Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts

Aaron Johnson
School of Meteorology, University of Oklahoma and Center for Analysis and Prediction of Storms, Norman, Oklahoma

Xuguang Wang
School of Meteorology, University of Oklahoma
and Center for Analysis and Prediction of Storms, Norman, Oklahoma

Fanyou Kong
Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

Ming Xue

School of Meteorology, University of Oklahoma
and Center for Analysis and Prediction of Storms, Norman, Oklahoma

Submitted to
Monthly Weather Review
January 16, 2013
Revised
April 21, 2013

Corresponding author address:
Dr. Xuguang Wang
School of Meteorology
University of Oklahoma
120 David L. Boren Blvd.
Norman, OK, 73072
xuguang.wang@ou.edu

Abstract

Forecasts generated with 1 and 4 km grid spacing using the WRF-ARW model (ARW1 and ARW4, respectively) during the 2009-2011 NOAA Hazadous Weather Testbed Spring Experiments by the Center for Analysis and Prediction of Storms are compared and verified using object-based measures, including average values of object attributes, the Object-based Threat Score (OTS) and the Median of Maximum Interest (MMI). Verification was first performed against observations at scales resolvable by each forecast model and then performed at scales resolvable by both models by remapping ARW1 to the ARW4 grid (ARW1to4). Thirty-hour forecasts of one-hour accumulated precipitation initialized at 0000 UTC on 22, 36 and 33 days during the spring of 2009, 2010 and 2011, respectively, are evaluated over a domain covering most of the central and eastern United States.

ARW1, ARW1to4 and ARW4 all significantly over-forecasted the number of objects during diurnal convection maxima. The over-forecasts by ARW1 and ARW1to4 were more pronounced than ARW4 during the first convection maximum at 1-h lead time. The average object area and aspect ratio were closer to observations for ARW1 and ARW1to4 than for ARW4. None of the models showed a significant advantage over the others for average orientation angle and centroid location. Increased accuracy for ARW1, compared to ARW4, was statistically significant for the MMI but not the OTS. However, ARW1to4 had similar MMI and OTS as ARW4 at most lead times. These results are consistent with subjective evaluations that the greatest impact of grid spacing is on the smallest resolvable objects.

1. Introduction

An advantage of convection-allowing forecasts is the more realistic appearance of convection than when a parameterization scheme is used (Bernardet et al. 2000; Done et al. 2004; Clark et al. 2007; Weisman et al. 2008; Clark et al. 2010; Coniglio et al. 2010). The advantages of decreasing grid-spacing beyond 4 km, relative to the disadvantage of increased computational expense, are still not fully understood and can be application dependent (Weisman et al. 1997; Petch et al. 2002; Adlerman and Droegemeier 2002; Bryan et al. 2003; Roebber et al. 2004; Kain et al. 2008; Lean et al. 2008; Roberts and Lean 2008; Schwartz et al. 2009; Bryan and Morrison 2012). Weisman et al. (1997) argued that explicit depiction of convection at 4 km grid spacing may be sufficient to resolve mesoscale convective features, even though storm-scale details are not fully resolved. Computational requirements currently preclude real-time operational forecasts with sub-kilometer grid spacing over a domain large enough to resolve meso- and synoptic-scale features as well. However, within the range of 1-4 km grid spacing Lean et al. (2008) found greater accuracy with 1 km than 4 km grid spacing. Roberts and Lean (2008) also found an acceptable level of skill was attained on smaller spatial scales with 1 km grid spacing than with 4 km grid spacing. However, Kain et al. (2008) and Schwartz et al. (2009) found little added value of 2 km, compared to 4 km, grid spacing for second day precipitation forecasts. Clark et al. (2012) showed examples of more realistic storm structures, in terms of simulated reflectivity for a squall line case and updraft helicity for a supercell case, with 1 km grid spacing than with 4 km grid spacing for ~1 day lead time. However, a systematic advantage of 1 km grid spacing was not subjectively reported by the forecasters.

The impact of grid spacing in the range of 1-4 km has been mainly evaluated using subjective and traditional grid point or neighborhood-based measures, with an exception of Kain

et al. (2008) which also verified the number of contiguous "reflectivity entities". Objective object-based methods (e.g., Ebert and McBride 2000; Davis et al. 2006a), where attributes such as size, shape and location of objects are evaluated, have also been proposed and applied to convection-allowing forecasts (e.g., Davis et al. 2006b, 2009; Ebert and Gallus 2009; Gallus 2010; Johnson et al. 2011ab; Johnson and Wang 2012, 2013). These studies have shown that the object-based methods can reduce the sensitivity to precise grid point location while also being sensitive to storm features and characteristics in a way that mimics subjective evaluation.

Here, the impact of 1 km vs. 4 km grid spacing in the 2009-2011 National Oceanic and Atmospheric Administration Hazardous Weather Testbed (NOAA HWT) Spring Experiment forecasts is evaluated using object-based methods. Two perspectives on the evaluation are considered. First, the averaged number, size, shape and location of forecast objects over the entire domain and forecast period are compared with that of the observed objects. Second, the accuracy of fields of forecast objects, compared to the corresponding fields of observed objects at the same time, is evaluated over the forecast period using two object-based verification measures. The forecast and observation data and the object-based methods are described in section 2. Results of the average object attribute verification are presented in section 3 and results of the object-based accuracy measures are presented in section 4. Section 5 contains a brief summary and conclusions.

2. Data and Methods

a. Forecast and verification data

During the 2009-2011 NOAA HWT Spring Experiments, the Center for Analysis and Prediction of Storms (CAPS) produced experimental real-time convection-allowing ensemble

forecasts over a near-CONUS (CONtinental United States) domain, initialized at 00 UTC (Xue et al. 2009; Kong et al. 2009; Weiss et al. 2010, 2011). The control member for the ensembles was run with the Advanced Research Weather Research and Forecasting (WRF-ARW, or ARW; Skamarock et al. 2005; 2008) model at 4 km grid spacing (i.e., ARW4), with Thompson et al. (2008) microphysics, Goddard shortwave radiation (Tao et al. 2003), Noah land surface model (Ek et al. 2003) and Mellor-Yamada-Janjic boundary layer (Mellor and Yamada 1982; Janjic´ 1994) parameterizations. There was no cumulus parameterization, sub-grid turbulent mixing or explicit computational mixing (Xue et al. 2009). The initial condition analysis background and the lateral boundary conditions were obtained from the operational North American Mesoscale (Rogers et al. 2009) model. Additional surface and radar data were assimilated using the Advanced Regional Prediction System 3DVAR and cloud analysis package (Gao et al. 2004; Xue et al. 2003; Hu et al. 2006). A separate high-resolution forecast, ARW1, was generated with an identical configuration as ARW4, except with 1 km grid spacing. Forecasts between the last week of April and the middle of June were generated on 22, 36 and 33 days in 2009, 2010 and 2011 respectively. The WRF-ARW version was 3.0.1, 3.1.1 and 3.2.1 in 2009, 2010 and 2011, respectively. The upgrade from version 3.0.1 to 3.1.1 in 2010 included an upgrade from the WRF version 2 Thompson microphysics scheme (Skamarock et al. 2005) to the WRF version 3 Thompson microphysics scheme (Skamarock et al. 2008). The forecast domain was enlarged in 2010 and 2011, compared to 2009, but a common verification domain shown in Fig. 3 is used in this study.

Quantitative precipitation estimates (QPEs) from the National Severe Storms Laboratory (NSSL) Q2 product (Zhang et al 2011) on a .01 degree (~1 km) grid are used as the verification data. Before generating the verifying objects (i.e., observed objects) for the ARW1 forecasts, the

verifying QPE data are bilinearly interpolated to the same ARW1 model grid. For the ARW4 forecasts, the verifying QPE data are smoothed with a 2-D Gaussian filter with 1.5 km[1] standard deviation before being bilinearly interpolated to the ARW4 model grid. The filtering is to avoid potential aliasing errors with the bilinear interpolation (Weaver 1983 pg. 288), although results are not sensitive to the filtering due to further smoothing when defining objects (section 2b, below). The purpose of interpolating the verification data to the corresponding forecast model grid is to avoid biasing the results towards the observation resolution. Therefore the method of obtaining forecast objects is identical to the method of obtaining the observed objects to which the forecast objects are compared.

