional confidence that rotation objects representing the low-level mesocyclone of each of the three supercells are being appropriately matched.

b. *DAS verification*

The ensemble-mean displacement and amplitude score values for each forecast are largely determined by the contribution of the amplitude component, which is generally larger in magnitude than the distance component (Fig. 11). Additionally, little variation in AMP

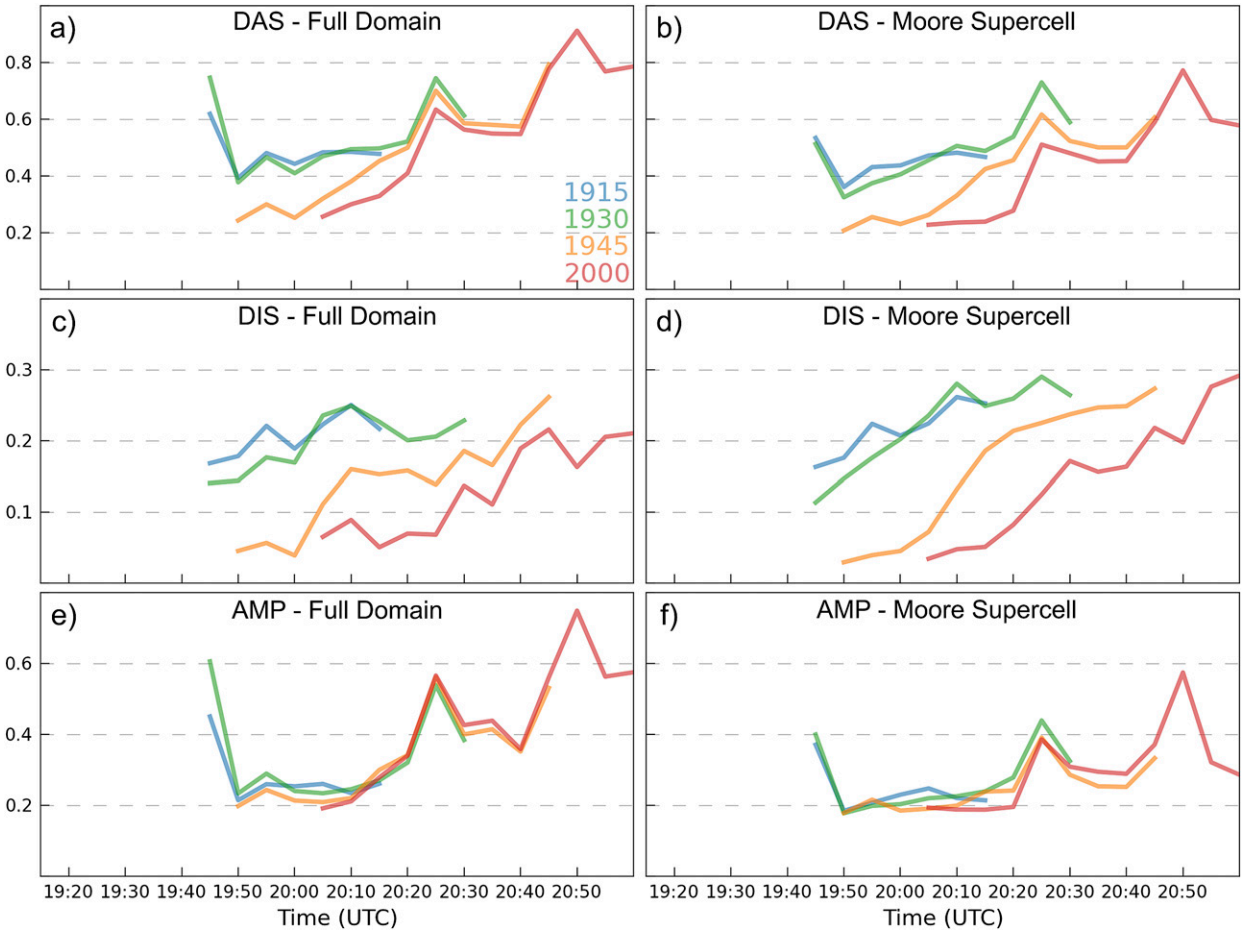

FIG. 11. As in Fig. 8, but for values of (a),(b) DAS, (c),(d) DIS, and (e),(f) AMP within the (left) full verification and (right) Moore supercell domains.

among different forecasts valid at the same time is noted, which contrasts large variations over the course of any given forecast (Figs. 11e,f). The similarity in AMP scores between different forecasts is similar to object-based and subjective intensity comparisons; however, the dramatic variation with time suggests that AMP values are being primarily driven by fluctuations in the observed azimuthal wind shear, which will impact both amplitude changes between the forecast and verification datasets as well as the characteristic intensity used to normalize the AMP score (Keil and Craig 2009).

The DIS component exhibits larger variety among different forecasts and generally follows the expected forecast evolution from subjective and object-based verification (Figs. 11c,d). A reduction in DIS, implying a more accurate spatial forecast, is present with each successive forecast and DIS generally increases with forecast time, indicating larger displacement between the forecast and observed regions of low-level rotation. Similarly to the object-based verification (Figs. 8f and 9a,c), more slowly increasing, or at times subtly decreasing, DIS values are evident during the latter portion of forecasts when calculated over the full verification domain (Fig. 11c). This improvement occurs as spurious regions of rotation develop trailing the two southern supercells, resulting in an apparent, but erroneous decrease in distance error. A more consistent increase in DIS with increasing forecast time is present when the subdomain containing the Moore supercell is considered (Fig. 11d).

