Object-Based Evaluation of a Storm-Scale Ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment

AARON JOHNSON AND XUGUANG WANG
School of Meteorology, University of Oklahoma, and Center for Analysis and Prediction of Storms, Norman, Oklahoma

(Manuscript received 10 May 2012, in final form 17 August 2012)

ABSTRACT

Object-based verification of deterministic forecasts from a convection-allowing ensemble for the 2009 NOAA Hazardous Weather Testbed Spring Experiment is conducted. The average of object attributes is compared between forecasts and observations and between forecasts from subensembles with different model dynamics. Forecast accuracy for the full ensemble and the subensembles with different model dynamics is also evaluated using two object-based measures: the object-based threat score (OTS) and the median of maximum interest (MMI).

Forecast objects aggregated from the full ensemble are generally more numerous, have a smaller average area, more circular average aspect ratio, and more eastward average centroid location than observed objects after the 1-h lead time. At the 1-h lead time, forecast objects are less numerous than observed objects. Members using the Advanced Research Weather Research and Forecasting Model (ARW) have fewer objects, more linear average aspect ratio, and smaller average area than members using the Nonhydrostatic Mesoscale Model (NMM). The OTS aggregated from the full ensemble is more consistent with the diurnal cycles of the traditional equitable threat score (ETS) than the MMI because the OTS places more weight on large objects, while the MMI weights all objects equally. The group of ARW members has higher OTS than the group of NMM members except at the 1-h lead time when the group of NMM members has more accurate maintenance and evolution of initially present precipitation systems provided by radar data assimilation. The differences between the ARW and NMM accuracy are more pronounced with the OTS than the MMI and the ETS.

1. Introduction

An advantage of high-resolution storm-scale forecasts, compared to forecasts that parameterize convection, is the ability to explicitly depict features of convective precipitation systems such as their size, shape, and organization (e.g., Xue et al. 2001; Roebber et al. 2002; Done et al. 2004; Clark et al. 2007; Weisman et al. 2008). The depictions of such features can be subjectively used as guidance for distinguishing convective storm modes such as discrete cells, line segments, and organized mesoscale systems, which can aid forecasts of specific severe weather hazards (e.g., Weiss et al. 2004; Gallus et al. 2008; Coniglio et al. 2010). A review of the advantages, limitations, challenges, and open questions related to both high-resolution and ensemble forecasts can be found in Roebber et al. (2004).

Traditional gridpoint-based verification measures (e.g., mean square error or equitable threat score) may not reflect the realism of forecast features at convective scales (Baldwin et al. 2001; Gilleland et al. 2009). Object-based verification measures are one of the nontraditional methods that have been proposed to compare features that may have slightly different spatial and/or temporal locations (Templet and Keenan 1982; Smith and Mullen 1993; Ebert and McBride 2000; Marchok 1999; Baldwin and Lakshminarahan 2003; Case et al. 2004; Marchak et al. 2005; Marzban and Sandgathe 2006, 2008; Davis et al. 2006a,b; Grams et al. 2006; Micheas et al. 2007; Gilleland et al. 2008, 2010). Object-based measures can yield results that are more consistent with subjective evaluations than gridpoint-based measures for high-resolution
precipitation forecasts (Davis et al. 2009; Johnson et al. 2011a,b). Object-based methods also allow for physically descriptive diagnoses of model deficiencies and errors by associating the errors with specific forecast features. Such diagnostic information can elucidate processes in the model that do not evolve as they do in the atmosphere, and thus aid model development. For a thorough review of object-based verification methods please see Gillem et al. (2009).

The Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma has generated storm-scale ensemble forecasts (SSEFs) over a near-CONUS (continental United States) domain during the National Oceanic and Atmospheric Administration Hazardous Weather Testbed (NOAA HWT) Spring Experiments (Kong et al. 2007, 2008, 2009, 2010; Xue et al. 2007, 2008, 2009, 2010). Verification of these ensemble forecasts from different perspectives can help answer different scientific questions related to SSEFs (Kong et al. 2007, 2008, 2009, 2010; Clark et al. 2008, 2009, 2010a,b; Coniglio et al. 2010; Gallus 2010; Kain et al. 2010; Schwartz et al. 2009, 2010; Xue et al. 2010; Schaffer et al. 2011; Johnson and Wang 2012). The 2009 ensemble contains two subensembles, sharing the Weather Research and Forecasting (WRF) Nonhydrostatic Mesoscale Model (NMM) and the Advanced Research WRF (ARW) dynamical cores, respectively. Each of the ARW and NMM subensembles comprise members with different combinations of physical parameterization schemes, in addition to initial and lateral boundary condition (IC–LBC) perturbations. The purpose of this study is to provide object-based evaluation of the 2009 ensemble to reveal the error characteristics and skill of the forecasts comprising both the full ensemble and the ARW and NMM subensembles, and to explore if and how the skill is dependent on different object-based verification scores. This study complements the clustering analysis (Johnson et al. 2011a,b) and the verification studies of the probabilistic forecasts derived from the same ensemble (Clark et al. 2011; Johnson and Wang 2012). The findings of this study can imply and facilitate the optimal design of a storm-scale ensemble. The following paragraphs describe the specific aspects that are considered in this study.

One object-based verification metric is a comparison of the distribution of the forecast and the observed object attributes. Such comparisons have demonstrated regionally dependent size and intensity biases and timing errors of precipitation systems in convection-allowing forecasts in midlatitudes (Davis et al. 2006a,b; Cai 2011; Lane et al. 2011). Similar methods have also been applied to tropical precipitation forecasting (Skok et al. 2010). Other object-based verification metrics include object-based scores of forecast accuracy (e.g., Ebert and McBride 2000; Davis et al. 2006a,b, 2009; Ebert and Gallus 2009; Gallus 2010; Skok et al. 2010; Case et al. 2011; Johnson et al. 2011a). Among these scores, some adopted a fuzzy logic approach instead of a crisp distinction between matched and unmatched objects. For example, the median of maximum interest (MMI), proposed by Davis et al. (2009), and the object-based threat score (OTS), proposed by Johnson et al. (2011a), have been used to measure the overall similarity of precipitation fields, while avoiding the challenge of objectively determining matched and unmatched objects. Both scores were shown to yield results consistent with subjective evaluations (Davis et al. 2009; Johnson et al. 2011a).

