Neighborhood- and Object-Based Probabilistic Verification of the OU MAP Ensemble Forecasts during 2017 and 2018 Hazardous Weather Testbeds

AARON JOHNSON, XUGUANG WANG, AND YONGMING WANG

School of Meteorology, University of Oklahoma, Norman, Oklahoma

ANTHONY REINHART

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

ADAM J. CLARK

NOAA/OAR/National Severe Storms Laboratory, and School of Meteorology, University of Oklahoma, Norman, Oklahoma

ISRAEL L. JIRAK

NOAA/NWS/Storm Prediction Center, Norman, Oklahoma

(Manuscript received 27 March 2019, in final form 25 November 2019)

ABSTRACT

An object-based probabilistic (OBPROB) forecasting framework is developed and applied, together with a more traditional neighborhood-based framework, to convection-permitting ensemble forecasts produced by the University of Oklahoma (OU) Multiscale data Assimilation and Predictability (MAP) laboratory during the 2017 and 2018 NOAA Hazardous Weather Testbed Spring Forecasting Experiments. Case studies from 2017 are used for parameter tuning and demonstration of methodology, while the 2018 ensemble forecasts are systematically verified. The 2017 case study demonstrates that the OBPROB forecast product can provide a unique tool to operational forecasters that includes convective-scale details such as storm mode and morphology, which are typically lost in neighborhood-based methods, while also providing quantitative ensemble probabilistic guidance about those details in a more easily interpretable format than the more commonly used paintball plots. The case study also demonstrates that objective verification metrics reveal different relative performance of the ensemble at different forecast lead times depending on the verification framework (i.e., object versus neighborhood) because of the different features emphasized by object- and neighborhood-based evaluations. Both frameworks are then used for a systematic evaluation of 26 forecasts from the spring of 2018. The OBPROB forecast verification as configured in this study shows less sensitivity to forecast lead time than the neighborhood forecasts. Both frameworks indicate a need for probabilistic calibration to improve ensemble reliability. However, lower ensemble discrimination for OBPROB than the neighborhood-based forecasts is also noted.

1. Introduction

The forecast problem that convection-allowing model (CAM) ensembles are uniquely suited to address is the determination of convective-scale details of the approximate initiation location, convective mode, and degree of upscale organization of specific convective systems. Such information can allow forecasters to anticipate the transition from a primary threat of tornado and very large hail associated with discrete supercells to a primary threat of organized straight-line winds and flash flooding that are more characteristic of mesoscale convective systems. Storm mode and upscale organization are difficult to infer from neighborhood-based products and therefore often require manual evaluation of the convective systems in each individual ensemble member. Even then, it may not be obvious to the forecaster how to quantify the relevant forecast probabilities

Corresponding author: Dr. Aaron Johnson, ajohns14@ou.edu

DOI: 10.1175/WAF-D-19-0060.1

implied by the ensemble. For similar reasons, the objective verification of CAM ensembles can also be challenging.

Traditional gridpoint-based and neighborhood-based verification methods can provide only a partial evaluation of the quality of CAM forecasts from an operational severe weather forecasting perspective (e.g., Skinner et al. 2018). Gridpoint-based metrics are highly sensitive to small spatial displacements of high amplitude features such as convective storms, even when such displacements are not relevant for the end user of the CAM forecast (Baldwin et al. 2001; Gilleland et al. 2009; Johnson et al. 2011a). Neighborhood methods (e.g., Ebert 2008) can alleviate this problem by eliminating the sensitivity to spatial displacement errors, but this can also smooth out much of the convective-scale detail that is generally assumed to be unpredictable. However, information about convective-scale details of storm morphology and convective organization, which can be influenced by the more predictable larger scales (Lilly 1990), are often specifically what the forecaster aims to obtain from the CAM. In contrast, object-based methods have been acknowledged in several past studies as one effective way to retain information about storm morphology and organization while objectively quantifying forecast attributes of interest in a way that mimics a subjective expert evaluation (e.g., Davis et al. 2006a,b, 2009; Johnson et al. 2011a,b, 2013; Wolff et al. 2014; Clark et al. 2014; Stratman and Brewster 2017). These studies were applied in the context of deterministic CAM forecasts.

The object-based framework has also been applied in the context of ensemble forecasts, treating each ensemble member as a deterministic forecast to be evaluated individually (Clark et al. 2012a; Johnson and Wang 2013; Pinto et al. 2015; Bytheway and Kummerow 2015; Skinner et al. 2016; Schwartz et al. 2017). One of the main motivations for ensemble forecasting is the ability to predict the uncertainty of the forecast, which requires a probabilistic approach to forecast evaluation. However, there have been only a few limited studies of object-based verification in a probabilistic context. For example, probability fields were used to define objects in both Gallus (2010) and Schwartz et al. (2017), but in a way that loses convective-scale morphology details because of calculating (neighborhood) probability first, before identifying the objects. Skinner et al. (2016) also used object-based methods in the context of ensemble probabilistic verification but focused on spatial probabilities of matched objects rather than forecasting the probability that an object will be matched. Ensemble forecasts have also been evaluated in a features-based framework that partitions skill into the structure, amplitude and location of convective features with each component error averaged over the ensemble, rather than evaluating probabilistic forecasts directly (Radanovics et al. 2018). Also of note, object-based probabilities were used outside of the ensemble forecasting context in Karstens et al. (2018), using radar-based nowcasting at short lead times.

For the purpose of severe weather forecasting, verification efforts must consider the convective mode of model storms (e.g., Gallus et al. 2008; Duda and Gallus 2010; Smith et al. 2012; Pettet and Johnson 2003), which has typically been evaluated subjectively (e.g., Carlberg et al. 2018). An alternative method of object-based ensemble probabilistic forecasting was introduced in Johnson and Wang (2012, hereafter JW12). In the JW12 approach, an ensemble control member is treated as a deterministic forecast within which each storm object is assigned a probability of occurring based on the similarity of objects in the other ensemble members. An advantage of this approach is that the probabilistic information content of the CAM ensemble can be objectively quantified in a way that mimics subjective evaluations based on the relevant features used by severe weather forecasters, such as storm mode and organization. The JW12 approach provides an additional perspective that ensemble verification research can use to identify CAM ensemble configuration optimizations that are most likely to directly translate into improved operational severe weather forecasts for the public. A second advantage of this approach from JW12 is that the object-based probabilistic forecast can quickly summarize information that would otherwise be obtained only by manually examining each available ensemble member. Such manual examination could be prohibitively time consuming in an operational setting, especially during rapidly evolving situations and when many CAM ensembles are available to the forecaster. Thus, the probabilistic object-based forecasts may also allow forecasters to better utilize existing numerical guidance.

Despite the apparent advantages of object-based probabilistic forecasts for evaluating ensemble guidance for severe weather, there has been limited ability to generalize broadly from previous object-based verification studies. One limitation is that object-based verification has typically relied on a large number of user-defined and tunable parameters within the object matching steps. For example, Davis et al. (2009) and Johnson and Wang (2013) both used piecewise linear functions for the similarity of each object attribute, as well as a weight and confidence value for combining object attribute similarities into an overall quantification of object similarity that was referred to as "total interest."

This limitation is addressed in the present study by simplifying the object matching procedure in a way that maintains the method's flexibility to suit different users' needs while reducing the tunable object matching parameters to just 3 key values with clear physical interpretations of storm spatial scale (object area), storm mode (e.g., linear versus cellular; aspect ratio), and location (centroid location). A second limitation of several past object-based verification studies, in terms of generalizing past work to the context of severe convective weather prediction, is a reliance on a single model variable to define and match objects. For example, in Johnson and Wang (2013), a threshold was applied to precipitation fields to identify objects and the object's attributes were defined in terms of the size, shape, and location of those precipitation objects. From a severe weather forecasting perspective, consideration of the relationships between variables (e.g., whether a particular reflectivity object is producing severe wind or hail or both) can be just as important as consideration of each variable in isolation. Although for some applications additional levels of complexity may be worthwhile (e.g., object merging as in Davis et al. 2009), in the present study objects are identified using composite reflectivity while other relevant variables are also used as object attributes that allow for emphasis on storms representing specific severe weather hazards.

The Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL) host an annual Spring Forecasting Experiment (SFE) in the Hazardous Weather Testbed (HWT) to use and evaluate innovative severe weather forecasting technologies and techniques (Clark et al. 2012b;, 2018; Gallo et al. 2017, 2018). During the 2017 and 2018 SFEs the University of Oklahoma Multiscale data Assimilation and Predictability (MAP) laboratory contributed a real time CAM ensemble initialized by an ensemble of analyses produced by the coupling of the GSI-based ensemble–variational (EnVar) and ensemble Kalman filter (EnKF) hybrid data assimilation (DA) system. Compared to other GSI-based EnVar hybrid systems applied to meso- or convective scales (e.g., Schwartz and Liu 2014; Hu et al. 2017), this hybrid DA system assimilates observations resolving a variety of scales including the capability of direct assimilation of both the radar radial velocity and radar reflectivity (Johnson et al. 2015; Wang and Wang 2017). More details of the system adopted by the MAP laboratory can be found in Duda et al. (2019) and Wang et al. (2019, manuscript submitted to *Wea. Forecasting*). Starting in 2018, the system also included an object-based probabilistic forecast interface, which was used and evaluated during the daily simulated operational forecasting activities in the HWT. This study therefore focuses on precipitation and other forecast variables related to severe weather that are of interest to SPC and HWT forecasters.

The purpose of this paper is to further develop the probabilistic object-based ensemble verification method of JW12 with an emphasis on contrasting the interpretation of the object-based and neighborhood-based probabilistic framework, then document the performance of the 2018 OU MAP ensemble forecasts during HWT using both neighborhood and object-based methods. The object-based method and parameters are developed and demonstrated using independent data from the 2017 OU MAP ensemble forecasts. Since several of the most high-impact cases from 2017 were used to develop the method and choose the object identification and matching parameters, only the 2018 data is used for the systematic verification to avoid overfitting the verification method to our subjective evaluations on 2017 cases.

The organization of the paper is as follows. The design of the ensemble DA and forecast system are presented in section 2. In section 3, we further develop the object-based probabilistic forecast framework of JW12 to better represent the distribution of convective storm morphologies in an ensemble forecast. The object-based, and a neighborhood-based, evaluation are then contrasted for a case study in section 4 in order to highlight the different aspects of forecast performance that each is sensitive to. Finally, the performance of the OU MAP ensemble during the 2018 HWT is then documented in section 5 using both the object and neighborhood-based frameworks. Section 6 contains the summary and conclusions.

2. Ensemble forecast and DA system

A similar strategy of DA and ensemble forecast for producing the OU MAP ensemble was adopted during the 2017 and 2018 HWT SFEs. Cycled WRF forecasts and 3D-EnVar DA were conducted from 1800 to 0000 UTC each day, with the first DA update occurring at 1900 UTC after a 1-h ensemble forecast, initialized at 1800 UTC. A GSI-based hybrid EnVar DA system interfaced with the Advanced Research version of the Weather Research and Forecasting (WRF-ARW; Skamarock et al. 2008) Model was employed to assimilate all conventional observations in the operational prepbufr stream, except for precipitable water (i.e., METARs, shipborne, buoy, aircraft, radiosonde, profiler), hourly from 1900 to 0000 UTC. This GSI-based hybrid DA system has also been further developed with the capability to directly assimilate radar

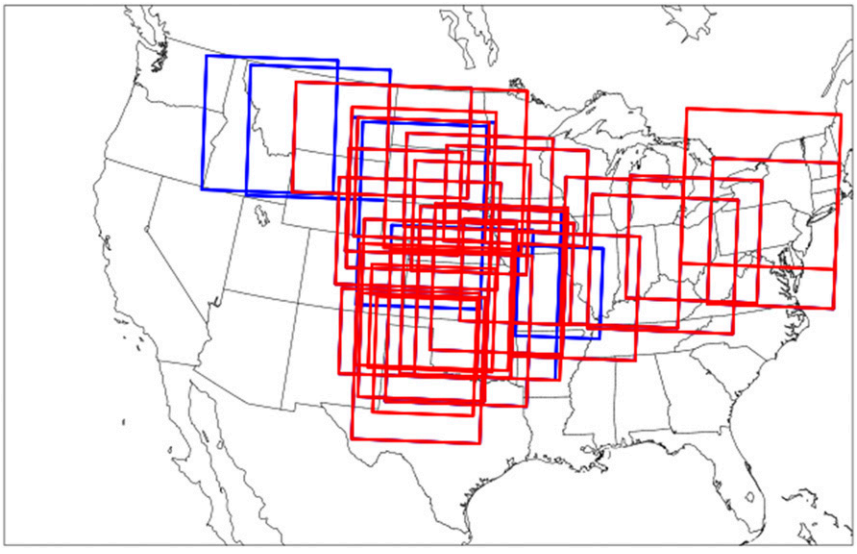

FIG. 1. Computational domain overlaid with daily verification domains for early (blue) and late (red) lead times. For many cases the red box for one day is exactly on top of the blue box for the next day. The domains differ in cases where there was a day in between forecast cases without a forecast being run (i.e., weekends).