An important difference between the object-based and DAS methods is that DAS is only calculated when nonzero values of both forecast and observed rotation are present. As a result, ensemble members that do not produce low-level vertical vorticity exceeding $\text{Fcst}_{\text{thresh}}$ are not included in the ensemble-mean calculation; in other words, DAS values do not account for missed events. Therefore, DIS values are best interpreted similarly to the total interest score, as a relative measure of

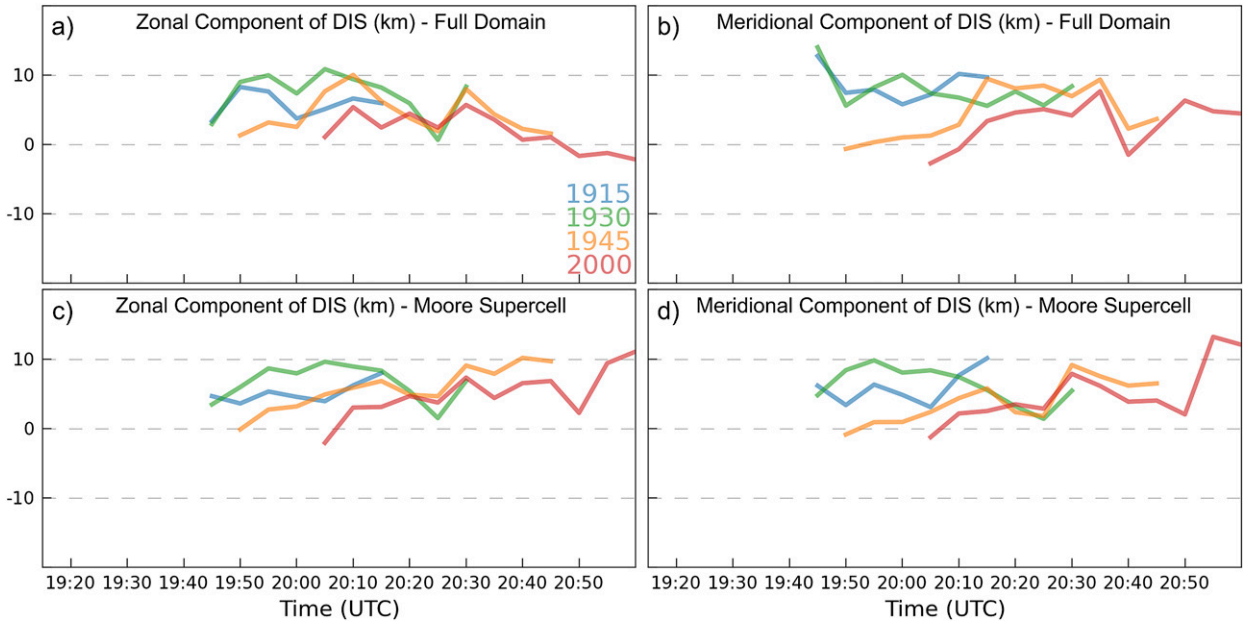

FIG. 12. As in Fig. 9, but for the optical flow-derived ensemble-mean (a),(c) zonal and (b),(d) meridional displacement (km).

forecast quality among members producing useful forecasts. An additional consequence of requiring nonzero values in both the forecast and observed datasets is that DIS scores are not available prior to 1945 UTC when observed values of twice the azimuthal wind shear first exceed Ob_{thresh} (Fig. 5). This unavailability accounts for the apparently large improvement from the 1930 UTC forecast, for which no DIS scores are available during the first 15 min of the forecast, and the 1945 UTC forecast (Figs. 11c,d). Additionally, the intermittent observations of low-level rotation within the southern two supercells result in greater temporal variability in DIS scores calculated for the full verification domain than in the subdomain including the Moore supercell.

Similarly to the object-based centroid displacement, the DIS score can be partitioned into zonal and meridional components (Fig. 12). This is accomplished by averaging the ensemble mean, nonnormalized magnitude of vectors used to morph the observation to the forecast field with the negative of the vectors used to morph the forecast to the observation field. Resulting values of zonal and meridional displacement do indicate generally positive biases to storm motion and a southward displacement increment with each successive forecast. However, these trends are subtle, do not indicate increasing forecast displacement with time, and do not match the subjective interpretation as well as the object-based measures (Figs. 6 and 9). The lack of clearly increasing displacement with forecast time in the DIS components is likely a result of the morphing process, in which vector orientation may vary in a counterintuitive manner over small distances in order to minimize the RMS error between the two fields (Keil and Craig 2009; their Figs. 2 and 4).

