The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program

CHRISTOPHER A. DAVIS, BARBARA G. BROWN, RANDY BULLOCK, AND JOHN HALLEY-GOTWAY

National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 8 December 2008, in final form 21 April 2009)

ABSTRACT

The authors use a procedure called the method for object-based diagnostic evaluation, commonly referred to as MODE, to compare forecasts made from two models representing separate cores of the Weather Research and Forecasting (WRF) model during the 2005 National Severe Storms Laboratory and Storm Prediction Center Spring Program. Both models, the Advanced Research WRF (ARW) and the Nonhydrostatic Mesoscale Model (NMM), were run without a traditional cumulus parameterization scheme on horizontal grid lengths of 4 km (ARW) and 4.5 km (NMM). MODE was used to evaluate 1-h rainfall accumulation from 24-h forecasts valid at 0000 UTC on 32 days between 24 April and 4 June 2005. The primary variable used for evaluation was a ''total interest'' derived from a fuzzy-logic algorithm that compared several attributes of forecast and observed rain features such as separation distance and spatial orientation. The maximum value of the total interest obtained by comparing an object in one field with all objects in the comparison field was retained as the quality of matching for that object. The median of the distribution of all such maximum-interest values was selected as a metric of the overall forecast quality.

Results from the 32 cases suggest that, overall, the configuration of the ARW model used during the 2005 Spring Program performed slightly better than the configuration of the NMM model. The primary manifestation of the differing levels of performance was fewer false alarms, forecast rain areas with no observed counterpart, in the ARW. However, it was noted that the performance varied considerably from day to day, with most days featuring indistinguishable performance. Thus, a small number of poor NMM forecasts produced the overall difference between the two models.

1. Introduction

Recent work by numerous authors has highlighted a series of novel methods for verifying the numerical prediction of highly localized, irregular fields such as precipitation. These novel methods are summarized in a companion article by Gilleland et al. (2009, manuscript submitted to *Wea. Forecasting*). Several methods fall under the heading of displacement verification methods, wherein spatial structures are examined objectively. Perhaps the most well-known methods of this type are those that make fluid flow analogies to map the forecast field to the observed field (Hoffman et al. 1995; Keil and Craig 2007) and features-based methods (Ebert and McBride 2000; Davis et al. 2006a,b, hereafter D06a,b) that treat forecast and observed precipitation regions as discrete features with identifying attributes (shape, location, intensity, etc.).

The present paper expands on the method summarized in D06a now referred to as the method for object-based diagnostic evaluation (MODE; documentation information available online at http://www.dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v1.0.pdf). MODE represents a class of spatial verification methods whose objective is to identify localized features of interest in scalar fields and compare features in two fields to identify which features best correspond to each other. When objects have been identified and categorized, statistics of the similarities of the objects in the two datasets are computed. In this sense, MODE can be considered a rudimentary algorithm for image processing and image matching, but developed for meteorological

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Christopher A. Davis, NCAR, P.O. Box 3000, Boulder, CO 80307.
E-mail: cdavis@ucar.edu

DOI: 10.1175/2009WAF2222241.1

applications. The degree of similarity between forecast and observed objects provides a measure of forecast quality. The philosophy behind the development of MODE has been to develop a procedure that mimics what a human expert would do to find features and decide whether a given feature in a forecast represents an analogous feature in the observations. The decision about matching a forecast and observed object is generally done from the perspective of a forecaster or user of the information. Because there are many different users, and because each user will bring a unique perspective to bear on the matching decision, it may be preferable not to demand a single, dichotomous outcome for matching. In some situations a binary decision about matching is necessary, but we will consider matching more generally as an inherently ''fuzzy'' process where it is more likely or less likely, but never certain. This simply reflects the fact that no forecast provides a perfect representation of an observed feature and there is always uncertainty in the observations.

As we will demonstrate, it is possible to derive statistics about forecast quality even if we do not make a binary decision about whether a forecast feature matches an observed feature. Once a decision about matching is made, however, additional metrics of the quality of the forecast can then be derived. These mainly involve geometric properties of the objects, but also involve intensity as determined within the present context by the distribution of rainfall accumulation within an object.

The goal of the present paper is a comparison, using MODE, of numerical forecasts made by two different models, the Nonhydrostatic Mesoscale Model (NMM) and the Advanced Research version of the Weather and Research Forecasting (WRF) model (ARW). Both models exist within the overarching WRF software framework, but for purposes of this study, they are considered independent models. These are described further in section 3. Results from the MODE-based evaluation will be compared with subjective impressions of differences in forecast quality.

Following a description of the object identification, matching, and merging procedures that MODE comprises, we will present results from idealized cases where objects have simple geometric properties and differences between synthetic forecast and observed features are prescribed. We will then present results from a nine-case sample obtained from the 2005 National Severe Storms Laboratory/Storm Prediction Center (NSSL/SPC) Spring Program. These are the same nine cases discussed in more detail by Ahijevych et al. (2009, hereafter A09). Then, we will consider the overall performance of the ARW model and the NMM for 32 days on which both models produced forecasts.

