Object-Based Verification of a Prototype Warn-on-Forecast System

PATRICK S. SKINNER,[a,b] DUSTAN M. WHEATLEY,[c] KENT H. KNOPFMEIER,[a,b]
ANTHONY E. REINHART,[a,b] JESSICA J. CHOATE,[a,b] THOMAS A. JONES,[a,b] GERALD J. CREAGER,[a,b]
DAVID C. DOWELL,[d] CURTIS R. ALEXANDER,[d] THERESE T. LADWIG,[d,e] LOUIS J. WICKER,[b]
PAMELA L. HEINSELMAN,[b] PATRICK MINNIS,[f] AND RABINDRA PALIKONDA[g]

[a] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[c] *Louisville, Kentucky*
[d] *NOAA/OAR/Earth System Research Laboratory, Boulder, Colorado*
[e] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[f] *NASA Langley Research Center, Hampton, Virginia*
[g] *Science Systems and Applications Inc., Hampton, Virginia*

(Manuscript received 9 February 2018, in final form 23 May 2018)

ABSTRACT

An object-based verification methodology for the NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e) has been developed and applied to 32 cases between December 2015 and June 2017. NEWS-e forecast objects of composite reflectivity and 30-min updraft helicity swaths are matched to corresponding reflectivity and rotation track objects in Multi-Radar Multi-Sensor system data on space and time scales typical of a National Weather Service warning. Object matching allows contingency-table-based verification statistics to be used to establish baseline performance metrics for NEWS-e thunderstorm and mesocyclone forecasts. NEWS-e critical success index (CSI) scores of reflectivity (updraft helicity) forecasts decrease from approximately 0.7 (0.4) to 0.4 (0.2) over 3 h of forecast time. CSI scores decrease through the forecast period, indicating that errors do not saturate during the 3-h forecast. Lower verification scores for rotation track forecasts are primarily a result of a high-frequency bias. Comparison of different system configurations used in 2016 and 2017 shows an increase in skill for 2017 reflectivity forecasts, attributable mainly to improvements in the forecast initial conditions. A small decrease in skill in 2017 rotation track forecasts is likely a result of sample differences between 2016 and 2017. Although large case-to-case variation is present, evidence is found that NEWS-e forecast skill improves with increasing object size and intensity.

1. Introduction

NOAA's Warn-on-Forecast (WoF) project is tasked with producing probabilistic, short-term $O(0–3)$-h guidance for thunderstorm hazards (Stensrud et al. 2009, 2013). In recent years, prototype WoF systems have demonstrated an ability to produce accurate ensemble forecasts in case studies of tornado-producing mesocyclones (e.g., Dawson et al. 2012; Yussouf et al. 2013, 2015; Supinie et al. 2017), severe hail (Snook et al. 2016; Labriola et al. 2017), and flash flooding (Yussouf et al. 2016). One system, the NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e), has provided ensemble forecasts in real time during the springs of 2016 and 2017 (Wheatley et al. 2015; Jones et al. 2016). In 2016 and 2017, NEWS-e forecasts were issued up to 17 times daily at 30-min intervals for a 750 km × 750 km domain where severe thunderstorms were expected. The large amount of forecast data produced during these real-time cases makes subjective verification, which has typically been employed for case studies, difficult and motivates the development of automated verification techniques for WoF guidance.

Automating verification of WoF guidance for thunderstorm hazards presents several challenges. First, forecasts are issued at convection-allowing scales, typically with ~3-km horizontal grid spacing, which requires the use of spatial verification methods (e.g., Gilleland et al. 2009, 2010) to avoid double penalties in point verification metrics associated with small displacement errors (Wilks 2011). Second, WoF is interested in predicting localized,

Corresponding author: Patrick S. Skinner, patrick.skinner@noaa.gov

DOI: 10.1175/WAF-D-18-0020.1

rare events occurring in convective storms. These events occur infrequently compared to quantities typically used in model verification, such as precipitation, even during widespread severe weather outbreaks (e.g., Yussouf et al. 2015). Finally, phenomena such as mesocyclones are not fully sampled by conventional observations, which requires the development of verification datasets from imperfect proxies of thunderstorm hazards (Sobash et al. 2011; Skinner et al. 2016; Sobash et al. 2016a,b; Dawson et al. 2017).

Verification techniques based on object identification and matching (e.g., Davis et al. 2006a,b; Ebert and Gallus 2009) are appealing for overcoming the challenges associated with the verification of WoF guidance. Object-based methods are designed to be applicable to noncontinuous and nontraditional features of interest (Davis et al. 2006a). Additionally, object identification and matching algorithms are adaptable to a variety of user needs. For example, objects may be matched according to user-defined total interest values (Davis et al. 2006a,b) and objects derived from different input fields can be used in verification provided they are consistently defined to isolate features of interest (Wolff et al. 2014). Finally, object-based methods provide extensive diagnostic information about forecast and observed objects, allowing specific error sources in forecasts to be quantified. These advantages have resulted in the extensive use of object-based methods for verification of convection-allowing model forecasts. Recent examples include the verification of quantitative precipitation estimates (Gallus 2010; Hitchens et al. 2012; Johnson and Wang 2012; Duda and Gallus 2013; Johnson and Wang 2013; Johnson et al. 2013; Clark et al. 2014; Schwartz et al. 2017), as well as specific features in radar (Burghardt et al. 2014; Pinto et al. 2015; Cai and Dumais 2015; Skinner et al. 2016; Sobash et al. 2016b; Burlingame et al. 2017; Jones et al. 2018), satellite (Griffin et al. 2017a,b), or damage (Clark et al. 2012, 2013; Stratman and Brewster 2017) proxies.

A final complication specific to the verification of WoF guidance is that accurate forecasts are needed on spatial and temporal scales typical of thunderstorm warning products issued by the National Weather Service. These small time and space scales limit the utility of local storm reports as a verification dataset (Sobash et al. 2011, 2016a,b) owing to errors in the timing, location, and reporting frequency of severe weather (e.g., Brooks et al. 2003; Doswell et al. 2005; Trapp et al. 2006). In contrast, proxies from Doppler radar observations can be matched to model output with minimal errors in time and space. Additionally, radar data can be used to verify WoF forecasts in real time, providing forecasters with rapidly updating measures of forecast performance. These attributes make radar proxies an attractive option for the verification of short-term forecasts of convective storm hazards (Yussouf et al. 2015; Skinner et al. 2016; Dawson et al. 2017).

This study adapts the object-based mesocyclone verification methodology developed by Skinner et al. (2016) for application to NEWS-e reflectivity and mesocyclone forecasts during 2016[1] and 2017. Verification statistics from 32 total cases are used to establish a baseline of skill for NEWS-e forecasts of general and severe thunderstorms. Beyond baseline verification statistics aggregated across all cases, forecast skill is compared for different cases, forecast initialization times, and object diagnostic properties in order to quantify system performance for differing storm modes and mesoscale environments. To the authors' knowledge, this study is the first examination of the skill of Warn-on-Forecast guidance across many cases spanning a variety of storm modes and mesoscale environments.

An object identification and matching strategy for NEWS-e general and severe thunderstorm forecasts is presented in section 2. Object-based verification metrics and diagnostic properties for 2016 and 2017 NEWS-e composite reflectivity and rotation track forecasts are presented in section 3, including comparisons between different cases, initialization times, and system configurations. Conclusions, limitations, and recommendations for future research are provided in section 4.

2. Methodology

a. *Description of the forecast dataset*

The NEWS-e is an on-demand, ensemble data assimilation and prediction system nested within the High-Resolution Rapid Refresh Ensemble (HRRRE; Dowell et al. 2016). NEWS-e comprises an ensemble of 36 WRF-ARW (Skamarock et al. 2008) members with diverse physical parameterizations (Table 1; Wheatley et al. 2015) run over a 750 km × 750 km domain with 3-km horizontal grid spacing (Fig. 1). Analyses are initialized at 1800 UTC daily with the initial and boundary conditions provided by the HRRRE (Fig. 2) and the domain location determined through collaboration with the Storm Prediction Center or as part of the Hazardous Weather Testbed Spring Forecast Experiment (Kain et al. 2003; Gallo et al. 2017). Following initialization, analyses are produced every 15 min via assimilation of satellite column-integrated liquid or ice water path (Minnis et al. 2011;

[1] A single case from 23 December 2015 is run using the 2016 system configuration and is considered part of the 2016 dataset.

TABLE 1. Physical parameterization options for NEWS-e forecast members during 2016 and 2017 (adapted from Wheatley et al. 2015, their Table 2). Planetary boundary layer (PBL) options include the Yonsei University (YSU), Mellor–Yamada–Janjić (MYJ), and Mellor–Yamada–Nakanashi–Niino (MYNN) schemes, which are paired with either Dudhia and Rapid Radiative Transfer Model (RRTM) or the Rapid Radiative Transfer Model for GCMs (RRTMG) parameterizations for shortwave and longwave radiation. All members utilize the RAP land surface parameterization. Physics options for NEWS-e analysis members 19–36 are repeated (e.g., member 19 would have the same options as member 1).

Member	PBL	Shortwave radiation	Longwave radiation
1	YSU	Dudhia	RRTM
2	YSU	RRTMG	RRTMG
3	MYJ	Dudhia	RRTM
4	MYJ	RRTMG	RRTMG
5	MYNN	Dudhia	RRTM
6	MYNN	RRTMG	RRTMG
7	YSU	Dudhia	RRTM
8	YSU	RRTMG	RRTMG
9	MYJ	Dudhia	RRTM
10	MYJ	RRTMG	RRTMG
11	MYNN	Dudhia	RRTM
12	MYNN	RRTMG	RRTMG
13	YSU	Dudhia	RRTM
14	YSU	RRTMG	RRTMG
15	MYJ	Dudhia	RRTM
16	MYJ	RRTMG	RRTMG
17	MYNN	Dudhia	RRTM
18	MYNN	RRTMG	RRTMG

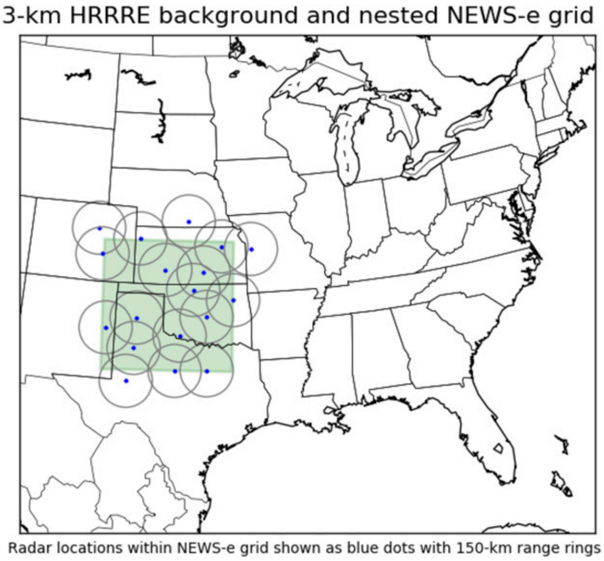

3-km HRRRE background and nested NEWS-e grid
Radar locations within NEWS-e grid shown as blue dots with 150-km range rings

FIG. 1. Example NEWS-e domain from 16 May 2017. The map shown corresponds to the HRRRE domain, with the nested NEWS-e domain shaded green. WSR-88D sites whose data are assimilated into NEWS-e are marked by blue dots with 150-km range rings drawn in gray.

Jones and Stensrud 2015; Jones et al. 2016), WSR-88D radar reflectivity and radial velocity data, and surface observations using an ensemble Kalman filter (EnKF).[2] Beginning at 1900 UTC, 18-member forecasts with a duration of 180 (90) min are issued at the top (bottom) of each hour until 0300 UTC (Fig. 2).

As both NEWS-e and HRRRE are experimental systems being actively developed, several configuration changes were introduced between 2016 and 2017 (Table 2). Differences can be divided into changes in model configuration, changes in HRRRE initial and boundary conditions, and changes in observation processing and assimilation. Model configuration changes from 2016 to 2017 include an upgrade from WRF-ARW version 3.6.1 to 3.8.1 and changing the microphysical parameterization from Thompson (Thompson et al. 2008) to the NSSL two-moment scheme (Mansell et al. 2010), which is expected to better represent storm-scale microphysical processes in supercells (Dawson et al. 2010, 2014). Changes to the HRRRE configuration include an expansion of the forecast domain (2017 version shown in Fig. 1), introduction of EnKF-based hourly assimilation of radar reflectivity data, and changes to the observation localization and posterior inflation methodologies (Ladwig et al. 2018). Additionally, 2017 initial conditions for NEWS-e were taken from a 1-h, 36-member HRRRE forecast initialized at 1700 UTC that provided each NEWS-e analysis member with a unique set of initial conditions. NEWS-e boundary conditions during 2017 were taken from a 9-member HRRRE forecast issued at 1500 UTC and repeated four times to populate the 36-member NEWS-e ensemble. In 2016, NEWS-e initial and boundary conditions were provided by a 3-h, 18-member HRRRE forecast initialized at 1500 UTC and identical initial and boundary conditions were used for 18 pairs of NEWS-e members. Ensemble spread across these member pairs was produced through diversity in the physics options (Table 1). Finally, assimilation of Automated Surface Observing Station (ASOS) observations was performed for 2017 NEWS-e cases 15 min past the top of each hour and the methodology for creating Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) superobservations of radar reflectivity data was changed from nearest-neighbor to Cressman (Cressman 1959) interpolation. Additional background and details of the NEWS-e system

[2] The specific EnKF technique is the ensemble adjustment Kalman filter (Anderson 2001) included in the Data Assimilation Research Testbed (DART; Anderson and Collins 2007; Anderson et al. 2009) software. For simplicity, the more general term EnKF is used for the remainder of this manuscript.

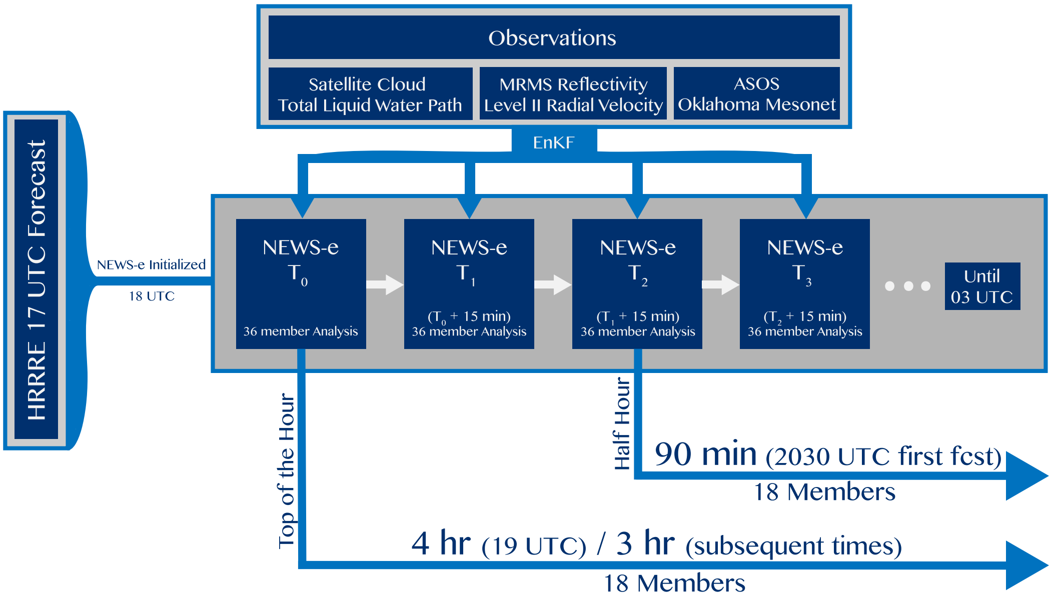


FIG. 2. Schematic of the NEWS-e system configuration for 2017.

configuration are available in Wheatley et al. (2015) and Jones et al. (2016).