c. *Comparison with additional NEWS-e cases*

The ability of the object-based and DAS methods to discriminate between forecast quality is evaluated using a small subset of low-level rotation forecasts from additional NEWS-e cases (Wheatley et al. 2015). In addition to the Moore supercell, verification scores are calculated for forecasts of supercells producing long-track tornadoes on 31 May 2013 (Bluestein et al. 2015) and 11 May 2014. These two additional cases were selected as they bracket the subjectively determined skill in probabilistic rotation swaths generated for events from the springs of 2013 and 2014 (Fig. 13). Forecasts of the 31 May 2013 storm produce high probabilities of low-level rotation over the observed tornado track an hour prior to tornado genesis and maintain high probabilities through the duration of the event (Figs. 13e–h). In contrast, moderate probabilities of low-level rotation are not predicted for the 11 May 2014 event until 30 min prior to tornado genesis, and the probabilistic swath is displaced to the north of the damage track in forecasts initialized with 30 and 15 min of lead time (Figs. 13j,k). High probabilities of low-level rotation are not predicted over the damage track until the final forecast (Fig. 13l). Forecasts for the Moore supercell represent

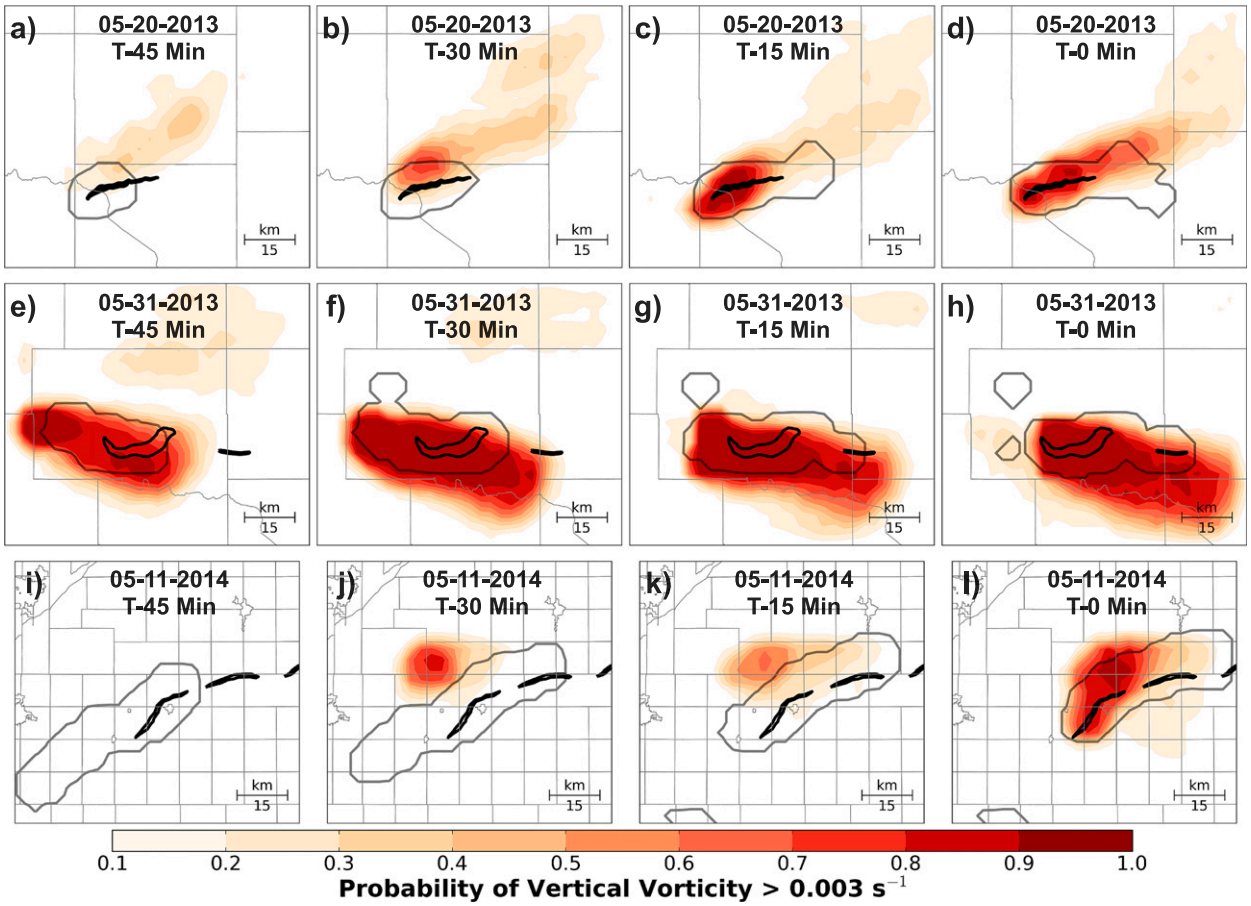

FIG. 13. NEWS-e probabilities of 500–2000-m average vertical vorticity exceeding 0.003 s^{-1} for 1-h forecasts of the (top) 20 May 2013 Moore, (middle) 31 May 2013, and (bottom) 11 May 2014 tornadic supercells. Forecasts are initialized (a),(e),(i) 45; (b),(f),(j) 30; (c),(g),(k) 15; and (d),(h),(l) 0 min prior to the time of tornado genesis. The Ob_{thresh} contour of twice the observed azimuthal wind shear (s^{-1}) is plotted in dark gray, and damage paths are marked in black for each case.

an intermediate level of quality, where probabilities of low-level rotation are higher and less displaced spatially than those for 11 May 2014, but are lower and with larger spatial errors than those for the 31 May 2013 case (Figs. 13a–d).