An overview of the domain average precipitation during the 2009-2011 Spring Experiments is shown in Fig. 1. All seasons exhibited a similar observed precipitation diurnal cycle with maxima at early lead times and at ~24 h (both around 0000 UTC), and minima at ~15 h (1500 UTC). The forecast precipitation was generally similar between ARW1 and ARW4 except for substantially greater precipitation in ARW1 at 1-3 h then slightly less precipitation at ~5-15 h. The difference at early lead times may be due to an enhanced impact of horizontal resolution during the model "spin-up" period when the precipitation systems introduced with radar data assimilation are adjusting to the model dynamics. For example, Lean et al. (2008) found reduced spin-up time for precipitation with a 1 km model compared to a 4 km model. Johnson and Wang (2013) also found the precipitation forecasts at early lead times to be sensitive to other aspects of the model configuration such as model dynamics core and microphysics scheme. There were some changes in forecast bias among the seasons. Under-forecasting during hours ~5-15 was more pronounced in 2010-2011 than in 2009, whereas over-

[1] 1.5 km was used to provide a cutoff wavelength (defined as a 1/4 power reduction in the filtered field at the cutoff wavelength; Sakmann and Neher 2009, pg. 485), of 2 times the grid spacing of the 4 km grid (i.e., 8 km).

forecasting during the 18-30 h diurnal cycle maximum was more pronounced in 2009 than 2010-2011. The differences in forecast biases among seasons may be related to the upgrades of the WRF model version and/or the different flow patterns characterizing each season.

b. Definition of objects and attributes

The Method for Object based Diagnostic Evaluation (MODE, available at http://www.dtcenter.org/met/users; Davis et al. 2006a) is used to identify forecast and observed objects in gridded fields of hourly accumulated precipitation. Features smaller than the effective resolution of the model are removed by first averaging over a 4 grid point radius (16 km for ARW4 and 4 km for ARW1), following Davis et al. (2006a,b). Each contiguous area in the smoothed field that exceeds a certain threshold is then defined as an object. Following Johnson et al. (2011a), a threshold of 6.5 mm h^{-1} is used to mimic subjective identification of distinct convective precipitation systems. This threshold and smoothing was used in Johnson et al. (2011a) in the similar context of determining the locations, modes and organization of resolvable convective features for the purpose of severe storm forecasting. Following Davis et al. (2006a,b), objects of less than a pre-specified area, here 16 (i.e., 4x4) grid points, are omitted to remove objects smaller than the effective resolution of the model forecasts (Skamarock 2004)[2]. This procedure is intended to only evaluate the features that the models can effectively resolve at their own resolution.

After defining the objects, attributes describing each object are then calculated. 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 mode, following Johnson et al. (2011a) and Johnson and Wang (2013). The specific attributes calculated for this study are

[2] The effect of this criterion is minimal because the 4 point averaging radius already removes most of such small objects.

centroid location, area, aspect ratio (the ratio of minor axis to major axis), and orientation angle (of major axis in degrees counter-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 also be found in Davis et al. (2006a).

c. Object-based forecast accuracy measures

Objects are compared using a fuzzy logic algorithm based on Total Interest which quantifies the similarity of paired forecast and observed objects (Davis et al. 2006a; 2009). The degree of similarity for each attribute of a pair of objects is quantified with an interest value, *f*. Attributes with little similarity between objects have a low interest value (see Fig. 3 in Johnson and Wang 2013). The interest values for all attributes are then combined into a weighted average, called the Total Interest, *I*, for the pair of objects:

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

In Eq. 1, *S* is the number of object attributes (here, 4) and c_s and w_s are the confidence and weight, defined below and in Table 2 of Johnson and Wang (2013), assigned to the interest value of the s^{th} attribute. 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 ("DEFAULT" in Table 1). The confidence values, described in greater detail in Johnson and Wang (2013) (their Table 2), can be thought of as an additional

weight that is not constant. Total Interest quantifies the overall degree of similarity between two objects with a fuzzy value between 0.0 and 1.0.

Whereas Total Interest, *I*, quantifies the similarity between two objects, two scores are adopted to quantify the similarity between two fields with many objects. The Median of Maximum Interest (MMI; Davis et al. 2009) is calculated by determining a maximum Total Interest for each object in the forecast and observed fields, then finding the median of all maximum Total Interests. The maximum Total Interest is the highest Total Interest that can be obtained by pairing the object with any object in the field to which it is being compared.

The second score adopted is the Object-based Threat Score (OTS; Johnson et al. 2011a):

$$OTS = \frac{1}{A_f + A_o} \left\{ \sum_{p=1}^{P} I^p (a_f^p + a_o^p) \right\} \quad (2)$$

The OTS is calculated by first determining pairs of corresponding objects in the forecast and observed fields. Then the summation over all *P* pairs of corresponding objects of the area of the paired objects (In Eq. 2, a_f and a_o for forecast and observed object, respectively) is calculated. The area of each paired object is weighted by the pair's Total Interest (I^P). The summation is normalized by the total area of all objects (A_f and A_o). Thus, the OTS is the fraction of object area contained in paired objects, weighted by the degree of similarity of those paired objects. Both the MMI and OTS have a value of 1.0 for perfect forecasts and a minimum value of 0.0. Unlike the MMI, a large object contributes to the OTS more than a small object.

Statistically significant differences are determined at the 95% confidence level using permutation resampling which does not require restrictive assumptions about the distribution of the test statistic (Wilks 2006; Hamill 1999). For the verification of average object attributes each

object attribute of area, aspect ratio and orientation angle is considered an independent sample, due to low correlation of nearby object attributes, while centroid location attributes are grouped together if they are within 800 km of another object, following Johnson and Wang (2013). For the 1 km grid spacing forecasts and observations the aspect ratio attributes for objects within 800 km of each other are also grouped together due to insufficiently low correlation of nearby object aspect ratios (not shown). For the verification with OTS, MMI and the total number of objects, all objects from the same forecast are grouped together as an independent sample, also following Johnson and Wang (2013). Further details of the resampling method can be found in Johnson and Wang (2013). Both the object-based accuracy measures and the attribute-based statistics are calculated at forecast lead times of 1, 3, 6, 12, 18, 24 and 30 hours, from forecasts all initialized at 0000 UTC.

3. Verification of averaged object attributes

In general, ARW1 has more objects with smaller average area and smaller (less circular) average aspect ratio than ARW4 (Fig. 2). Smaller and more irregular features with finer resolution are subjectively apparent in Fig. 3 which is representative of the objectively identified differences between ARW1 and ARW4 described below.

Compared to their own verifying observations, both forecast models over-forecast the number of objects at the 3, 18, 24 and 30 h lead times (valid at 0300, 1800, 0000 and 0600 UTC, respectively; Fig. 2a). Thus the over-forecasting is most pronounced and most significant during the diurnal convective maxima. The over-forecasting during the first convective maximum around 1 h lead time is more pronounced for ARW1 than ARW4 with the former showing a statistically significant difference whereas the latter does not show a significant difference from observations (Fig. 2a). While both ARW1 and ARW4 forecast objects have significantly smaller

average area than observed, the difference is less pronounced for ARW1 than ARW4 (Fig. 2b). ARW4 objects are consistently and significantly more circular than the corresponding observed objects at most lead times (Fig. 2c). In contrast, the difference between ARW1 and OBS1 average aspect ratio is significant only at the 1 and 12 h lead times. The too circular (linear) average ARW1 aspect ratio at the 1 h (12 h) lead time is consistent with the over- (under-) forecasting of the number of objects (Fig. 2a) being associated mainly with smaller objects which tend to be more circular than larger objects (e.g., Fig. 4). The more circular average shape of ARW4 objects than ARW1 objects, relative to their own corresponding observations, is likely a result of 4 km grid spacing being too coarse to resolve the observed irregular shapes of many of the smaller objects. The differences in orientation angle and centroid location are generally not statistically significant at most lead times (Fig. 2d,e,f). However, at early lead times (i.e., 1 h in the zonal direction and 1-3 h in the meridional direction) there is a southeastward bias in average centroid location for both ARW1 and ARW4. The bias is only slightly reduced by the finer grid spacing of ARW1 and remains significant in the meridional direction.