radial velocity and reflectivity (Johnson et al. 2015; Wang and Wang 2017). The GSI EnVar hybrid solvers for the system are described in Wang (2010). The same GSI EnVar method is used for the global system (e.g., Wang et al. 2013) and regional system (e.g., Schwartz and Liu 2014; Lu et al. 2017a, b; Hu et al. 2017). Radar data are here assimilated every 20 min during the 2300–0000 UTC period. The system produces 41 distinct analyses, consisting of one analysis from the EnVar component of the hybrid DA system, and 40 perturbed member analyses from the ensemble square root filter (EnSRF) component. The GFS deterministic analysis and subsequent forecast were used to provide the initial conditions and lateral boundary conditions (ICs/LBCs) for the EnVar deterministic forecast. The ICs/LBCs for the 40-member ensemble forecasts were produced by adding to the GFS deterministic analysis 40 perturbations, which were extracted from the Global Ensemble Forecast System (GEFS) and the Short-Range Ensemble Forecast (SREF) of the National Centers for Environmental Prediction (NCEP). The GEFS/SREF perturbations are added to the deterministic GFS analysis at 1800 UTC, before running the first set of background forecasts. After each EnVar DA cycle, and before the DA begins, the analyzed perturbations were recentered around the EnVar analysis to reset the EnKF ensemble around the EnVar analysis. The recentering procedure consists of replacing each member's analysis with its difference from the ensemble mean analysis, added to the EnVar analysis. Beginning at 0000 UTC, a 36-h, 10-member ensemble forecast was initialized using the final analysis from the EnVar and the first 9 EnKF members.

The forecast domain approximately covers the contiguous United States (Fig. 1) using a 3-km grid and was advanced between DA cycles and during the free forecast using different models in 2017 and 2018 HWT. The Nonhydrostatic Multiscale Model on the B grid (NMMB; Janjić 2004) was used in 2017 and the WRF ARW was used in 2018. All members in each year adopted their own same physics parameterizations. Different models were used because the 2017 HWT ensemble was intended to evaluate developments for the NMMB-based North American Model Rapid Refresh (NAMRR), while the 2018 HWT ensemble was intended to evaluate developments for the ARW-based High-Resolution Rapid Refresh (HRRE). The NMMB in 2017 used Ferrier–Aligo microphysics (Aligo et al. 2018), the Mellor–Yamada–Janjić (MYJ; Janjić 2002) planetary boundary layer (PBL) scheme, and the Noah land surface model (LSM, Mitchell et al. 2005). The physics schemes in 2018 for WRF-ARW were Thompson microphysics (Thompson et al. 2004, 2008), the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004, 2006) PBL scheme, and the Noah LSM (Mitchell et al. 2005). A web-based interface to OBPROB (and other) products from the OU MAP ensemble is provided to HWT participants in real time during the Spring Forecasting

Experiments (e.g., http://weather.ou.edu/~map/dev/obprob_2019.php).

3. Description of the OBPROB forecast method

Before describing the object-based probabilistic (OBPROB) method in detail, it is necessary to distinguish the object-based framework from the gridpoint-based framework that many readers will be more familiar with. In a gridpoint framework, the model forecast represents a set of boxes that are spatially distributed in a predictable manner (i.e., on a regular grid, usually) and each grid point in this set contains values with some physical interpretation (e.g., the average temperature within the grid box, the average *u* component of wind within the grid box, etc.). In the object-based framework the model forecast represents a set of objects (i.e., convective systems in our case) that is different in each forecast, unlike the set of grid boxes which represent the same thing in every forecast. Each object in this set also contains values with some physical interpretation (e.g., the size of the object, the location of the object, the intensity of the object based on some metric, etc.). It has been argued above that a key advantage of CAMs over coarser resolution models is their ability to be subjectively interpreted in terms of how they represent distinct meteorological features (i.e., objects), regardless of exactly which grid box they fall into. While the completely different framework for discretizing the forecast atmospheric states creates some challenges from an objective verification perspective, it also creates an opportunity to evaluate aspects of CAM ensemble forecasts that are important to users but not necessarily easily verified in the traditional gridpoint-based framework, even when using a neighborhood approach.

a. Object definition

Objects in this study are identified in a similar way as done previously with the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2006a) tool. After applying a Gaussian convolution with a 6-km (2 grid point) radius to the composite reflectivity field, a threshold of 35 dBZ is applied to identify a set of discrete objects. The two gridpoint convolution radius is much smaller than previous studies focused on mesoscale precipitation (e.g., Davis et al. 2006a; Johnson et al. 2011a). The smaller radius is intended to minimize gridscale noise while retaining the resolved and partially resolved convective-scale features. Objects with an area less than 42 grid points are then omitted in order to focus only on robust, established convection. This value is chosen based on the effective model resolution of about seven grid points (Skamarock 2004). The minimum object area is slightly larger than the area of a hypothetical circular object with a diameter right at the effective resolution of 7 grid points. The area of this hypothetical object would be 38.5 grid points. Attributes related to the size, shape, and location of these objects are then calculated. Additional nonreflectivity variables are also included as object attributes. These variables include the average of the within-object 90th–100th percentile of 2–5-km hourly maximum updraft helicity as a measure of storm rotation, hourly maximum 10-m wind speed as a measure of straight-line wind severity, hourly accumulated precipitation as a measure of flash flooding potential, and column/hourly maximum hail size. Hail size is quantified differently for forecasts and observations because the forecast data that were saved from the real time forecasts do not have an exactly analogous observation dataset. For forecasts, hail size is diagnosed directly by the WRF Thompson microphysics scheme. For observations, hail size is calculated using the Maximum Expected Size of Hail (MESH; Smith et al. 2016) algorithm.

In this study, the additional nonreflectivity attributes are used to limit consideration to objects meeting criteria related to specific severe weather hazards. The forecaster is then able to either focus only on storm mode by ignoring these additional attributes or further limit the focus to storms producing specific hazards such as large hail or strong rotation by omitting from consideration all objects that fail to meet a threshold value of the corresponding attribute. Thus, these multivariable ''intensity'' attributes are used to limit consideration to storm objects that are intense enough to represent a specific severe weather threat during the object-identification, rather than in the object-matching procedure.

b. Object matching

The large number of subjectively determined parameter values that must be specified has arguably limited the broader acceptance of, and generalization from, many object-based postprocessing and verification studies (e.g., as noted in Skinner et al. 2016). The object matching procedure in this study is loosely based on our earlier work, but with a greater emphasis on simplicity. In particular, object matching in many previous studies (e.g., Davis et al. 2009; JW12) was based on a total interest (i.e., overall object similarity) value determined from the weighted average of the objects' similarity in terms of each attribute. In the commonly used MODE utility, a piecewise linear similarity function is defined for each attribute and the user specifies a weight and confidence value for each similarity function

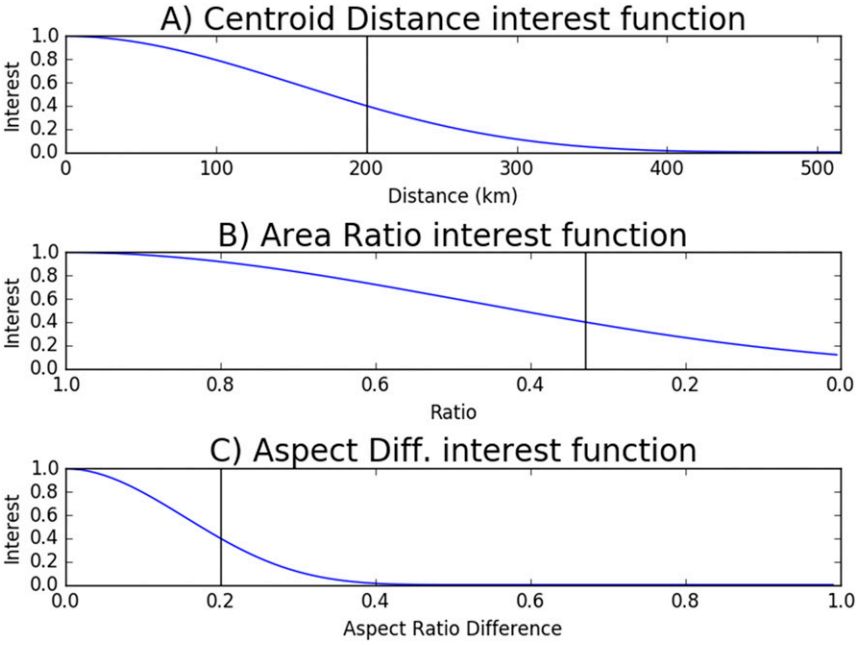

FIG. 2. Interest functions to compare a pair of objects in terms of their (a) centroid distance, (b) area ratio, and (c) aspect ratio difference. Vertical black lines are placed at the *e*-folding distance.

(e.g., Davis et al. 2009). The purpose of the confidence values is to change the weight given to certain attributes depending on the similarity of other attributes. For example, if two objects are spatially very far apart then the weight given to the similarity of their area is reduced because it could already be inferred that they do not correspond to each other so having a similar area is meaningless (JW12). Here, the piecewise linear similarity functions of MODE are replaced by the Gaspari and Cohn (1999) function that is commonly used in DA and can be defined in terms of a single *e*-folding distance. The Gaspari and Cohn function is approximated using Eq. (4.10) of Gaspari and Cohn (1999), which is approximately Gaussian in shape, but goes exactly to zero at a specified distance. Furthermore, the weighted average is here replaced with a simple product of each attribute's similarity between the two objects being compared:

$$I = f_{a_1} \times f_{a_2} \times f_{a_3}. \quad (1)$$

In Eq. (1), I is the total interest between the two objects and f_{a_i} is the similarity (or interest) between the two objects only in terms of attribute i. The three attributes used for matching are centroid distance, area ratio, and aspect ratio difference (Fig. 2). Based on subjective evaluations of the first two authors, as well as forecaster feedback during the HWT SFE, *e*-folding scales for the similarity functions are set to 200 km, 0.33 (technically, encoded as 1.0–0.33 since greater area ratio means more similarity not more error, unlike centroid distance or aspect ratio difference), and 0.2, respectively. The sensitivity of the systematic verification to changing these values is discussed further in section 6. While MODE is not the only option for object-based verification, and even simpler matching criteria have also been used, we emphasize here the importance of keeping some of the flexibility of more complex methods while still maintaining as much simplicity as possible.

Equation (1) obviates the need for the confidence values and weights that were used in previous studies. In this study, we interpret the "weight" on a particular attribute to be effectively increased (decreased) by simply increasing (decreasing) the *e*-folding distance of the interest function for that attribute. When total interest is a product, rather than an average, the confidence values are unnecessary because any attribute that has near-zero similarity will cause the total interest to also be near zero. We note that Skinner et al. (2016) also combined multiplicative and additive components of a total interest by including a product of centroid distance interest and temporal distance interest in the average. Objects are considered a match if their total interest is at least 0.2. We found this threshold to correspond well to subjective determinations of matched and unmatched objects in the test cases from the 2017 dataset.

c. Object probabilities

JW12 proposed a probabilistic approach to object-based forecasting that is appropriate for ensemble forecast verification. In short, the approach consisted of considering one ensemble member as the "control" forecast, which theoretically should be the most likely forecast. Probabilities are then assigned to each object in the control forecast based on how many of the other ensemble members are forecasting a sufficiently similar object to be considered a match. Therefore, control forecast storms that are similar to storms in most other ensemble members have high probability to occur and outlier storms in the control member have low probability to occur. In the OU MAP multiscale hybrid DA and ensemble forecast system, the forecast initialized from the EnVar analysis is expected to have lower analysis and forecast errors than the other ensemble members on average (Wang and Wang 2017; Wang et al. 2019, manuscript submitted to *Wea. Forecasting*) making it the obvious candidate for the object-based control member.