Weighted OTS and DIS values are calculated for each of the three cases for forecasts initialized with 45, 30, 15, and 0 min of lead time (Fig. 14). Each event is verified using a 0.8° latitude × 1.0° longitude subset of the full forecast domain that encompasses the rotation track of each supercell. Generation of the forecast and verification low-level rotation fields are identical for each case, with the exception that only KTLX and the Hastings, Nebraska (KUEX), radial velocities are used for the 31 May 2013 and 11 May 2014 events, respectively.

Forecasts of the 31 May 2013 event issued 45 min prior to tornado genesis produce dramatically higher weighted OTS scores than those for the 11 May 2014 and Moore events (Fig. 14a).[7] Additionally, DIS values for the 31 May 2013 forecast are well below those for the other two events, indicating less displacement between the observed and forecast rotation tracks (Fig. 14b). OTS scores for the 31 May 2013 forecast issued with 30 min of lead time remain well above the other two events throughout the duration of the forecast. However, large increases in OTS are apparent for both the 11 May 2014 and Moore forecasts, reflecting the development of higher probabilities of low-level rotation near the observed rotation tracks. Additionally, higher values of OTS are present during the early portions of the

[7] Smaller initial OTS values for the Moore supercell are produced compared to the full verification domain (Fig. 8a), indicating that much of the forecast skill at that time was associated with predictions of the central domain supercell. These values are consistent with low probabilities of rotation along the Moore track in the 1915 UTC forecast (Fig. 13a).

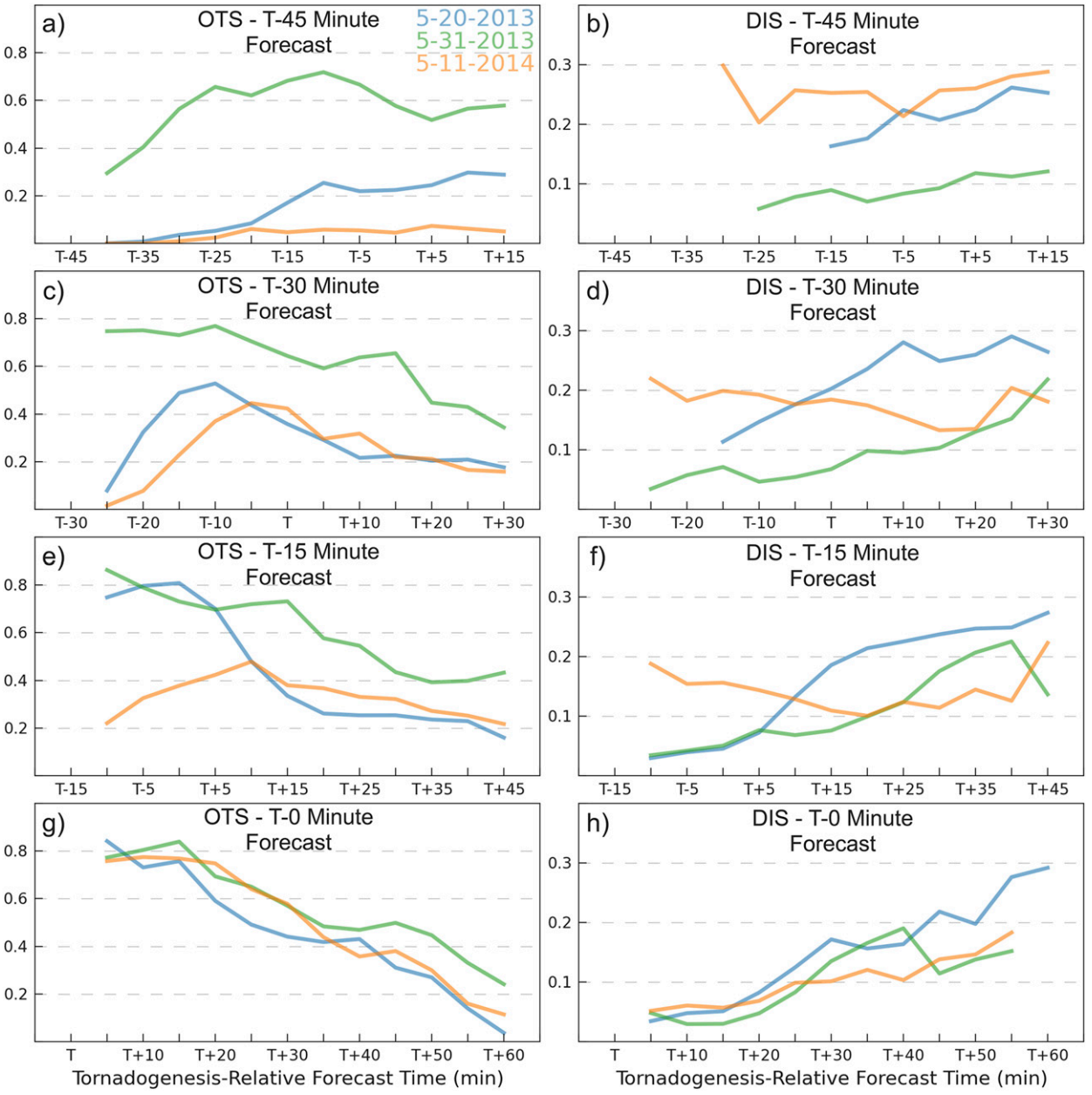

FIG. 14. Time series plots of ensemble-mean (left) weighted OTS and (right) DIS for forecasts initialized (a),(b) 45, (c),(d) 30, (e),(f) 15, and (g),(h) 0 min prior to genesis of the strongest tornado observed during the 20 May 2013 Moore supercell (blue), on 31 May 2013 (green), and on 11 May 2014 (orange). Each forecast score is calculated for only the target supercell domain.