The total number of objects and average aspect ratio are also evaluated separately for objects of different sizes at the 24 h lead time which is representative of other lead times as well (Fig. 4). The larger number of objects in ARW1 is primarily due to a larger number of small objects that are not resolved by ARW4 (Fig. 4a). The increase in the number of small objects is balanced by a decrease in the number of larger objects such that the total area of objects in ARW1 and ARW4 is similar (not shown), consistent with Kain et al. (2008). For both ARW1 and ARW4, the over-prediction of the number of objects occurs at all sizes and peaks around 32-128 km^2 for ARW1 and around 512-2048 km^2 for ARW4 (Fig. 4a,b). The maximum number of objects occurs at approximately the same grid-relative size of 32-128 km^2 (i.e., 32-128 grid

squares) for ARW1 compared to 512-2048 km^2 (i.e., 32-128 grid squares) for ARW4. The ARW1 objects are less circular than the ARW4 objects. These effects of grid spacing on forecasts extend to objects larger than 4000 km^2 (Fig. 4). Similar characteristics are also seen in the corresponding ARW1 and ARW4 verifying observations (Fig. 4 dashed lines), which justifies comparing the forecast objects with the observed objects defined using the corresponding grid spacing (section 2b).

The sensitivity of these results to the use of a 6.5 mm h^{-1} precipitation threshold to define objects was evaluated by calculating the identical statistics as in Fig. 2 using precipitation thresholds of 12.7 mm h^{-1} and 2.54 mm h^{-1} (not shown). The values of the attributes were affected by the threshold but the relative differences between ARW1 and ARW4 and their corresponding observations were similar to Fig. 2. This consistency suggests that the comparison of ARW1 and ARW4 object attributes shown in Fig. 2 is robust to the choice of precipitation threshold. The relative differences between ARW1 and ARW4 and between the forecasts and the verifying observations in object attribute statistics were also generally consistent during the 3-season period of study.

In the previous comparison, ARW1 and ARW4 are verified against corresponding observed objects defined on the same grid as the forecast objects in order to emphasize spatial scales resolvable by each forecast model. To further compare ARW1 and ARW4 on scales that both can resolve, the object attributes are also evaluated after first remapping the ARW1 forecasts to the same 4 km grid as ARW4 (i.e., ARW1to4) (Fig. 5). The remapping consists of taking the average value of the 16 ARW1 grid boxes in the same area as each ARW4 grid box. The attributes of average object area and average aspect ratio are more similar to the observed objects for ARW1to4 than for ARW4 (Fig. 5b,c). Like ARW1 (Fig. 2a), ARW1to4 shows

greater over-forecasting of the number of objects than ARW4 during the diurnal convective maxima, especially at the 1 h lead time. Although the ARW1to4 objects were on average farther south-east than the ARW4 objects (Fig. 5e,f) due to the over-forecasting of the number of objects being most common in the southeast part of the domain (Johnson and Wang 2013), neither ARW1to4 nor ARW4 was consistently or statistically significantly more similar to the observations for orientation angle and centroid location at most lead times (Fig. 5d,e,f).

4. Verification of forecast accuracy

In contrast to the systematic agreement of forecast and observed objects considered in the previous section (i.e., forecast realism), this section evaluates the day-to-day agreement between forecast and observed objects (i.e., forecast accuracy) using the object-based verification measures OTS and MMI. For the OTS, there is not a statistically significant difference between ARW1 and ARW4 (Fig. 6a). However, the ARW1 OTS is greater than the ARW4 OTS at all but the 12 h lead time (Fig. 6a), when significant under-forecasting of the total number of objects is found for ARW1 only (Fig. 2a). For the MMI, ARW1 is significantly better than ARW4 at all lead times (Fig. 6b). As defined in section 2c, the MMI is impacted equally by all objects while the OTS is impacted more strongly by the accuracy of large objects than the accuracy of small objects. Therefore, the greater magnitude and statistical significance of the differences in MMI than the differences in OTS indicate that the impact of horizontal grid spacing mainly affects the accuracy of the smaller objects.

An example of the greater accuracy of the smallest objects for ARW1 than ARW4 can be seen in Figure 3. As also noted subjectively on several other cases (not shown), there are often large regions over which the observation corresponding to ARW4 shows few or no objects but the observation corresponding to ARW1 shows some very small objects (e.g., the Ohio River

Valley region in Fig. 3). These observed objects are absent with 4 km grid spacing because they are smaller than the effective resolution of the ARW4 forecasts (see also section 2b). In such cases the ARW1 forecasts are able to resolve similar storms at the same scale as the observed storms (e.g., Fig. 3a,b) but the ARW4 forecasts release the static instability on an unrealistically large spatial scale (e.g., Fig. 3c,d). In general, the maximum Total Interest of the relatively small ARW1 objects is larger, on average, than that of the similar sized ARW4 objects while little difference is found for the larger objects (not shown).

To further evaluate the impact of increasing the grid spacing from 4 km to 1 km on the forecast accuracy for the spatial scales resolved by both forecast models, the ARW1 forecasts are again remapped to the ARW4 grid (i.e., ARW1to4). ARW1to4 and ARW4 are then both compared to the observations on the grid with 4 km grid spacing. The relative accuracy of the forecast models on scales that are resolvable by both was then evaluated by comparing the ARW4 OTS and MMI to ARW1to4 instead of ARW1. The OTS and MMI are both not significantly different between ARW4 and ARW1to4 at most lead times (Fig. 7). However, there is a significantly higher OTS and MMI for ARW4 than ARW1to4 at the 1 h lead time (Fig. 7a,b) and for the MMI at the 12 h lead time (Fig. 7b). The 1 h lead time corresponds to greater over-forecasting of precipitation amount and total number of objects for ARW1 than ARW4 (Fig. 1 and 2a). The 12 h lead time corresponds to significantly greater under-forecasting of the number of objects for ARW1 than ARW4 (Fig. 2a). The loss of the MMI advantage for ARW1 when remapping the forecasts to the 4 km grid therefore further supports the suggestion of Kain et al. (2008) that the advantage of increasing grid spacing beyond 4 km for convection forecasts is primarily on scales that are not fully resolvable with 4 km grid spacing while having little impact on the larger and more predictable scales.

The OTS and MMI are not evaluated at other thresholds because the Total Interest parameters, chosen for the specific application described in section 2b, were determined to mimic a subjective evaluation of object similarity at the 6.5 mm h^{-1} threshold only. Instead of precipitation threshold, the sensitivity to the weighting factors applied to each attribute for the Total Interest calculation is evaluated for the OTS and MMI results (Table 1). The weighting factors were perturbed by giving more weight to area and less weight to location (test 1), more weight to location and less weight to area (test 2), more weight to aspect ratio and less weight to orientation angle (test 3), and more weight to orientation angle and less weight to aspect ratio (test 4). The OTS and MMI were affected by these changes (Fig. 8) but the impact was similar for ARW1 and ARW4 so that the main results comparing the accuracy of ARW1 and ARW4 are not sensitive to these similar but different choices of the MODE parameters. The comparison of ARW1to4 and ARW4 showed similarly low sensitivity to the same variations of weighting factors (not shown). The relative accuracy of ARW1 and ARW4 was consistent from season to season (not shown), although the times of statistical significance depended on the season and sample size.