While the above object-based scores have been applied to verify and compare individual (as opposed to ensemble) forecasts in previous studies, this study focuses on applying the scores to verify the full ensemble and subensembles and to evaluate differences between the subensembles. In other words, the overall quality of the ensemble and its subensembles (as opposed to the quality of individual forecasts) will be evaluated. Specifically, verification scores aggregated from ensemble members comprising the full ensemble and the subensembles with different model dynamics (ARW and NMM) will be calculated. The verification of the ARW and NMM subensembles is motivated by Johnson et al. (2011b), which showed that the model dynamic cores had a dominant impact on the similarity of the object-based precipitation forecasts. Therefore the ARW and NMM subensembles may have different forecast error characteristics. An earlier study by Davis et al. (2009) applied the object-based metric to compare deterministic ARW and NMM model forecasts. In that study, both the ARW and NMM models have a single configuration of physical parameterization schemes. In the current study, each of the ARW and NMM subensembles has various configurations of the physical parameterization schemes (Table 1). The comparison of the ARW and NMM models in the current study is not limited to a particular configuration of the physical parameterization schemes.

Another aspect of the current study is to explore if and how the skill of the SSEFs depend on different object-based scores. While MMI and OTS have been applied to evaluate storm-scale forecasts in previous separate studies, the current study explores if and how the performance will be dependent on the choice of the object-based score. Both the OTS and MMI are applied and compared as measures of the accuracy of the 2009 ensemble, including both the full ensemble and subensembles.
In section 2, the object-based methods used for verification are described. In section 3 the realism of the forecasts from the full ensemble and subensembles is evaluated in terms of the object attribute distributions. In section 4 the accuracy of the forecasts from the full ensemble and subensembles is evaluated using the MMI and OTS. Conclusions and a discussion are presented in section 5.

2. Data and methods

a. Ensemble configuration and observation data

The NOAA HWT is a collaborative effort between the Storm Prediction Center (SPC), the National Severe Storms Laboratory (NSSL), and the Norman, Oklahoma, National Weather Service forecast office to facilitate development and transition to operations of new forecast technologies (Weiss et al. 2009). Since 2000 the HWT has hosted an annual spring experiment to provide model developers, research scientists, and operational forecasters an opportunity to interact, while evaluating and providing feedback on developing technologies in a simulated operational forecasting environment (Weiss et al. 2009). For the 2009 NOAA HWT Spring Experiment, CAPS produced an experimental real-time convection-allowing ensemble, 5 days a week for 6 weeks, over a near-CONUS domain (Kong et al. 2009; Xue et al. 2009).

The 2009 CAPS Spring Experiment ensemble consists of 20 members, with 10 members from the ARW model (Skamarock and Klemp 2008), 8 members from the NMM model (Janjić 1994), and 2 members from the CAPS Advanced Regional Prediction System (ARPS; Xue et al. 2000, 2003). All members use WRF-NMM, em members use WRF-ARW (i.e., Eulerian mass core), etaKF members useEta Model with Kain–Fritsch cumulus parameterization, and etaBMJ use Eta Model with Betts–Miller–Janjić cumulus parameterization. The N1 refers to the first negative bred perturbation from each SREF model.

In section 2, the object-based methods used for verification are described. In section 3 the realism of the forecasts from the full ensemble and subensembles is evaluated in terms of the object attribute distributions. In section 4 the accuracy of the forecasts from the full ensemble and subensembles is evaluated using the MMI and OTS. Conclusions and a discussion are presented in section 5.

2. Data and methods

a. Ensemble configuration and observation data

The NOAA HWT is a collaborative effort between the Storm Prediction Center (SPC), the National Severe Storms Laboratory (NSSL), and the Norman, Oklahoma, National Weather Service forecast office to facilitate development and transition to operations of new forecast technologies (Weiss et al. 2009). Since 2000 the HWT has hosted an annual spring experiment to provide model developers, research scientists, and operational forecasters an opportunity to interact, while evaluating and providing feedback on developing technologies in a simulated operational forecasting environment (Weiss et al. 2009). For the 2009 NOAA HWT Spring Experiment, CAPS produced an experimental real-time convection-allowing ensemble, 5 days a week for 6 weeks, over a near-CONUS domain (Kong et al. 2009; Xue et al. 2009).

The 2009 CAPS Spring Experiment ensemble consists of 20 members, with 10 members from the ARW model (Skamarock and Klemp 2008), 8 members from the NMM model (Janjić 2003), and 2 members from the CAPS Advanced Regional Prediction System (ARPS; Xue et al. 2000, 2001, 2003). All members have 4-km grid spacing and no cumulus parameterization. The integration domain and verification domain are shown in Fig. 1. The same ensemble forecast data used in Johnson