Moore forecast than for 11 May 2014, which is indicative of the smaller spatial displacement apparent in the DIS values (Figs. 14c,d). Both the OTS and DIS scores indicate further improvement in quality for the first half of the 1945 UTC ($T - 15$ min) Moore forecast, with maximum (minimum) OTS (DIS) values equaling those for the 31 May 2013 forecast (Figs. 14e,f). In contrast, OTS values remain comparatively low for the 11 May 2014 forecast, with larger DIS values reflecting the northward spatial displacement in the probabilistic swath at this time. High predicted probabilities of low-level rotation collocated with the observed damage tracks are produced for all three events in forecasts initialized at the time of tornado genesis (Figs. 13d,h,l). This apparent skill is indicated by the objective verification scores, which produce similarly high (low) values of OTS (DIS) for each forecast throughout the forecast period (Figs. 14g,h).

5. Summary and discussion

Two spatial verification methods, a time-weighted, object-based technique based on the MODE algorithm (Davis et al. 2006a,b) and the optical-flow-based displacement and amplitude score (Keil and Craig 2009), have been applied to convective-scale ensemble forecasts of low-level rotation. Forecast and verification low-level rotation fields are created by postprocessing predicted vertical vorticity and observed single-Doppler azimuthal wind shear in an effort to isolate low-level mesocyclones. The spatial verification techniques are assessed for their ability to reproduce subjective interpretations of forecast quality and evolution.

The object-based verification method is found to produce results consistent with subjective interpretations of each forecast (Figs. 4–10). Correspondence between forecast and observed low-level rotation objects is calculated according to a total interest score weighted primarily on spatial and temporal displacement. Matched objects, representing the highest available total interest score between forecast and observed rotation objects, are consistent with subjectively interpreted locations of maximum predicted vertical vorticity for the 20 May 2013 case (Figs. 4, 6, and 7). Additionally, utilization of ensemble-mean values of the object-based threat score (Johnson et al. 2011; Johnson and Wang 2013) as a summary score reproduces apparent improvements in quality between successive forecasts, as well as degradation in quality with increasing forecast time (Figs. 4 and 6–8). Partitioning the OTS into individual components provides further information on the factors determining changes in quality (Fig. 8). Examples include a binary calculation of OTS, which provides an area-weighted measure analogous to a threat score, and the total interest value, which provides a measure of spatiotemporal proximity between objects that may be further decomposed to isolate the contributions of spatial and temporal displacement. Additionally, bulk measures of matched objects can quantify specific forecast errors; for example, changes in the ensemble-mean centroid displacement among matched objects captures a positive bias in storm motion in the 20 May forecasts (Fig. 9).

Variation in the distance component of the displacement and amplitude score is consistent with both subjective and object-based measures of spatial displacement in the 20 May forecasts (Fig. 11). However, application of the DIS component to forecasts of low-level rotation is limited by a requirement that nonzero values of forecast and observed rotation are present. The relatively sparse and transient nature of low-level mesocyclones compared to other forecast products such as precipitation results in a reduced ability for DIS to account for rapid changes in observations. Additionally, DIS scores are limited to periods when low-level mesocyclones were observed, and ensemble members that do not produce strong low-level rotation are not incorporated into the DIS score. It is therefore recommended that DIS be utilized as a relative measure of spatial displacement among members producing strong low-level rotation, similar to the distance component of the total interest.

The weighted OTS and DIS scores are additionally found to reproduce subjective evaluations of probabilistic low-level rotation forecasts issued across a small subset of tornadic events during the springs of 2013 and 2014 (Figs. 13 and 14). Discrimination in forecast quality between different cases in both OTS and DIS values matches the subjective interpretation of probabilistic rotation swaths for three events identified as relatively poor, intermediate, and good forecast quality (Fig. 13; Wheatley et al. 2015).

Both the object-based and DAS methods produce minimal variability in maximum intensity measures between different forecasts (Figs. 10 and 11). This result is unsurprising considering the imperfect verification dataset and limitations in model resolution. Utilizing single-Doppler azimuthal wind shear as a proxy for vertical vorticity requires an assumption of solid-body rotation, and the effective resolution of observations will differ from those in the model, despite being interpolated onto the same grid. Additionally, the horizontal grid spacing of 3 km utilized by the NEWS-e is not sufficient to represent many storm-scale dynamic processes within supercells, resulting in only partial resolution of simulated low-level mesocyclones (Potvin and Flora 2015). These discrepancies are apparent in the different representations of low-level mesocyclones in forecast and verification fields (Fig. 1) and make an assessment of ensemble skill in predicting the intensity of low-level rotation prohibitively difficult at the current time.