5. Conclusion and discussion

The convection-allowing forecasts at 1 km and 4 km grid spacing (ARW1 and ARW4, respectively) produced by CAPS during the 2009-2011 NOAA HWT Spring Experiments were verified and compared using two object-based methods. The comparisons were performed for observations at scales resolvable by each forecast model (e.g., ARW1 vs OBS1 and ARW4 vs OBS4) as well as at the scales resolvable by both models (e.g., ARW1to4 vs OBS4 and ARW4 vs OBS4). The former allows a comparison that retains all resolvable features in the forecast model without biasing the results towards the resolution of the observation data. The latter

attempts to answer the question of whether the effects of moving from 4 km to 1 km grid spacing are limited to the scales that are unresolvable at 4 km grid spacing or whether there are also effects on scales resolved with both grid spacings.

First, a comparison of forecast object attributes was used to assess the impact of grid spacing on the average realism of precipitation forecast features of interest to severe weather forecasting. Both ARW1 and ARW4 over-forecasted the total number of objects during the diurnal convective maxima. For the first convective maximum around the 1 h lead time, the over-forecasting by ARW1 was more pronounced than by ARW4. In contrast to the total number of objects, the attributes of average area and average aspect ratio predicted by ARW1 were more similar to the corresponding observations than those predicted by ARW4. Both models showed a southeastward bias at very early lead times. However, neither ARW1 nor ARW4 was more similar to their verifying observations than the other. When evaluating only the scales that both models can resolve, ARW1to4 showed greater over-forecasting of the total number of objects than ARW4 during the convective maxima, especially at the 1 h lead time. However, ARW1to4 more realistically forecasted the average area and aspect ratio of objects than ARW4.

Second, object-based accuracy measures, OTS and MMI, were used to evaluate the relative accuracy of individual forecasts, instead of the average characteristics of all objects. There was not a significant difference between the OTS of ARW1 and ARW4 but the MMI was significantly better for ARW1 than ARW4 at all lead times. This difference between OTS and MMI was caused by the greater relative sensitivity of MMI than OTS to the accuracy of small objects. When only objects resolvable by both forecast models were evaluated, the differences in MMI between ARW1to4 and ARW4 were reduced compared to the differences between ARW1

and ARW4. ARW1to4 was actually significantly worse than ARW4 at the 1 h lead time for the OTS and MMI and at the 12 h lead time for MMI. These differences correspond to significant over-forecasting (under-forecasting) at the 1 h (12 h) lead time for ARW1 only. Therefore the advantages of 1 km vs 4 km grid spacing, in terms of the accuracy of precipitation forecast features of interest to severe weather forecasters, occur primarily on the scales resolved with 1 km grid spacing but not with 4 km grid spacing.

This study provides systematic and automated quantitative analysis of forecast characteristics relevant to our particular application. The object-based method was applied in the context of forecasting the locations, modes and organization of resolvable convective features. While the ability to choose parameters appropriate for this application can be considered an advantage of the object-based approach, it also limits the ability to generalize the results to other applications. For example, hydrology users may be interested in the maximum intensity attribute of objects, while forecasters interested in convective initiation may not be interested in the smoothing or object area criteria and may use OBS1 when verifying ARW4. However, our results were found to be relatively insensitive to our choices of object-based parameters. The object attribute results were similar at different thresholds and the forecast accuracy results were similar using different weighting factors in the Total Interest calculation (Eq. 1).

Acknowledgments.

This study was supported by the NSF AGS-1046081 award. Computing resources at the OU Supercomputing Center for Education & Research (OSCER) were used for this study. Jian Zhang and Youcun Qi of NSSL are acknowledged for making the QPEs available. Kevin

Thomas, Yunheng Wang, Keith Brewster, and Jidong Gao of CAPS are thanked for their efforts in producing the ensemble forecasts, mostly with support from grant NA17RJ1227 of the NOAA CSTAR program and NSF grant AGS-0802888, using computing resources at the National Institute of Computational Science (NICS). The authors are also grateful for the helpful comments and suggestions from three anonymous reviewers.

References

Adlerman, E. J., K. K. Droegemeier, 2002: The sensitivity of numerically simulated cyclic mesocyclogenesis to variations in model physical and computational parameters. *Mon. Wea. Rev.*, **130**, 2671–2691.

Bernardet, L. R., L. D. Grasso, J. E. Nachamkin, C. A. Finley, and W. R. Cotton, 2000: Simulating convective events using a high-resolution mesoscale model. *J. Geophys. Res.,* **105,** 14 963–14 982.

Bryan, G. H., J. C. Wyngaard, J. M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Mon. Wea. Rev.*, **131**, 2394–2416.

Bryan, G. H. and H. Morrison, 2012: Sensitivity of a simulated squall line to horizontal resolution and parameterization of microphysics. *Mon. Wea. Rev.* **140**, 202-225.

Clark, A. J., W. A. Gallus, T. Chen, 2007: Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Mon. Wea. Rev.*, **135**, 3456–3473.

Clark, A. J., W. A. Gallus, M. Xue, F. Kong, 2010: Convection-allowing and convection-parameterizing ensemble forecasts of a mesoscale convective vortex and associated severe weather environment. *Wea. Forecasting*, **25**, 1052–1081.

Clark, Adam J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74.

Coniglio M. C., K. L. Elmore, J. S. Kain, S. J. Weiss, M. Xue, M. L. Weisman, 2010: Evaluation of WRF model output for severe weather forecasting from the 2008 NOAA Hazardous Weather Testbed Spring Experiment. *Wea. Forecasting* **25**:2, 408-427.

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

Davis, C. A., B. G. Brown, R. Bullock, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795.

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

Done, J., C. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using weather research and forecast (WRF) model. *Atmos. Sci. Lett.*, **5,** 110–117.

Ebert, E., and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.*, **239,** 179–202.

Ebert, E. E., W. A. Gallus, 2009: Toward better understanding of the Contiguous Rain Area (CRA) method for spatial forecast verification. *Wea. Forecasting*, **24**, 1401–1415.

Ek, M. B., K. E. Mitchell, Y. Lin, P. Grunmann, E. Rogers, G. Gayno, and V. Koren, 2003: Implementation of the upgraded Noah land-surface model in the NCEP operational mesoscale Eta model. *J. Geophys. Res.,* **108,** 8851.

Gallus, W. A., 2010: Application of object-based verification techniques to ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 144–158.

Gao, J.-D., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Ocean. Tech.*, **21,** 457-469.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167.

Hu, M., M. Xue, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D level-II data for the prediction of Fort Worth tornadic thunderstorms. Part I: Cloud analysis and its impact. *Mon. Wea. Rev.*, **134,** 675-698.

Janjic´, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945.

Johnson, A., X. Wang, F. Kong, M. Xue, 2011a: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of object-oriented cluster analysis method for precipitation fields. *Mon. Wea. Rev.,* **139**, 3673-3693.

Johnson, A., X. Wang, M. Xue, and F. Kong, 2011b: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part II: Season-long ensemble clustering and implication for optimal ensemble design. *Mon. Wea. Rev.,* **139**, 3694-3710.

Johnson, A., 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.

Johnson, A., X. Wang, 2013: Object-based evaluation of a storm scale ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment. *Mon Wea Rev.,* 141, 1079-10983

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952.

Kong, F., M. Xue, K.W. Thomas, J. Gao, Y. Wang, K. Brewster, K.K. Droegemeier, J. Kain, S. Weiss, D. Bright, M. Coniglio, and J. Du, 2009: A real-time storm-scale ensemble forecast system: 2009 Spring Experiment, *10th WRF Users' Workshop, NCAR Center Green Campus*, Boulder, CO, June 23-26, 2009, Paper 3B.7.

Lean, H. W., P. A. Clark, M. Dixon, N. M. Roberts, A. Fitch, R. Forbes and C. Halliwell, 2008: Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Wea. Rev*., **136**, 3408-3424.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys*., **20**, 851-875.

Petch, J. C., A. R. Brown, and M. E. B. Gray, 2002: The impact of horizontal resolution on the simulations of convective development over land, *Q. J. R. Meteorol. Soc*. **128**, 2031–2044.