<table>
<thead>
<tr>
<th>Member</th>
<th>IC</th>
<th>LBC</th>
<th>R</th>
<th>MP</th>
<th>PBL</th>
<th>SW</th>
<th>LSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARW CN</td>
<td>CN</td>
<td>NAMf</td>
<td>Y</td>
<td>Thompson</td>
<td>MYJ</td>
<td>Goddard</td>
<td>Noah</td>
</tr>
<tr>
<td>ARW C0</td>
<td>NAMa</td>
<td>NAMf</td>
<td>N</td>
<td>Thompson</td>
<td>MYJ</td>
<td>Goddard</td>
<td>Noah</td>
</tr>
<tr>
<td>ARW N1</td>
<td>CN – em</td>
<td>em N1</td>
<td>Y</td>
<td>Ferrier</td>
<td>YSU</td>
<td>Goddard</td>
<td>Noah</td>
</tr>
<tr>
<td>ARW N2</td>
<td>CN – nmm</td>
<td>nmm N1</td>
<td>Y</td>
<td>Thompson</td>
<td>MYJ</td>
<td>Dudhia</td>
<td>RUC</td>
</tr>
<tr>
<td>ARW N3</td>
<td>CN – etaKF</td>
<td>etaKF N1</td>
<td>Y</td>
<td>Thompson</td>
<td>YSU</td>
<td>Goddard</td>
<td>Noah</td>
</tr>
<tr>
<td>ARW N4</td>
<td>CN – etaBMJ</td>
<td>etaBMJ N1</td>
<td>Y</td>
<td>WSM6</td>
<td>MYJ</td>
<td>Dudhia</td>
<td>RUC</td>
</tr>
<tr>
<td>ARW P1</td>
<td>CN + em</td>
<td>em N1</td>
<td>Y</td>
<td>WSM6</td>
<td>MYJ</td>
<td>Dudhia</td>
<td>RUC</td>
</tr>
<tr>
<td>ARW P2</td>
<td>CN + nmm</td>
<td>nmm N1</td>
<td>Y</td>
<td>WSM6</td>
<td>YSU</td>
<td>Dudhia</td>
<td>Noah</td>
</tr>
<tr>
<td>ARW P3</td>
<td>CN + etaKF</td>
<td>etaKF N1</td>
<td>Y</td>
<td>Ferrier</td>
<td>MYJ</td>
<td>Dudhia</td>
<td>Noah</td>
</tr>
<tr>
<td>ARW P4</td>
<td>CN + etaBMJ</td>
<td>etaBMJ N1</td>
<td>Y</td>
<td>Thompson</td>
<td>YSU</td>
<td>Goddard</td>
<td>RUC</td>
</tr>
<tr>
<td>NMM CN</td>
<td>CN</td>
<td>NAMf</td>
<td>Y</td>
<td>Ferrier</td>
<td>MYJ</td>
<td>GFDL</td>
<td>Noah</td>
</tr>
<tr>
<td>NMM C0</td>
<td>NAMa</td>
<td>NAMf</td>
<td>N</td>
<td>Ferrier</td>
<td>MYJ</td>
<td>GFDL</td>
<td>Noah</td>
</tr>
<tr>
<td>NMM N1</td>
<td>CN – nmm</td>
<td>nmm N1</td>
<td>Y</td>
<td>Ferrier</td>
<td>YSU</td>
<td>GFDL</td>
<td>Noah</td>
</tr>
<tr>
<td>NMM N3</td>
<td>CN – etaKF</td>
<td>etaKF N1</td>
<td>Y</td>
<td>WSM6</td>
<td>YSU</td>
<td>GFDL</td>
<td>Noah</td>
</tr>
<tr>
<td>NMM N4</td>
<td>CN – etaBMJ</td>
<td>etaBMJ N1</td>
<td>Y</td>
<td>WSM6</td>
<td>MYJ</td>
<td>Dudhia</td>
<td>RUC</td>
</tr>
<tr>
<td>NMM P1</td>
<td>CN + em</td>
<td>em N1</td>
<td>Y</td>
<td>WSM6</td>
<td>MYJ</td>
<td>GFDL</td>
<td>RUC</td>
</tr>
<tr>
<td>NMM P2</td>
<td>CN + nmm</td>
<td>nmm N1</td>
<td>Y</td>
<td>Thompson</td>
<td>YSU</td>
<td>GFDL</td>
<td>RUC</td>
</tr>
<tr>
<td>NMM P4</td>
<td>CN + etaBMJ</td>
<td>etaBMJ N1</td>
<td>Y</td>
<td>Ferrier</td>
<td>YSU</td>
<td>Dudhia</td>
<td>RUC</td>
</tr>
<tr>
<td>ARPS CN</td>
<td>CN</td>
<td>NAMf</td>
<td>Y</td>
<td>Lin</td>
<td>TKE</td>
<td>Two layer</td>
<td>Noah</td>
</tr>
<tr>
<td>ARPS C0</td>
<td>NAMa</td>
<td>NAMf</td>
<td>N</td>
<td>Lin</td>
<td>TKE</td>
<td>Two layer</td>
<td>Noah</td>
</tr>
</tbody>
</table>

* By design, this configuration is not optimized for forecast quality, but is intended to balance the sometimes competing goals of maximizing available resources, providing useful forecasts, and allowing for postseason analysis by multiple studies with different objectives. For example, some members have identical LBC perturbations and some members withhold radar data.
et al. (2011b) are used in the current study. Further details of the other physics and IC–LBC perturbations can be found in Johnson et al. (2011b) and Table 1.

The ensemble forecasts were generated each weekday for 6 weeks between 30 April 2009 and 6 June 2009. Two days are excluded because of missing forecast data and 2 days are excluded because of missing observation data resulting in 26 days of data used in this study. Radar-derived quantitative precipitation estimates from the NSSL Q2 product (Zhang et al. 2005) are the verification data, referred to as observations.

b. Object and attribute identification

The Method for Object-based Diagnostic Evaluation (MODE; http://www.dtcenter.org/met/users; Davis et al. 2006a) is used to identify objects in gridded fields of hourly accumulated precipitation. Skamarock (2004) used kinetic energy spectra to show that the effective resolution of a model is about seven grid points. Using this as an approximate scale of meaningful forecast features, all forecast and observed fields are smoothed with a 4-gridpoint (16 km) averaging radius before defining objects, following Davis et al. (2006a). The purpose of the smoothing is to retain the subjectively important features, while deemphasizing features with a diameter smaller than the effective resolution of the model. Each contiguous area in the smoothed field that exceeds a threshold is then defined as an object. Johnson et al. (2011a) found a 6.5-mm threshold to result in objects that were similar to the authors’ subjective interpretations of convective storms during multiple independent events from the 2009 Spring Experiment. The same threshold is therefore used in this study. The focus of this study on convective scale (both spatial and temporal) systems motivates the use of hourly, instead of longer, accumulation periods, following Davis et al. (2009). Also, following Davis et al. (2006a,b), objects of less than a given area, here 16 grid points, are omitted to reduce the impact on the results from objects smaller than the model’s effective resolution. The effect of this criterion is minimal because the 16-km averaging radius already removes most of such small objects. The effect of each step is demonstrated in Fig. 2. Objects are identified from the raw forecasts, without bias correction, because one of the goals of this study is to provide diagnostic information about the error characteristics that can be used to improve the configuration of individual ensemble members directly.