NEWS-e forecasts of composite reflectivity and updraft helicity (Kain et al. 2008) in the 2–5- and 0–2-km layers above ground level (AGL) are examined in this study. These products were selected to test NEWS-e skill in forecasting all thunderstorms (composite reflectivity) and severe thunderstorms (updraft helicity). Examination of updraft helicity calculated over different vertical layers is used to determine if NEWS-e can accurately identify storms producing low-level mesocyclones, which have been found to be the best proxy for tornado occurrence (Trapp et al. 2005). Updraft helicity swaths aggregated over a 30-min period centered on each 5-min NEWS-e forecast output time are used as the final mesocyclone forecast product.

b. *Description of the verification dataset*

Verification of NEWS-e forecasts requires proxies for thunderstorm and mesocyclone occurrence to be derived from WSR-88D data. These proxies are developed using output from the MRMS system, which provides composite WSR-88D observations across the continental United States in real time (Smith et al. 2016).

As composite reflectivity observations are available through MRMS, they are an obvious choice for verification of NEWS-e composite reflectivity forecasts. However, even though the same field is available in both the forecast and verification datasets, it is not an identical, "apples to apples" comparison. Differences between the simulated and observed composite reflectivity will arise through the model microphysical parameterization, radar sampling differences, and interpolation of radar data to the model grid. As a result of these differences, simulated and observed composite reflectivities are treated as different quantities in determining thresholds used for object identification (see section 2c).

The verification dataset for mesocyclone forecasts is developed using rotation tracks derived from MRMS azimuthal wind shear data (Miller et al. 2013). Specifically, maximum range-corrected MRMS cyclonic azimuthal wind shear (Smith and Elmore 2004; Newman et al. 2013; Mahalik et al. 2016) in the 0–2- and 2–5-km layers AGL is calculated every 5 min over the NEWS-e domain. Following quality control and interpolation onto the NEWS-e grid, these azimuthal wind shear data are aggregated to produce 30-min rotation tracks for verification of NEWS-e updraft helicity swaths.

A challenge in using azimuthal wind shear rotation tracks as a verification dataset is that spurious observations for rarely occurring phenomena, such as mesocyclones, can have a large impact on verification metrics. Therefore, extensive quality control is applied to the MRMS azimuthal wind shear fields to mitigate the impact of erroneous observations. Initial quality control is applied prior to the calculation of the azimuthal wind shear, with nonmeteorological returns removed by a neural net trained using polarimetric data (Lakshmanan et al. 2014). Radial velocity data are dealiased using a modified method of Jing and Weiner (1993) that incorporates near-storm-environment soundings from the Rapid Refresh (RAP) model. MRMS azimuthal wind

TABLE 2. Changes in the NEWS-e system configuration between 2016 and 2017. Additional changes to the HRRRE configuration are discussed in section 2a.

	2016	2017
WRF-ARW version	3.6.1	3.8.1
Microphysics	Thompson	NSSL two-moment scheme
Initial conditions	3-h HRRRE 1500 UTC forecast (18 members)	1-h HRRRE 1700 UTC forecast (36 members)
Boundary conditions	HRRRE 1500 UTC forecast (18 members)	HRRRE 1500 UTC forecast (9 members)
ASOS assimilation	No	Hourly
Reflectivity superobservations	Nearest-neighbor interpolation	Cressman objective analysis

shear is then calculated only where the quality controlled reflectivity is greater than 20 dBZ and blended onto a grid with 0.01° (2016) or 0.005° (2017) latitude–longitude grid spacing. Interpolation of azimuthal wind shear data onto the NEWS-e grid is performed using a Cressman analysis scheme with a 3-km radius of influence. To be included in the objective analysis, azimuthal wind shear data must be cyclonic[3] and occur within 20 km of at least eight MRMS composite reflectivity observations greater than 45 dBZ. At least four azimuthal wind shear observations must meet these criteria for the grid box to be retained in the final analysis. The criteria for being retained in the objective analysis of azimuthal wind shear field are stricter than in past studies (Miller et al. 2013) and have been chosen to minimize spurious values in the output. Finally, regions less than 5 km or greater than 150 km from the nearest WSR-88D site are removed to mitigate range-related impacts.

c. *Object identification*

The methodology for object identification in composite reflectivity or rotation track fields is adapted from the Method for Object-Based Diagnostic Evaluation (MODE) software (Davis et al. 2006a,b) available in the Model Evaluation Tools provided by the National Center for Atmospheric Research. Thunderstorms and mesocyclones are typically sparse, contiguous maxima in both forecast and observation fields, so simple intensity thresholds are used to define object boundaries. However, defining these thresholds is complicated owing to differences in the forecast and verification fields. For example, values that best discriminate mesocyclones in azimuthal wind shear data will be different from the best discriminators in updraft helicity data. To define intensity thresholds that can consistently identify objects in different fields, we assume that a perfect forecast should produce an identical areal footprint in both the forecast and verification fields. This assumption allows percentile thresholds (e.g., Mittermaier and Roberts 2010; Dawson et al. 2017) to be used for object identification.

Percentile thresholds are determined using climatologies of forecast and verification fields (Sobash et al. 2016a). These climatologies are sensitive to changes in the system configuration, so separate climatologies are constructed for the 2016 and 2017 cases (Fig. 3). Each climatology is constructed by aggregating nonzero gridpoint values greater than the domain-wide 99th percentile from each output time a NEWS-e forecast or interpolated MRMS field is available. These extreme percentile values are used to match thresholds in the forecast and verification fields. The 99.95th percentile value is chosen as a threshold for rotation track objects, which corresponds to 2–5-km updraft helicity and azimuthal wind shear values between 50 and 65 $m^2 s^{-2}$ and between 0.0035 and 0.005 s^{-1}, respectively. These updraft helicity values are similar to intensity thresholds used for mesocyclone identification in prior studies (e.g., Kain et al. 2008; Clark et al. 2012; Dawson et al. 2017).

Despite their general similarities, clear differences in the climatologies of updraft helicity and azimuthal wind shear are present between 2016 and 2017 (Figs. 3b,c). These differences are attributable to changes in model configuration and the relatively small sample of cases, which results in different yearly distributions of storm mode and intensity (Tables 3 and 4). As updraft helicity is an integrated product of vertical velocity and vertical vorticity, it is sensitive to changes in the magnitude or alignment of the two input fields. Comparison of cases run using both Thompson and NSSL two-moment microphysics has revealed that slightly higher values of updraft helicity are produced by the NSSL two-moment scheme (not shown). However, comparison of storm mode, highest SPC risk, and reported tornadoes between 2016 and 2017 cases (Tables 3 and 4) indicates that the 2016 cases include more supercells, which are expected to be associated with higher updraft helicities.

[3] NEWS-e has produced qualitatively accurate mesoanticyclone forecasts (Jones and Nixon 2017); however, only mesocyclone forecasts are considered in this study.

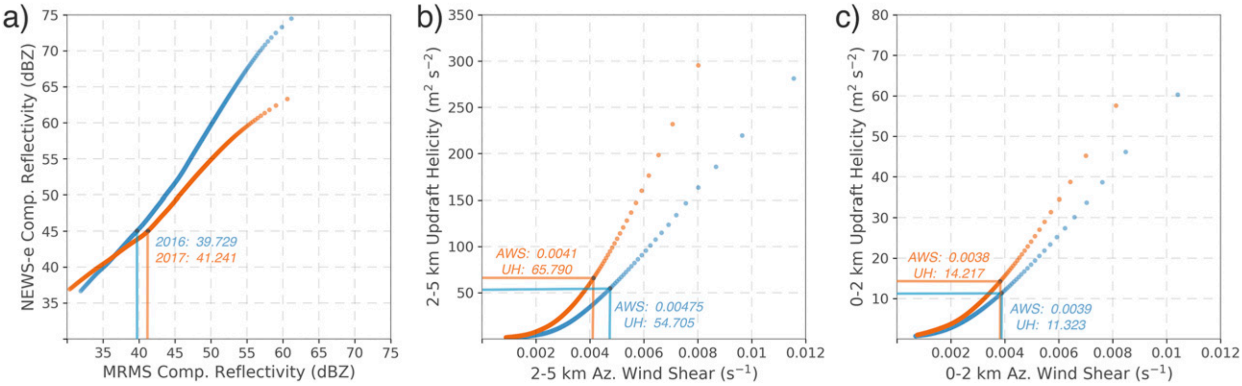


FIG. 3. Climatologies of forecast and verification datasets for the 2016 (blue) and 2017 (orange) cases. Scatterplots show the 99.1–99.98 percentile values for (a) composite reflectivity (dBZ), (b) 2–5-km updraft helicity (m^2 s^{-2}) or azimuthal wind shear (AWS; s^{-1}), and (c) 0–2-km updraft helicity or AWS. Thresholds used for object identification are annotated.

The larger proportion of supercell cases in 2016 is likely additionally reflected in a higher percentile threshold value of MRMS 2–5-km azimuthal wind shear (Fig. 3b).[4]

A composite reflectivity threshold of 45 dBZ is used for NEWS-e output for both 2016 and 2017, and the MRMS threshold is set according to the corresponding percentile (Fig. 3a). As with rotation track output, variation in the composite reflectivity climatology is apparent between 2016 and 2017. Though the MRMS climatology is slightly lower in 2017 than 2016, most of the differences between the two years are attributable to changes in NEWS-e configuration. Examination of vertical profiles of simulated reflectivity between cases run with both the Thompson and NSSL two-moment microphysics reveals that the Thompson scheme produces stronger values of simulated reflectivity above roughly 7 km (Lappin et al. 2018), resulting in much larger maximum NEWS-e composite reflectivity values in 2016 than 2017. While these differences are most pronounced for NEWS-e values above ~50 dBZ, the MRMS percentile corresponding to 45 dBZ is similar for both 2016 (99.292%) and 2017 (99.374%).

The changes in climatologies from year to year illustrate difficulties in establishing an adaptable object identification methodology for proxy variables such as composite reflectivity or rotation tracks. The large numbers of tunable parameters, from quality control of observations through object identification and matching, are a limitation of object-based verification techniques. Thresholds used in object identification and matching in this study have been determined through trial and error and have been consistently applied in order to compare between different fields and system configurations. Changes to thresholds used for object identification result in different numerical values of verification metrics, but little qualitative change in comparisons between 2016 and 2017 (see the appendix).

Prior to matching the forecast and verification objects, a final series of quality control measures are applied in order to minimize the retention of spurious objects (Fig. 4). A size threshold of 100 (144) km^2 is applied to rotation track (composite reflectivity) objects. Additionally, rotation track objects are subjected to a continuity threshold of 15 min, which requires tracks to consist of input from at least three separate times. Finally, objects with a minimum spatial displacement of less than 10 km are merged into a single object.

d. Object matching and verification

Objects in the forecast and verification fields, as well as their associated diagnostic properties, are extracted using the Scikit-image python library (Van der Walt et al. 2014). Forecast and verification objects are then matched according to a total interest score (Davis et al. 2006a,b), adapted from Skinner et al. (2016), using the centroid and minimum spatial displacement and time displacement between object pairs as inputs:

$$\mathrm{TI} = \left\{ \frac{\left[\dfrac{(\mathrm{cd}_{\max} - \mathrm{cd})}{\mathrm{cd}_{\max}}\right] + \left[\dfrac{(\mathrm{md}_{\max} - \mathrm{md})}{\mathrm{md}_{\max}}\right]}{2} \right\} \left[\frac{(t_{\max} - t)}{t_{\max}}\right], \tag{1}$$

where TI is the total interest score, cd is the centroid distance between an object pair, md is the minimum distance between an object pair, and t is the time

[4] MRMS azimuthal wind shear results were merged onto a coarser grid in 2016 than 2017; however, differences attributable to MRMS grid spacing are largely smoothed out during interpolation to the NEWS-e grid.

TABLE 3. Summary of 2016 NEWS-e cases. For each date the available forecast period, satellite data availability (DA), maximum SPC risk from the 1630 UTC outlook within the NEWS-e domain, number of SPC-archived tornado reports within the domain and forecast period, primary states affected, and predominant storm mode are provided.