There are several limitations to the methods presented herein. Primarily, a large amount of tunable parameters are required in the development of the low-level rotation datasets and in calculation of the verification scores. Averaging, convolution, and thresholding applied to both the forecast and verification datasets may be altered to produce different representations of low-level rotation. Parameters chosen for this study are similar to prior methods used in probabilistic rotation forecasts (e.g., Dawson et al. 2012; Yussouf et al. 2013; Jones et al. 2016; Wheatley et al. 2015; Yussouf et al. 2015) and have been chosen to best isolate the features of interest, observed and forecast low-level

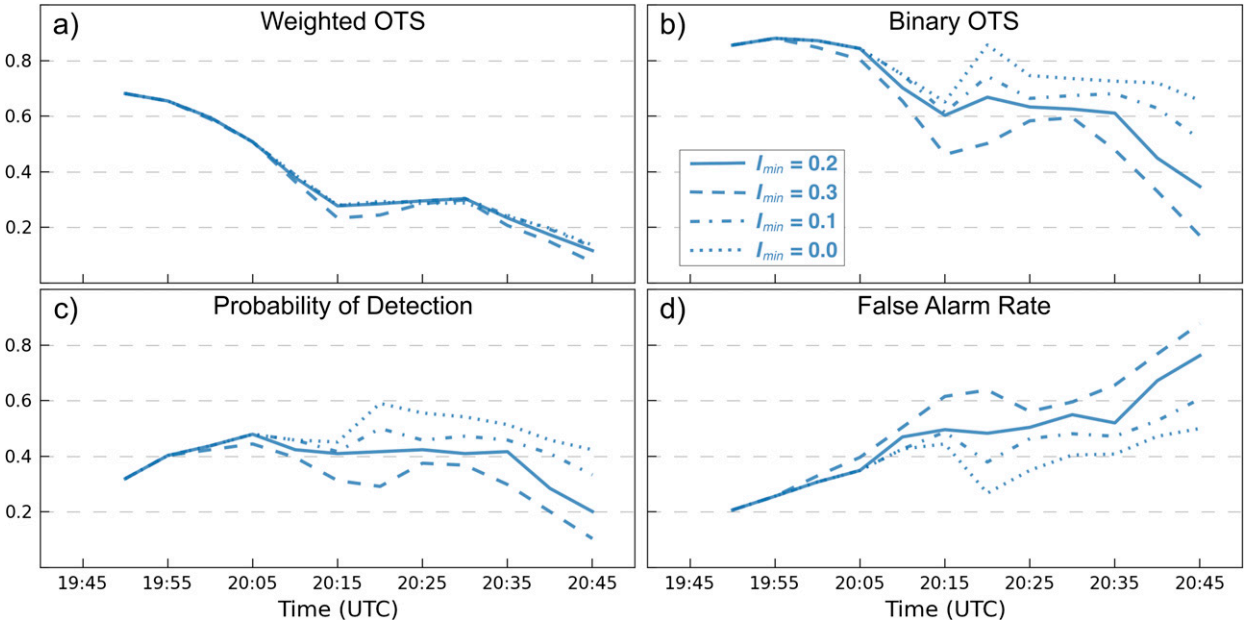

FIG. A1. Ensemble-mean time series of (a) weighted object-based threat score, (b) binary object-based threat score, (c) probability of matching observed objects (probability of detection), and (d) percentage of unmatched forecast rotation objects (false alarm rate) for the full verification domain of the forecast initialized at 1945 UTC 20 May 2013. Plots are included for varying total interest thresholds I_{min} of 0.2 (solid), 0.3 (dashed), 0.1 (dashed–dotted), and 0.0 (dotted), in which no threshold was applied.

mesocyclones, in the three cases considered (Wolff et al. 2014). However, application of these methods to a much larger database of probabilistic rotation forecasts is necessary to determine their ability to consistently isolate areas of intense low-level rotation in varying convective modes and large-scale environments. Changes to tunable parameters used by the spatial verification methods, such as maximum allowable space offsets for object matching or image morphing, will result in variations in subsequent skill scores. However, these variations are consistent across different forecasts, resulting in changes in the absolute magnitude of skill scores but not relative changes in scores between different forecasts.

Only ensemble forecasts of low-level rotation, used as an indication of tornado likelihood, are considered in this study. As Warn-on-Forecast is envisioned as a total hazard short-term prediction system (Stensrud et al. 2009, 2013), extension of the methodologies presented herein to probabilistic forecasts of severe hail, straight-line winds, and flash flooding will be necessary. This extension will require further development and refinement of observational proxies for thunderstorm hazards or the development of high-resolution numerical analyses of convective storms (Gao et al. 2013) to serve as verification datasets.

We do not recommend that the spatial verification methods presented herein be used in place of qualitative assessments of probabilistic rotation tracks for individual case studies. Rather, we envision the objective scores being useful for quantifying differences across a large dataset of forecasts, as will be created as prototype Warn-on-Forecast systems approach operational implementation. Automation of spatial verification methods will allow forecasts run with variations in model and data assimilation methodology to be intercompared in order to assist in determining best practices for convective-scale numerical weather prediction. Additionally, performance of probabilistic forecasts of low-level rotation across different storm modes and environments is largely unknown, and may be quantified through comparison of spatial verification skill scores.

Acknowledgments. Dr. Christian Keil is gratefully acknowledged for providing code for calculating the displacement and amplitude score. Additionally, we thank Dr. Thomas Jones and Mr. Gerry Creager for supporting NEWS-e development, Dr. Corey Potvin for many helpful conversations, and Drs. Jeff Snyder and Chris Karstens for assisting in the production of rotation tracks. We appreciate thoughtful reviews provided by Dr. Adam Clark and three anonymous reviewers, which resulted in an improved manuscript. The Warning Decision Support System–Integrated Information and Observational Processing and Wind Synthesis software were utilized in processing WSR-88D data and the

freely provided enthought python build and SciPy, matplotlib, Basemap, and scikit-image python libraries were used to create the analyses herein. PSS was supported by a National Research Council Research associateship and additional funding was provided by NOAA's Warn-on-Forecast project.