In the context of the HWT Spring Experiment we focus on attributes relevant for severe weather forecasting, such as shape, size, and area, which can indicate storm morphology or mode (Johnson et al. 2011a). The specific attributes calculated for this study are centroid location, area, aspect ratio (the ratio of minor axis to major axis), and orientation angle (of major axis in degrees clockwise from zonal). Objects with an aspect ratio of 1.0 are circular and objects with decreasing aspect ratio are increasingly linear. The choice of attributes is application dependent and may not be optimal for other applications. Further details about object identification with MODE can be found in Davis et al. (2006a).

In section 3 the object attributes are averaged over all objects to compare the characteristics of forecast and observed objects and forecast objects from different subensembles. The total number of forecast objects is calculated as the average number of objects per member.

c. Quantification of similarity of objects

Objects are compared using the fuzzy logic algorithm described in Davis et al. (2006a, 2009). The degree of similarity for each attribute of a pair of objects is quantified with an interest value $f$ shown in Fig. 3. Attributes with little similarity between objects have a low interest value (Fig. 3). The interest values for all attributes are then combined into a weighted average, called the total interest $I$ for the pair of objects:

$$I = \frac{\sum_{s=1}^{S} c_s w_s f_s}{\sum_{s=1}^{S} c_s w_s}. \quad (1)$$

In Eq. (1), $S$ is the number of object attributes and $c_s$ and $w_s$ are the confidence and weight, respectively, defined below, assigned to the interest value of the $s$th
attribute. Total interest quantifies the overall degree of similarity between two objects with a fuzzy value between 0.0 and 1.0. It is considered fuzzy because a total interest threshold is not applied to make a crisp distinction between matched and unmatched objects. The fuzzy approach has been shown to work well by earlier studies (Davis et al. 2009; Johnson et al. 2011a).

In Eq. (1), each interest value is assigned a constant weight $w$ and a variable confidence value $c$ (Table 2). The weights are equally assigned as 2.0 each to size (area ratio), location (centroid distance), and shape. The weight for shape is further divided into 1.0 each for aspect ratio and orientation angle. Confidence for shape attributes is proportional to centroid distance interest (CDI) and area ratio (AR) because there is little confidence that objects with very different location and/or size represent the same feature. In such instances their shape is irrelevant. Orientation angle is deemphasized for nearly circular objects through a low confidence. Confidence for centroid distance and area ratio is also reduced for objects that are different in size or far away, respectively, for the same reason.

d. Quantification of object-based forecast accuracy

Two measures are used to quantify the similarity between all objects in corresponding forecast and observed fields (i.e., forecast accuracy): the MMI (Davis et al. 2009) and the fuzzy OTS (Johnson et al. 2011a). As described below, the OTS and MMI are sensitive to different aspects of forecast accuracy. Both scores are presented because different users may be interested in model performance from different perspectives.

The MMI is calculated by first determining a maximum total interest [Eq. (1)] for each object in the forecast and the observed fields, when compared to any object in the opposing field. The median of such maximum total interests is then used as an overall measure of the similarity of the two fields (Davis et al. 2009). Thus, all objects contribute equally to the MMI.

The OTS is calculated by first determining pairs of corresponding objects in the forecast and observed fields. Unlike the MMI, the OTS is based on a one-to-one correspondence between forecast and observed objects. Thus, some objects will not have a corresponding object if the forecast and the observed fields have a different number of objects. Corresponding objects are determined by their total interest [Eq. (1)]. A pair of corresponding objects is identified beginning with the most similar (i.e., highest total interest) pair of objects. Those two objects are then removed from consideration and the next most similar pair of objects is identified as the next pair of corresponding objects. The process is repeated until no objects remain. The OTS is then calculated as the summation over all $P$ pairs of corresponding objects of the area of the paired objects ($a_f$ and $a_o$ for forecast and observation object, respectively), weighted by their similarity as quantified by total interest $I$ [Eq. (1)] divided by the total area of all unpaired objects ($A_f$ and $A_o$):
In other words, the OTS is defined as the fraction of the area of all objects that is contained in matched objects, weighted by their degree of similarity. Both the MMI and the OTS have a value of 1.0 for perfect forecasts and a minimum value of 0.0. Unlike the MMI, large objects contribute to the OTS more than small objects [Eq. (2)]. Also unlike the MMI, over- (under) forecasting the number of objects decreases the OTS by resulting in unpaired forecast (observed) objects that contribute to the denominator of Eq. (2) but not the summation in the numerator.

Results are presented using an aggregated MMI and OTS. The MMI is aggregated over multiple forecasts and/or ensemble members by combining the distribution

### Table 2. Attributes and parameter values used for MODE fuzzy matching algorithm [centroid distance (CD), centroid distance interest (CDI), area ratio (AR), and $T$ denotes aspect ratio].