Date	Forecast period (UTC)	Satellite DA	SPC outlook	No. of tornado reports	Primary states affected	Primary storm mode
23 Dec 2015	1900–0100	No	Moderate	24	AL, MS, TN	Supercell
31 Mar 2016	1900–0130	No	Enhanced	24	AL, MS, TN	Mixed
10 Apr 2016	1900–0300	No	Enhanced	0	OK, TX	Linear
29 Apr 2016	1900–2330	No	Slight	0	AL, MS	Linear
7 May 2016	1900–0300	Yes	Slight	15	CO, KS	Mixed
8 May 2016	1900–0300	Yes	Enhanced	9	KS, OK	Supercell
9 May 2016	1900–0100	Yes	Enhanced	16	AR, KS, OK	Supercell
10 May 2016	1900–0300	Yes	Enhanced	19	IL, IN, KY	Mixed
16 May 2016	1900–0300	Yes	Enhanced	10	OK, TX	Linear
17 May 2016	1900–0300	Yes	Enhanced	1	TX	Mixed
22 May 2016	1900–0300	Yes	Enhanced	38	KS, OK, TX	Supercell
23 May 2016	1900–0300	Yes	Enhanced	5	OK, TX	Supercell
24 May 2016	1900–0300	Yes	Enhanced	29	CO, KS, NE, OK	Supercell
25 May 2016	1900–0300	Yes	Slight	14	KS, OK	Supercell

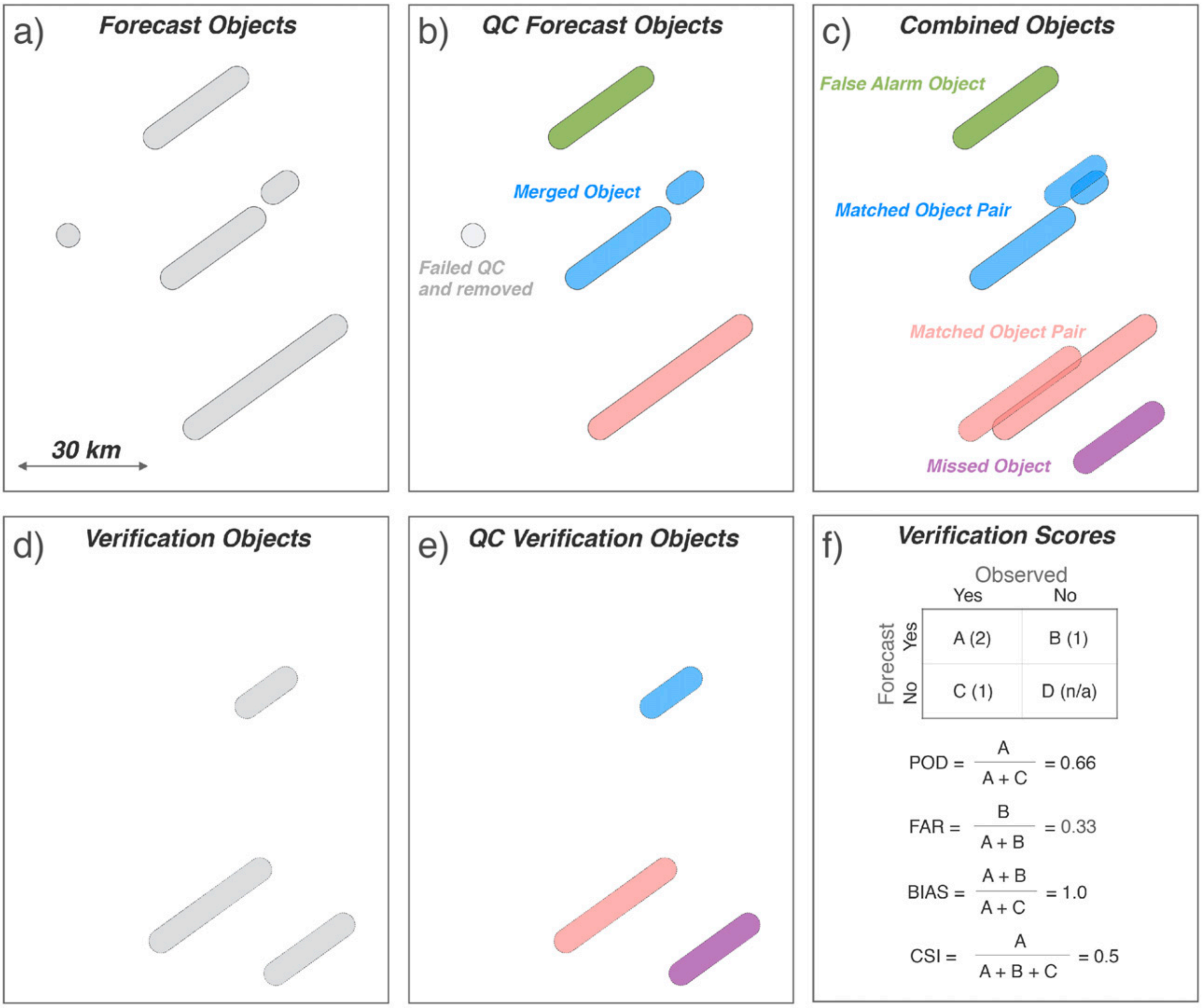
difference between an object pair. The max subscript indicates the maximum allowable threshold for object matching and is set to 40 km for the centroid and minimum distances and 25 min for time displacement. Total interest scores are calculated for each possible pair of forecast and verification objects, with matched pairs requiring a total interest score greater than 0.2, as in Skinner et al. (2016). In cases where multiple forecast objects are matched to a single verification object, only the forecast object with the highest total interest is retained as a match, while other objects are reclassified as unmatched.

Calculation of the total interest for this study uses fewer input properties than are typically used in MODE. This simplification is made possible by the generally sparse and contiguous objects in both forecast and verification fields, which allows representative object matching using a small number of input measures (Schwartz et al. 2017). The mean of the two measures of spatial displacement is used as a single input to the final total interest in order to allow matching of objects that may largely overlap but have centroid displacements greater than the allowable threshold, which often occurs for reflectivity objects associated with mesoscale

TABLE 4. As in Table 3, but for 2017 cases.

Date	Forecast period (UTC)	Satellite DA	SPC outlook	No. of tornado reports	Primary states affected	Primary storm mode
1 May	1900–0300	Yes	Enhanced	6	NY, PA	Linear
2 May	1900–0300	Yes	Slight	0	OK, TX	Supercell
3 May	1900–0300	Yes	Enhanced	2	LA, TX	Linear
4 May	1900–0300	Yes	Marginal	11	GA, SC	Mixed
8 May	1900–0300	Yes	Slight	1	CO, NM	Supercell
9 May[a]	1900–0300	Yes	Slight	6	NM, TX	Supercell
11 May	1900–0300	Yes	Enhanced	11	AR, LA, OK, TX	Mixed
15 May	1900–0300	Yes	Slight	0	CO, KS, NE	Mixed
16 May[a]	1900–0300	Yes	Moderate	26	KS, OK, TX	Supercell
17 May[a]	1900–0300	Yes	Enhanced	17	IA, IL, MN, WI	Mixed
18 May[a]	1900–0300	Yes	High	34	KS, OK, TX	Supercell
19 May	1900–0300	Yes	Enhanced	4	OK, TX	Mixed
22 May	1900–0300	Yes	Slight	0	NM, TX	Supercell
23 May[a]	1900–0300	Yes	Slight	0	TX	Mixed
25 May	1900–0300	Yes	Slight	2	CO, KS	Supercell
26 May	1900–0300	Yes	Slight	8	CO, KS	Supercell
27 May[a]	1900–0300	Yes	Moderate	8	AR, MO, OK	Mixed
30 May	1900–0300	Yes	Slight	1	MD, PA, VA	Mixed

[a] Cases reproduced using Thompson microphysics.



	Observed Yes	Observed No
Forecast Yes	A (2)	B (1)
Forecast No	C (1)	D (n/a)

$$\mathrm{POD} = \frac{A}{A+C} = 0.66$$

$$\mathrm{FAR} = \frac{B}{A+B} = 0.33$$

$$\mathrm{BIAS} = \frac{A+B}{A+C} = 1.0$$

$$\mathrm{CSI} = \frac{A}{A+B+C} = 0.5$$

FIG. 4. Schematic depicting the object matching and verification process. Initial thresholded fields from the (a) forecast from a single ensemble member and (d) observations are subjected to size and continuity quality control thresholds prior to (b),(e) object identification. (c) Forecast objects are matched to verification objects according to prescribed spatiotemporal displacement thresholds with matched pairs being considered hits, unmatched forecast objects false alarms, and unmatched verification objects misses. This classification of objects allows the (f) standard contingency-table metrics POD, FAR, BIAS, and CSI to be calculated to quantify forecast skill.

convective systems. As with object identification thresholds, verification scores are sensitive to the maximum allowable offsets in space and time, but qualitative comparisons between datasets remain similar (see the appendix).

Object matching allows matched object pairs to be classified as ''hits,'' unmatched forecast objects as ''false alarms,'' and unmatched verification objects as ''misses'' (Fig. 4). These classifications allow the contingency-table-based probability of detection (POD), false alarm ratio (FAR), frequency bias (BIAS), and critical success index (CSI) to be used to quantify the skill of NEWS-e reflectivity and mesocyclone forecasts. Given that object matching does not produce a quantity analogous to correct negatives in the contingency table, verification metrics are limited to those that consider only hits, misses, and false alarms. Additionally, missed verification objects are calculated as the residual of the number of observed objects and the number of matched forecast objects at each time step. This approach results in infrequent occurrences where observed objects are incorrectly classified owing to forecast objects matched across time steps.

A limitation to verifying NEWS-e forecasts using contingency-table-based metrics is that they provide deterministic measures of forecast quality. This deterministic verification framework does not provide a measure of skill for probabilistic guidance, which is a

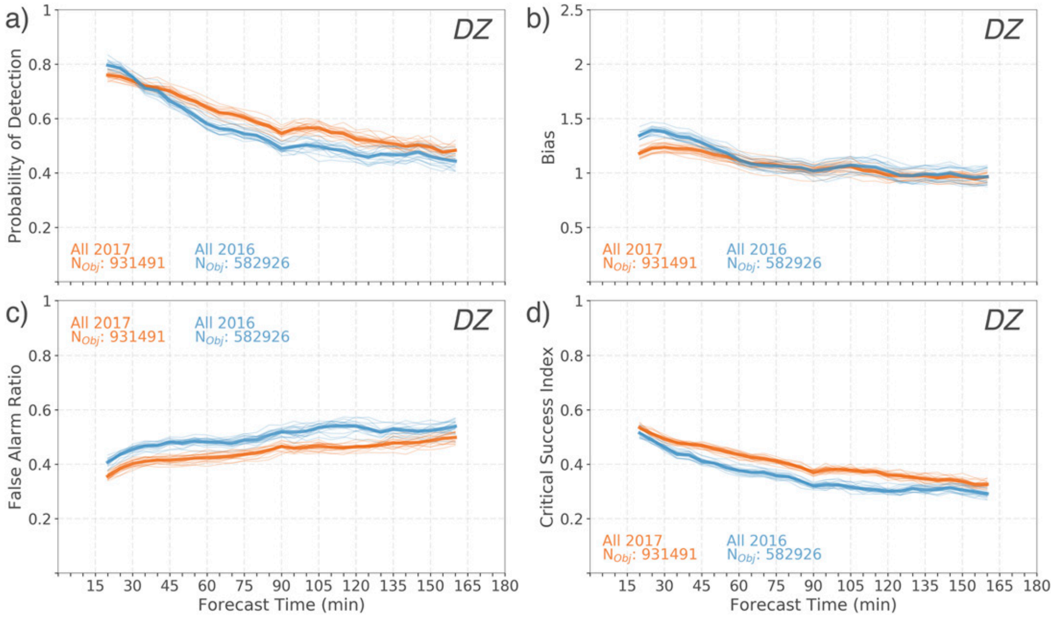


FIG. 5. Time series of object-based (a) POD, (b) BIAS, (c) FAR, and (d) CSI for composite reflectivity forecasts from 2016 (blue) and 2017 (orange). Individual ensemble members are plotted with thin lines, and the ensemble mean is shown in bold. Ensemble means are calculated as the mean of verification metrics from each ensemble member. The first and last 20 min of the forecast are masked so that only forecast times when objects could be matched in time as well as space are considered. The total number of objects from each year is annotated.

fundamental aspect of the Warn-on-Forecast project (Stensrud et al. 2009). Despite this limitation, contingency-table metrics provide familiar and intuitive measures of forecast skill that are attractive for producing an initial baseline measure of forecast quality that can be compared to probabilistic verification measures in future research.

Beyond bulk contingency-table verification measures, diagnostic features associated with objects allow specific forecast errors to be identified (Wolff et al. 2014). Specifically, object area, maximum intensity, and centroid displacement are used in this study to identify variations in forecast skill for different storm modes and intensities and potential phase and storm motion biases, respectively.

3. Object-based verification of NEWS-e forecasts

a. *Comparison of 2016 and 2017 composite reflectivity forecasts*

NEWS-e forecasts were produced for a total of 14 cases during 2016 and 18 cases during 2017 across a variety of geographic locations, storm modes, and storm environments (Tables 3 and 4). Variation in cases between years prevents direct comparison of the impacts of the NEWS-e system configuration changes on forecast skill; however, bulk verification metrics for the two years can be qualitatively compared. Baselines of NEWS-e composite reflectivity forecast skill for 2016 and 2017 have been produced by aggregating all object hits, misses, and false alarms from each case and ensemble member, then calculating the POD, FAR, BIAS, and CSI at each available forecast time (Fig. 5).

The ability of rapidly cycling assimilation of radar and satellite data to accurately initialize individual thunderstorms is evident in the verification metrics as a high POD and low FAR in the NEWS-e composite reflectivity forecasts (Figs. 5a,c). NEWS-e POD 20 min into the forecast is over 0.7 (0.8) for 2017 (2016), with corresponding false alarm ratios of approximately 0.4 for both years. The initial bulk skill, as represented by CSI, decreases with increasing forecast time, but does not level off before the end of the forecast period indicating that forecast errors do not saturate through 3 h of forecast time. The POD remains above the FAR for approximately 75 min of forecast time for both 2016 and 2017.

Despite the generally skillful composite reflectivity forecasts for both years, clear differences are apparent between 2016 and 2017 (Fig. 5). A positive bias is present during the first forecast hour for both years, but is more pronounced in the 2016 forecasts. This positive bias in 2016 forecasts results in a higher POD during the first 30 forecast minutes, but 2017 forecasts have a higher

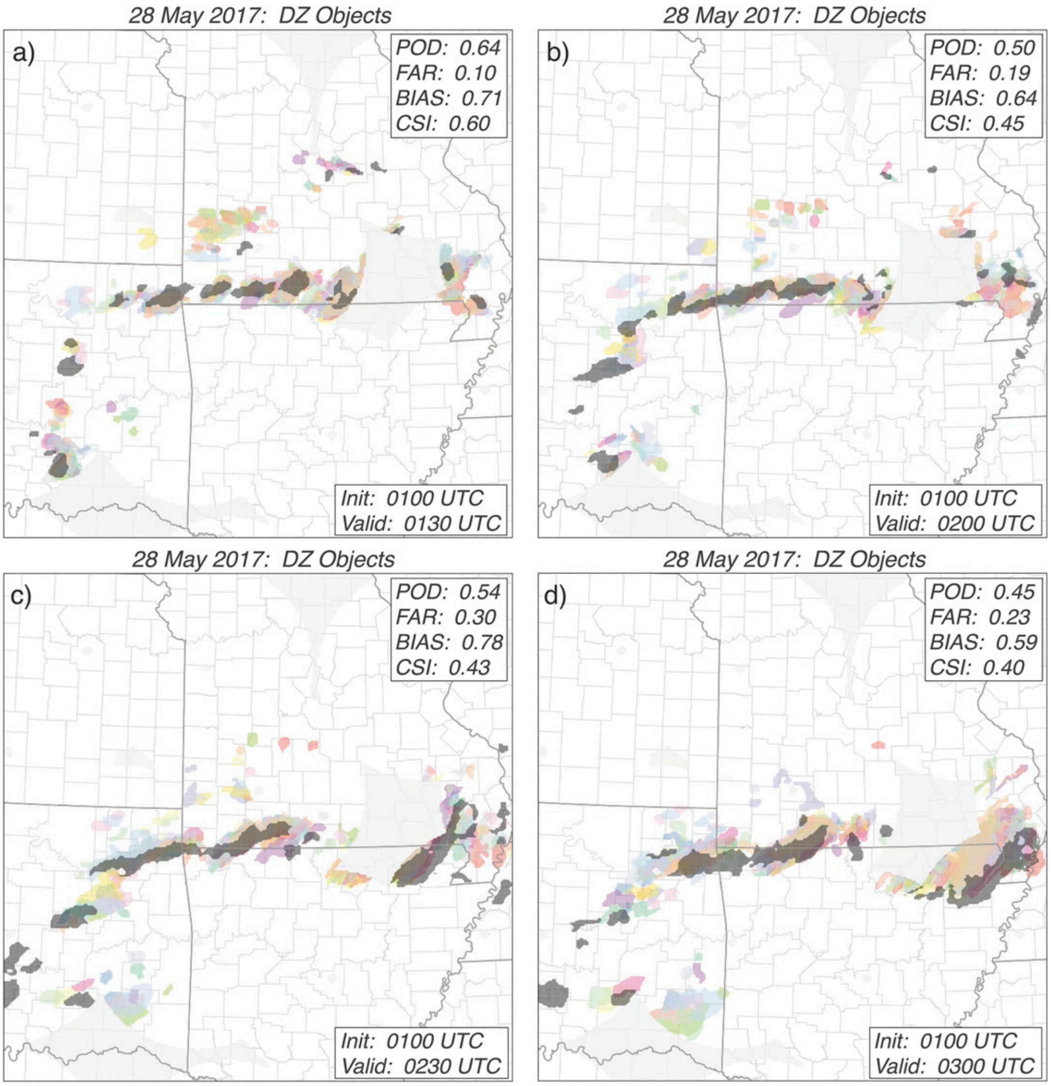


FIG. 6. Paintball plots of composite reflectivity objects (a) 30, (b) 60, (c) 90, and (d) 120 min into forecasts initialized at 0100 UTC 28 May 2017. Colored shading indicates NEWS-e member forecast objects, with different colors assigned to each ensemble member, and dark gray shading showing observed objects. Regions shaded light gray are less than 5 km or greater than 150 km from the nearest WSR-88D and are not considered in the verification. Ensemble mean POD, FAR, BIAS, and CSI scores are provided in the top right of each panel.