One issue with this approach is that the predefined control member may not be the most representative of the center of the ensemble distribution for any given forecast, regardless of the average behavior over many cases. Also, the method may not easily generalize to ensembles with equally likely members. Therefore, we now consider some other alternative approaches. Another option would be to use the object-based threat score (OTS; Johnson et al. 2011a) to identify the case dependent "representative" member. This is done here by finding the member with the largest average OTS in comparison to all other ensemble members, where the OTS measures overall similarity between two ensemble members i and j as follows:

$$\mathrm{OTS}_{ij} = \frac{1}{A_i + A_j}\left[\sum_{p=1}^{P} I^p(a_i^p + a_j^p)\right]. \qquad (2)$$

In Eq. (2), A is the total area of all objects from that member, P is the number of paired objects, I^p is the total interest [i.e., similarity; Eq. (1)] for the the pth pair of objects, and a_i and a_j are the areas of the pth pair of objects. Following Johnson et al. (2011a), objects are paired iteratively by first pairing the two objects with greatest total interest, then removing the paired objects from consideration and resorting the list of all potential pairings based on total interest until one of the members runs out of objects. The representative member is taken as the member with the greatest value of OTS when its OTS relative to all other ensemble members is averaged. Other methods of choosing a representative member, such as the member closest to the ensemble mean (Schwartz et al. 2014), have also been considered. However, quantifying differences between members depends on the spatial scales and features of interest (e.g., Dey et al. 2014). Our past work with OTS suggests that it is suitable for quantifying the similarity of CAM ensemble members in a way that mimics subjective comparisons (Johnson et al. 2011a).

The "representative" member is intended to represent the center of the ensemble distribution and should therefore be more skillful than any other member over the entire domain when averaged over enough forecasts in an unbiased ensemble, but even then may *locally* still not be the member that is most representative of the ensemble distribution for a particular convective system. Therefore, a third option is also introduced by taking the objects from all ensemble members and constructing a hypothetical ensemble pseudomember using the objects that are locally most representative of the ensemble distribution. These objects are obtained according to the following steps:

1) Make a list of all objects in the forecast ensemble, together with the objects' probabilities, calculated from the percentage of ensemble members with a matching (i.e., total interest > 0.2) object.
2) Sort all of the objects by probability, breaking ties according to the average total interest with all the objects from other ensemble members that it matched to.
3) Add the highest probability object to the object list of the pseudomember.
4) Remove from consideration the added object, as well as all matching objects in other members that contributed to the probability of the added object, leaving a new, smaller list of objects.
5) Repeat from step 2 until no objects remain in the list of ensemble forecast objects.

When plotted, the result of this process is conceptually similar to ensemble "paintball" plots that HWT forecasters are accustomed to using to summarize CAM ensemble forecasts (e.g., Roberts et al. (2019). In the paintball plot a threshold is applied to a field such as composite reflectivity (i.e., a rudimentary object identification), and the resulting objects are all overlaid on the same plot with a different color for each ensemble member.

In the pseudomember object-based probabilistic forecast, all objects in the ensemble are also represented on the same plot. However, the clutter of the plot is reduced compared to a paintball plot, which facilitates interpretation. This is because many objects contribute to the plot by increasing the probability of a similar object that is already plotted, rather than being explicitly overlaid.

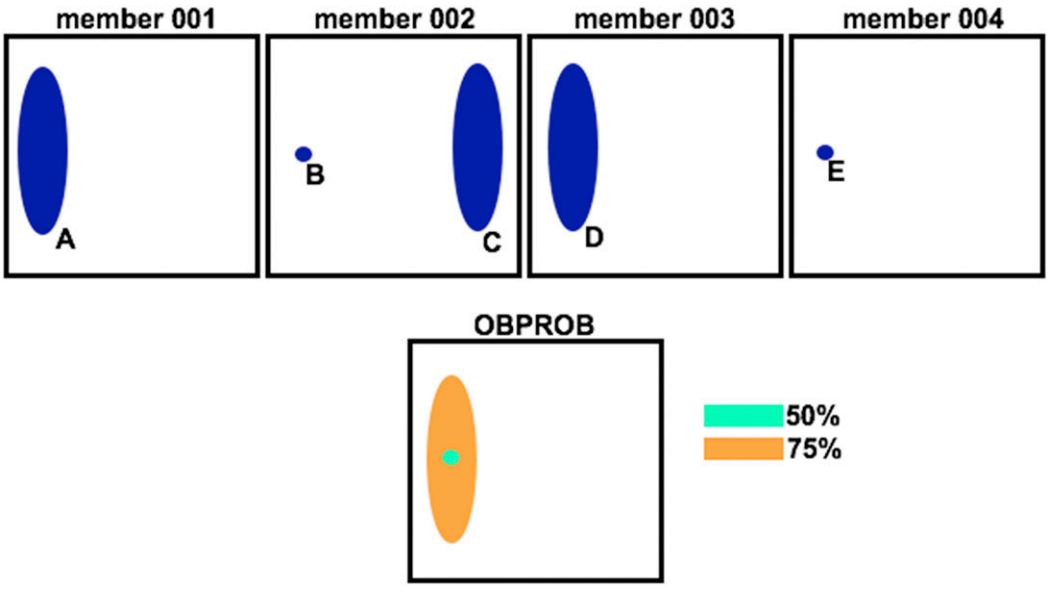

FIG. 3. Illustrative example of the procedure of generating the OBPROB forecast. (top) Hypothetical objects within a 4-member ensemble forecast. (bottom) The resulting OBPROB forecast.

Since the objects now have an associated probability, care must be taken when interpreting what exactly they represent a probability of. Simply stated, they reflect the probability that the convective system represented by the plotted object will occur. The matching procedures described above are used to quantify whether the represented convective system indeed occurs. Specifically, matching to observation objects occurs when the plotted object and the observation object have a total interest greater than 0.2. In the remainder of this paper, the object-based probabilistic forecasting method using the pseudomember objects is abbreviated as OBPROB.

The steps 1–5 above, and the interpretation of the resulting probabilistic forecast is conceptually illustrated in Fig. 3. In Fig. 3 we consider a simplified example of a 4-member ensemble with forecast objects labeled as A, B, C, D, and E. For illustration purposes, we assume that objects appearing to have the same shape, size, or location have an interest value of 1.0 for that attribute, while objects appearing to have different shape, size, or location have an interest value of 0.333 for that attribute. The overall similarity [i.e., total interest, Eq. (1)] between any two objects is the product of the interest value for each of shape, size, and location. So, the total interest between objects A and B, $I_{AB} = I_{location} \times I_{shape} \times I_{size}$, is $1.0 \times 0.33 \times 0.33 = 0.11$ because they have the same centroid location but different shape and size. Similarly, the total interest between objects A and C is $0.33 \times 1 \times 1 = 0.33$ because they have the same shape and size but different location. Table 1 shows the Total Interest between all pairs of objects from different ensemble members in this example. If we use a matching threshold of 0.2, we find that there are three of the four ensemble members with an object matching object A (including itself). Object A is therefore interpreted as representing a convective system with a 75% chance of occurring. The probability of occurrence, and average similarity to matched objects, for each object is summarized in Table 2. There are three objects (A, C, and D) with 75% probability, so we use the average similarity to matched objects to break the tie and select either object A or D (since they are exactly identical in this simple example, it doesn't matter which one we choose) to first add to the final OBPROB plot. After plotting the orange object (object A) with 75% probability in Fig. 3 and removing object A and the objects that it was matched to (objects C and D) from consideration, we recalculate the total interests among the two remaining objects B and E. Since these objects

TABLE 1. Total interest values among the objects A through E from Fig. 3. Duplicate entries are blacked out from the table. Here "N/A" is used to indicate a comparison between objects in the same ensemble member forecast, which are not calculated. Total interest values between "matched" objects are highlighted with bold font.

	Object A	Object B	Object C	Object D	Object E
Object A	**1.0**				
Object B	0.11	**1.0**			
Object C	**0.33**	N/A	**1.0**		
Object D	**1.0**	0.11	**0.33**	**1.0**	
Object E	0.11	**1.0**	0.0	0.11	**1.0**

TABLE 2. Object probability and average total interest to matched objects, for objects A through E from Table 1.

Object	Probability	Avg total interest to matched objects
A	75%	0.777
B	50%	1.0
C	75%	0.556
D	75%	0.777
E	50%	1.0

are identical, they perfectly match each other but do not match any remaining object in the other two ensemble members, so we choose one of them to add to the OBPROB plot with a 50% probability (Fig. 3).

The interpretation of the OBPROB product in this contrived example is that there is a 75% probability of a large elliptical object to occur, and its most likely location is indicated by the plotted orange object. Independently, there is also a 50% probability of a much smaller, circular object to occur, and its most likely location also happens to be the same as the most likely location of the orange object. This outcome illustrates a potential source of confusion in qualitative interpretation of the OBPROB forecasts. A naïve interpretation of this OBPROB plot would be that the total probability of one of these two storm types occurring at this location is greater than 100%. However, a more complete consideration of how the probabilities are generated would focus on the probability of specific convective morphologies to occur, rather than focusing on the probability of convection in terms of the exact location of the object. In particular, it should be noted that (i) there are some scenarios (i.e., member 002) where both objects will occur, and (ii) spatial location is just one of three equally weighted attributes of the plotted objects, which have minimal ''overlap'' in their sizes and shapes.

While we acknowledge this subjective limitation that the OBPROB plots are interpreted differently than how forecasters are accustomed to interpreting probability map plots, this approach has the strength of representing the full ensemble forecast distribution, while also eliminating issues related to choosing a ''control'' member. This is important because, as suggested by Schwartz et al. (2014), it has yet to be definitively shown that a case dependent best-guess member can be chosen a priori to provide better systematic performance than a randomly chosen member. Furthermore, verification of the OBPROB forecasts can quantitively evaluate how well the ensemble distribution of storm morphologies represents the uncertainty in the storm morphology forecast. Further discussion of how the object-based probabilities might be differently displayed for the purpose of operational forecasting applications is continued in section 6.

4. Demonstration of OBPROB using 27–28 May 2017 case study

Case studies from 2017 were used to subjectively identify suitable object identification and object matching parameters for the object-based probabilistic verification applied to severe weather hazard prediction. The details of the method are demonstrated in this section using a representative case study from the severe weather outbreak taking place in the southern Great Plains at ~0000–0600 UTC 28 May 2017. The observed radar reflectivity mosaic at 0000 UTC shows convection initiating in southern Oklahoma, as well as multicellular convection with a loosely linear organization in northeastern Oklahoma (Fig. 4a). In Oklahoma, multicell clusters start to form within 1–3 h (Figs. 4b–d), growing upscale into a more linear system at ~0400 UTC (Fig. 4e). A southeastward propagating storm also emerged from convection in southeastern Colorado and moved into northern Oklahoma (Figs. 4c–f). This case study is investigated over the limited domain shown in Fig. 4 for ease of interpretation. First, we use this case study to demonstrate the application of OBPROB and qualitatively compare the quantitative information that it provides about the ensemble forecast performance to a neighborhood-based approach. In particular, we emphasize that, unlike the NMEP forecasts, the OBPROB forecasts are especially sensitive to convective system morphology.

a. Subjective evaluation

The 10-member ensemble forecast of composite reflectivity for the ensemble forecast initialized at 0000 UTC 27 May 2017, and valid 30 h later at 0600 UTC 28 May 2017, is shown in Fig. 5. Also shown in Fig. 5 are two common methods of summarizing ensemble forecasts in an operational setting, a paintball plot and a neighborhood maximum ensemble probability (NMEP; Schwartz and Sobash 2017; Roberts et al. 2019) plot. For NMEP, we here use a 40-km neighborhood, selected based on the 40-km radius used by SPC in probabilistic outlooks. Since NMEP can result in abrupt discontinuities in the probability field, a Gaussian spatial convolution with a sigma value of 40 km is also applied to the final NMEP probability fields. Manual examination of the individual ensemble member forecasts reveals that all members are predicting convection in eastern Oklahoma at this time. Many of the members are also predicting a linear east–west-oriented mesoscale organization of the convection (Figs. 5a–j). The paintball plot (Fig. 5k) summarizes this information in a single figure, making it a useful product