<table>
<thead>
<tr>
<th>Attribute diff between objects</th>
<th>Weight $w$</th>
<th>Confidence $c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centroid distance</td>
<td>2.0</td>
<td>AR</td>
</tr>
<tr>
<td>Area ratio</td>
<td>2.0</td>
<td>1.0 if CD ≤ 160 km</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1 - [(CD - 160)/640]$ if 160 &lt; CD &lt; 800 km</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0 if CD ≥ 800 km</td>
</tr>
<tr>
<td>Aspect ratio diff</td>
<td>1.0</td>
<td>CDI $\times$ AR</td>
</tr>
<tr>
<td>Orientation angle diff</td>
<td>1.0</td>
<td>CDI $\times$ AR $\times$ $\sqrt{a^2 + b^2}$</td>
</tr>
</tbody>
</table>

$a, b = [(T - 1)^2/(T^2 - 1)]^{0.3}$ for the two objects being compared.
of maximum total interests from all forecast days and ensemble members of interest, at the same forecast lead time, before calculating the median. Likewise, the OTS is aggregated over multiple forecasts and/or ensemble members by first calculating the pairs of corresponding objects from each forecast day and member, and then calculating the summation in Eq. (2) over all such pairs. For example, consider two hypothetical ensemble members, both having 10 forecast objects and 10 observed objects. Their individual MMI and OTS are each calculated as a median of 20 maximum total interests and a summation over ten pairs of corresponding objects, respectively. Their aggregated MMI and OTS are calculated as a median of 40 maximum total interests and a summation over 20 pairs of corresponding objects, respectively, instead of averaging the individual scores. Aggregating verification scores instead of averaging the daily scores reduces the sensitivity of the result to small perturbations on days with a small number of forecast objects, respectively. Their aggregated MMI and OTS are each calculated as a median of 40 maximum total interests and a summation over 20 pairs of corresponding objects, respectively. Their aggregated MMI and OTS are each calculated as a median of 20 maximum total interests and a summation over ten pairs of corresponding objects.

### e. Statistical significance tests

The average object attributes are compared between the forecasts and the observations and between the different subensembles. Statistically significant differences are determined at the 95% confidence level using permutation resampling (Wilks 2006; Hamill 1999). Permutation resampling is selected because it is non-parametric and does not require restrictive assumptions about the distribution of the test statistic (Wilks 2006). Each of 1000 resamples is obtained to be consistent with the null hypothesis that there is no difference between the two groups. Each resample is accomplished by randomly assigning each object to one group or the other, as further detailed in Hamill (1999). If the difference between the actual values is larger than 95% of the resampled differences, then the null hypothesis is rejected. Many earlier studies using permutation resampling have grouped all forecasts (e.g., all grid points) from the same day together to ensure independence of the samples (Hamill 1999; Wilks 2006). Such grouping may be excessively stringent when evaluating the statistical significance of average object attributes. For example, Wang and Bishop (2005) combined nearby stations into three separate groups on each day to triple the number of samples for resampling without violating the independence of the samples. Furthermore, unlike traditional gridded forecasts it is not necessarily the case that the attributes of nearby objects from the same forecast are correlated. To determine if this is the case, the correlations of attribute values between 10 000 randomly selected forecast objects and the spatially closest object within the same forecast are calculated and shown in Table 3. The square of the correlation coefficient is the proportion of variability in the attribute of one of the paired objects that is linearly dependent on the attribute of the other paired object (Murphy 1995). Correlation coefficients less than 0.224, therefore, indicate that attributes from nearby objects in the same forecast are at least 95% linearly independent. This criterion is satisfied for attributes of orientation angle, area, and aspect ratio, with the exception of aspect ratio at the 12-h lead time with a correlation of 0.258. However, centroid location has high correlation coefficients as a result of explicitly comparing nearby objects. Table 3 suggests that a less stringent resampling for the attributes of orientation angle, area, and aspect ratio can be obtained by considering each object as an independent sample.

The above resampling method is invalid for location attributes because of the correlation of nearby objects. However, grouping all objects on the same day together may still be unnecessarily restrictive. Clusters of objects are identified by grouping objects that are within 800 km of each other into a common cluster. The correlation of the average centroid location of nearby clusters is below the 0.224 threshold except for zonal centroid location at the 6-h lead time, which has a correlation of 0.338 (Table 4). For centroid location attributes each cluster is therefore considered an independent sample for the permutation resampling. Since the total number of objects is an attribute of the entire field, this attribute is resampled using each day as a single sample.

Statistically significant differences between the aggregated MMI and OTS of different groups of members are also determined using permutation resampling (section 2d; Wilks 2006; Hamill 1999). The accuracy of different objects in the same forecast may be more strongly correlated than their attributes so samples are grouped by day for these tests.

Statistical significance of differences in the frequency that each subensemble (ARW and NMM) contained the highest MMI and OTS is assessed using a binomial test (Panofsky and Brier 1968). The binomial test is chosen...
because it can be performed analytically. The binomial test is based on the null hypothesis that all members have an equal chance of being the best on any given day. Under the null hypothesis there is a 10/18 (8/18) chance that the ARW (NMM) subensemble will contain the best member since there are 18 members considered, 10 ARW, and 8 NMM. Thus, the binomial distribution [Jolliffe 2007, his Eq. (3)] gives the probability that a subensemble will contain the best member on a given number of days \( n \) under the null hypothesis. If such probability, integrated over \( n \geq x \) (or \( n \leq x \)) where \( x \) is the number of days that the group actually had the best member, is less than 2.5\% (for a two-sided test) then the null hypothesis is rejected with 95\% confidence. Both the ARW and NMM subensembles are tested for having a significantly high or low frequency of having the best score, compared to what is expected from random variability.

3. Comparison of object attribute distributions

Object attribute distributions are first compared between all forecasts and observations in section 3a then compared between the ARW and NMM subensembles in section 3b.

a. Forecasts versus observations

1) QUALITATIVE COMPARISON

Figure 4 shows the distributions of the forecast object attributes from all 20 ensemble members and of the observed object attributes. Only the 24-h forecast time is shown because results at other times are qualitatively similar. The distributions are also qualitatively similar for the ARW and NMM subensembles (not shown). Qualitative similarity between the forecast and observed distributions suggests a generally realistic depiction of forecast features.