POD for all following times after biases between the years become similar. Furthermore, 2017 forecasts have a lower FAR through the duration of the forecast, which combined with the higher POD at later forecast times, results in higher CSI scores at all forecast times.

Examples of the composite reflectivity object distribution from a single forecast with similar CSI scores to the 2017 ensemble mean are provided in Fig. 6. These ‘‘paintball’’ plots illustrate the accuracy of a NEWS-e reflectivity forecast with CSI scores similar to the yearly mean, with most ensemble members correctly predicting the position of thunderstorms within a developing MCS along the Missouri and Arkansas border. In this example, most of the forecast error is driven by missed objects along the western extent of the domain in eastern Oklahoma. Although some ensemble members correctly predict the locations of these thunderstorms, many do not, particularly for developing convection during the second hour of the forecast (Figs. 6c,d). Several false alarm objects are also present, mainly in southern Missouri and southeastern Oklahoma; however, these false alarm objects occur in only a few ensemble members, resulting in low ensemble mean false alarm ratios. Finally, phase errors are apparent in the

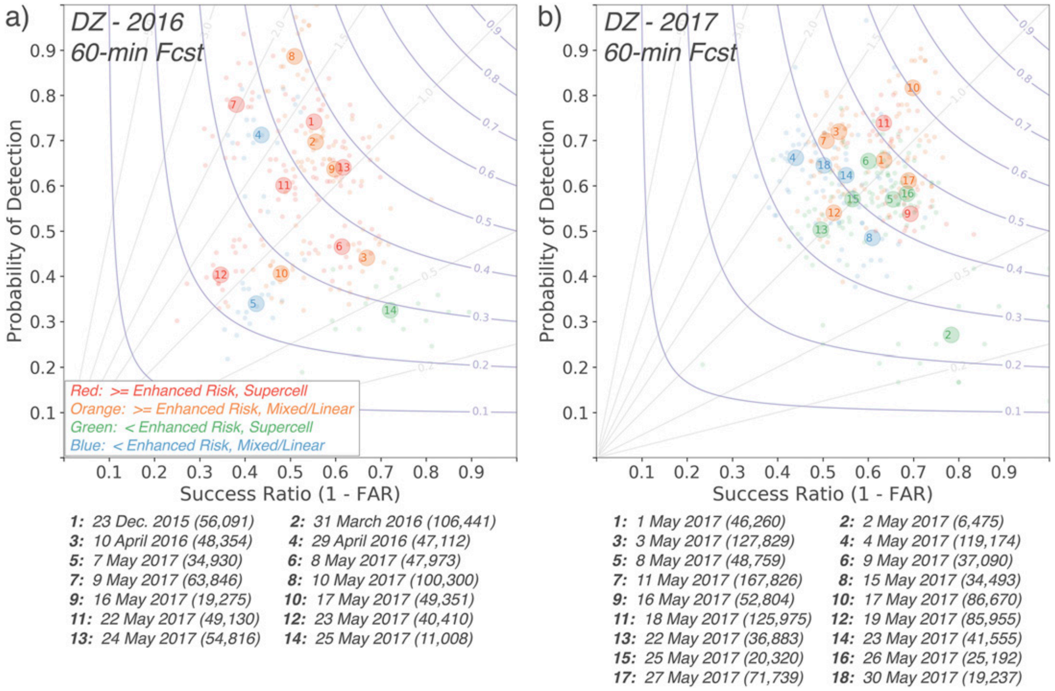


FIG. 7. Performance diagrams (Roebber 2009) for 60-min composite reflectivity forecasts from each case during (a) 2016 and (b) 2017. Diagonal (curved) lines in the diagram represent lines of constant BIAS (CSI). Small circles indicate scores of individual ensemble members, and large circles represent the ensemble mean from each case. Cases are numbered according to the legend provided below each plot and color coded according to the maximum SPC risk in the NEWS-e domain and storm mode. The total number of objects identified for each case is provided following each date in the legend.

forecast of the MCS along the eastern Missouri and Arkansas borders, with NEWS-e predictions lagging the observed evolution 2 h into the forecast (Fig. 6d). Despite these phase errors, many of the ensemble member objects are classified as matches owing to minimum and centroid distance displacements lower than the prescribed thresholds. This example was selected to illustrate what a NEWS-e forecast that produces CSI values roughly similar to the 2017 mean *can* look like. Many combinations of POD, FAR, and BIAS can produce similar CSI values and variation is observed between cases, forecasts within a single case, and within the evolution of a single forecast.[5]

Object contingency-table elements may be aggregated across each day NEWS-e was run instead of forecast output time to provide a measure of case-to-case variations in skill. These variations, as well as differences between 2016 and 2017 NEWS-e composite reflectivity forecasts, are apparent comparing performance diagrams (Roebber 2009) 60 min into the forecast of each available case (Fig. 7). With the exception of one outlier, the 2017 cases are more clustered, with ensemble mean CSI and frequency bias values between roughly 0.3 and 0.6 and between 0.75 and 1.5, respectively. The one outlier case, 2 May 2017, featured isolated storms that initiated after 0100 UTC, resulting in the fewest forecast and observed reflectivity objects from either year. In contrast, more case-to-case variation is present in the 2016 forecasts, with CSI and BIAS values of approximately 0.2–0.5 and 0.5–2.0, respectively.

Changes in NEWS-e performance for different storm modes and environments are examined by categorizing each case according to the maximum SPC 1630 UTC day 1 categorical risk within the NEWS-e domain and subjectively determined primary storm mode (Tables 3 and 4). SPC categorical risk provides an imperfect measure of environmental favorability as it is influenced by storm coverage as well as environment. However, categorical risk does provide a measure of the likelihood of severe weather for a given case, which allows NEWS-e skill to be compared for cases with limited potential for severe weather (e.g., 29 April 2016) to those with high potential (e.g., 24 May 2016). Clear stratification of composite

[5] At the time of writing, NEWS-e forecast graphics and verification statistics from each case are archived online (www.nssl.noaa.gov/projects/wof/news-e/realtime).

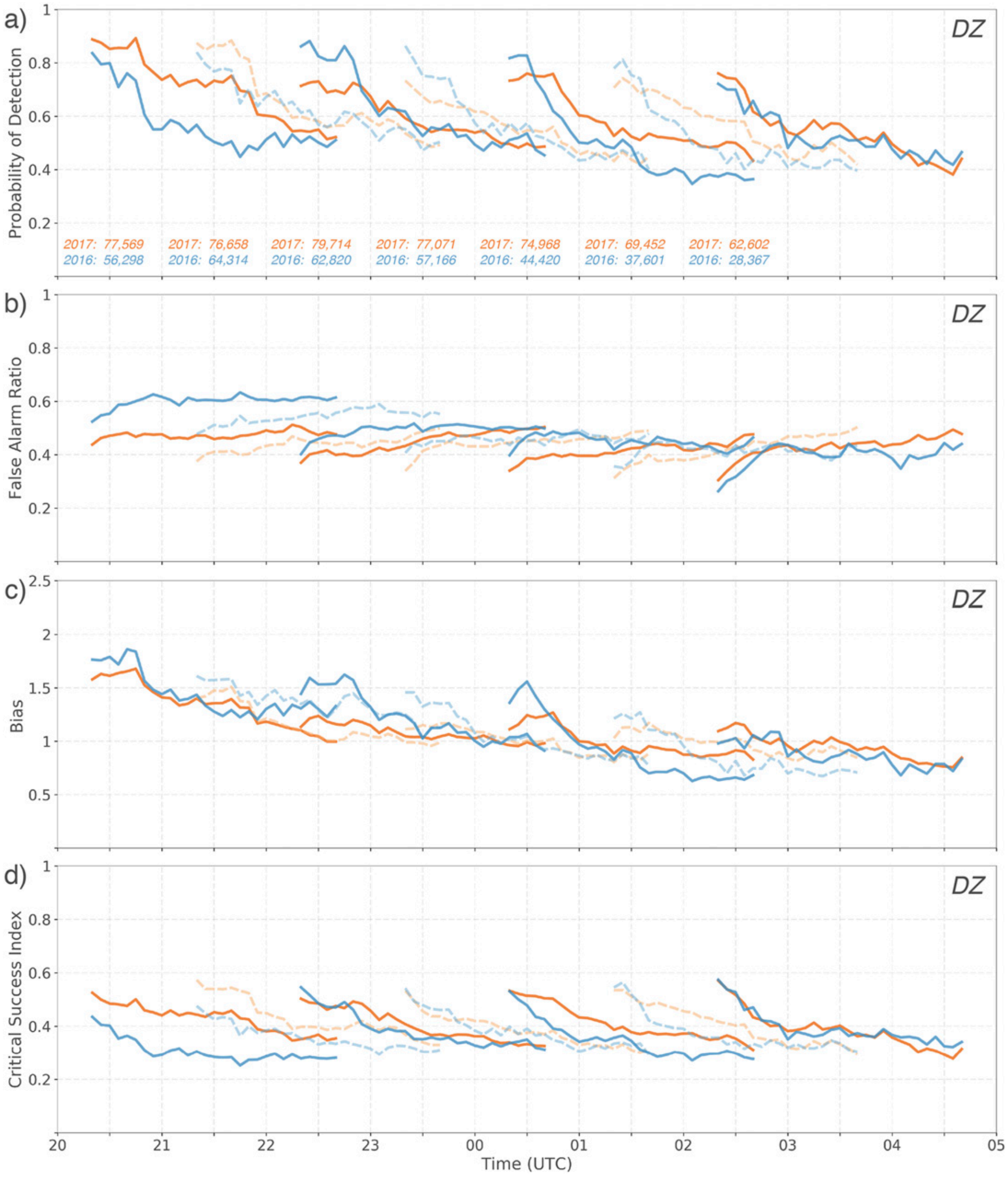


FIG. 8. Time series of the object-based ensemble mean (a) POD, (b) FAR, (c) BIAS, and (d) CSI for composite reflectivity forecasts aggregated for each forecast initialization hour between 2000 and 0200 UTC. Scores from 2017 (2016) forecasts are plotted in orange (blue), and every other forecast is plotted using lighter, dashed lines in order to improve readability. As in Fig. 5, the first and last 20 min of each forecast are masked. The total number of objects for each forecast initialization hour is annotated in (a).

reflectivity CSI scores by SPC categorical risk is not apparent in either year but subtle variation is present in 2017, where an enhanced risk or higher was present for six of the highest nine scoring cases and a slight risk or lower for seven of the lowest nine scoring cases. No clear differences in skill are apparent between cases classified as supercellular or mixed/linear storm mode in either 2016 or 2017.

Temporal variation in NEWS-e composite reflectivity forecasts is examined by aggregating objects across cases for each hourly forecast initialization time (Fig. 8). A decrease in BIAS and FAR with increasing forecast initialization time is evident in both the 2016 and 2017 cases. These decreases are coupled with a decrease in POD at later initialization times; however, this decrease is smaller than the decreases in FAR, resulting in a net increase in CSI. These changes with forecast initialization time likely arise in part through ensemble ‘‘spin up’’ of thunderstorms with cycled data assimilation. Initialization times in the late afternoon coincide with the most likely time of convection initiation (CI) and several data assimilation cycles are required to produce an accurate analysis of these thunderstorms in NEWS-e (e.g., Yussouf and Stensrud 2010). Additionally, the potential for spurious convection in NEWS-e is highest during these times owing to imbalances introduced during data assimilation or the erroneous prediction of CI. The combination of these two factors contributes to a higher BIAS and FAR during earlier initialization times and the decrease at later times indicates NEWS-e is producing a more accurate analysis of most thunderstorms within the domain. Additionally, less widespread CI during the evening is likely responsible for the slight decrease in POD with increasing forecast initialization time.

Though variation between the 2016 and 2017 verification metrics is present for all different initialization times, the largest differences are for forecasts initialized at 2000 and 2100 UTC (Fig. 8). The CSI of 2017 forecasts at these times is notably higher, at times greater than 0.1, than of the 2016 forecasts. We surmise that this improvement is likely primarily attributable to upgrades in the HRRRE between 2016 and 2017, which include the hourly ensemble assimilation of radar reflectivity observations and alterations to the observation localization and posterior inflation methodologies (Ladwig et al. 2018). These improvements provide NEWS-e forecasts with an improved set of storm and mesoscale initial conditions that translates to improved NEWS-e performance for early forecast periods.

Beyond changes in skill during earlier forecasts, 2016 composite reflectivity forecasts generally have a higher frequency bias than 2017 forecasts, particularly during the first hour of each forecast (Fig. 8). This positive bias is additionally evident in bulk (Fig. 5) and case-to-case (Fig. 7) comparisons of 2016 and 2017 forecasts and is primarily a function of the different microphysical parameterizations utilized in 2016 and 2017 (Table 2). The sensitivity of frequency bias to microphysical parameterization is demonstrated by reproducing six cases from 2017 with an identical configuration except that Thompson[6] microphysics is used in place of the NSSL two-moment scheme. Cases rerun with Thompson microphysics all exhibit higher frequency biases in 60-min composite reflectivity forecasts than those run with NSSL two-moment scheme (Fig. 9). Despite the consistent differences in frequency bias, compensating variations in POD and success ratio (1 − FAR) occur between the two sets of experiments, resulting in small and variable changes to the CSI. Composite reflectivity objects are identified in the Thompson experiments using the 2016 NEWS-e reflectivity climatology (Fig. 3). Since only cases from 2017 were compared, biases will be impacted by differences in the observed reflectivity climatology between 2016 and 2017. However, the increase in frequency bias for the Thompson runs is exacerbated if either the 2016 or 2017 climatology is applied to both sets of experiments (not shown) and the results match subjective member-by-member comparisons between the two sets of experiments, providing confidence that the two schemes produce differing biases of thunderstorm coverage.