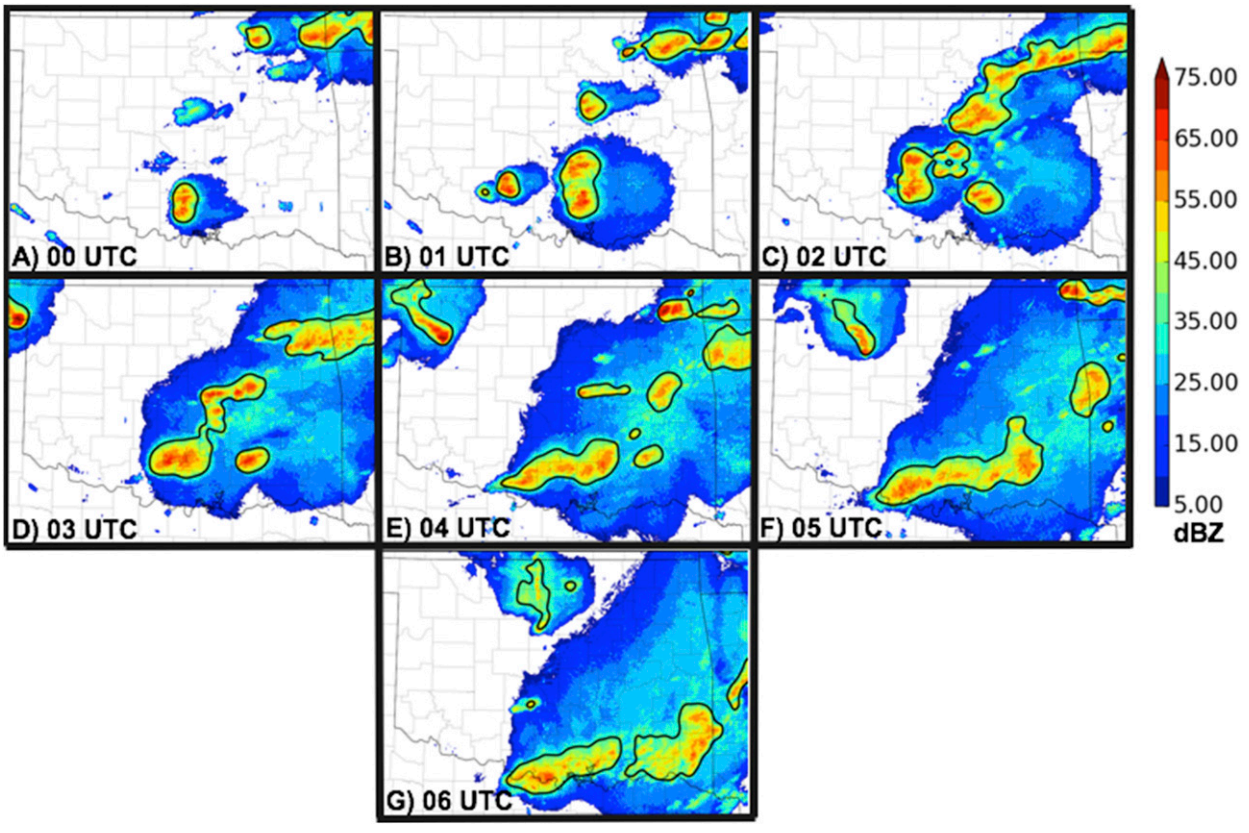

FIG. 4. MRMS composite reflectivity for 28 May 2017 at (a) 0000, (b) 0100, (c) 0200, (d) 0300, (e) 0400, (f) 0500, and (g) 0600 UTC. The 35-dBZ contour is highlighted in black.

for an operational severe weather forecaster. However, the multiple overlapping objects make it difficult to pick out specific predicted storms or to quantify the probability of specific storm modes or attributes. The NMEP (Fig. 5l) provides a more quantitative probabilistic forecast, but because of the use of a mesoscale neighborhood radius the convective-scale details that forecasters often wish to obtain from CAM ensembles are lost, demonstrating the need for the OBPROB method.

The OBPROB forecasts at the 30-h lead time using the three different methods of choosing the object-based control member are shown in Fig. 6. The EnVar analysis member (Fig. 6a) is clearly not representative of the ensemble distribution of storm modes and types of organization that are subjectively seen in Fig. 5 for this case. Although the plotted object correctly portrays our inference from the ensemble members that it has a low probability of occurring, information about the storms that are more likely to occur is missing from such a plot. The ensemble representative member based on Eq. (2) (Fig. 6b) shows a storm object that is consistent with how a forecaster might interpret the most likely convective organization given the ensemble forecast in Fig. 5, with a probability of occurring objectively quantified at 50%. This object is also qualitatively similar to the observation objects described below in Fig. 8d. However, this product is also incomplete because a forecaster would likely still want to know about other less probable outcomes in this region or the most probable outcome in a completely different region of the forecast domain which is not well shown by this same member (not shown). Only the pseudomember method quantifies the most likely linear mesoscale organization, as well as the less likely cellular modes and the low-probability disorganized cluster of convection closer to central Oklahoma (Fig. 6c). Therefore, the pseudomember method is used for calculating objective verification statistics in this study.

The data used for verification consists of Multi-Radar Multi-Sensor (MRMS) data. MRMS blends single quality controlled and processed WSR-88D data into a suite of seamless CONUS-wide products. MRMS data are output every 2 min over a CONUS domain with 0.01° horizontal resolution (~1 km) for MESH and composite reflectivity and at a 0.005° horizonal resolution (~500 m) for azimuthal shear (AzShear).

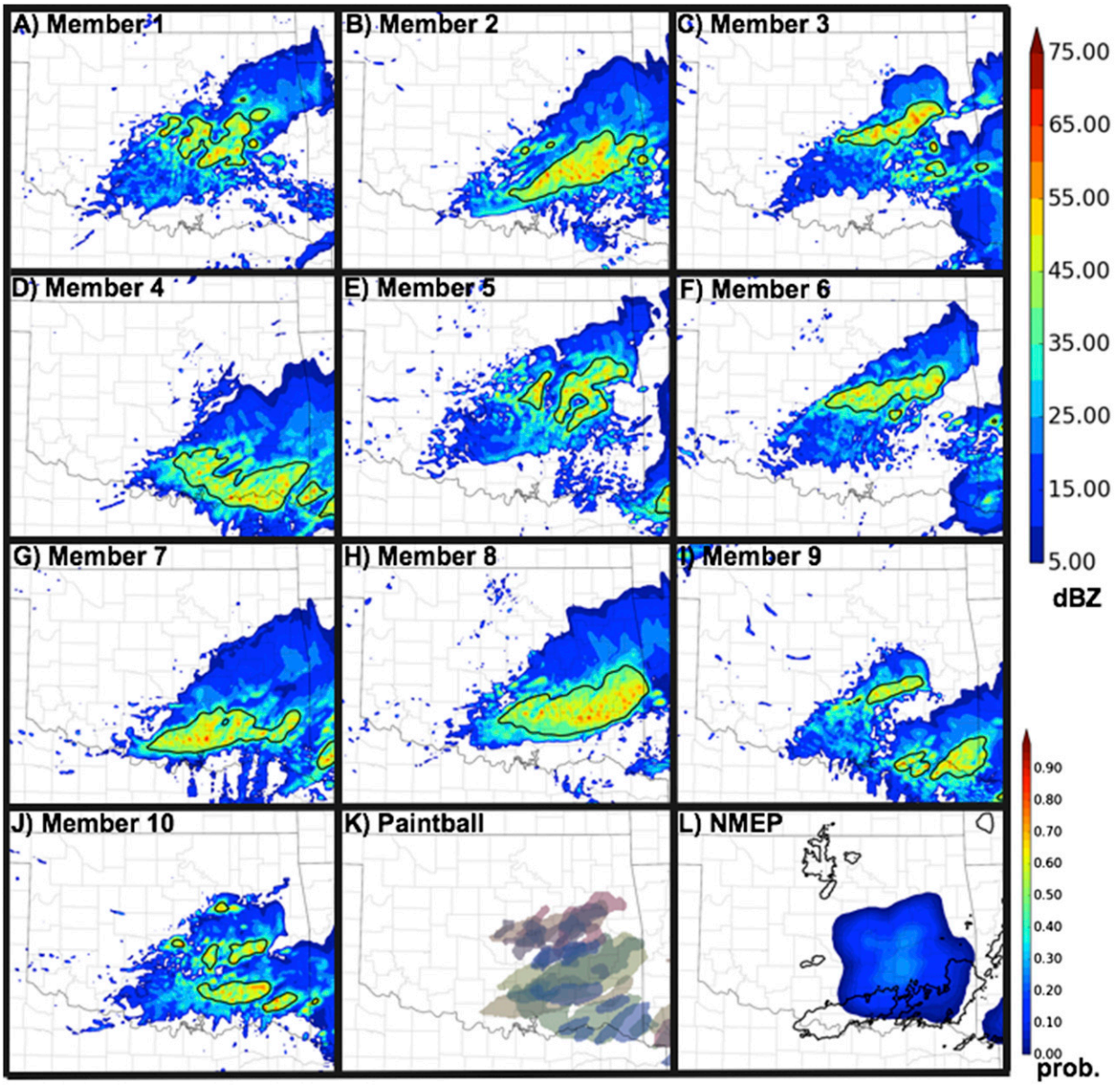

FIG. 5. Ensemble forecast initialized at 0000 UTC 27 May 2017 and valid at 0600 UTC 28 May. (a)–(j) The 10 member ensemble of composite reflectivity with object outlines contoured in black (member 1 is the EnVar member). Also shown are (k) the corresponding paintball plot and (l) the corresponding neighborhood maximum ensemble probability (using a 40-km search radius and 40-km Gaussian smoothing) plot for a 35-dBZ threshold. The different colors of objects in (k) indicate that the objects are from different ensemble members. The black line in (l) is the 35-dBZ observation reflectivity contour.

See Smith et al.(2016) and Zhang et al. (2016) and references therein for details on the MRMS operational system. AzShear is computed using a linear least squares derivative that is taken on the quality controlled velocity field and the azimuthal gradient represents the rotation in the field (Mahalik et al. 2019). For this project, AzShear data was reprocessed to use the updated version of the AzShear code for better consistency and results across the dates of this study. Accumulated precipitation forecasts were verified against the gauge-corrected radar quantitative precipitation estimates from MRMS (Qi and Martinaitis 2016; Zhang et al. 2016).

b. Objective evaluation

The Brier score (BS; Brier 1950) and Brier skill scores (BSS; Wilks 2006) are calculated for this case at forecast hours 24–30 (i.e., 0000 UTC 28 May–0600 UTC 28 May). For OBPROB BS, observation objects that are not

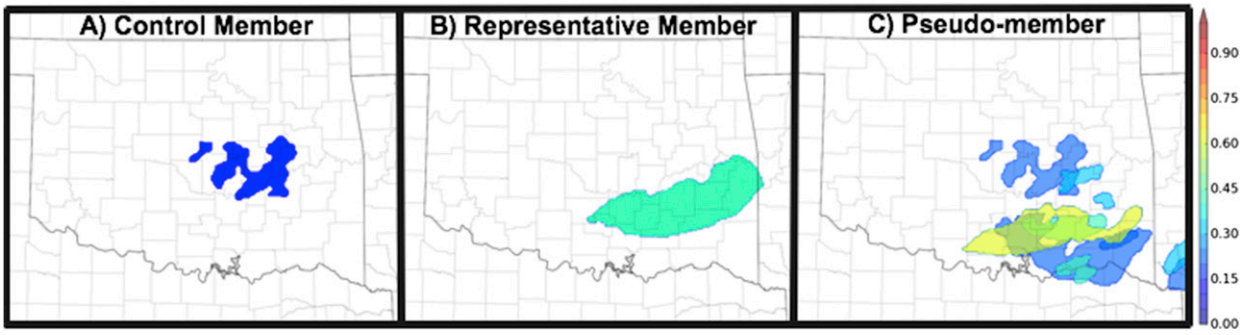

FIG. 6. Example OBPROB forecast for the ensemble forecast shown in Fig. 5 using three different methods of selecting which objects to assign probabilities to, including (a) the ensemble control member, (b) the ensemble representative member, and (c) the combined pseudomember. Panel (a) corresponds to member 1 in Fig. 5 and (b) corresponds to member 8 in Fig. 5. The corresponding observation objects can be found in Fig. 8f.

matched to any of the probabilistic forecast objects are taken to have been forecast with 0% probability. Unlike the gridpoint-based framework, the OBPROB method does not otherwise produce 0% probabilities so there are no "correct null" events contributing to the BS. This is advantageous because the OBPROB BS is not sensitive to large regions of correct null forecast on some cases, which would lower the BS, artificially appearing to be a particularly skillful forecast. In the neighborhood framework, the correct nulls are controlled for by defining a reference BS and presenting the BSS with respect to that reference forecast. Here, the reference forecast is the domain average observation event (e.g., reflectivity threshold exceedance) frequency. For the object-based framework, it is not clear what reference forecast should be used, so we plot the BSS with respect to a worst possible reference BS of 1.0. The use of a worst possible forecast as the reference score follows the approach in the commonly used fractions skill score of Roberts and Lean (2008). Due to the lack of correct null forecasts in the OBPROB context, the result should not be very sensitive to the choice of reference BS. While the NMEP performance decreases during this 6-h period (Fig. 7; thick green line), the OBPROB performance increases at the end of the 6-h period compared to the beginning of the period (Fig. 7; blue line). The general trend of decreasing performance later in the forecast period for NMEP is also seen in the BS without regard to a reference BS (Fig. 7; thin green line).