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

Roebber, P. J., D. M. Schultz, B. A. Colle, D. J. Stensrud, 2004: Toward improved prediction: high-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949.

Rogers, E., and Coauthors, 2009: The NCEP North American Mesoscale modeling system: Recent changes and future plans. Preprints, 23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction, Omaha, NE, Amer. Meteor. Soc., 2A.4.

Sakmann, B. and E. Neher, 2009: *Single-Channel Recording*. Springer-Verlag New York. 730 pp.

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

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

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the advanced research WRF version 2. NCAR Tech Note NCAR/TN-468_STR, 88 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO 80307.].

Skamarock, W. C., and coauthors, 2008: A description of the advanced research WRF version 3. NCAR Tech Note NCAR/TN-475_STR, 113 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO 80307.].

Tao, W.-K., and Coauthors, 2003: Microphysics, radiation, and surface processes in the Goddard Cumulus Ensemble (GCE) model. *Meteor. Atmos. Phys.*, **82**, 97–137.

Thompson, G., 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.

Weaver, H. J., 1983: *Applications of Discrete and Continuous Fourier Analysis*. J. Wiley and Sons, 375 pp.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125,** 527-548.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23**, 407–437.

Weiss, S. J., and Coauthors, 2010: NOAA Hazardous Weather Testbed Experimental Forecast Program Spring Experiment 2010: Program overview and operations plan. NOAA, 64 pp. [Available online at http://hwt.nssl.noaa.gov/Spring_2010/ Spring_Experiment_2010_ops_plan_21May.pdf]

Weiss, S. J., and Coauthors, 2011: NOAA Hazardous Weather Testbed Experimental Forecast Program Spring Experiment 2011: Program overview and operations plan. NOAA, 62 pp. [Available online at http://hwt.nssl.noaa.gov/Spring_2011/ Spring_Experiment_2011_ops_plan_13May_v5.pdf]Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences: An Introduction*. 2nd ed. Academic Press, 467 pp.

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

Xue, M., D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Physics*, **82,** 139-170.

Xue, M., F. Kong, K. W. Thomas, J. Gao, Y. Wang, K. Brewster, K. K. Droegemeier, X. Wang, J. Kain, S. Weiss, D. Bright, M. Coniglio, and J. Du, 2009: CAPS realtime 4-km multi-

model convection-allowing ensemble and 1-km convection-resolving forecasts from the NOAA Hazardous Weather Testbed 2009 Spring Experiment. *Extended Abstract, 23rd Conf. Wea. Anal. Forecasting/19th Conf. Num. Wea. Pred. Amer. Meteor. Soc.,* Paper 16A.2.

Zhang, J., and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) system?Description, results and future plans. Bull. Amer. Meteor. Soc., 92, 1321-1338.

List of Figures

FIG. 6. Forecast accuracy of ARW1 (dashed) and ARW4 (solid) as measured by the (a) OTS and (b) MMI, aggregated over all 91 forecasts. Statistically significant differences between ARW1 and ARW4 are indicated with asterisks along the bottom horizontal axis.

FIG. 7. As in Figure 6, except for ARW4 (solid) and ARW1to4 (dashed).

FIG 8. As in Figure 6, except for the alternate choices of attribute weights defined in Table 1.

TABLE 1. Alternate values of attribute weight used for calculating Total Interest when testing the sensitivity of the OTS and MMI results to the choice of weights. Each row shows the weighting factors and times of statistically significant differences between ARW1 and ARW4 for the default parameters used for the presented results and four sensitivity tests. The first column labels each test for consistency with Fig. 8, the next four columns indicate the weight applied to each attribute, the sixth column lists the lead times at which the OTS is significantly different between ARW1 and ARW4 for each test, and the seventh column is like the sixth, but for the MMI.

Experiment	Area ratio	Aspect ratio dif.	Angle dif.	Centroid dist.	OTS sig.	MMI sig.
DEFAULT	2	1	1	2	none	1-30
TEST1	3	1	1	1	30	1-30
TEST2	1	1	1	3	none	1-30
TEST3	2	0	2	2	none	1-30
TEST4	2	2	0	2	none	1-6,18-30

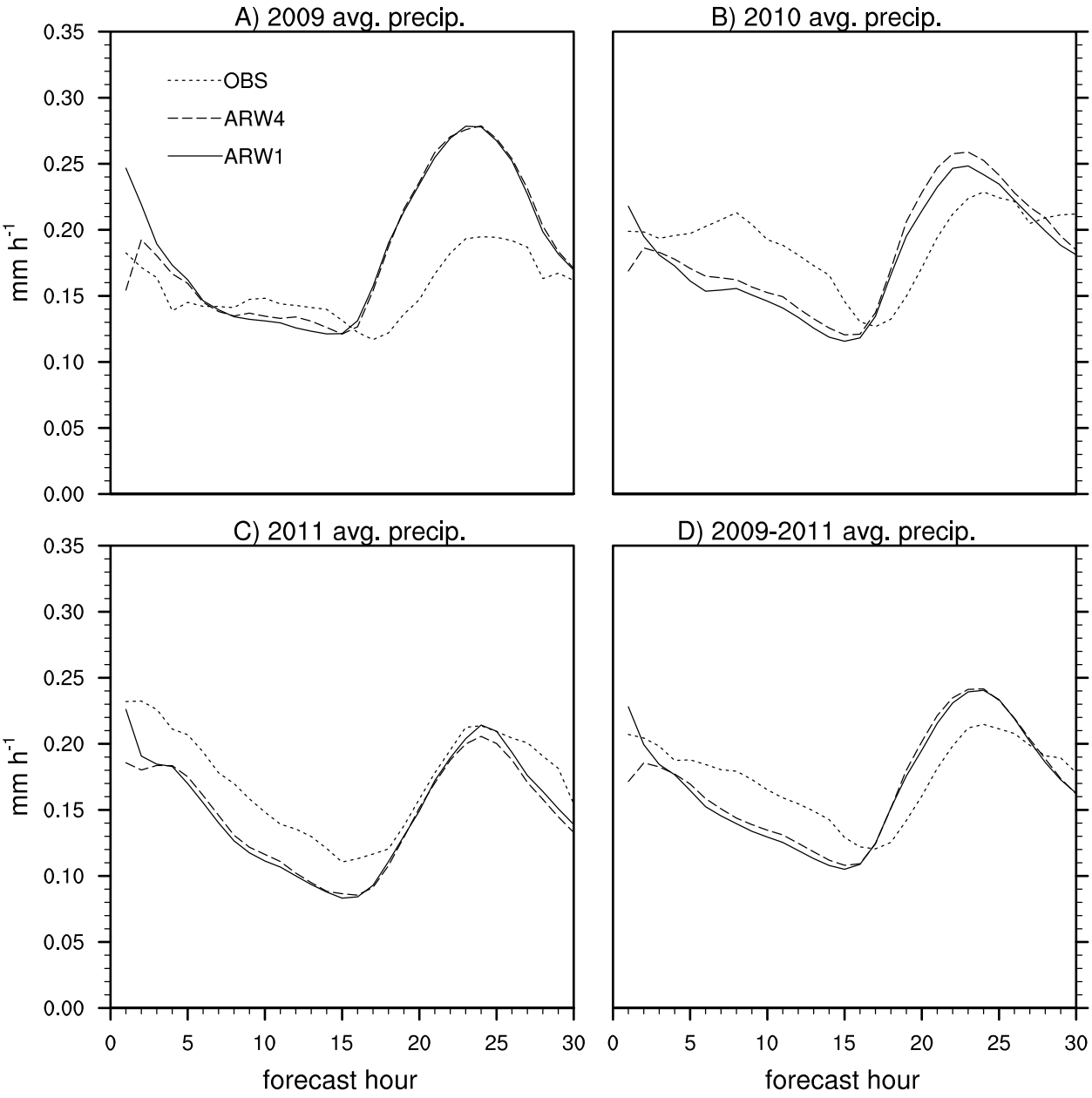

FIG. 1. Domain average accumulated precipitation during (a) 22 days from the 2009 Spring Experiment, (b) 36 days from the 2010 Spring Experiment, (c) 33 days from the 2011 Spring Experiment and (d) 91 days from the 2009-2011 Spring Experiments.