Four main features of the observed distributions are present in the forecast distributions (Fig. 4). First, the number of objects decreases rapidly with increasing area (Figs. 4a,b). Second, most of the objects have an aspect ratio between 0.4 and 0.9 with the distribution skewed toward more circular objects (Figs. 4c,d). Third, a positive orientation angle (southwest–northeast) is more common than a negative orientation angle (northwest–southeast; Figs. 4e,f). Subjective evaluation of the forecasts suggested that for relatively large objects, the positive angle is consistent with the typical orientation of synoptic-scale cold fronts, as also discussed by Davis et al. (2006a). The more numerous small objects were, however, typically located away from such fronts. Their orientation angle may be at least partly explained by storm propagation during the 1-h accumulation period, which was often toward the northeast in advance of upper-level troughs. For example, at the 12-h lead time, objects identified from instantaneous (i.e., not affected by propagation) reflectivity exceeding 35 dBZ had an average angle of 3.7° and 7.7° for forecasts and observations, respectively, instead of 9.1° and 15.3° for accumulated precipitation. Fourth, there are more objects in the eastern (Figs. 4g,h) and southern (Figs. 4i,j) portions of the domain than the western and northern portions of the domain. A preference for the southern and eastern portions of the domain in both forecast and observed distributions is consistent with an abundance of moisture over the southeast United States resulting from proximity to the Gulf of Mexico and Atlantic Ocean (Hagemeyer 1991).

Although the main features of the observed distributions are found in the forecast distributions, there is a disproportionately large amount of objects in the eastern part of the domain in the forecasts (Fig. 4g) compared to the observations. This apparent bias and other, less obvious, differences are discussed further below using a more quantitative evaluation.

2) QUANTITATIVE COMPARISONS

The average values of each attribute for the forecasts and the observations, along with the total number of objects, are shown as a function of lead time in Fig. 5. Statistically significant differences between the forecast and observed values are indicated with an asterisk along the horizontal axis.

Figure 5a shows the number of objects in the verification domain for each forecast lead time, averaged over the 20 ensemble members and the whole experiment period. The number of objects is significantly overforecast after the 1-h lead time, especially at the 24-h lead time, during the diurnal convective maximum at 0000 UTC (Fig. 5a). Unlike later lead times, at the 1-h
lead time the total number of objects is underforecast. The underforecasting is also seen when only considering the members with radar data assimilation (Table 1) for the full ensemble and both subensembles (not shown). The underforecasting of objects at the 1-h lead time may be a result of suboptimal radar data assimilation requiring a spinup period for storms to fully develop according to the model attractor. Grid spacing also plays a role since the control member was found to overforecast the number of objects at 1-h lead time when grid spacing was reduced from 4 to 1 km (not shown).

Figures 5b–f show the average object area, aspect ratio, orientation angle, zonal (i.e., east–west) centroid location, and meridional (i.e., north–south) centroid location.
location, respectively, of all the forecast and the observed objects at a particular forecast lead time. The forecast objects on average have a smaller area than the observed objects (Fig. 5b). The forecast objects on average are also more circular, except at the 1-h lead time, than the observed objects (Fig. 5c). The orientation angle of the forecast objects have no statistically significant difference from that of the observed objects.
except at the 3-h lead time where the forecast objects have a more zonal average orientation angle than the observed objects (Fig. 5d). Consistent with the eastward bias of the distribution of forecast objects in Figs. 4g,h, which is at the 24-h lead time, the eastward bias is reflected at all lead times, although the difference is only significant at the 1-, 6-, and 30-h lead times (Fig. 5e). For average meridional centroid location, the differences between the forecast and the observed objects are neither statistically significant nor consistent across lead times (Fig. 5f).

In summary, the forecast objects are too numerous (after the 1-h lead time), too small, too circular, and too far east. Overforecasting of small circular precipitation areas was frequently noted subjectively in the eastern and southeastern part of the domain (e.g., Fig. 6). The overforecasting of such objects explains the general overforecasting of the number of objects, while their smaller-than-average size accounts for the smaller average area in forecast objects than observed objects. Since the overforecasted objects tend to be more circular and farther east than average, this also explains the more circular and farther east average forecast centroid location at some lead times. The tendency of the overforecasting to occur in the eastern part of the domain is likely due to the fact that this is where most of the precipitation systems and moisture were located. Previous studies suggest several possible hypotheses for the source of this overforecasting. The case study of Bryan and Morrison (2012) suggests that 4-km grid spacing may be too coarse to allow for sufficient entrainment of dry midlevel air into incipient convective cells. This might cause small circular areas of forecast precipitation where in reality storms should have dissipated. However, further diagnostics using the ARW control (CN) member (Table 1) with 1-km grid spacing showed even greater overforecasting of the number of objects (not shown). A more likely hypothesis is that this feature is a consequence of model physics and dynamics errors, consistent with the dominant impact of model dynamics and physics on forecast clustering in Johnson et al. (2011b). Davis et al. (2009) have also suggested improperly tuned numerical dissipation as a source of excessive small-scale variability, which could contribute to excessive small-scale convection. Additional simulations and sensitivity tests are needed to test these hypotheses, which are beyond the scope of this study.

b. ARW versus NMM members

Johnson et al. (2011b) found clustering of the 2009 SSEFs to primarily correspond to differences in model
dynamics. Therefore, the subensembles of the ARW and NMM members are compared (Fig. 7). Members with the ARW model have fewer objects on average than members with the NMM model except at the 1-h lead time (Fig. 7a). The difference is statistically significant except at the 1- and 18-h lead times. The ARW objects on average have a significantly smaller area and significantly more linear aspect ratio than the NMM objects at most lead times (Figs. 7b,c). The ARW objects have a more southwest-to-northeast average orientation angle than the NMM objects across lead times, but with no statistical significance (Fig. 7d). There are no significant differences between the average ARW and NMM objects in both the meridional and zonal centroid locations (Figs. 7e,f).