The cause of the different relative performance between forecast hours 24 and 30 for OBPROB and NMEP for this case are illustrated in Fig. 8. The NMEP forecast at the 24-h lead time (i.e., 0000 UTC 28 May) shows a lot of overlap between the areas of high neighborhood probability and observed reflectivity, as well as large areas of high probability far (i.e., >40 km) from areas of observed precipitation (Fig. 8a). At forecast hour 30 (0600 UTC), the area of high probability is focused too far north and east and there are observed precipitation areas where there was near-zero probability forecast (Fig. 8d). The subjective evaluation is consistent with the slightly worse objective performance of the NMEP forecast in Fig. 7 at forecast hour 30. In contrast, the OBPROB forecast at the 24-h lead time (Fig. 8b) correctly indicates a high (70%) probability of a storm in southern Oklahoma, but also incorrectly indicates an even higher (90%) probability of a very linear storm in northeastern Oklahoma. The forecast also indicates a moderate probability (50%) of cellular convection in north Texas that did not materialize in the observations. At the 30-h lead time (Fig. 8e), the OBPROB forecast correctly predicts that convection in southeastern Oklahoma is most likely to have grown upscale and have an east–west linear orientation. Thus, when evaluating the ensemble forecast in terms of storm morphology and organization, the trend in the

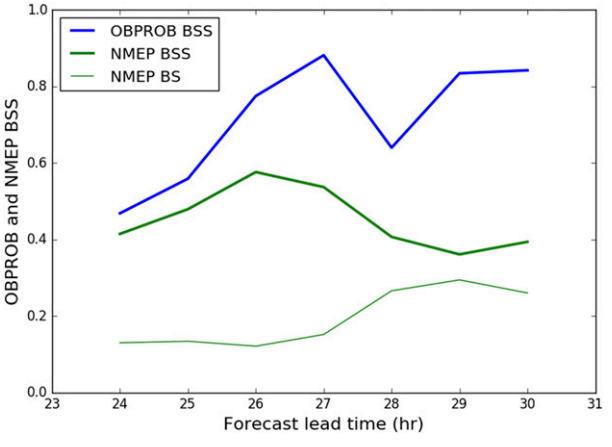

FIG. 7. Brier skill score of NMEP (green) and Brier score of OBPROB (blue) during forecast hours 24–30 of the 27–28 May 2017 case study.

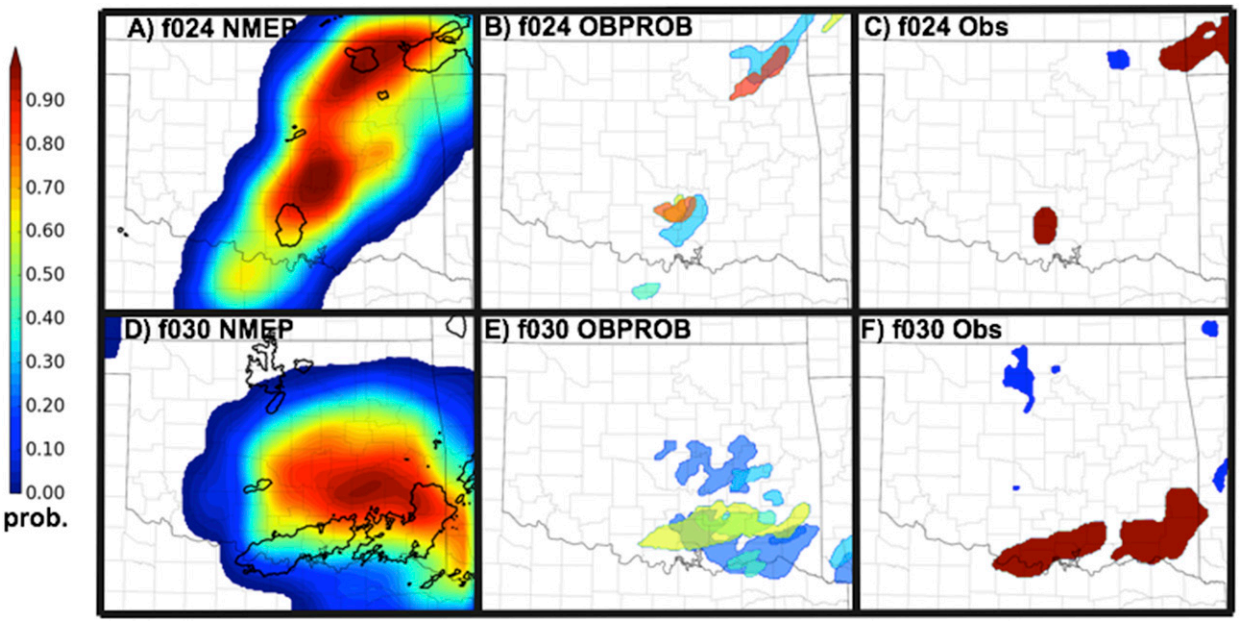

FIG. 8. Comparison of ensemble performance at (top) forecast hour 24 and (bottom) forecast hour 30, showing (a),(d) neighborhood maximum ensemble probability, (b),(e) OBPROB, and (c),(f) observation objects with matched objects shaded red and unmatched objects shaded blue.

OBPROB verification metric in Fig. 7 is more consistent with a subjective evaluation than the trend in the NMEP verification metric. This case study emphasizes that the NMEP and OBPROB verification frameworks provide distinct and useful diagnostic information about the performance of CAM ensemble forecast performance. Thus, neither framework can alone provide a complete objective evaluation of a CAM ensemble system.

The OBPROB technique can also make use of the multivariable storm attributes to focus on probabilities of specific types of storms, such as those producing high rain rates (here, 25.4 mm h^{-1}), strong updraft rotation (here, 2–5-km UH exceeding 100 $m^2 s^{-2}$), or strong surface wind speeds (here, 20 m s^{-1}) (e.g., Fig. 9).[1] At 0000 UTC, the probability of storm objects with strong rotation (Fig. 9b) is very similar to the overall storm probability (Fig. 8b). This suggests that the storms at this time in central Oklahoma are very likely to be supercellular or, in the case of larger objects representing convective clusters or MCSs in northeast Oklahoma, containing embedded supercellular structures. It should be noted that the OBPROB representation is able to distinguish between these two possibilities while a simple neighborhood probability of the UH field could not (not shown). At this time there are also moderate probabilities of storms with high rain rates and strong surface winds in central and northeastern Oklahoma (Figs. 9a,c). At 0600 UTC, both heavy precipitation and strong rotation are given a 40% probability of occurring within an east–west-oriented MCS in southeastern Oklahoma (Figs. 9d,e), while no strong wind producing storms are predicted in this area (Fig. 9f). The corresponding observation proxies (radar derived quantitative precipitation estimate, radar azimuthal shear, and severe wind reports) are also shown in Fig. 9. The comparison to observations suggest that the forecast for this case had much room for improvement in terms of predicting the occurrence and timing of threats for specific severe weather hazards. For example, while the forecast probability of objects with strong rotation is much higher at 0000 UTC (Fig. 9b) than 0600 UTC (Fig. 9e), the observed values of azimuthal shear in central and southern Oklahoma are generally higher at 0600 UTC (Fig. 9k) than 0000 UTC (Fig. 9h). Also, the probability of strong wind producing forecast storms becomes nonexistent at 0600 UTC (Fig. 9f), but this is actually the time when the widespread severe wind reports actually occurred (Fig. 9l).

5. Systematic verification of 2018 OU MAP ensemble

The OU MAP real time forecast ensembles from the 2018 HWT SFE are now verified systematically in

[1] MESH was not saved as an output variable in the 2017 OU MAP ensemble.

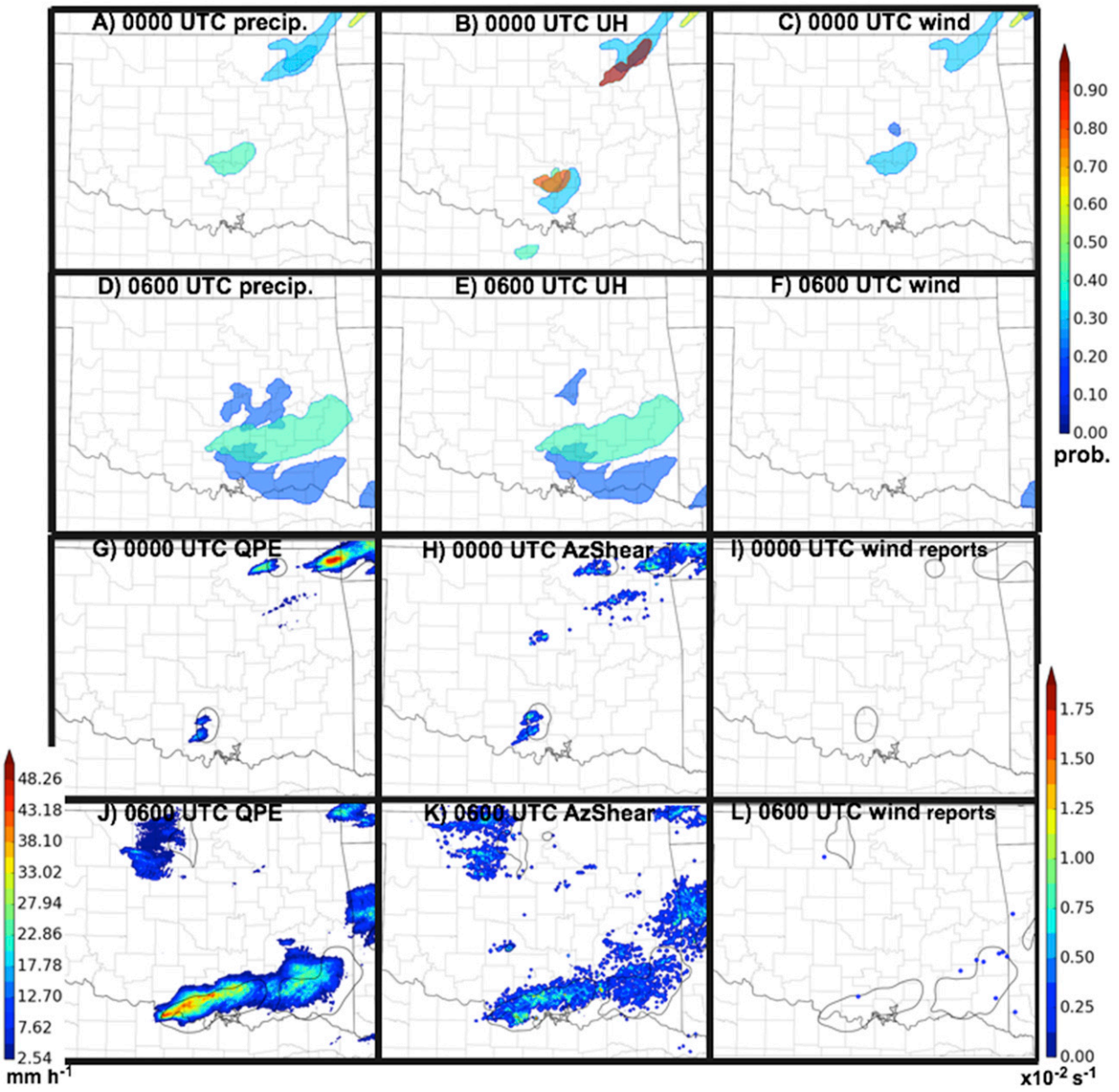

FIG. 9. OBPROB forecasts for objects meeting within-object thresholds of different severe weather hazards at (a)–(c) the 24-h forecast time and (d)–(f) the 30-h forecast time. A threshold of 25.4 mm h^{-1} of hourly accumulated precipitation is used in (a) and (d), (b) and (e) use 100 $m^2 s^{-2}$ of updraft helicity, and (c) and (f) use 20 m s^{-1} of 10-m wind speed. Also shown are (g)–(l) the observation proxies (quantitative precipitation estimate, azimuthal shear, and severe wind reports, respectively) corresponding to (a)–(f), with observation object outlines contoured in black. Shown in (g) and (j) is the hourly accumulated quantitative precipitation estimate (mm h^{-1}), (h) and (k) show hourly maximum azimuthal shear ($\times 10^{-2}$ s^{-1}), and (i) and (l) show reports of severe [>50 kt (1 kt ≈ 0.51 m s^{-1})] wind within 1 h of the plotted time. Black contours in (g)–(l) are the outlines of the corresponding observation objects.

the context of both OBPROB and NMEP forecasts. The systematic verification is conducted over regional subdomains selected based on expectations of potentially hazardous convective weather by forecasters at the lead time of interest (Fig. 1). Results are shown for both early (forecast hours 1–6, corresponding to 0100–0600 UTC) and late (forecast hours 21–27, corresponding to 2100–0300 UTC) lead times. These verification periods emphasize the most convectively active periods of the diurnal cycle (late afternoon and evening). The binning of multiple forecast lead times to be verified together is motivated by the much

TABLE 3. List of forecast initialization dates (time is 0000 UTC for all forecasts) and corresponding verification subdomain center points for both early (forecast hour 1–6) and late (forecast hour 21–27) lead times, as well as the width (NX) and height (NY) in grid points of each subdomain, for the systematic evaluation of the 2018 OU MAP ensemble forecasts.