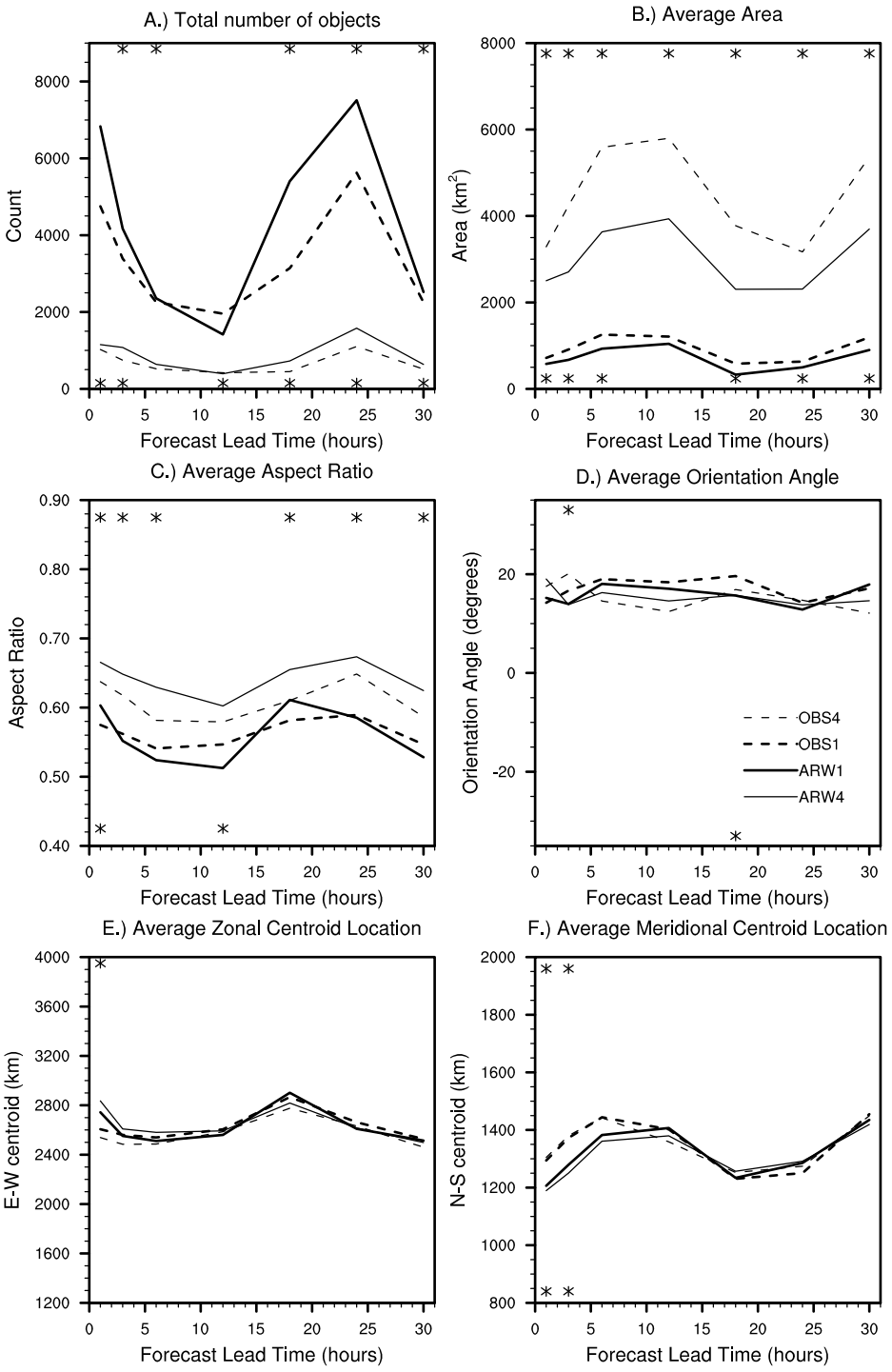

FIG. 2. Average attribute values of forecast (solid) and observed (dashed) objects on ARW1 (thick lines) and ARW4 (thin lines) grids, as a function of forecast lead time for (a) total number of objects, (b) object area, (c) aspect ratio, (d) orientation angle, (e) zonal grid point of centroid, and (f) meridional grid point of centroid. Asterisks indicate statistically significant difference

between forecasts and corresponding observations at the 95% confidence level for ARW1 and ARW4 on the bottom and top horizontal axes, respectively. Larger values in (e) and (f) are farther east and north, respectively.

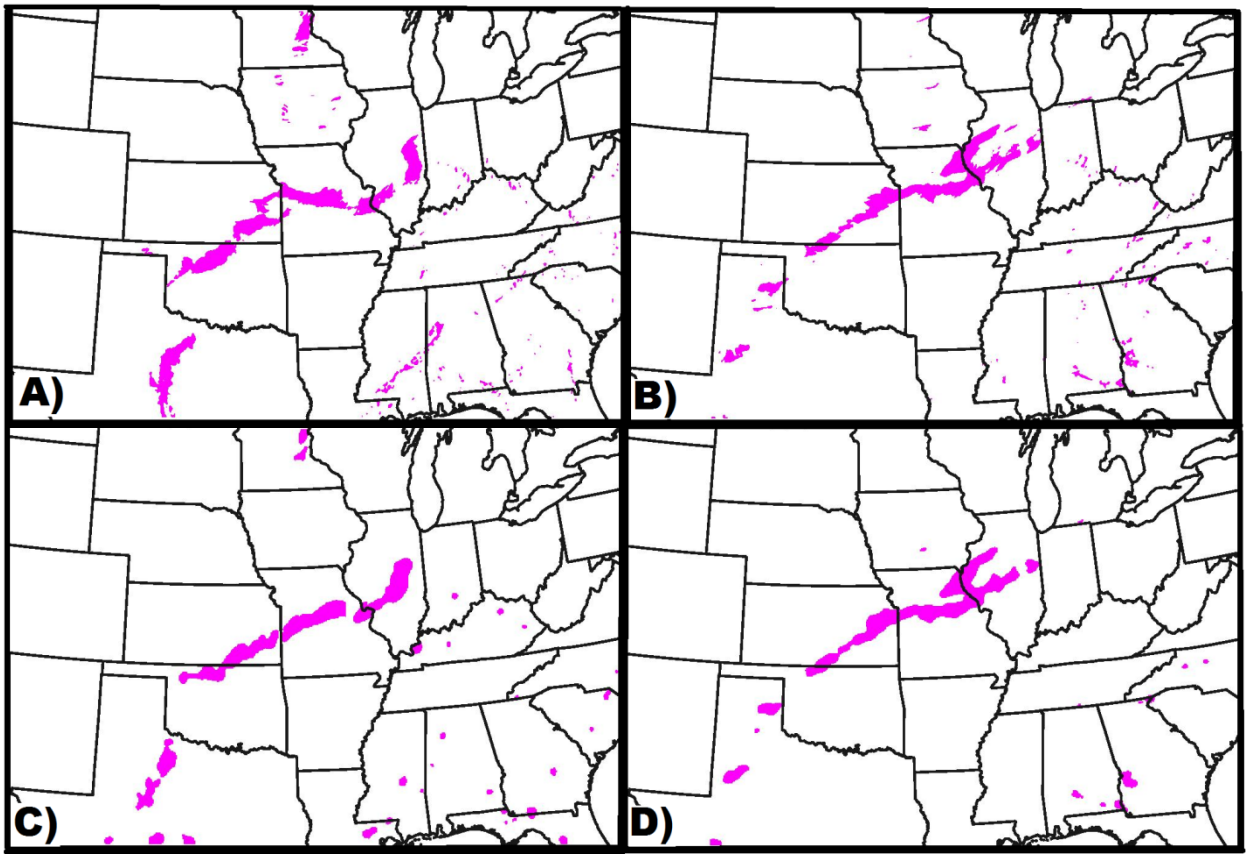

FIG. 3. A representative case of forecast [ARW1 in (a) and ARW4 in (c)] and corresponding observed [OBS1 in (b) and OBS4 in (d)] objects. The forecasts were initialized at 00 UTC 15 May 2009 and valid at 00 UTC 16 May 2009.