APPENDIX

Sensitivity of Object-Based Verification to the Total Interest Threshold

Verification measures used with the object-based method will be sensitive to changes in tunable parameters such as the maximum allowable offsets in space d_{max}, time t_{max}, and the total interest threshold required for matching objects. A representative example of this sensitivity using the total interest threshold is provided in Fig. A1. As would be expected, a reduction of the total interest threshold results in more matched objects (Fig. A1c), fewer false alarms (Fig. A1d), and a resulting increase in the binary OTS (Fig. A1b), mostly during the latter portion of the forecast when rotation objects are near the prescribed maximum space and time radii. As the variation in matching frequency occurs in objects with relatively low total interest scores, the weighted OTS is minimally affected by changing the threshold (Fig. A1a). Additionally, assessment of the relative quality across forecasts is mostly insensitive to changing tunable parameters, as they will induce consistent changes in verification measures across different forecasts.

Variation of the maximum time and space offsets for object matching will produce larger changes to weighted OTS values, since both the relative frequency of the matched objects and the total interest scores of the matches will change. However, as with variation of the total interest threshold, these changes will occur in a consistent manner, preserving the ability to intercompare different forecasts.

REFERENCES

Anderson, J. L., and N. Collins, 2007: Scalable implementations of ensemble filter algorithms for data assimilation. *J. Atmos. Oceanic Technol.*, **24**, 1452–1463, doi:10.1175/JTECH2049.1.

——, T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The Data Assimilation Research Testbed: A community facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296, doi:10.1175/2009BAMS2618.1.

Atkins, N. T., K. M. Butler, K. R. Flynn, and R. M. Wakimoto, 2014: An integrated damage, visual, and radar analysis of the 2013 Moore, Oklahoma, EF5 tornado. *Bull. Amer. Meteor. Soc.*, **95**, 1549–1561, doi:10.1175/BAMS-D-14-00033.1.

Bluestein, H. B., J. C. Snyder, and J. B. Houser, 2015: A multiscale overview of the El Reno, Oklahoma, tornadic supercell of 31 May 2013. *Wea. Forecasting*, **30**, 525–552, doi:10.1175/WAF-D-14-00152.1.

Burgess, D., and Coauthors, 2014: 20 May 2013 Moore, Oklahoma, tornado: Damage survey and analysis. *Wea. Forecasting*, **29**, 1229–1237, doi:10.1175/WAF-D-14-00039.1.

Burghardt, B. J., C. Evans, and P. J. Roebber, 2014: Assessing the predictability of convection initiation in the high plains using an object-based approach. *Wea. Forecasting*, **29**, 403–418, doi:10.1175/WAF-D-13-00089.1.

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

Clark, A. J., W. A. Gallus II, and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, doi:10.1175/2010WAF2222404.1.

——, and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, doi:10.1175/2010MWR3624.1.

——, J. S. Kain, P. T. Marsh, J. Correia Jr., M. Xue, and F. Kong, 2012: Forecasting tornado pathlengths using a three-dimensional object identification algorithm applied to convection-allowing forecasts. *Wea. Forecasting*, **27**, 1090–1113, doi:10.1175/WAF-D-11-00147.1.

——, J. Gao, P. T. Marsh, T. Smith, J. S. Kain, J. Correia Jr., M. Xue, and F. Kong, 2013: Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea. Forecasting*, **28**, 387–407, doi:10.1175/WAF-D-12-00038.1.

——, 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, doi:10.1175/WAF-D-13-00098.1.

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

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, doi:10.1175/MWR3146.1.

——, ——, ——, and J. H. Gotway, 2009: The Method for Object-based Diagnostic Evaluation (MODE) applied to WRF forecasts from the 2005 Spring Program. *Wea. Forecasting*, **24**, 1252–1267, doi:10.1175/2009WAF2222241.1.

Dawson, D. T., II, L. J. Wicker, E. R. Mansell, and R. L. Tanamachi, 2012: Impact of the environmental low-level wind profile on ensemble forecasts of the 4 May 2007 Greensburg, Kansas, tornadic storm and associated mesocyclones. *Mon. Wea. Rev.*, **140**, 696–716, doi:10.1175/MWR-D-11-00008.1.

Ebert, E. E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510, doi:10.1175/2009WAF2222251.1.

——, and W. A. Gallus Jr., 2009: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. *Wea. Forecasting*, **24**, 1401–1415, doi:10.1175/2009WAF2222252.1.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10 143–10 162, doi:10.1029/94JC00572.

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

Gao, J., and Coauthors, 2013: A real-time weather-adaptive 3DVAR analysis system for severe weather detections and warnings. *Wea. Forecasting*, **28**, 727–745, doi:10.1175/WAF-D-12-00093.1.

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

——, ——, ——, and ——, 2010: Verifying forecasts spatially. *Bull. Amer. Meteor. Soc.*, **91**, 1365–1373, doi:10.1175/2010BAMS2819.1.

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077, doi:10.1175/MWR-D-11-00356.1.