Initialization date	"Early" center lat	"Early" center lon	"Early" NX	"Early" NY	"Late" center lat	"Late" center lon	"Late" NX	"Late" NY
29 Apr 2018	46.01°N	109.63°W	274	256	46.10°N	104.80°W	344	216
30 Apr 2018	46.10°N	104.80°W	344	216	42.41°N	97.83°W	227	216
1 May 2018	42.41°N	97.83°W	227	216	39.22°N	95.76°W	260	288
2 May 2018	39.22°N	95.76°W	260	288	38.47°N	96.33°W	277	331
3 May 2018	38.47°N	96.47°W	277	331	39.60°N	94.52°W	217	200
4 May 2018	39.60°N	94.52°W	217	200	42.83°N	76.93°W	312	304
7 May 2018	45.95°N	113.31°W	258	263	44.28°N	101.59°W	258	233
8 May 2018	44.28°N	101.59°W	258	233	43.61°N	94.64°W	277	188
9 May 2018	43.61°N	94.64°W	277	188	40.44°N	87.24°W	231	288
10 May 2018	40.44°N	87.24°W	231	288	41.40°N	101.49°W	317	210
11 May 2018	41.40°N	101.49°W	317	210	37.01°N	100.14°W	244	203
14 May 2018	37.20°N	98.35°W	278	346	36.53°N	98.34°W	278	296
15 May 2018	36.53°N	98.34°W	278	296	40.41°N	76.67°W	257	286
16 May 2018	40.41°N	76.67°W	257	286	34.33°N	100.79°W	252	287
17 May 2018	34.33°N	100.79°W	252	287	43.72°N	101.52°W	283	315
18 May 2018	43.45°N	101.50°W	283	315	37.61°N	100.05°W	300	239
21 May 2018	38.52°N	91.67°W	166	178	39.66°N	84.53°W	286	263
22 May 2018	39.66°N	84.53°W	286	263	40.77°N	82.19°W	258	241
23 May 2018	40.77°N	82.19°W	258	241	43.00°N	102.73°W	222	202
24 May 2018	43.00°N	102.73°W	222	202	44.26°N	97.39°W	287	184
25 May 2018	44.26°N	97.39°W	287	184	34.98°N	100.43°W	198	228
28 May 2018	42.38°N	101.36°W	244	359	37.40°N	100.66°W	249	349
29 May 2018	37.40°N	100.66°W	249	349	38.09°N	98.46°W	219	302
30 May 2018	38.09°N	98.46°W	219	302	36.11°N	101.54°W	271	168
31 May 2018	36.11°N	101.54°W	271	168	38.27°N	90.65°W	249	244
1 Jun 2018	38.27°N	90.65°W	249	244	44.90°N	99.13°W	270	318

smaller number of objects (order of 10–100) at a given lead time on a given case, compared to the number of grid points (order of 10 000–100 000). The forecast initialization times and verification subdomains are listed in Table 3 for reference. The object identification (convolution radius, reflectivity threshold, minimum object area) and matching (*e*-folding distance for object shape, size and location, and total interest threshold) parameters were selected based on maximizing agreement between objective results and subjective evaluations in the 2017 case studies, as described above. These determinations were based on the ~1 day forecast lead time since this is most consistent with how the ensemble was used in real time during the HWT SFEs. Given the general expectation for errors to increase with increasing forecast lead time, we also apply the OBPROB verification to shorter lead time forecasts and compare the change in skill with lead time in the OBPROB and NMEP frameworks. This comparison demonstrates that the OBPROB and NMEP frameworks are indeed sensitive to different aspects of forecast performance and are therefore complementary. While a comparison between OBPROB and similar object-based verification methods may also be of interest, the focus of this section is to comprehensively (i.e., using both neighborhood- and object-based frameworks) verify the OU MAP real time ensemble forecasts from the 2018 HWT SFE using an appropriate neighborhood-based method (NMEP) and an appropriate object-based method (OBPROB).

The overall verification statistics are summarized in Table 4. We select 95% as the confidence level (i.e., p value < 0.05) for statistical significance. The NMEP BSS is calculated for reflectivity, UH, MESH, and hourly precipitation. The OBPROB BSS is calculated for all objects (for comparison to reflectivity NMEP), objects exceeding the UH threshold, objects exceeding the MESH threshold, and objects exceeding the hourly precipitation threshold. For OBPROB, there is not a statistically significant (based on p values calculated with permutation resampling, following Johnson and Wang 2012) difference in skill between early and late lead times, except for UH which actually increases in skill at the later time. The difference for UH objects is likely due to a difficulty in spinning up the storm rotation at early lead times. For NMEP, the skill generally decreases significantly at the later lead time

TABLE 4. Overall BSS of OBPROB and NMEP forecasts at early (first and third rows) and late (second and fourth rows) lead times. For OBPROB forecasts, the first column corresponds to all reflectivity objects while the second through fourth columns exclude all forecast and observed objects that do not have a 90th–100th percentile average UH of at least 65 $m^2 s^{-2}$ (0.005 s^{-1} for observed azimuthal shear), 25.4-mm MESH, or 25.4 mm h^{-1} precipitation, respectively. For NMEP forecasts, the columns correspond to forecasts for 35 dBZ, 65 $m^2 s^{-2}$ (0.005 s^{-1} for observed azimuthal shear), 25.4-mm MESH, or 25.4 mm h^{-1} precipitation, respectively. The number in parentheses for the "late" scores is the *p* value of the difference from the corresponding "early" score, based on the permutation resampling method described in Hamill (1999) and Johnson and Wang (2012).

	dBZ	UH	MESH	Precipitation
Early OBPROB	0.708	0.549	0.767	0.771
Late OBPROB	0.687 (0.174)	0.692 (0.004)	0.758 (0.698)	0.788 (0.415)
Early NMEP	0.410	0.015	−3.183	0.066
Late NMEP	0.150 (0.001)	−0.016 (0.44)	−1.919 (0.013)	−0.014 (0.087)

compared to the earlier lead time, with the exceptions of UH and precipitation which have a difference that is not statistically significant, although the precipitation difference would be significant at the 90% confidence level (Table 4). The stronger dependence of skill on forecast lead time for NMEP than OBPROB may suggest that OBPROB is relatively more sensitive to forecast errors that are not growing substantially between the early (1–6 h) and late (21–27 h) lead times. We speculate that these errors are resulting from biases in the model and/or physics configuration, rather than growing IC errors, because the convective-scale IC errors saturate on the time scale of about an hour (Surcel et al. 2015).

Reliability diagrams (Wilks 2006) for the NMEP forecasts are shown in Fig. 10. In perfectly reliable forecasts, the forecast probability (horizontal axis) would always equal the observed relative frequency (vertical axis). Thus, the diagonal line indicates perfect reliability, the region above the diagonal indicates underconfident forecast probabilities and the region below the diagonal indicates overconfident forecast probabilities. At the early lead times (Fig. 10a), reflectivity is the most reliably forecasted variable. MESH and precipitation are overforecast, while UH is underforecast. The UH underforecasting is consistent with the hypothesis that it suffered from a spin up period at the early lead times. At later lead times, UH underforecasting is reduced, reflectivity reliability is a little worse, and the MESH and precipitation overforecasting remains. The large departures of many of the reliability curves in Fig. 10 from the diagonal suggest that the OU MAP ensemble performance will likely benefit from simple calibration techniques, in addition to tuning of the ensemble physics configuration, such as using a multiphysics or stochastic physics configuration. Such calibration could improve both biases in each member's forecast of specific variables as well as deficiencies in ensemble spread.

Reliability diagrams for OBPROB are shown in Fig. 11. Similar to NMEP, objects producing large hail or heavy precipitation are not predicted with particularly good reliability, further emphasizing a need for calibration of these variables. While OBPROB reliability for all objects (black lines) is far from perfect, even at early lead times, limiting consideration to strongly rotating objects (red lines) improves reliability at the early lead times compared to all objects (Fig. 11). However, the relatively small sample size of rotating objects at early lead times and high forecast probabilities (Fig. 11c) limits the ability to generalize from this result. receiver operating characteristic (ROC; Wilks 2006) curves are used to evaluate the ability of probabilistic forecasts to discriminate departures from the climatological event frequency with an area under the curve (AUC) of >0.7 a reasonable discriminator for useful forecasts (Buizza et al. 1999). For NMEP forecasts at early lead times (Fig. 12a), precipitation and hail have good discrimination, despite their poor reliability seen above, and reflectivity has the best discrimination. UH has relatively poor discrimination at early lead times. At later lead times (Fig. 12b), the AUC has decreased for reflectivity, hail, and precipitation while UH discrimination becomes closer to the other variables. The AUC values around 0.7 (Fig. 12 legend) provide further optimism that calibration to improve the reliability of the NMEP forecasts will lead to improved skill, since the discrimination is already at an acceptable level. For OBPROB (Fig. 13), the discrimination is not as good as for NMEP, especially for all objects (black lines). The OBPROB discrimination is actually a little better when only considering objects that have strong rotation (i.e., red line is above black line in both Figs. 13a and 13b). The poor discrimination of the OBPROB forecasts suggests that more advanced calibration methods (e.g., machine learning techniques) may be needed, since discrimination is more challenging to improve with simple postprocessing than reliability.

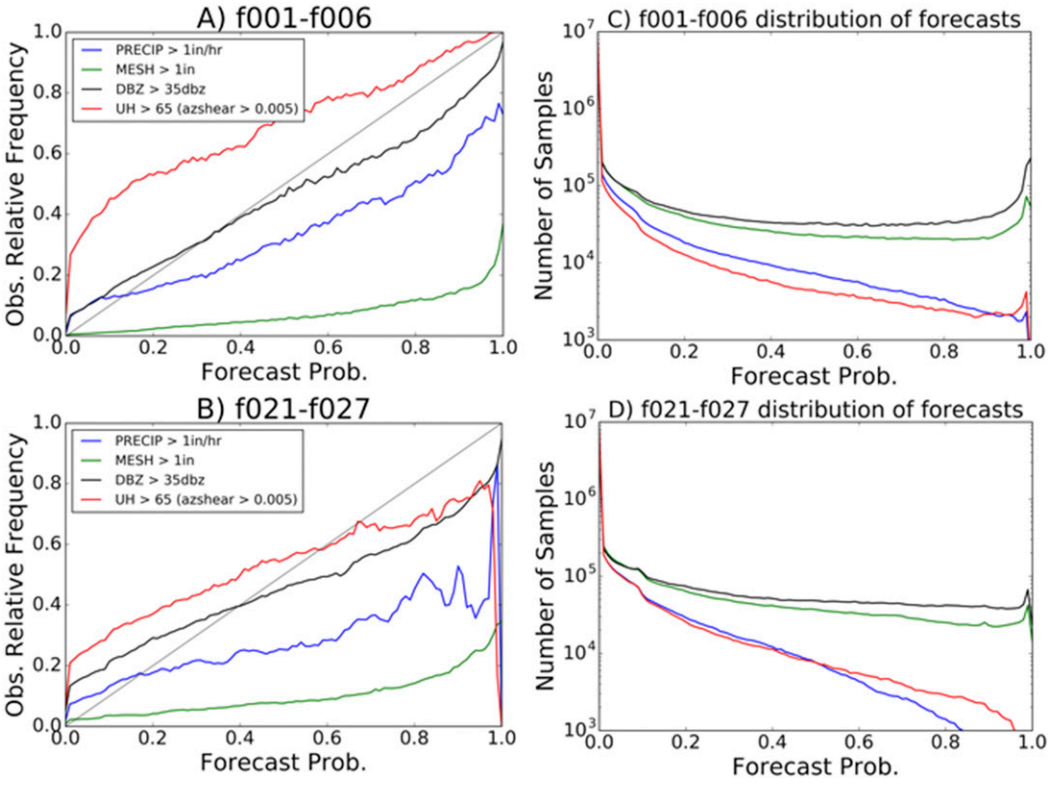

FIG. 10. Reliability diagrams of all forecasts from the 2018 HWT period (26 cases) for NMEP forecasts of precipitation exceeding 25 mm h^{-1} (blue), MESH exceeding 25.4 mm (green), reflectivity exceeding 35 dBZ (black), and updraft helicity exceeding 65 $m^2 s^{-2}$ (red; using 0.005 s^{-1} azimuthal shear as observation proxy) for (a) early forecast lead times and (b) late forecast lead times. (c),(d) The corresponding frequency of forecasts in each probability bin (i.e., grid points).