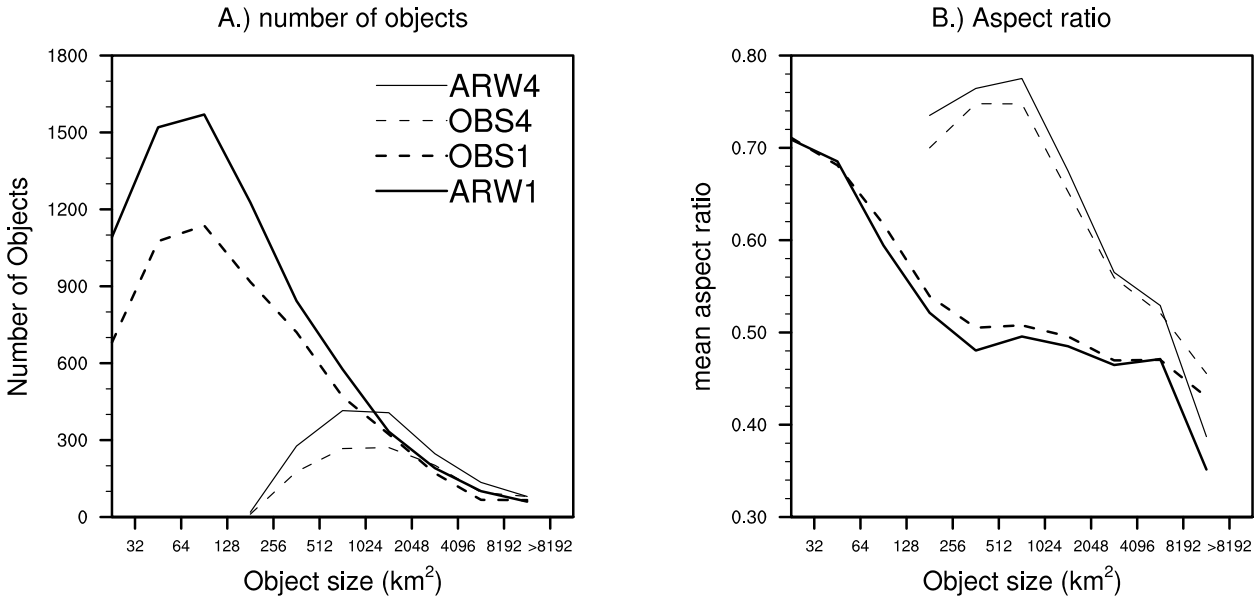

FIG. 4. 24 h lead time forecast (solid) and observed (dashed) (a) total number and (b) average aspect ratio of objects with area ≤ 32 km^2, 32 km^2 < area ≤ 64 km^2, 64 km^2< area ≤ 128 km^2, 128 km^2< area $\leq$256 km^2, 256 km^2< area ≤ 512 km^2, 512 km^2< area ≤ 1024 km^2, 1024 km^2< area $\leq$ 2048 km^2, 2048 km^2< area ≤ 4096 km^2, 4096 km^2< area ≤ 8192 km^2, and > 8192 km^2, for ARW1 (thick) and ARW4 (thin).

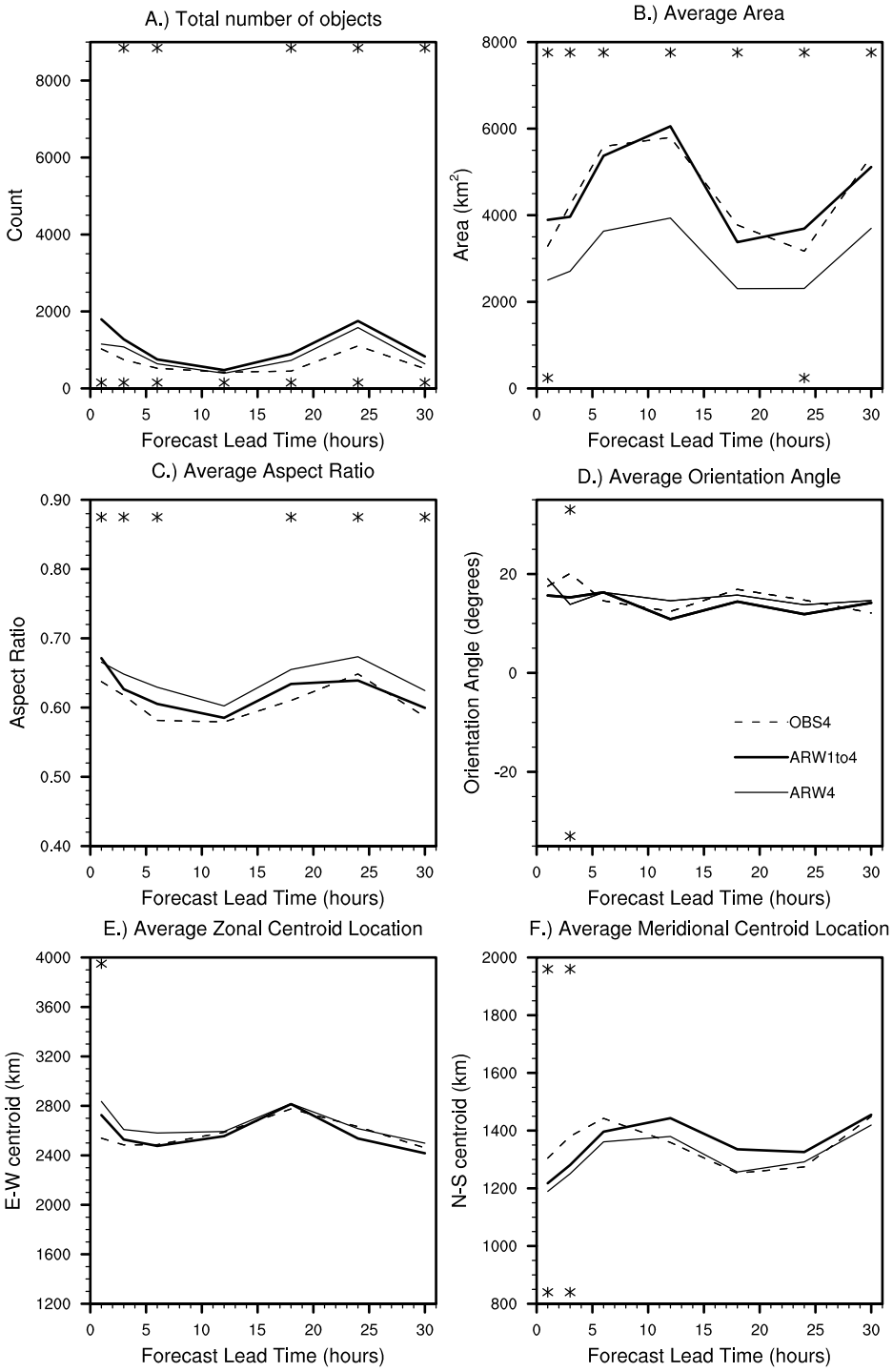

FIG. 5. As in Figure 2, except for ARW4 (thin lines) and ARW1to4 (thick lines). Asterisks along the bottom and top axes denote statistical significance for ARW4 and ARW1to4, respectively.