——, and ——, 2013: Object-based evaluation of a storm-scale ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment. *Mon. Wea. Rev.*, **141**, 1079–1098, doi:10.1175/MWR-D-12-00140.1.

——, ——, F. Kong, and M. Xue, 2011: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of the object-oriented cluster analysis method for precipitation fields. *Mon. Wea. Rev.*, **139**, 3673–3693, doi:10.1175/MWR-D-11-00015.1.

——, ——, ——, and ——, 2013: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts. *Mon. Wea. Rev.*, **141**, 3413–3425, doi:10.1175/MWR-D-13-00027.1.

Jones, T. A., K. Knopfmeier, D. Wheatley, G. Creager, P. Minnis, and R. Palikondo, 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, doi:10.1175/WAF-D-15-0107.1.

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, doi:10.1175/WAF2007106.1.

Keil, C., and G. C. Craig, 2007: A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Wea. Rev.*, **135**, 3248–3259, doi:10.1175/MWR3457.1.

——, and ——, 2009: A displacement and amplitude score employing an optical flow technique. *Wea. Forecasting*, **24**, 1297–1308, doi:10.1175/2009WAF2222247.1.

Lakshmanan, V., C. Karstens, J. Krause, and L. Tang, 2014: Quality control of weather radar data using polarimetric variables. *J. Atmos. Oceanic Technol.*, **31**, 1234–1249, doi:10.1175/JTECH-D-13-00073.1.

Lange, H., and G. C. Craig, 2014: The impact of data assimilation length scales on analysis and prediction of convective storms. *Mon. Wea. Rev.*, **142**, 3781–3808, doi:10.1175/MWR-D-13-00304.1.

Marzban, C., and S. Sandgathe, 2010: Optical flow for verification. *Wea. Forecasting*, **25**, 1479–1494, doi:10.1175/2010WAF2222351.1.

——, ——, H. Lyons, and N. Lederer, 2009: Three spatial verification techniques: Cluster analysis, variogram, and optical flow. *Wea. Forecasting*, **24**, 1457–1471, doi:10.1175/2009WAF2222261.1.

Meng, Z., and F. Zhang, 2011: Limited-area ensemble-based data assimilation. *Mon. Wea. Rev.*, **139**, 2025–2045, doi:10.1175/2011MWR3418.1.

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

Miller, P. A., M. F. Barth, L. A. Benjamin, R. S. Artz, and W. R. Pendergrass, 2007: MADIS support for UrbaNet. Preprints, *14th Symp. on Meteorological Observation and Instrumentation/16th Conf. on Applied Climatology*, San Antonio, TX, Amer. Meteor. Soc., JP2.5. [Available online at https://ams.confex.com/ams/pdfpapers/119116.pdf.]

Newman, J. F., V. Lakshmanan, P. L. Heinselman, M. B. Richman, and T. M. Smith, 2013: Range-correcting azimuthal shear in Doppler radar data. *Wea. Forecasting*, **28**, 194–211, doi:10.1175/WAF-D-11-00154.1.

Pinto, J. O., J. A. Grim, and M. Steiner, 2015: Assessment of the High-Resolution Rapid Refresh model's ability to predict mesoscale convective systems using object-based evaluation. *Wea. Forecasting*, **30**, 892–913, doi:10.1175/WAF-D-14-00118.1.

Potvin, C. K., and M. L. Flora, 2015: Sensitivity of idealized supercell simulations to horizontal grid spacing: Implications for Warn-on-Forecast. *Mon. Wea. Rev.*, **143**, 2998–3024, doi:10.1175/MWR-D-14-00416.1.

Putnam, B. J., M. Xue, Y. Yung, N. A. Snook, and G. Zhang, 2014: The analysis and prediction of microphysical states and polarimetric variables in a mesoscale convective system using double-moment microphysics, multinetwork radar data, and the ensemble Kalman filter. *Mon. Wea. Rev.*, **142**, 141–162, doi:10.1175/MWR-D-13-00042.1.

Schwartz, C. S., and Coauthors, 2009: Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, doi:10.1175/2009MWR2924.1.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032, doi:10.1175/MWR2830.1.

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

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

Snook, N., M. Xue, and J. 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, doi:10.1175/MWR-D-11-00117.1.

——, ——, and ——, 2015: Multiscale EnKF assimilation of radar and conventional observations and ensemble forecasting for a tornadic mesoscale convective system. *Mon. Wea. Rev.*, **143**, 1035–1057, doi:10.1175/MWR-D-13-00262.1.

Sobash, R. A., and L. J. Wicker, 2014: Ensemble forecasts of the 17 November 2013 Illinois tornado outbreak using EnKF radar and surface data assimilation. *Proc. 27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 8A.6. [Available online at https://ams.confex.com/ams/27SLS/webprogram/Paper255618.html.]

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, doi:10.1175/2009BAMS2795.1.

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

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

Van der Walt, S., and Coauthors, 2014: Scikit-image: Image processing in Python. *PeerJ*, **2**, e453, doi:10.7717/peerj.453.

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, doi:10.1175/WAF-D-15-0043.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Elsevier, 627 pp.

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, doi:10.1175/WAF-D-13-00135.1.

Yussouf, N., E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013: The ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma City tornadic supercell storms using single- and double-moment microphysics schemes. *Mon. Wea. Rev.*, **141**, 3388–3412, doi: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, doi:10.1175/MWR-D-14-00268.1.