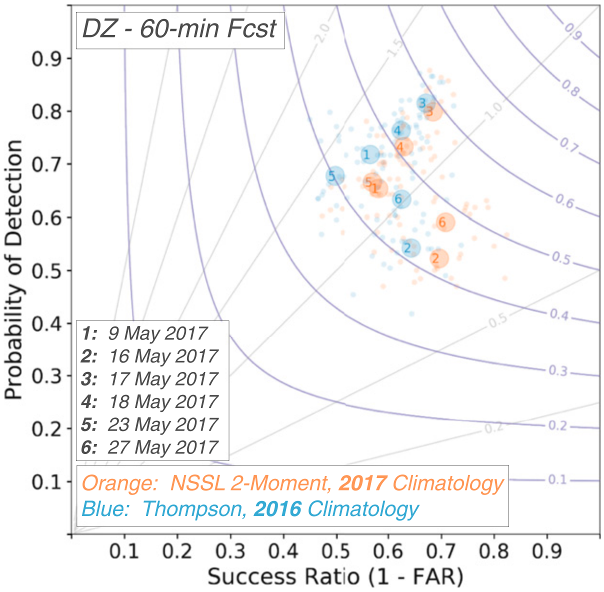


FIG. 9. As in Fig. 7, but for 60-min composite reflectivity forecasts from six cases in 2017 (orange) and the same six cases rerun using Thompson microphysics (blue). The 2016 reflectivity climatology was used to identify objects in the forecasts using Thompson microphysics.

b. Comparison of 2016 and 2017 updraft helicity forecasts

In general, object-based verification scores are lower for mesocyclone forecasts than reflectivity forecasts (Figs. 10 and 11). The CSI for NEWS-e 2–5-km updraft helicity swath forecasts decreases from approximately 0.35–0.45 to 0.2 over the course of a 3-h forecast during both 2016 and 2017, a reduction of about 0.1 from CSI scores for composite reflectivity forecasts (Fig. 5). This

[6] An updated, aerosol-aware version of the Thompson scheme (Thompson and Eidhammer 2014) was used for these experiments, which is different than the version used for the 2016 cases. The impact of the changes within the Thompson scheme on NEWS-e forecasts is not known and beyond the scope of this paper.

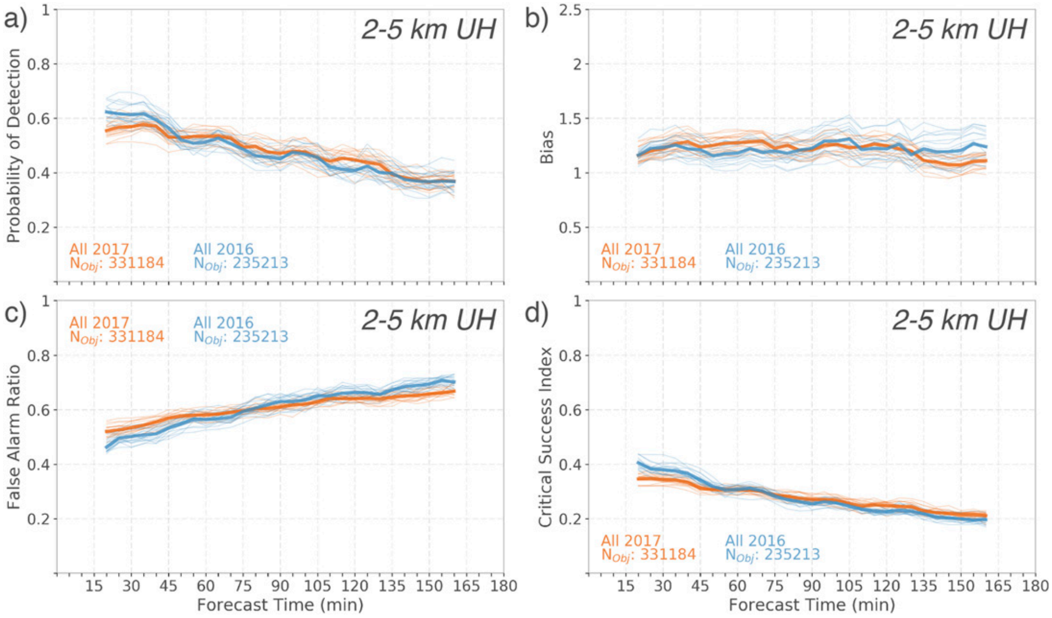


FIG. 10. As in Fig. 5, but for 2–5-km updraft helicity forecasts.

reduction in CSI is primarily driven by a higher FAR in updraft helicity forecasts and corresponds to a small positive frequency bias at all forecast times. The positive frequency bias and increased FAR for 2–5-km rotation track objects indicate that NEWS-e overpredicted midlevel mesocyclone development in thunderstorms in both 2016 and 2017, especially given nearly unbiased reflectivity forecasts following the first forecast hour (Fig. 5). Despite generally lower scores than reflectivity forecasts, the CSI of the rotation track forecasts decreases through the entirety of the 3-h forecast, suggesting that forecast errors do not saturate during the period.

Verification scores for NEWS-e 0–2-km updraft helicity forecasts are generally similar, although slightly lower, than scores for 2–5-km updraft helicity forecasts

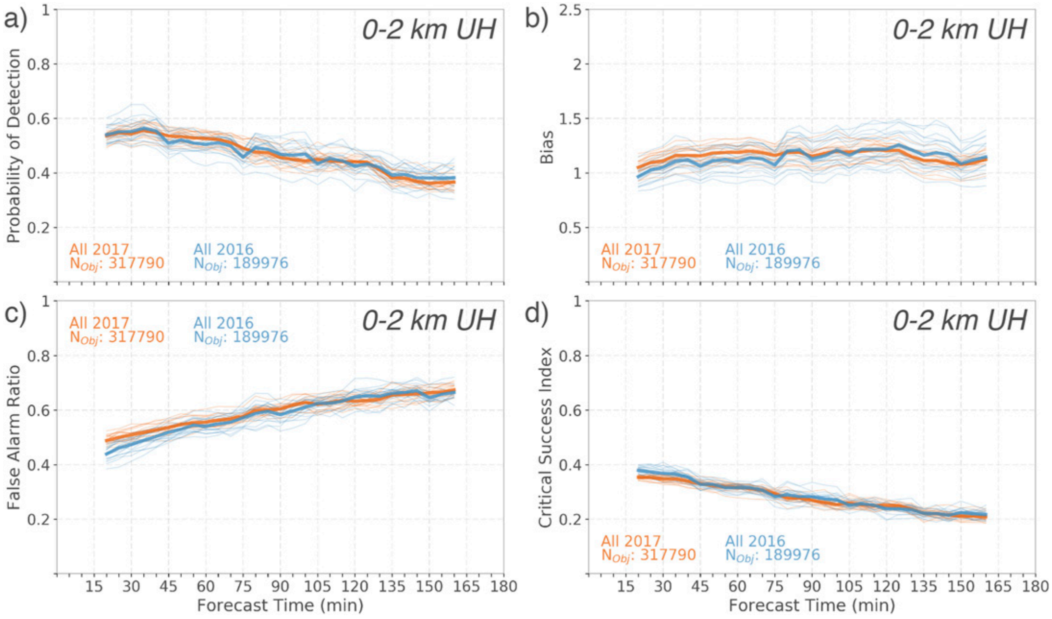


FIG. 11. As in Fig. 5, but for 0–2-km updraft helicity forecasts.

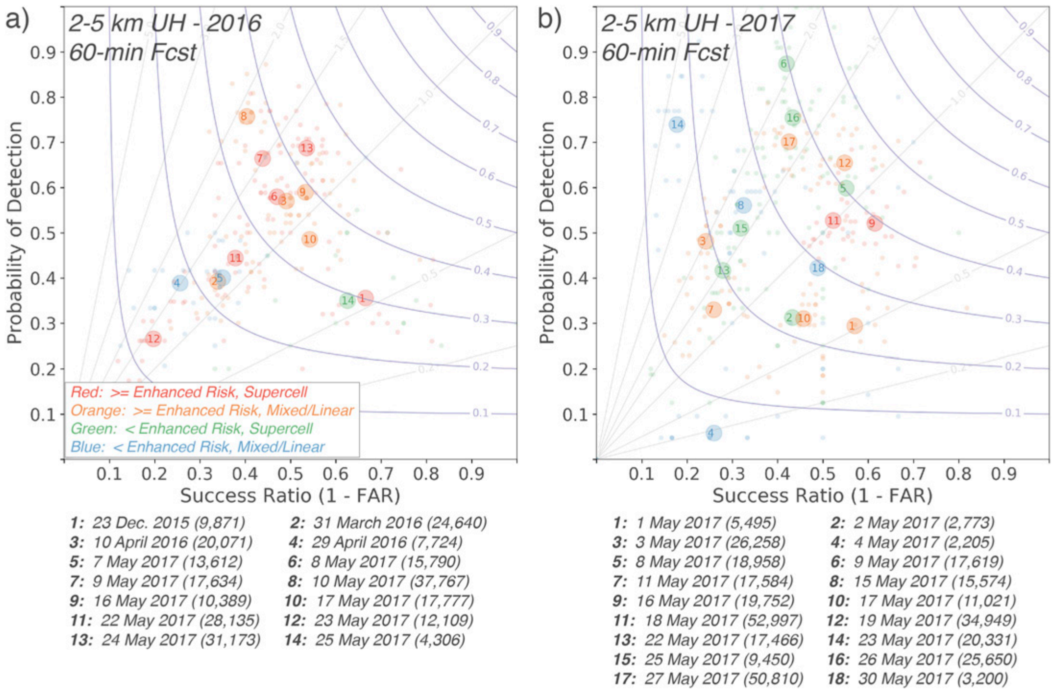


FIG. 12. As in Fig. 7, but for 60-min 2–5-km updraft helicity forecasts.

(Fig. 11). The number of rotation track objects in the 0–2-km layer is about 5% (20%) lower in 2017 (2016), resulting in a smaller overprediction bias and reduced POD and FAR. Though fewer low-level rotation track objects are identified, the strong similarities in verification scores are similar to the findings of Sobash et al. (2016a) and suggest that NEWS-e forecasts are generally not discriminating between low- and midlevel mesocyclone development. This lack of discrimination is consistent with prior studies that have found that horizontal grid spacing of 1 km or less is needed to resolve storm-scale processes responsible for low-level mesocyclogenesis (e.g., Potvin and Flora 2015).

NEWS-e 2–5-km updraft helicity forecasts performed slightly better in 2016 than 2017 during the first hour of the forecast (Fig. 10), exhibiting both a higher POD and lower FAR. However, there is large case-to-case variability in forecast performance at 60 min for both 2016 and 2017 (Fig. 12), with CSI and BIAS values ranging from less than 0.1 to greater than 0.4 and roughly 0.25 to greater than 4.0, respectively. Similarly to composite reflectivity forecasts, consistent variation of forecast skill across different storm modes or SPC categorical risks is not apparent. However, despite the lack of a strong relationship with forecast skill, there are clear differences in the case-to-case distribution of storm mode and categorical risk between 2016 and 2017, which suggests that sample differences between the two years[7] may contribute to bulk performance differences.

Comparison of 2–5-km updraft helicity swath forecast verification metrics from the six cases reproduced using the Thompson microphysics (Fig. 13) suggests variation in skill between the 2016 and 2017 forecasts is attributable to sampling differences. Changing the microphysical parameterization results in small, inconsistent changes to POD, FAR, BIAS, and CSI across the six cases. Furthermore, using the 2016 climatological threshold for object identification results in poor scores and large positive biases greater than 2.0 for all six cases, regardless of microphysical parameterization. This reduction in skill using the 2016 climatology confirms that changes in the updraft helicity climatology between 2016 and 2017 are primarily driven by differences in the observations, as opposed to changes in the composite reflectivity climatology, which are primarily driven by the microphysical parameterization (Figs. 3 and 9).

In addition to case-to-case variations in the verification scores for rotation track forecasts, some cases exhibit

[7] Though large samples of individual forecast objects are available, many of these objects will be highly correlated owing to the ensemble and high-frequency forecast output in NEWS-e. Therefore, sample diversity is better represented by the number of different cases rather than the total number of objects.

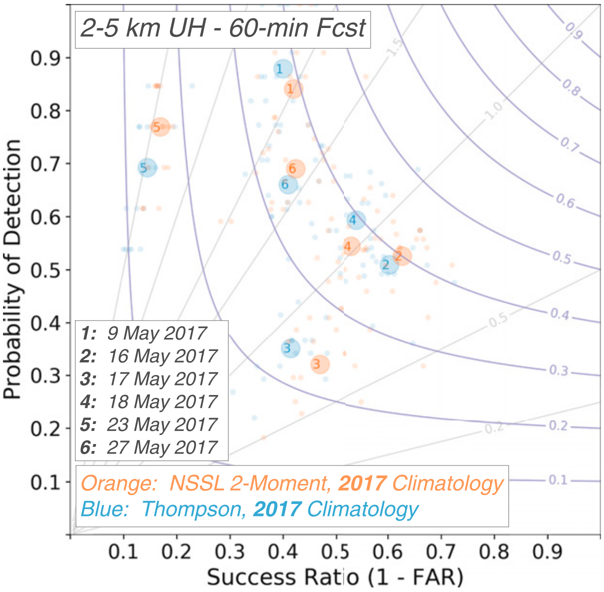


FIG. 13. As in Fig. 9, but for 2–5-km updraft helicity forecasts. Note that the 2–5-km updraft helicity climatology for 2017 is used to define the rotation track objects in both the Thompson and NSSL two-moment experiments.

large differences between the performance of composite reflectivity and 2–5-km updraft helicity forecasts (cf. Figs. 7 and 12). In these cases NEWS-e produces generally accurate predictions of composite reflectivity objects, but less skillful predictions of rotation tracks. Many cases with the largest reductions in CSI (greater than 0.2) in updraft helicity forecasts are characterized by predominantly mixed-mode or linear convection, and include 31 March 2016, 3 May 2017, 11 May 2017, 17 May 2017, and 23 May 2017. The reduced performance in rotation track forecasts in these cases is typically attributable to either underforecasts of mesocyclones embedded in mesoscale convective systems or overforecasts of mesocyclones in cellular convection. Examples of the two error sources are provided in Fig. 14, where, despite accurate composite reflectivity forecasts, most ensemble members miss mesocyclone development within an MCS in Iowa (Fig. 14b) or dramatically overpredict mesocyclone development within mixed-mode storms in Texas (Fig. 14d).