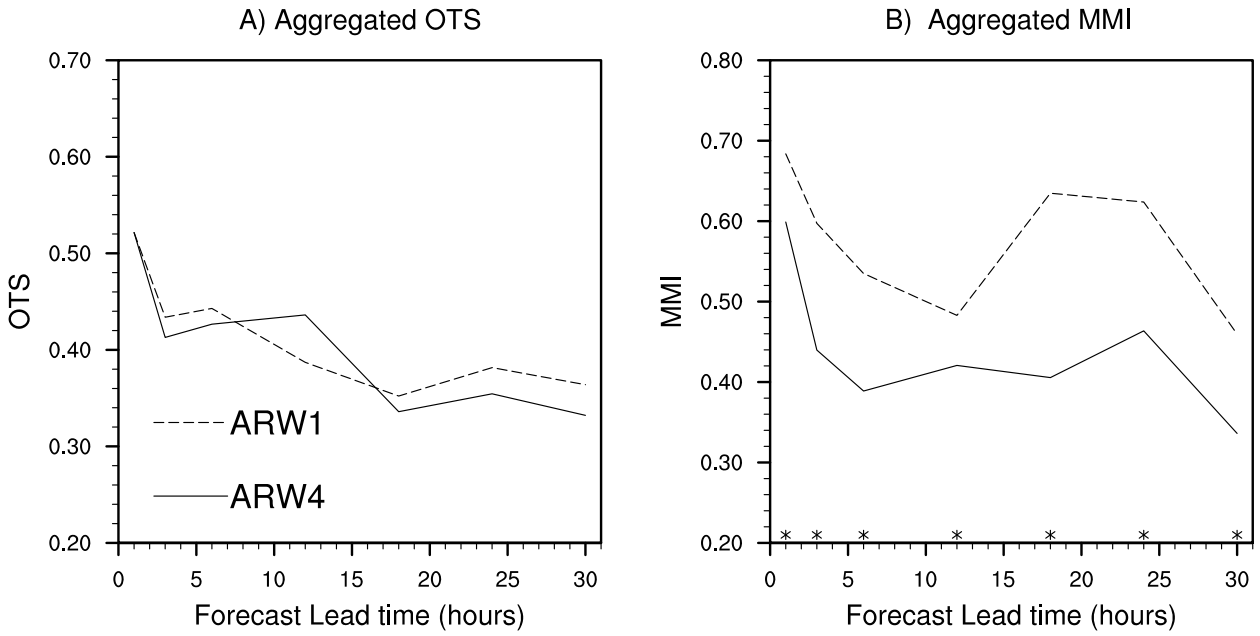

FIG. 6. Forecast accuracy of ARW1 (dashed) and ARW4 (solid) as measured by the (a) OTS and (b) MMI, aggregated over all 91 forecasts. Statistically significant differences between ARW1 and ARW4 are indicated with asterisks along the bottom horizontal axis.

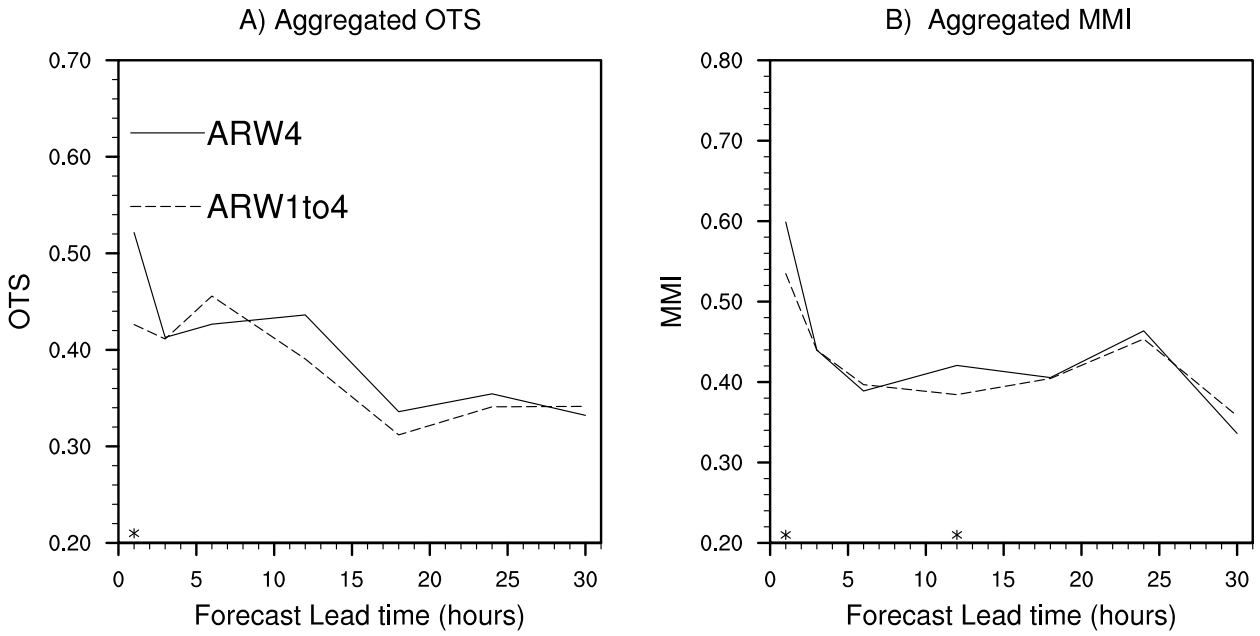

FIG. 7. As in Figure 6, except for ARW4 (solid) and ARW1to4 (dashed).

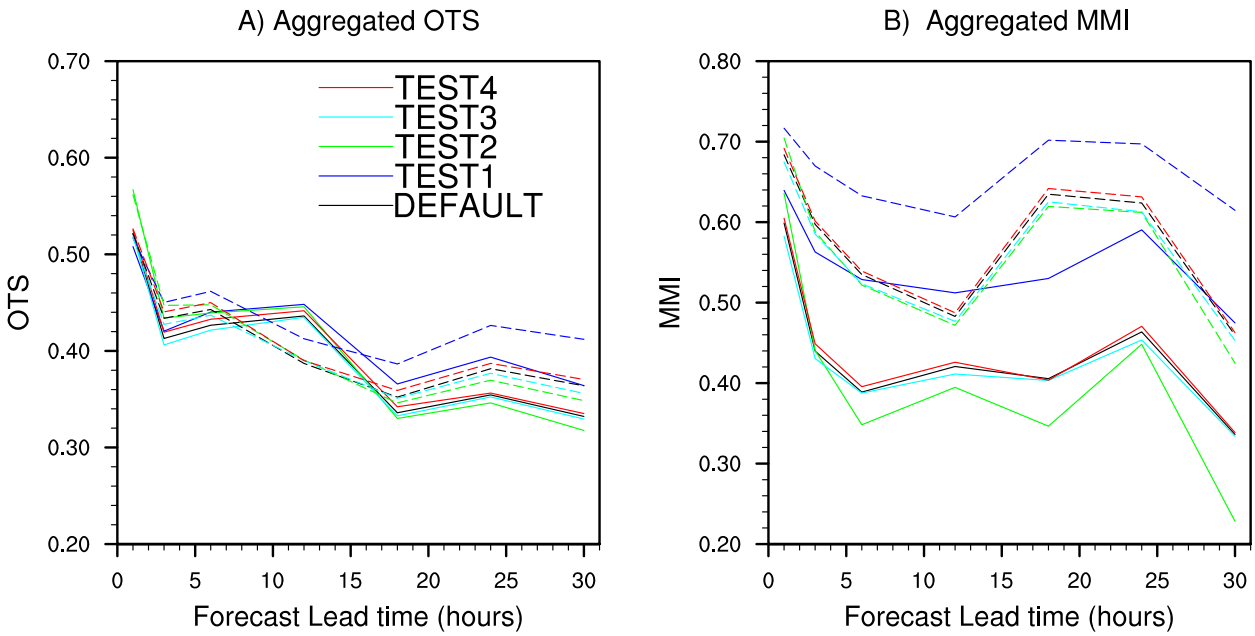

FIG 8. As in Figure 6, except for the alternate choices of attribute weights defined in Table 1.