However, the number of cases needed to successfully train machine learning models with the object attributes may be prohibitively large. We leave this topic for future work.

6. Summary and conclusions

This paper describes further development of an object-based probabilistic (OBPROB) ensemble forecasting and verification framework that was initially proposed in JW12. There are three main developments to the OBPROB framework relative to JW12. First, the object matching procedure has been simplified. Second, multivariable intensity attributes are used during the object identification procedure. Third, a pseudomember is constructed to more fully represent the full ensemble forecast distribution. The pseudomember is constructed by calculating a probability for each storm object in the forecast ensemble and, starting with the most probable object, only plotting objects that were not already implicitly included through their impact on a previously plotted object's probability. A case study demonstrates that this method can concisely convey the full forecast ensemble "envelope," including both the most-probable convective-scale details of storm mode and associated severe hazards, with quantified uncertainty, as well as lower-probability possibilities.

The object-matching was simplified compared to several past object-based verification studies, but may still have some sensitivity to parameter choices. We chose parameters in this study based on the subjective correspondence between the OBPROB products and manual evaluation of forecast ensembles for several independent case studies from 2017. Therefore, we have confidence that the results are representative of how a forecaster would manually interpret "matching" objects. However, we also show the sensitivity of these parameter choices in Fig. 14 by repeating the OBPROB reliability diagrams using 27 parameter permutations (Table 5). Figure 14 shows qualitatively similar comparisons among the types of objects in groups of lines with parameter perturbations (thin lines in Fig. 14) as were found for the parameter settings subjectively chosen based on the 2017 case studies (thick lines in Fig. 14).

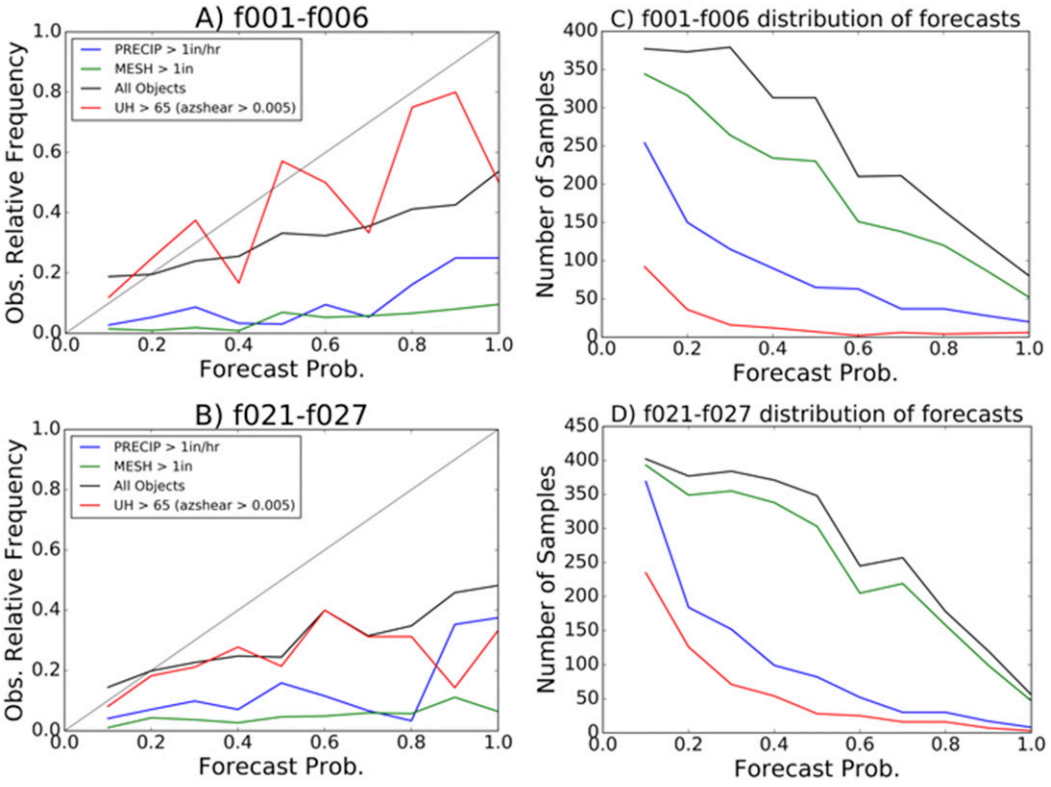

FIG. 11. Reliability diagrams of all forecasts from the 2018 HWT period (26 cases) for OBPROB forecasts of all objects (black), objects meeting the precipitation criterion (blue), objects meeting the MESH criterion (green), and objects meeting the updraft rotation criterion (red; using $0.005\,s^{-1}$ azimuthal shear as observation proxy) for (a) early forecast lead times and (b) late forecast lead times. (c),(d) The corresponding frequency of forecasts in each probability bin (i.e., objects).

Comparison to the neighborhood-based NMEP forecasts reveals the unique forecast features that OBPROB verification is most sensitive to. However, we emphasize that the purpose of the comparison is not to say one framework is better or worse than the other. Rather, we see them both as complementary tools that can be used as part of a comprehensive verification of convection permitting ensemble forecast systems. Overall, the OBPROB systematic verification was much less sensitive to the choice of early (forecast hours 1–6) or late (forecast hours 21–27) lead times than the neighborhood maximum ensemble probability (NMEP). Since OBPROB is designed to be sensitive to the convective morphology (i.e., shape and size) and severe weather hazards of objects with approximately similar locations, we hypothesize that model and/or physics related biases in these attributes are being reflected in the verification metrics. Ongoing future work will focus on using OBPROB to diagnose and improve different multimodel and multiphysics ensemble configurations for severe weather forecasting. While initial condition errors should also affect the object-based verification (e.g., through the centroid distance attribute), features such as location must only be *approximately* similar for objects to match in this study. We speculate that longer lead time (e.g., several days) forecasts may be needed for large-scale IC errors to grow large enough to affect the approximate similarity of object location. Another reason for the generally similar forecast performance at early and late lead times in the OBPROB forecast is likely the use of ~1 day lead time forecasts to tune the object identification and matching parameters with the 2017 case studies. Users interested in CAM guidance for warning time-scale guidance (e.g., Skinner et al. 2018) would likely require much greater precision in shape, size, and location of individual forecast storms in order to consider objects to match. Future work should therefore incorporate lead-time-dependent object matching parameters.

Reliability diagrams for both OBPROB and NMEP forecasts revealed biases that should be corrected with calibration in future iterations of the OU MAP ensemble forecasts in the HWT SFEs. For OBPROB, the

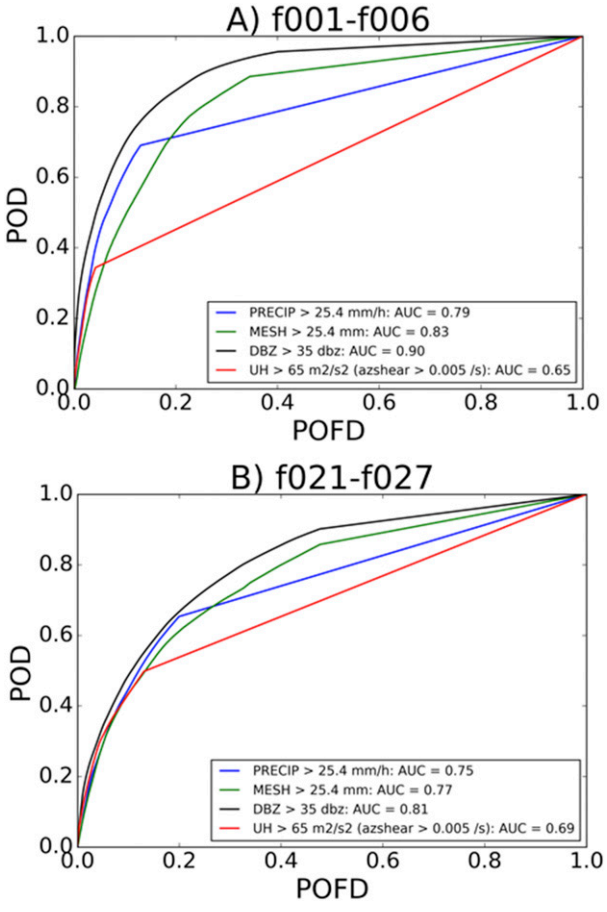

FIG. 12. ROC curves for all forecasts from the 2018 HWT period (26 cases) for NMEP forecasts of precipitation exceeding 25 mm hr^{-1} (blue), MESH (green) exceeding 25.4 mm, reflectivity exceeding 35 dBZ (black), and updraft helicity exceeding 65 $m^2 s^{-2}$ (red; using 0.005 s^{-1} azimuthal shear as observation proxy) for (a) early forecast lead times and (b) late forecast lead times. The ROC curves show the probability of detection (POD) vs probability of false detection (POFD) for a range of probabilities used to delineate a categorical forecast that the event will or will not occur.

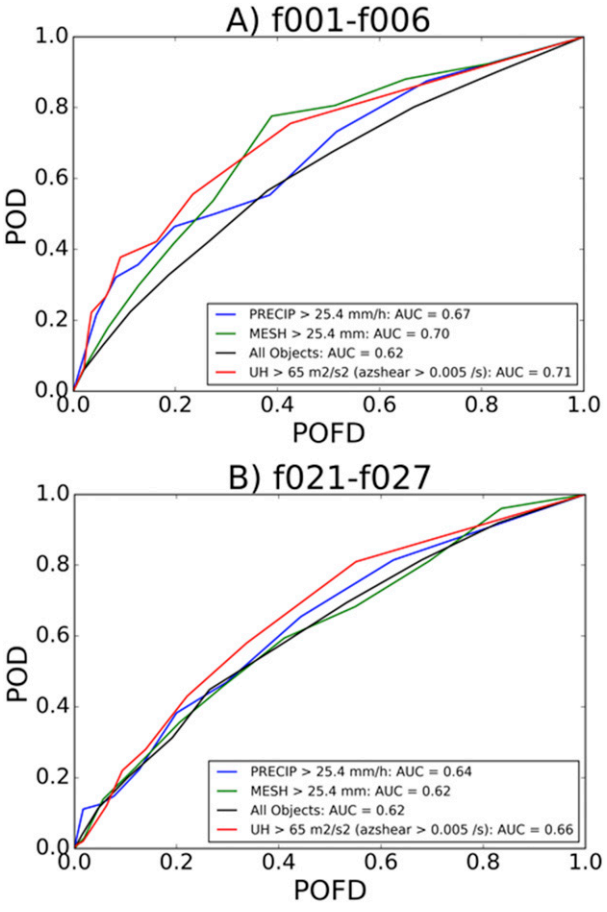

FIG. 13. ROC curves for all forecasts from the 2018 HWT period (26 cases) for OBPROB forecasts of all objects (black), objects meeting the precipitation criterion (blue), objects meeting the MESH criterion (green), and objects meeting the updraft rotation criterion (red; using 0.005 s^{-1} azimuthal shear as observation proxy) for (a) early forecast lead times and (b) late forecast lead times.

ROC diagrams indicated slightly improved discrimination when focusing on objects with strong rotation or large hail, compared to using all objects without consideration of associated severe hazards. However, there is still much room for improvement. Future work will explore machine learning methods for calibration of the OBPROB forecasts, based on past success in improving forecast discrimination with machine learning (e.g., Gagne et al. 2014), and the improved discrimination found in this study when using more than one variable (e.g., reflectivity and updraft helicity) to generate the OBPROB forecast.

The forecasts in this study could be described as emphasizing the watch to convective outlook time/space scales, in contrast to the warning time/space scales emphasized in Skinner et al. (2016). However, we would expect the underlying technique to apply similarly at very short lead times of a few hours or at longer lead times of 2–3 days, both of which are important applications of CAM ensemble forecasts for severe weather, if the object matching parameters are defined appropriately for the lead time of interest. Similar methods may also be applicable to larger-scale systems such as fronts or cyclones at even longer lead times. It is also worth noting that the objective evaluation of CAM ensembles, which also resolve meso- and synoptic scales of motion, could likely be further improved by also incorporating evaluation methods that have been applied to synoptic-scale forecasts such as the scenario-based method of Zheng et al. (2019).