Each subensemble is also compared to the observations (Fig. 7). For the total number of objects and average aspect ratio, the ARW objects are more similar to the observations than the NMM objects. However, for the average area, the NMM objects are more similar to the observations than the ARW objects. Neither group is consistently closer to the observations for average orientation angle and centroid location.

The overforecasting of small circular objects noted in section 3a was subjectively more pronounced in the NMM members than the ARW members (not shown), explaining the significantly worse total number and average aspect ratio of the NMM forecast objects. However, the significantly larger average area of NMM than ARW forecast objects is the opposite of what would be expected from this explanation. Johnson et al. (2011a) showed that the total area of NMM objects was on average 56% greater than that of ARW objects at the 24-h lead time (their Table 2). Meanwhile, the total number of objects at the 24-h lead time is on average only 42% greater for NMM objects than ARW objects (Fig. 7a). Therefore, the greater consistency between forecast and observed average area for the NMM subensemble is the result of offsetting errors where greater overforecasting of the size of large objects counteracts greater overforecasting of the number of small objects.

4. Object-based measures of forecast accuracy

a. All 20 ensemble members: OTS versus MMI

Figure 8 shows the aggregated OTS and MMI, as well as the traditional equitable threat score (ETS; Rogers et al. 1995) for comparison. All three measures show rapidly decreasing accuracy in the first 3-h of forecast time as the benefit of radar data assimilation diminishes. All three measures also show a local maximum in accuracy at the 12-h lead time, valid during the diurnal minimum of convective activity in regions with a strong diurnal cycle (Wallace 1975). The OTS is more similar than the MMI to the traditional ETS in that the 12-h lead time OTS–ETS maximum is more pronounced than the MMI maximum and is followed by a skill minimum during the diurnal maximum of convective activity at about the 24-h lead time. Unlike the ETS and the OTS, the MMI shows a maximum in accuracy at the 24-h lead time.

Further diagnostics were conducted to understand why the OTS and the MMI show opposite accuracy at the 24-h lead time, relative to earlier lead times. As described in section 2d, the OTS is weighted by object area. The OTS maximum at the 12-h lead time (valid at 1200 UTC) in Fig. 8 corresponds to a higher total interest of the relatively large forecast objects (i.e., >8192 km$^2$) than at the 24-h lead time (valid at 0000 UTC; Fig. 9a, dotted lines). Since the OTS is weighted by object area, it is dominated by these larger objects, which account for more total area than the smaller objects (Fig. 9b, dotted lines). Because of the small number of days with large objects during the diurnal convective minimum at 1200 UTC the pronounced OTS maximum is not statistically significantly greater than the OTS at the 6- or 18-h lead times.

Different from the OTS, the MMI is calculated with equal weight given to all objects, regardless of size. Subjective examination of forecasts with higher MMI at the 24-h lead time than at the 12-h lead time revealed that the afternoon MMI maximum corresponds to an increase in the number of small objects with high total interest during the afternoon convective maximum. While there were also more large objects at the 24-h lead time than at the 12-h lead time, the increase was more pronounced for smaller objects (Fig. 9b; solid and dashed; note log scale) and the increase in total interest was more consistent and pronounced for the smaller objects (Fig. 9a). The high total interest of small objects at the 24-h lead time is consistent with the subjective appearance of well-forecasted mesoscale regions of environments supportive of discrete cellular convection (not shown). Because of the smaller total area of the more numerous small objects (Fig. 9b) the well-forecasted small objects do not contribute to the OTS as much as to the MMI. The different results obtained with the MMI and OTS show that the most appropriate object-based score for verification depends on the interests of a particular user.

---

1 The ETS is calculated for the 6.5-mm threshold, for consistency with the threshold used to define objects.
Fig. 7. As in Fig. 5, but only for ensemble members using ARW (solid), NMM (dotted), and observed (dashed). Asterisks indicate a significant difference between ARW and NMM.
b. ARW versus NMM members

As suggested by Johnson et al. (2011b) and supported by Fig. 7 there are different average characteristics of the forecast objects from the ARW and NMM members. The ARW and NMM subensembles are therefore also compared in terms of forecast accuracy (Fig. 10). Except at the 1-h lead time, the ARW subensemble has a higher OTS than the NMM subensemble (Figs. 10a,c). In contrast, at the 1-h lead time the NMM subensemble has the higher OTS (Figs. 10a,c). The statistically significantly better performance of the NMM for the first hour is also found when measured by the traditional ETS (Fig. 10e). Subjective examination (e.g., Fig. 11) of several forecasts reveals that for the first hour, the NMM members on average did a better job than the ARW members to retain and evolve the storms produced by the radar data assimilation through the three-dimensional variational data assimilation (3DVAR) and cloud analysis methods (Gao et al. 2004; Xue et al. 2003; Hu et al. 2006). The ARW members with Ferrier microphysics performed particularly poorly in this regard on many cases (e.g., Figs. 11b,h). At later lead times, it is hypothesized that as model biases begin to dominate the effect of the data assimilation, the greater biases in the NMM members dominate the advantage gained at early lead times by the superior maintenance of the assimilated storms. This hypothesis was supported by an additional calculation of the OTS for both subensembles at the 24-h lead time using the bias-adjusted thresholds from Johnson et al. (2011a; their Table 2) to define the MODE objects. The effect of the bias adjustment was to reduce the difference in OTS from 0.062 to 0.016, which was no longer significant, even at the 80% confidence level.