Similarly to composite reflectivity forecasts, cycled data assimilation results in a reduction of the BIAS and FAR, and an increase in CSI with later forecast initialization times in mesocyclone forecasts (Fig. 15). Differences between 2016 and 2017 are inconsistent and at times highly variable across successive forecasts. However, it appears that 2017's CSI is improved in the 2000 and 2100 UTC forecasts, though to a lesser extent than the composite reflectivity forecasts. Additionally, CSI scores for 2016 are higher during the first 30–90 min of each forecast from 2200 UTC onward, indicating the improved skill in the first hour of bulk comparisons (Fig. 10) is consistent across most initialization times. Finally, 2016 forecasts initialized at 0200 UTC perform much better than 2017 forecasts. This improvement is not present in the 0200 UTC reflectivity forecasts (Fig. 8) and the reasons for the improvement are not clear. However, 4 of the 14 cases from 2016 did not issue forecasts at 0200 UTC (Table 3), which results in far fewer rotation track objects in 2016 than 2017 and will amplify the sampling differences between the years.

c. *Comparison of object diagnostic properties between 2016 and 2017*

Variation of NEWS-e performance with storm characteristics is examined by comparing differences between the size and maximum intensity of the matched and false alarm forecast objects. Differences between these diagnostic properties are visualized using scatterplots of composite reflectivity and rotation track objects aggregated from 60-min NEWS-e forecasts (Fig. 16). Kernel density estimation (KDE) is then used similarly to the approach employed by Anderson-Frey et al. (2016) to highlight regions within the size and maximum intensity parameter space where object properties occur most often. The KDE technique implemented here applies a Gaussian kernel with a smoothing bandwidth determined from a general optimization algorithm (Scott 1992) to each point within the parameter space. Kernels for each point are summed to provide a measure of the density of points and quantify the differences between the distribution of false alarms and matched objects.

Comparison of the size and maximum intensity of NEWS-e reflectivity objects reveals that larger and more intense objects were more likely to be matched to observations in both the 2016 and 2017 forecasts (Figs. 16a,b). This result is unsurprising as larger thunderstorms will be better resolved by the 3-km grid spacing employed by NEWS-e and more Doppler radar and satellite observations will be available for assimilation, likely resulting in a more accurate ensemble analysis. In addition to the differences between the size and intensity of matched and false alarm objects, differences between the object characteristics in 2016 and 2017 are apparent. As in the model climatologies (Fig. 3), much higher maximum composite reflectivity values are produced by the Thompson microphysical parameterization, with the strongest storms exhibiting values between 70 and 76 dBZ, compared to 58–64 dBZ in NSSL two-moment forecasts. Additionally, a small secondary peak in the 2017 object maximum intensity distribution is apparent at roughly 46 dBZ (Fig. 16b). This peak is produced by

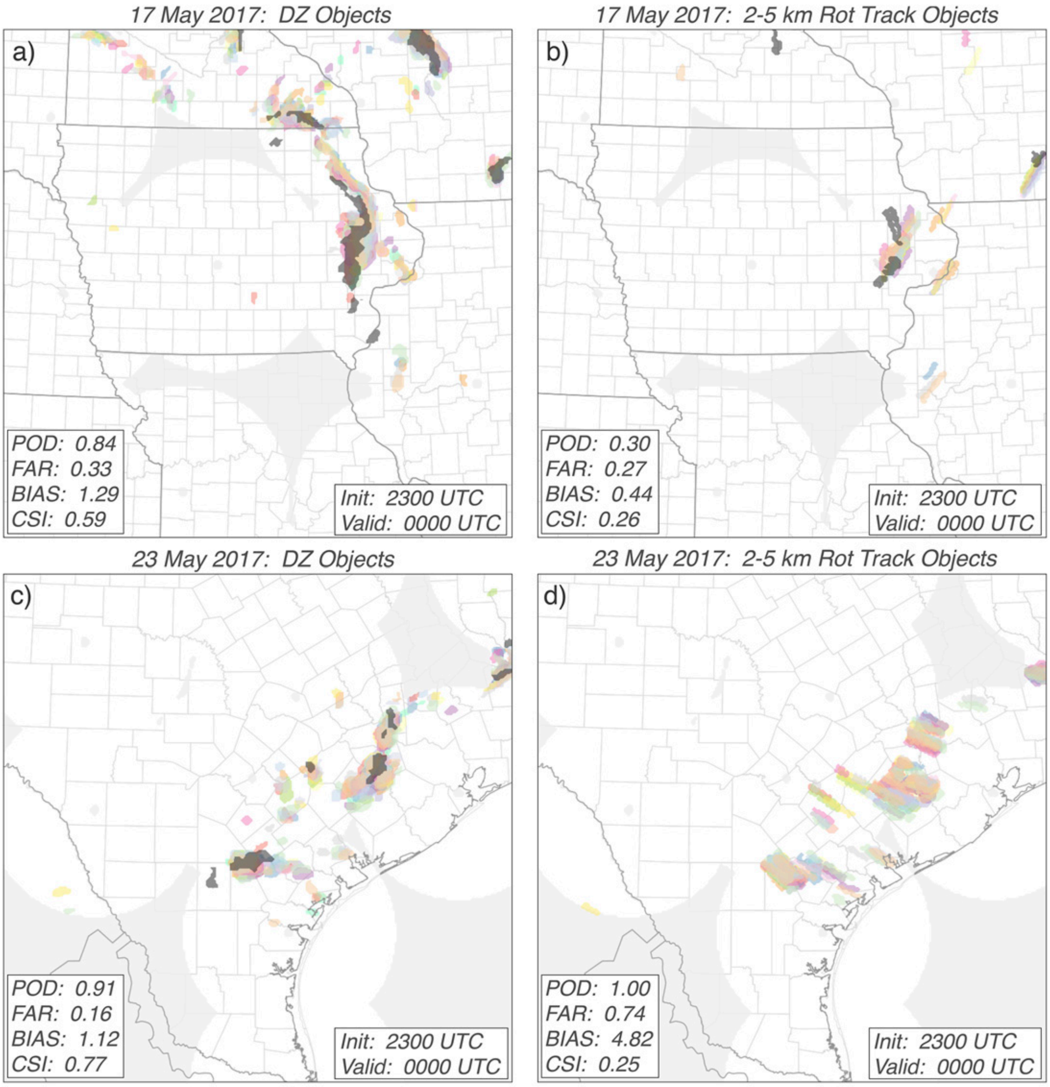


FIG. 14. As in Fig. 6, but for (left) composite reflectivity and (right) rotation track objects 60 min into the forecasts initialized at 2300 UTC (a),(b) 17 May and (c),(d) 23 May 2017. POD, FAR, BIAS, and CSI scores for each forecast are provided in the bottom left of each panel. Note that some forecast rotation track objects in (d) are matched to observed objects at different times, resulting in FAR of less than 1.0 despite no observed objects being present at the forecast time plotted.

misidentified objects within the stratiform region of mesoscale convective systems. These spurious objects represent less than 5% of the total number of reflectivity objects in the 60-min forecasts and will minimally impact the verification scores, but their presence in the NSSL two-moment forecasts provides another example of the challenges in identifying appropriate thresholds for object-based comparisons of different system configurations.

Similarly to reflectivity objects, larger and more intense rotation track objects were more likely to be matched in the 2016 forecasts (Fig. 16c), but smaller differences between the distribution of matched and false alarm objects are present in the 2017 forecasts (Fig. 16d). However, if the 2017 cases are split according to the subjectively defined primary storm mode (Table 4), supercell cases behave similarly to the 2016 forecasts, with larger and more intense objects being more likely to be matched to the observations (Fig. 16e). In addition, the ratio of matched to false alarm objects in supercell cases from 2017 is similar to the ratio from the 2016 cases. Rotation track objects from linear or mixed-mode cases in 2017 are typically larger than supercell cases and have a lower match to false alarm ratio

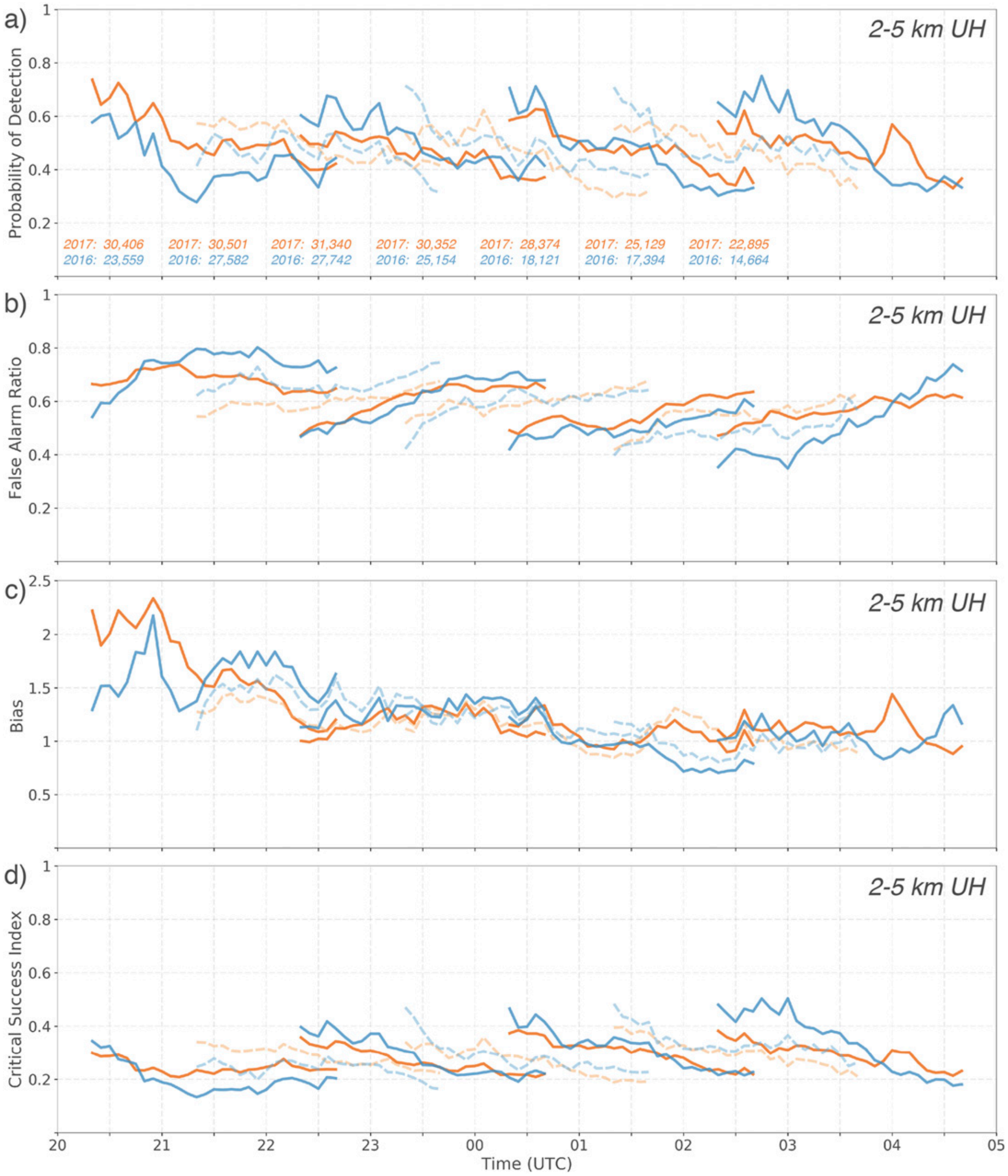


FIG. 15. As in Fig. 8, but for 2–5-km updraft helicity forecasts.

(Fig. 16f). Larger linear and mixed-mode objects likely arise through a combination of broad UH swaths along the gust front of mesoscale convective systems and faster storm motion. The apparent dependence of performance on storm mode provides further evidence that the increased skill during the first hour of the updraft helicity forecasts during 2016 is a product of sampling differences between the years rather than changes in model configuration.

Finally, centroid displacement in matched objects is examined to identify potential positive storm motion biases, which have been noted in previous prototype WoF forecasts (Yussouf et al. 2013; Wheatley et al. 2015; Yussouf et al. 2015; Skinner et al. 2016). In contrast with prior studies that found consistent, positive biases in storm speed for forecasts of discrete supercells, large variation in the centroid displacement of matched objects is present in 30-min NEWS-e forecasts of composite reflectivity and updraft helicity (Fig. 17). Much of this variation results from the inclusion of several cases with varying storm modes and coverage. Despite the larger total variation in centroid displacement, north

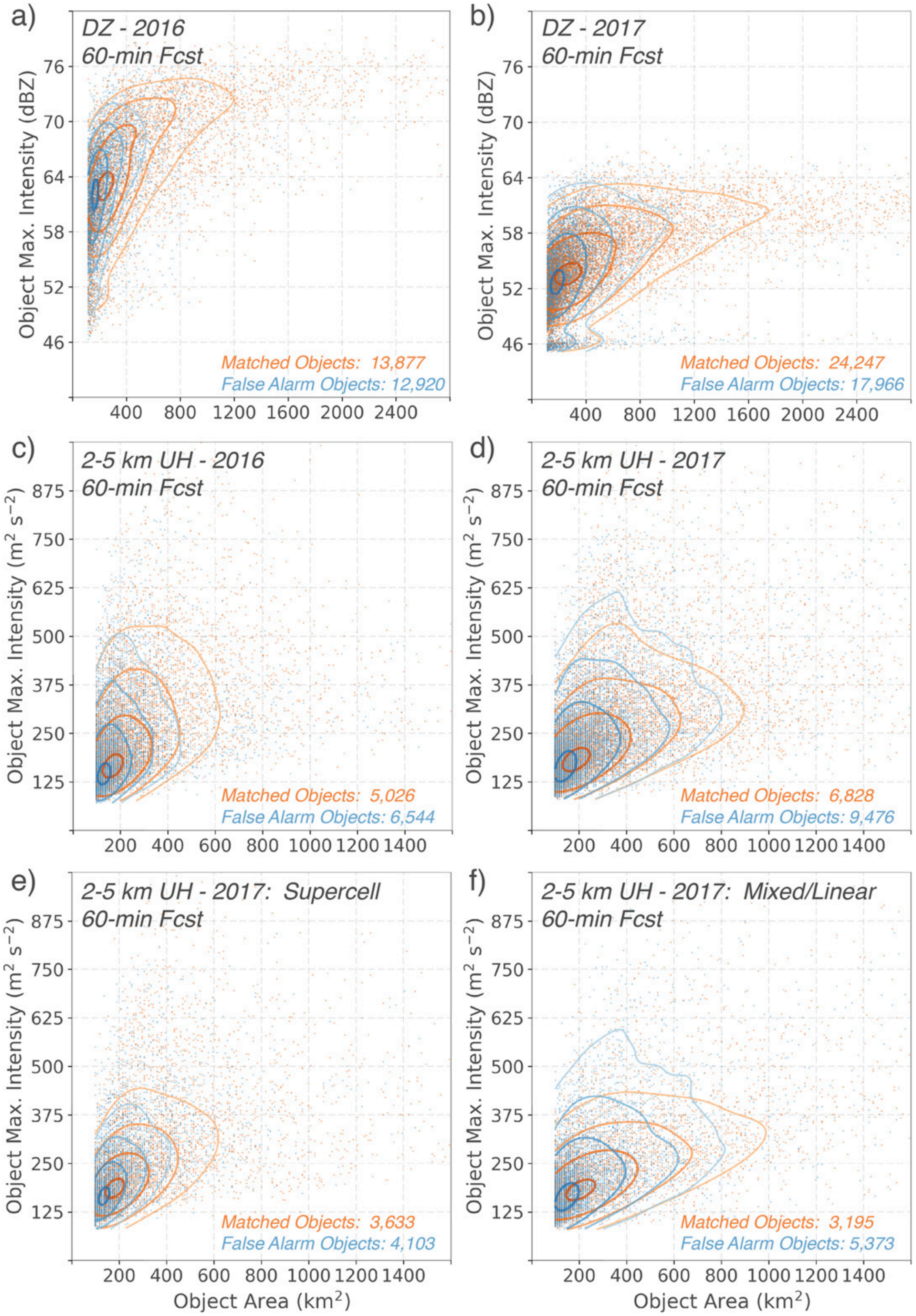


FIG. 16. Scatterplots of the parameter space of the object area and maximum intensity for 60-min NEWS-e forecasts of (a),(b) composite reflectivity (dBZ) and (c)–(f) 2–5-km updraft helicity (m^2 s^{-2}) during 2016 in (a) and (c), 2017 in (b) and (d), and 2017 cases classified as supercells in (e) or mixed/linear mode in (f). Matched objects are plotted in orange, and false alarm objects are in blue with the total number of objects in each category listed at the bottom right. KDE contours of the 95, 97.5, 99, and 99.9 percentile values of each distribution are overlain to illustrate differences between the matched and false alarm distributions. Every third reflectivity object is plotted to improve clarity.