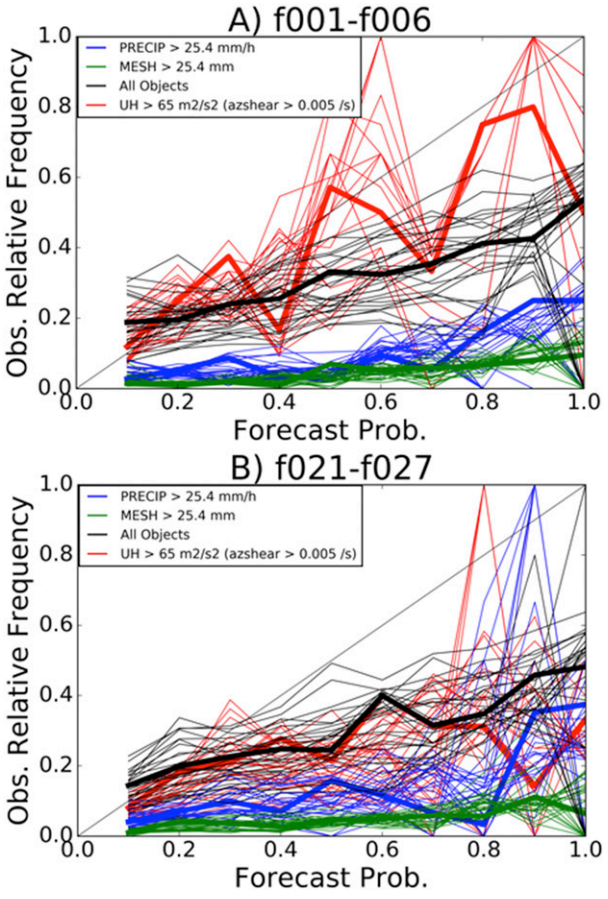

FIG. 14. As in Fig. 11 (thick lines), shown are reliability diagrams of all forecasts from the 2018 HWT period (26 cases) for OBPROB forecasts of all objects (black), objects meeting the precipitation criterion (blue), objects meeting the MESH criterion (green), and objects meeting the updraft rotation criterion (red; using 0.005 s^{-1} azimuthal shear as observation proxy) for (a) early forecast lead times and (b) late forecast lead times. This figure additionally shows the same result repeated for the 27 different combinations of object matching parameters listed in Table 5 (thin lines).

Despite the quantitative diagnostic information provided by verifying the OBPROB forecasts, there are limitations of the technique in terms of qualitatively interpreting the forecast in the operational forecasting environment. Since the forecast products ultimately provided to the public are more similar to the gridpoint-based framework, work is ongoing to further modify the OBPROB technique to first classify the forecast objects according to predefined storm modes or spatial scales of organization, then produce probabilistic plots for a specific type of convective system that can be interpreted analogously as an operational convective outlook. This work is ongoing and will be reported in future studies.

TABLE 5. Permutations of *e*-folding distances for object matching that were used for sensitivity tests in Fig. 14. The row in bold corresponds to the parameter settings used in the other shown results.

Centroid distance (km)	Aspect ratio difference	Area ratio
150	0.1	0.25
150	0.1	0.33
150	0.1	0.4
150	0.2	0.25
150	0.2	0.33
150	0.2	0.4
150	0.3	0.25
150	0.3	0.33
150	0.3	0.4
200	0.1	0.25
200	0.1	0.33
200	0.1	0.4
200	0.2	0.25
200	**0.2**	**0.33**
200	0.2	0.4
200	0.3	0.25
200	0.3	0.33
200	0.3	0.4
250	0.1	0.25
250	0.1	0.33
250	0.1	0.4
250	0.2	0.25
250	0.2	0.33
250	0.2	0.4
250	0.3	0.25
250	0.3	0.33
250	0.3	0.4

Acknowledgments. This work is primarily supported by NOAA Awards NA17OAR4590187, NA15OAR4590193, and NA16OAR4590236. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation Grant ACI-1053575. Some of the computing for this project was also performed at the OU Supercomputing Center for Education and Research (OSCER) at the University of Oklahoma (OU). The manuscript was greatly improved by the comments of three anonymous reviewers.

REFERENCES

Aligo, E. A., B. Ferrier, and J. R. Carley, 2018: Modified NAM microphysics for forecasts of deep convective storms. *Mon. Wea. Rev.*, **146**, 4115–4153, https://doi.org/10.1175/MWR-D-17-0277.1.

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2001: Verification of mesoscale features in NWP models. Preprints, *Ninth Conf. on Mesoscale Processes*, Ft. Lauderdale, FL, Amer. Meteor. Soc., 255–258.

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

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

Bytheway, J. L., and C. D. Kummerow, 2015: Toward an object-based assessment of high-resolution forecasts of long-lived convective precipitation in the central U.S. *J. Adv. Model. Earth Syst.*, **7**, 1248–1264, https://doi.org/10.1002/2015MS000497.

Carlberg, B. R., W. A. Gallus Jr., and K. J. Franz, 2018: A preliminary examination of WRF ensemble prediction of convective mode evolution. *Wea. Forecasting*, **33**, 783–798, https://doi.org/10.1175/WAF-D-17-0149.1.

Clark, A. J., J. S. Kain, P. T. Marsh, J. Correia Jr., M. Xue, and F. Kong, 2012a: 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.

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

——, R. 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.

——, and Coauthors, 2018: The 2018 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *29th Conf. on Severe Local Storms*, Stowe, VT, Amer. Meteor. Soc., 14B.8, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345519.html.

Davis, C. A., B. G. Brown, and R. 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.

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

Dey, S. R. A., G. Leoncini, N. M. Roberts, R. S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection permitting ensembles. *Mon. Wea. Rev.*, **142**, 4091–4107, https://doi.org/10.1175/MWR-D-14-00172.1.

Duda, J. D., and W. A. Gallus Jr., 2010: Spring and summer Midwestern severe weather reports in supercells compared to other morphologies. *Wea. Forecasting*, **25**, 190–206, https://doi.org/10.1175/2009WAF2222338.1.

——, X. Wang, and Y. Wang, 2019: Comparing the assimilation of radar reflectivity using the direct GSI-based ensemble-variational (EnVar) and indirect cloud analysis methods in convection-allowing forecasts over the continental United States. *Mon. Wea. Rev.*, **147**, 1655–1678, https://doi.org/10.1175/MWR-D-18-0171.1.

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

Gagne, D. J., II, A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

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.

——, and Coauthors, 2018: Spring Forecasting Experiment 2018 conducted by the Experimental Forecast Progam of the NOAA Hazardous Weather Testbed: Program overview and operations plan. Internal Tech. Doc., 46 pp., https://hwt.nssl.noaa.gov/sfe/2018/docs/HWT_SFE2018_operations_plan.pdf.

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.

——, N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the Midwest as a function of convective mode: A preliminary study. *Wea. Forecasting*, **23**, 101–113, https://doi.org/10.1175/2007WAF2006120.1.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, https://doi.org/10.1002/qj.49712555417.

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

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Hu, M., S. G. Benjamin, T. T. Ladwig, D. C. Dowell, S. S. Weygandt, C. R. Alexander, and J. S. Whitaker, 2017: GSI three-dimensional ensemble-variational hybrid data assimilation using a global ensemble for the regional rapid refresh model. *Mon. Wea. Rev.*, **145**, 4205–4225, https://doi.org/10.1175/MWR-D-16-0418.1.

Janjić, Z. I., 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.

——, 2004: The NCEP WRF core. Preprints, *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 12.7, https://ams.confex.com/ams/84Annual/techprogram/paper_70036.htm.

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

——, ——, M. Xue, and F. Kong, 2011b: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment.

Part II: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139**, 3694–3710, https://doi.org/10.1175/MWR-D-11-00016.1.

——, ——, 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.

——, ——, J. R. Carley, L. J. Wicker, and C. Karstens, 2015: A comparison of multiscale GSI-based EnKF and 3DVar data assimilation using radar and conventional observations for midlatitude convective-scale precipitation forecasts. *Mon. Wea. Rev.*, **143**, 3087–3108, https://doi.org/10.1175/MWR-D-14-00345.1.

Karstens, C. D., and Coauthors, 2018: Development of a human-machine mix for forecasting severe convective events. *Wea. Forecasting*, **33**, 715–737, https://doi.org/10.1175/WAF-D-17-0188.1.

Lilly, D. K., 1990: Numerical prediction of thunderstorms- Has its time come? *Quart. J. Roy. Meteor. Soc.*, **116**, 779–798, https://doi.org/10.1002/qj.49711649402.

Lu, X., X. Wang, Y. Li, M. Tong, and X. Ma, 2017a: GSI-based ensemble-variational hybrid data assimilation for HWRF for hurricane initialization and prediction: Impact of various error covariances for airborne radar observation assimilation. *Quart. J. Roy. Meteor. Soc.*, **143**, 223–239, https://doi.org/10.1002/qj.2914.

——, ——, M. Tong, and V. Tallapragada, 2017b: GSI-based, fully cycled, dual resolution hybrid ensemble-variational data assimilation system for HWRF: System description and experiment with Edouard (2014). *Mon. Wea. Rev.*, **145**, 4877–4898, https://doi.org/10.1175/MWR-D-17-0068.1.

Mahalik, M. C., B. R. Smith, K. L. Elmore, D. M. Kingfield, K. L. Ortega, and T. M. Smith, 2019: Estimates of gradients in radar moments using a linear least squares derivative technique. *Wea. Forecasting*, **34**, 415–434, https://doi.org/10.1175/WAF-D-18-0095.1.

Mitchell, K. E., and Coauthors, 2005: The Community Noah Land Surface Model (LSM)—User's guide (v2.7.1). NCEP, 26 pp., https://ral.ucar.edu/sites/default/files/public/product-tool/unified-noah-lsm/Noah_LSM_USERGUIDE_2.7.1.pdf.

Nakanishi, M., and H. Niino, 2004: An improved Mellor–Yamada level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31, https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

——, and ——, 2006: An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, https://doi.org/10.1007/s10546-005-9030-8.

Pettet, C. R., and R. H. Johnson, 2003: Airflow and precipitation structure of two leading stratiform mesoscale convective systems determined from operational datasets. *Wea. Forecasting*, **18**, 685–699, https://doi.org/10.1175/1520-0434(2003)018<0685:AAPSOT>2.0.CO;2.

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.

Qi, Y., and S. Martinaitis, 2016: A real-time automated quality control of hourly rain gauge data based on multiple sensors in MRMS system. *J. Hydrometeor.*, **17**, 1675–1691, https://doi.org/10.1175/JHM-D-15-0188.1.

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

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

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

Schwartz, C. S., and Z. Liu, 2014: Convection-permitting forecasts initialized with continuously cycling limited area 3DVAR, ensemble Kalman filter, and "hybrid" variational-ensemble data assimilation system. *Mon. Wea. Rev.*, **142**, 716–738, https://doi.org/10.1175/MWR-D-13-00100.1.

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

——, G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1.

——, ——, 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.

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

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

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

——, and Coauthors, 2018: Object-based verification of a prototype warn-on-forecast system. *Wea. Forecasting*, **33**, 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm classification and climatology. *Wea. Forecasting*, **27**, 1114–1135, https://doi.org/10.1175/WAF-D-11-00115.1.

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

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.

Surcel, M., I. Zawadski, and M. K. Yau, 2015: A study on the scale dependence of the predictability of precipitation patterns. *J. Atmos. Sci.*, **72**, 216–235, https://doi.org/10.1175/JAS-D-14-0071.1.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

——, 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.

Wang, X., 2010: Incorporating ensemble covariance in the gridpoint statistical interpolation variational minimization: A mathematical framework. *Mon. Wea. Rev.*, **138**, 2990–2995, https://doi.org/10.1175/2010MWR3245.1.

——, D. Parrish, D. Kleist, and J. S. Whitaker, 2013: GSI 3DVar-based ensemble-variational hybrid data assimilation for NCEP Global Forecast System: Single resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117, https://doi.org/10.1175/MWR-D-12-00141.1.

Wang, Y., and X. Wang, 2017: Direct assimilation of radar reflectivity without tangent linear and adjoint of the nonlinear observation operator in the GSI-based EnVar system: Methodology and experiment with the 8 May 2003 Oklahoma City tornadic supercell. *Mon. Wea. Rev.*, **145**, 1447–1471, https://doi.org/10.1175/MWR-D-16-0231.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences: An Introduction*. 2nd ed. Academic Press, 467 pp.

Wolff, J. K., M. Harrold, T. Fowler, J. Halley-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.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.

Zheng, M., E. K. Chang, and B. A. Colle, 2019: Evaluating U.S. East Coast winter storms in a multimodel ensemble using EOF and clustering approaches. *Mon. Wea. Rev.*, **147**, 1967–1987, https://doi.org/10.1175/MWR-D-18-0052.1.