For the MMI, there is no statistically significant difference between the ARW and NMM subensembles at the first hour (Figs. 10c,d). This is explained by the fact that differences in the average maximum and paired total interests are seen only for relatively large objects (not shown). The relative emphasis of the MMI on the smaller objects, therefore, results in the more similar accuracy of the subensembles at the 1-h lead time. The differences in forecast accuracy between the ARW and
Fig. 10. (a) Aggregated OTS for ARW and NMM subensembles, (b) number of days each group contained the member with the highest OTS, (c) aggregated MMI for ARW and NMM subensembles, (d) number of days each subensemble contained the member with the highest MMI, (e) aggregated ETS for ARW and NMM subensembles, and (f) number of days each subensemble contained the member with the highest ETS. Asterisks indicate statistical significance at the 95% confidence level.
NMM subensembles are also more pronounced and more frequently statistically significant with the OTS than with the MMI and ETS at later lead times (Fig. 10). For both the MMI and ETS, the accuracy of the ARW and NMM subensembles shows no statistically significant difference after the 6-h lead time. The results of Figs. 10a,c,e are further confirmed with another measure of forecast accuracy: the number of days that each subensemble contained the most accurate forecast (Figs. 10b,d,f). The better performance of the ARW subensemble than the NMM subensemble is also consistent with the results of Davis et al. (2009) who found a particular ARW model configuration to outperform a particular NMM model configuration in terms of the MMI.

5. Conclusions and discussion

Deterministic forecasts from a convection-allowing ensemble produced by CAPS for the 2009 NOAA HWT Spring Experiment, as well as subensembles configured with different model dynamics (ARW and NMM), are verified in terms of the realism of the averaged object attributes and forecast accuracy. The verification uses the object-based MODE algorithm to emphasize specific forecast features that may not be reflected in traditional gridpoint-based measures. The average realism of objects is quantified by the average value of the attributes of object area, aspect ratio, orientation angle, and centroid location, in addition to the total number of objects. Forecast accuracy is quantified with the
object-based threat score (OTS) and the median of maximum interest (MMI). The main findings of the study are summarized below.

First, the ensemble underforecasts the number of objects at the 1-h lead time. At later lead times, the ensemble overforecasts the number of objects, underforecasts the average object area, overforecasts the average aspect ratio, and has an eastward bias in average location, all of which are explained by an overforecasting of small circular objects, primarily in the eastern part of the domain. Second, the NMM forecast objects were greater in total number and average aspect ratio than the ARW forecast objects. This corresponds to the overforecasting of small circular objects being more pronounced in the NMM members than the ARW members. The ARW objects were on average more similar to the observed objects in terms of these attributes. This is consistent with the fewer false-alarm objects for an ARW model than an NMM model in Davis et al. (2009) where a single physical parameterization scheme, rather than an ensemble of schemes, was used for each of the ARW and NMM models. The NMM forecast objects were larger on average than the ARW forecast objects. This corresponds to a greater precipitation bias in the NMM members that dominates the greater number of small objects. In terms of average area, the NMM objects were more similar than the ARW objects to the observed objects, which was due to the counteracting errors in the size of large objects and the number of small objects for the NMM members. The different relative performance between subensembles, compared to the other attributes, emphasizes that the choice of ensemble member configurations could depend on what attributes are most important for the intended ensemble application. The counteracting errors affecting the area attribute also illustrate the importance of using multiple attributes and verification measures to get a complete and physically descriptive diagnosis of forecast differences.

Third, the MMI and OTS reveal different diurnal cycles of forecast accuracy of the full ensemble, with the OTS suggesting a pronounced maximum at the 12-h lead time and the MMI suggesting a pronounced maximum at the 24-h lead time. Further diagnostics suggest that the difference is because the OTS places more weight on small objects, while the MMI equally weights small and large objects. The different result from different scores emphasizes that the choice of verification score should depend on the intended use of the forecasts and a particular user’s conception of what makes an accurate forecast. For example, if the SSEFs are to be used for point forecasts of severe weather then the MMI may be of greater interest than the OTS. This is because a relatively small supercell storm can have as much, or more, impact on a given location than a larger mesoscale area of convection. However, if the SSEFs are to be used for quantitative precipitation or hydrological forecasting over broad areas then the OTS may be of greater interest than the MMI. This is because objects covering greater area can produce more total precipitation within a particular watershed as well as directly affect a larger number of people.

Fourth, at the 1-h lead time the NMM subensemble on average was more accurate than the ARW subensemble in terms of the OTS. Further diagnostics suggested that the inferior performance of the ARW subensemble at this lead time was mainly due to the ARW members with the Ferrier microphysics, which performed poorly in terms of maintaining the assimilated storms. Cautions need to be taken when including this combination of model dynamics and physics in future SSEF configurations. This difference between the ARW and NMM, measured by OTS, was not present in the MMI.

Finally, at later lead times the ARW members were more accurate than the NMM members. This is consistent with the greater overforecasting of small circular objects in the NMM objects and greater overall precipitation forecast bias for the NMM members. The varied relative performance at different lead times suggests that the optimal design of the SSEFs may depend on the forecast lead time of interest.

Statistical significance tests suggest that our sample size of 26 forecasts is sufficient to identify several differences between forecast and observed objects, and between ARW and NMM member forecasts. These results should not be generalized to other seasons characterized by fundamentally different weather phenomena (e.g., winter cyclones). Furthermore, caution is warranted when extrapolating our results from the 2009 spring season to other seasons because other Spring Experiment ensembles did not use the same configuration of ensemble members and were therefore not evaluated.

Acknowledgments. This research was supported by University of Oklahoma faculty Start-up Award 122-792100 and NSF Grant AGS-1046081. The CAPS real-time forecasts were produced at the Pittsburgh Supercomputing Center (PSC), and the National Institute of Computational Science (NICS) at the University of Tennessee, and were mainly supported by the NOAA CSTAR Program (NA17RJ1227). Fanyou Kong, Ming Xue, Kevin Thomas, Yunheng Wang, Keith Brewster, and Jidong Gao of CAPS are thanked for the production of the ensemble forecasts. Some of the computing for this project was also supported by the National Science Foundation and the Pittsburgh Supercomputing Center.
performed at the OU Supercomputing Center for Education and Research (OSCER) at the University of Oklahoma (OU). The authors are also grateful to NSSL for the QPE verification data, NCAR for making MODE source code available, and the three anonymous reviewers for their valuable comments.

REFERENCES