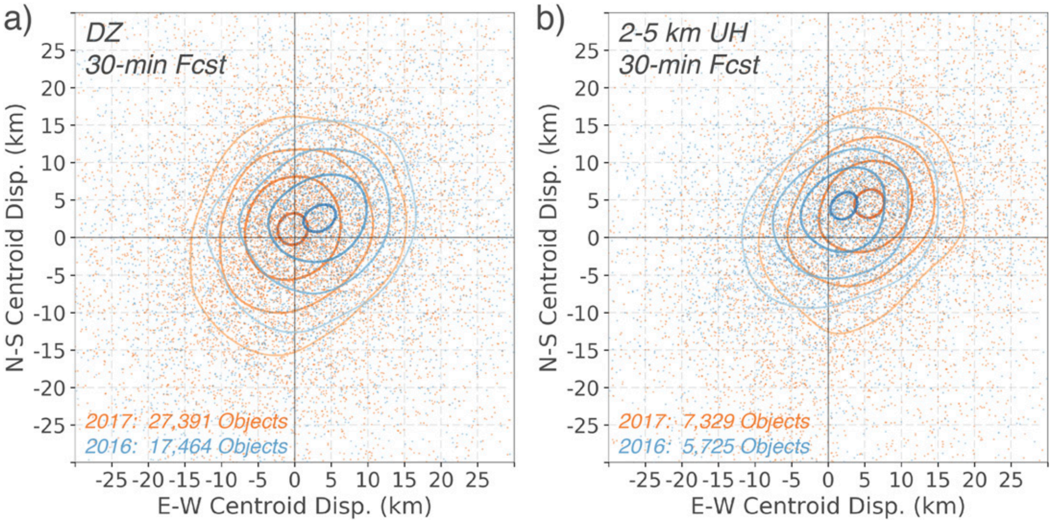


FIG. 17. Scatterplots of the east–west and north–south centroid displacements (km) of matched objects for 30-min NEWS-e forecasts of (a) composite reflectivity (dBZ) and (b) 2–5-km updraft helicity ($m^2 s^{-2}$). Objects from 2016 (2017) are plotted in blue (orange), and the total number of objects for each year is listed in the bottom left. KDE contours are overlain as in Fig. 16, and every third reflectivity object is plotted to improve clarity.

and eastward biases in centroid displacement, consistent with a positive bias in storm speed, are present in the 2016 reflectivity forecasts and updraft helicity forecasts from both 2016 and 2017. Though this apparent storm motion bias is consistent with past results and the subjective assessment of NEWS-e forecasts, centroid displacement biases can also arise through differences in simulated storm structure (Potvin et al. 2018). For example, changes to the reflectivity or rotation track object size with different physical parameterizations will induce changes to the object centroid positions and displacement from an observed object. As variation in the distribution of object sizes is noted between 2016 and 2017 for both reflectivity and rotation track objects (Fig. 14), it is unclear to what extent biases in centroid displacement are attributable to errors in storm motion or the storm and rotation track structure.

4. Conclusions and future work

An object-based strategy for verifying Warn-on-Forecast guidance has been presented and applied to 32 cases from 2016 and 2017. Composite reflectivity and updraft helicity swath forecasts from the NSSL Experimental Warn-on-Forecast System for ensembles are verified against corresponding observations in Multi-Radar Multi-Sensor products on time and space scales typical of National Weather Service warnings. Forecast and verification objects are classified as matched pairs, false alarms, and misses (Fig. 4), allowing contingency-table-based metrics to be used to establish a baseline of WoF performance for general and severe thunderstorms. Bulk verification scores from NEWS-e forecasts support the following conclusions:

- Percentile thresholds derived from model climatologies provide a method for prescribing appropriate object identification thresholds to different forecast and verification fields, for example, rotation tracks derived from predicted updraft helicity and observed azimuthal wind shear (Fig. 3).
- Cycled assimilation of Doppler radar and satellite cloud liquid water path observations every 15 min will accurately initialize individual thunderstorms within the NEWS-e domain, resulting in POD values greater than 0.7 and FAR values below 0.4 in NEWS-e 30-min forecasts of composite reflectivity (Fig. 5).
- Critical success index scores of NEWS-e composite reflectivity and updraft helicity forecasts decrease through the entirety of the 3-h forecast time, indicating that forecast errors do not saturate during the forecast period (Figs. 5 and 10).
- NEWS-e composite reflectivity forecasts are more accurate than updraft helicity forecasts, with CSI scores ~0.1 higher throughout the forecast period. This reduced performance in updraft helicity forecasts is primarily a result of overforecasting mesocyclone occurrence (Fig. 10)
- Little difference in NEWS-e forecast skill is evident when considering updraft helicity in the 0–2- or 2–5-km vertical layers (Figs. 10 and 11), indicating that NEWS-e horizontal grid spacing is too coarse to

resolve storm-scale processes responsible for the development of low-level mesocyclones.

Additionally, the following differences are observed between varying system configurations, storm modes, and storm environments:

- Improvement in composite reflectivity forecasts was noted from 2016 to 2017 and primarily driven by a lower FAR (Fig. 5). The improved performance is attributable to upgrades to the HRRRE, which provides a more accurate set of initial conditions to NEWS-e and results in more accurate early forecasts (Fig. 8) and to implementing the NSSL two-moment microphysical parameterization, which reduces a positive frequency bias during the first hour of the forecasts (Fig. 9).
- Updraft helicity forecasts during 2016 are more accurate than those in 2017 during the first hour of forecast time, with a higher POD and lower FAR (Fig. 10). Inconsistent changes in CSI for 2017 cases rerun with the Thompson microphysics (Fig. 13), and a similar skill to 2016 forecasts in 2017 cases with a primarily cellular storm mode (Fig. 16), suggest that more skillful 2016 forecasts are driven by sampling differences between the two years.
- There is tentative evidence that NEWS-e forecasts perform better for larger and more intense storms, as evidenced by larger and more intense reflectivity and rotation track objects being more likely to be matched to the observations (Fig. 16).

This study has demonstrated the utility of object-based verification for providing a bulk assessment of skill in Warn-on-Forecast guidance, comparing performance across different cases and system configurations, and providing information on specific forecast errors through the examination of object diagnostic properties. However, there are many limitations to the object-based approach for short-term, ensemble forecasts of thunderstorm hazards. Object-based verification is highly customizable, with user-defined thresholds required for object identification and matching (Davis et al. 2006a). While this flexibility permits the application of object-based verification to a wide variety of forecast problems, care must be taken to ensure that appropriate thresholds are used for consistent object identification and matching in different datasets, particularly for the verification of rare events where small differences in the number of objects identified can dramatically alter the verification scores (Fig. 4). A limitation to the contingency-table-based metrics employed here is that they only provide measures of skill for deterministic forecasts. Future work will incorporate additional metrics, such as the Brier skill score and reliability diagrams (Wilks 2011), in order to evaluate probabilistic NEWS-e guidance.

The primary limitation of object-based verification specific to this study is in the limited sample diversity across a relatively small number of available cases. Though large numbers of objects are identified, the ensemble and high-frequency nature of the NEWS-e forecasts results in a strong correlation across forecast objects, and variations in the model and observation climatologies complicate comparisons between the 2016 and 2017 forecasts (Fig. 3). We expect that more regular generation of real-time NEWS-e guidance, as is planned in 2018, will provide a larger sample size of cases and allow the baseline verification metrics presented here to be refined. Additionally, expanded computational resources will allow NEWS-e configuration testing across a large sample of prior cases, permitting hypothesis testing of forecast skill.

A final note is that while object-based verification of thunderstorm guidance can provide useful bulk measures of forecast skill, it does not discriminate between the intensities of the different thunderstorms. For example, a marginally severe supercell producing a weak rotation track object will influence the verification scores as much as an object associated with a violent tornado. Given the large numbers of thunderstorms typically present within the NEWS-e domain (e.g., Figs. 6 and 14), changes in forecast quality for the most significant storms for a given case may be masked by changes to storms that produce limited impacts on life and property. Therefore, future research will examine methods for weighting the rotation track objects by impact through incorporation of NWS warning products and local storm reports. The challenges associated with the verification of cases producing multiple storms with varying impacts underscore the importance of subjective verification for the assessment of forecast skill in individual case studies.

Acknowledgments. This research was funded by NOAA's Warn-on-Forecast project with additional funding provided by the VORTEX-SE project through Grant NA16OAR4320115. Partial funding for this research was also provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, under the U.S. Department of Commerce. This paper benefitted from thoughtful reviews by Dr. Burkely Gallo and three anonymous reviewers, which greatly improved the final draft. Drs. Adam Clark, Corey Potvin, and Nusrat Yussouf provided many helpful conversations over the course of this research, and Dr. Darrel Kingfield and Karen Cooper are thanked for their assistance with

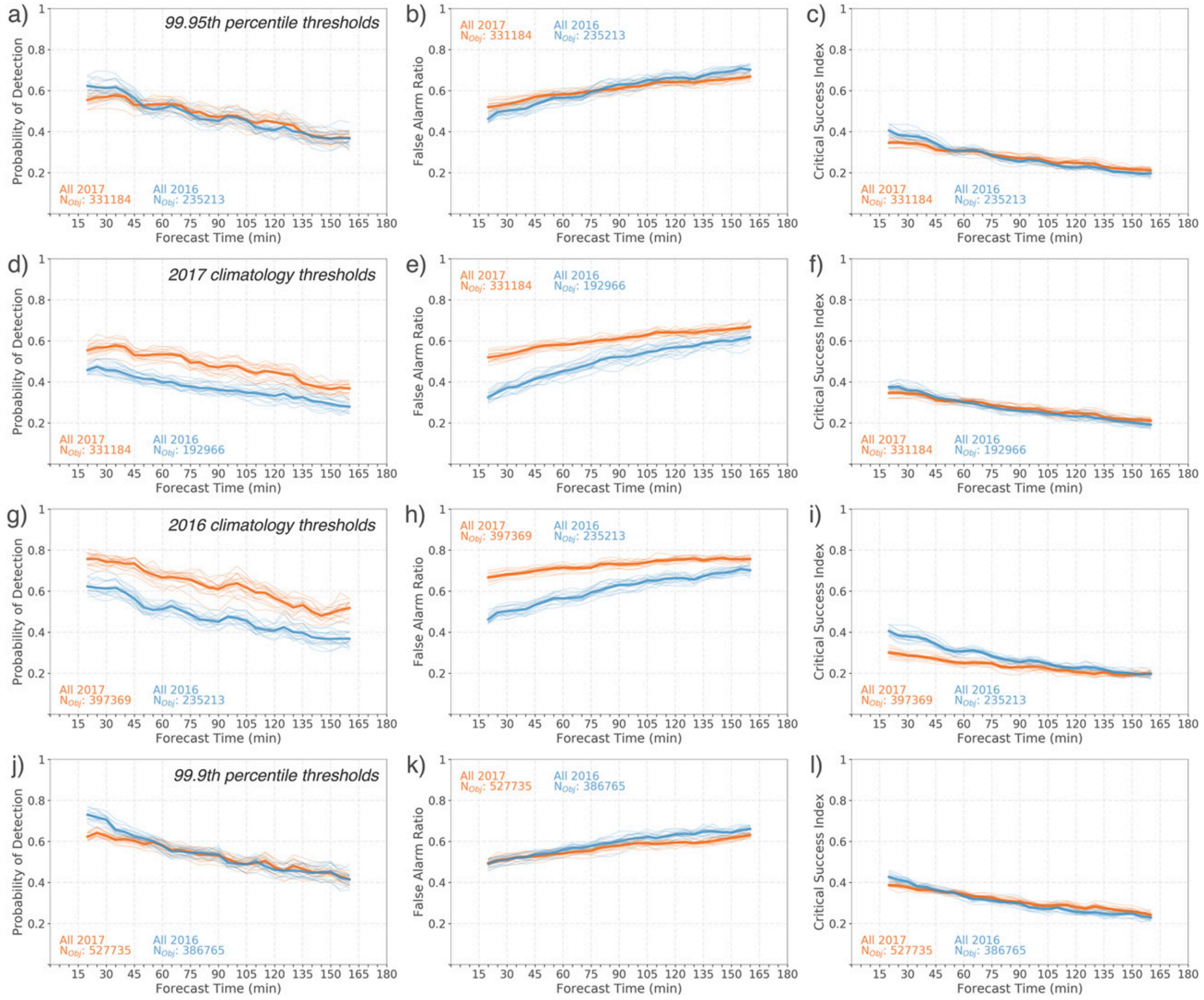


FIG. A1. Time series of (left) POD, (center) FAR, and (right) CSI for NEWS-e 2–5-km rotation track forecasts. The intensity threshold used to identify forecast and observed rotation track objects is varied between the (a)–(c) 99.95 percentile from each year's climatology (as in Fig. 7), (d)–(f) the 99.95 percentile from the 2017 climatology only, (g)–(i) the 99.95 percentile from the 2016 climatology only, and (j)–(l) the 99.9 percentile from each year's climatology. Individual ensemble member scores are plotted in thin orange (blue) lines with thick orange (blue) lines representing the ensemble mean for the 2017 (2016) NEWS-e forecasts.

MRMS processing. All analyses and visualizations were produced using the freely provided Anaconda Python distribution and SciPy, Matplotlib, basemap, netcdf4, sharppy, scikit-image, and scikit-learn libraries.

APPENDIX

Verification Score Sensitivity to Object Identification and Matching Thresholds

The highly configurable nature of object-based verification measures results in sensitivities of skill scores to user-defined thresholds. The impact of varying the user-defined intensity threshold for object identification and distance threshold for object matching is examined in Figs. A1 and A2.

Variation of the intensity threshold for object identification does result in differences in the POD and FAR, including changes in comparisons between scores for 2016 and 2017 forecasts (Fig. A1). However, the relative score changes between 2016 and 2017 are attributable to changes in the frequency bias, which produce contrasting changes in POD and FAR that generally result in little net change to the critical success index. An exception is applying the 2016 intensity threshold to the

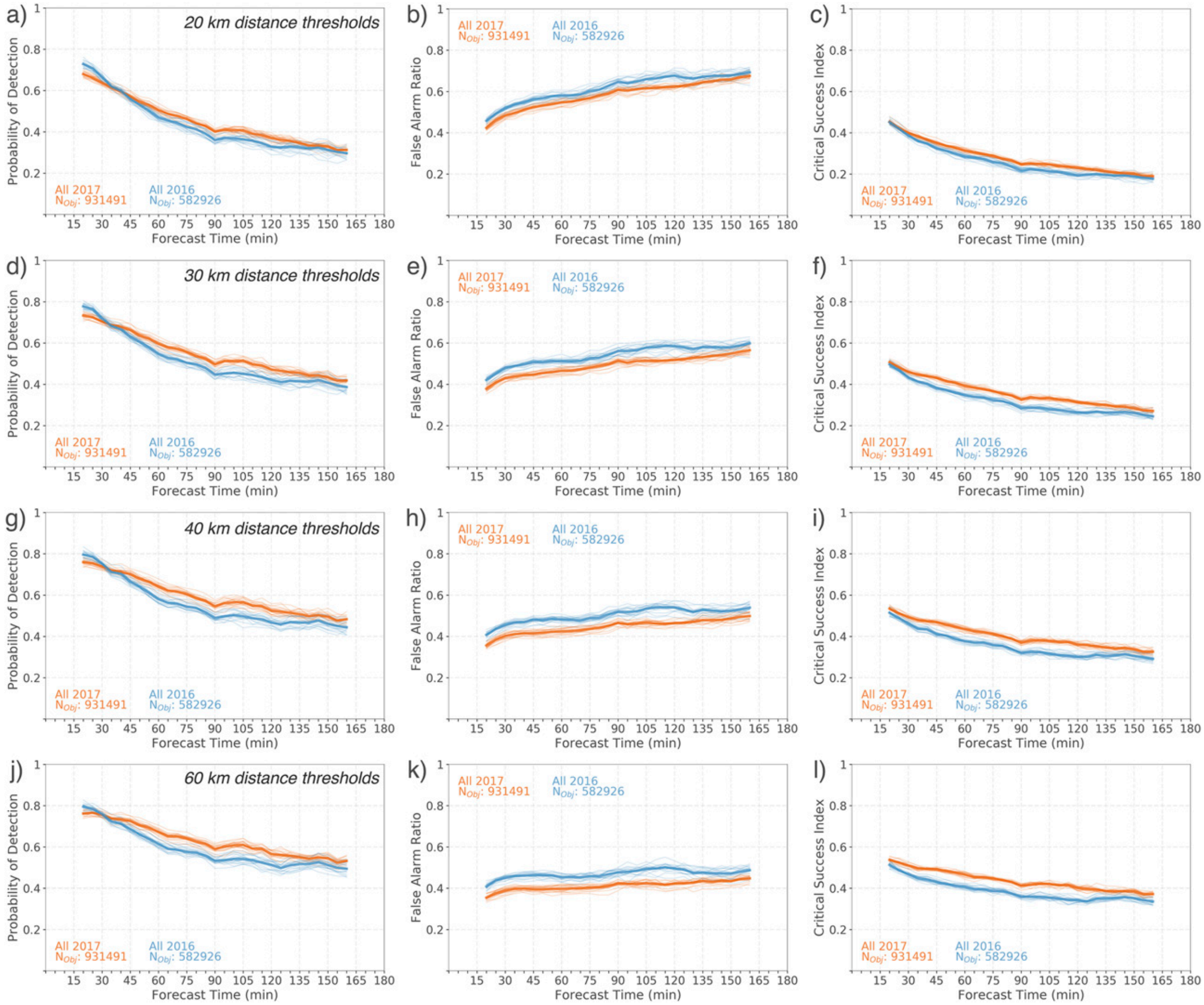


FIG. A2. As in Fig. A1, but for the composite reflectivity objects and the maximum distance threshold for object matching is varied between (a)–(c) 20, (d)–(f) 30, (g)–(i) 40 (as in Fig. 5), and (j)–(l) 60 km.

2017 forecasts (Figs. A1g–i). Using a lower value of updraft helicity for object identification results in approximately 60 000 more rotation track objects in the 2017 forecasts that are predominately false alarms, lowering the CSI scores throughout the forecast period. This sensitivity illustrates the importance of considering model climatologies to define representative object identification thresholds when comparing forecast systems with different configurations. Small changes to the percentile threshold produce little relative variation in skill scores between 2016 and 2017 (Figs. A1j–l), and composite reflectivity forecasts are relatively insensitive to changes in the object identification threshold (not shown), likely owing to small differences between the 2016 and 2017 climatologies below ~50 dBZ (Fig. 3a).

As would be expected, increasing the distance threshold for object matching results in corresponding decreases to the FAR and increases to the POD and CSI, particularly during the latter portions of the forecast period (Fig. A2). However, there is little relative change between the 2016 and 2017 forecasts in any verification metric for either composite reflectivity or rotation track forecasts.

REFERENCES

Anderson, J. L., 2001: An ensemble adjustment 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.

——, and N. Collins, 2007: Scalable implementations of ensemble filter algorithms for data assimilation. *J. Atmos. Oceanic Technol.*, **24**, 1452–1463, https://doi.org/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, https://doi.org/10.1175/2009BAMS2618.1.

Anderson-Frey, A. K., Y. P. Richardson, A. R. Dean, R. L. Thompson, and B. T. Smith, 2016: Investigation of near-storm environments for tornado events and warnings. *Wea. Forecasting*, **31**, 1771–1790, https://doi.org/10.1175/WAF-D-16-0046.1.

Brooks, H. E., C. A. Doswell III, 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.

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, https://doi.org/10.1175/WAF-D-13-00089.1.

Burlingame, B. M., C. Evans, and P. J. Roebber, 2017: The influence of PBL parameterization on the practical predictability of convection initiation during the Mesoscale Predictability Experiment (MPEX). *Wea. Forecasting*, **32**, 1161–1183, https://doi.org/10.1175/WAF-D-16-0174.1.

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

Clark, A. J., 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, https://doi.org/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, https://doi.org/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, https://doi.org/10.1175/WAF-D-13-00098.1.

Cressman, G. P., 1959: An operational objective analysis system. *Mon. Wea. Rev.*, **87**, 367–374, https://doi.org/10.1175/1520-0493(1959)087<0367:AOOAS>2.0.CO;2.

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, https://doi.org/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, https://doi.org/10.1175/MWR3146.1.

Dawson, D. T., II, M. Xue, J. A. Milbrandt, and M. K. Yau, 2010: Comparison of evaporation and cold pool development between single-moment and multimoment bulk microphysics schemes in idealized simulations of tornadic thunderstorms. *Mon. Wea. Rev.*, **138**, 1152–1171, https://doi.org/10.1175/2009MWR2956.1.

——, 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, https://doi.org/10.1175/MWR-D-11-00008.1.

——, E. R. Mansell, Y. Jung, L. J. Wicker, M. R. Kumjian, and M. Xue, 2014: Low-level ZDR signatures in supercell forward flanks: The role of size sorting and melting of hail. *J. Atmos. Sci.*, **71**, 276–299, https://doi.org/10.1175/JAS-D-13-0118.1.

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

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.

Duda, J. D., and W. A. Gallus Jr., 2013: The impact of large-scale forcing on skill of simulated convective initiation and upscale evolution with convection-allowing grid spacings in the WRF. *Wea. Forecasting*, **28**, 994–1018, https://doi.org/10.1175/WAF-D-13-00005.1.

Ebert, E. E., 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, https://doi.org/10.1175/2009WAF2222252.1.

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

Gallus, W. A., Jr., 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. Brown, and E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

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

Griffin, S. M., J. A. Otkin, C. M. Rozoff, J. M. Sieglaff, L. M. Cronce, and C. R. Alexander, 2017a: Methods for comparing simulated and observed satellite infrared brightness temperatures and what do they tell us? *Wea. Forecasting*, **32**, 5–25, https://doi.org/10.1175/WAF-D-16-0098.1.

——, ——, ——, ——, ——, ——, T. L. Jensen, and J. K. Wolff, 2017b: Seasonal analysis of cloud objects in the High-Resolution Rapid Refresh (HRRR) model using object-based verification. *J. Appl. Meteor. and Climatology*, **56**, 2317–2334, https://doi.org/10.1175/JAMC-D-17-0004.1.

Hitchens, N. M., M. E. Baldwin, and R. J. Trapp, 2012: An object-oriented characterization of extreme precipitation-producing convective systems in the midwestern United States. *Mon. Wea. Rev.*, **140**, 1356–1366, https://doi.org/10.1175/MWR-D-11-00153.1.

Jing, Z., and G. Weiner, 1993: Two-dimensional dealiasing of Doppler velocities. *J. Atmos. Oceanic Technol.*, **10**, 798–808, https://doi.org/10.1175/1520-0426(1993)010<0798:TDDODV>2.0.CO;2.

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, https://doi.org/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, https://doi.org/10.1175/MWR-D-12-00140.1.

——, ——, 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., and D. J. Stensrud, 2015: Assimilating cloud water path as a function of model cloud microphysics in an idealized simulation. *Mon. Wea. Rev.*, **143**, 2052–2081, https://doi.org/10.1175/MWR-D-14-00266.1.

——, and C. Nixon, 2017: Short-term forecasts of left-moving supercells from an experimental Warn-on-Forecast system. *J. Oper. Meteor.*, **5**, 161–170, https://doi.org/10.15191/nwajom.2017.0513.

——, 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, https://doi.org/10.1175/WAF-D-15-0107.1.

——, X. Wang, P. S. Skinner, A. Johnson, and Y. Wang, 2018: Assimilation of *GOES-13* imager clear-sky water vapor (6.5 μm) radiances into a Warn-on-Forecast system. *Mon. Wea. Rev.*, **146**, 1077–1107, https://doi.org/10.1175/MWR-D-17-0280.1.

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

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

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.

Ladwig, T. T., and Coauthors, 2018: Development of the High-Resolution Rapid Refresh Ensemble HRRRE toward an operational convective-allowing ensemble data assimilation and forecast system. *Sixth Symp. on the Weather, Water, and Climate Enterprise*, Austin, TX, Amer. Meteor. Soc., TJ1.2, https://ams.confex.com/ams/98Annual/webprogram/Paper334565.html.

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, https://doi.org/10.1175/JTECH-D-13-00073.1.

Lappin, F. M., D. M. Wheatley, K. H. Knopfmeier, and P. S. Skinner, 2018: An evaluation of changes to the NSSL Experimental Warn-on-Forecast System for ensembles (NEWS-e) in spring 2017. *22nd Conf. on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface*, Austin, TX, Amer. Meteor. Soc., 170, https://ams.confex.com/ams/98Annual/webprogram/Paper332842.html.

Mahalik, M. C., B. R. Smith, D. M. Kingfield, K. L. Ortega, T. M. Smith, and K. L. Elmore, 2016: Improving NSSL azimuthal shear calculations using an updated derivation and range-based corrections. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 182, https://ams.confex.com/ams/28SLS/webprogram/Paper301510.html.

Mansell, E. R., C. L. Ziegler, and E. C. Bruning, 2010: Simulated electrification of a small thunderstorm with two-moment bulk microphysics. *J. Atmos. Sci.*, **67**, 171–194, https://doi.org/10.1175/2009JAS2965.1.

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.

Minnis, P., and Coauthors, 2011: CERES edition-2 cloud property retrievals using TRMM VIRS and Terra and Aqua MODIS data—Part I: Algorithms. *IEEE Trans. Geosci. Remote Sens.*, **49**, 4374–4400, https://doi.org/10.1109/TGRS.2011.2144601.

Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, https://doi.org/10.1175/2009WAF2222260.1.

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, https://doi.org/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, https://doi.org/10.1175/WAF-D-14-00118.1.

Potvin, C. A., 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, https://doi.org/10.1175/MWR-D-14-00416.1.

——, J. R. Carley, A. J. Clark, L. J. Wicker, J. S. Kain, A. R. Reinhart, and P. S. Skinner, 2018: Inter-model storm-scale comparisons from the 2017 HWT Spring Forecasting Experiment. *Eighth Conf. on Transition of Research to Operations*, Austin, TX, Amer. Meteor. Soc., 3B.5, https://ams.confex.com/ams/98Annual/webprogram/Paper332705.html.

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., G. S. Romine, K. R. Fossell, R. A. Sobash, and M. L. Weisman, 2017: Toward 1-km ensemble forecasts over large domains. *Mon. Wea. Rev.*, **145**, 2943–2969, https://doi.org/10.1175/MWR-D-16-0410.1.

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

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

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

——, 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**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

——, 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: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, https://doi.org/10.1175/2009BAMS2795.1.

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

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.

Supinie, T. A., N. Yussouf, Y. Jung, M. Xue, J. Cheng, and S. Wang, 2017: Comparison of the analyses and forecasts of a tornadic supercell storm from assimilating phased-array radar and WSR-88D observations. *Wea. Forecasting*, **32**, 1379–1401, https://doi.org/10.1175/WAF-D-16-0159.1.

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.*, **71**, 3636–3658, https://doi.org/10.1175/JAS-D-13-0305.1.

——, P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

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.

Van der Walt, S., J. L. Schonberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, 2014: Scikit-image: Image processing in Python. *PeerJ*, **2**, e453, https://doi.org/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, https://doi.org/10.1175/WAF-D-15-0043.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 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, https://doi.org/10.1175/WAF-D-13-00135.1.

Yussouf, N., and D. J. Stensrud, 2010: Impact of phased-array radar observations over a short assimilation period: Observing system simulation experiments using an ensemble Kalman filter. *Mon. Wea. Rev.*, **138**, 517–538, https://doi.org/10.1175/2009MWR2925.1.

——, 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, 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.