2. MODE and idealized examples

The first step in MODE is to identify objects on a two-dimensional field such as precipitation rate or precipitation accumulation. The procedure is described in detail by D06a. Object identification is done through the application of a convolution operator, governed by a convolution radius (R) and a threshold (T) on the intensity of the field. The convolution step is effectively a smoothing operation. The parameter R is expressed in units of grid increments, and T carries units of the field being evaluated (mm in the case of rainfall). These steps serve two purposes: 1) to make areas more contiguous than in the original field and 2) to filter out small or weak features of precipitation if the user is not interested in them. Once the forecast and observed objects are defined using the convolution and thresholding procedures, the original intensity of the field is restored at points within objects.

The convolution and thresholding operations effectively select the portion of the field that is of greatest interest to the user of the method, and therefore there is not necessarily a universally optimal choice for these parameters. Minimal smoothing and a very low threshold will result in a large number of objects, many of them small. Heavy smoothing and a high threshold will result in very few, intense rain areas. We identify R with the minimum spatial scale of interest and T with the minimum rainfall intensity of interest. These parameters are user defined, but we will show that the dependence of MODE-derived metrics of the forecast quality on these parameters is a useful diagnostic for understanding the performance of models. That is, for what spatial scales and intensities does a given model perform best, or where does it perform better than a competitor?

The primary addition to MODE beyond what was described by D06a includes a more sophisticated algorithm for matching and merging objects in the forecast and observed fields. While D06a included simple matching rules based only on the distance separating the centroids of the forecast and observed objects, MODE employs a fuzzy-logic system that considers numerous attributes in identifying a match between forecast and observed objects. The present application of MODE also recognizes that the choice of a threshold can affect whether objects are contiguous or discrete. The overarching goal is to mimic the decision process that a human exercises in concluding whether a match has occurred. However, it is clear that no two humans will use the same criteria applied exactly the same way. For this reason, we will not rely entirely on a binary decision about matching that may rest on somewhat arbitrary choices. We will consider the likelihood that a given pair of objects constitutes a

match and utilize the distribution of such likelihood values among all possible pairs of objects to derive an assessment of relative model skill.

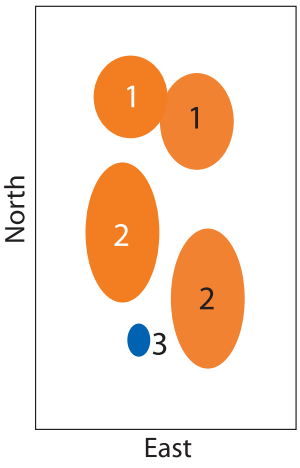
For any forecast and observed object pair, a fuzzy-logic algorithm is used to derive a value of what we will call total interest. Total interest for the jth object pair is defined as

$$I_j = \frac{\sum_{i=1}^{M} c_i w_i F_{i,j}}{\sum_{i=1}^{M} c_i w_i}, \quad (1)$$

where F is the interest function that prescribes, on a scale from 0 to 1 with 1 being perfect, how closely a forecast attribute matches the observed attribute. The coefficient w is the weight assigned to that interest function and c is a function of attributes that describes the confidence in a partial interest value obtained from $w_i F_{i,j}$. Total interest comprises M interest functions that compare attributes of each object pair. The attributes we consider are (a) centroid distance separation, (b) minimum separation distance of object boundaries, (c) orientation angle difference, (d) area ratio, and (e) intersection area. The orientation angle is the angle that the long axis of the object makes with respect to the grid direction (i.e., the x direction, which is nearly east–west in the present study). The area ratio is the area of the smaller of the two objects divided by the area of the larger; hence, it is forced to lie between 0 and 1. The intersection area is the fraction of overlap area normalized by the average of the areas of the two objects and is also bounded by 0 and 1. The confidence (c_i), which represents how well a given attribute describes the forecast error, is unity for all attributes except the orientation angle and centroid separation. Nontrivial confidence functions are described in appendix A.

The overall assumption is that the more alike the two objects are as measured by the above attributes, the more likely the forecast object is a numerical representation of the observed member of the pair (and not some other feature in the observed field). Total interest assumes a value between 0 and 1, and may be interpreted as the likelihood of a match, but it is not strictly a probability.

In the schematic shown in Fig. 1, there are two observed rain areas and three forecast areas, one of which is notably smaller than the others. Also shown is a hypothetical total interest value for each forecast–observed object pair. In this example, we have assumed that a match occurs when the total interest is 0.7 or greater. This tunable value has been chosen to represent mesoscale systems, but a greater value might be selected in consideration of rainfall over local watersheds or urban areas, for example, where a closer match between forecast and observed objects is necessary for a forecast to be considered of high quality. Using this threshold, forecast object 1 matches observed objects 1 and 2. Forecast object 2 matches observed object 2. Forecast object 3 is unmatched, primarily owing to its small size relative to other objects. Because a "double match" exists with forecast area 1, observed areas 1 and 2 are viewed as a "cluster." In addition, because one of the objects in this cluster also matches forecast area 2, forecast areas 1 and 2 also form a cluster. Forecast area 3 matches no observed object and is not part of any cluster.



Interest Matrix

Forecast \ Observed	1	2
1	0.90	0.75
2	0.50	0.80
3	0.40	0.55

FIG. 1. Schematic showing hypothetical forecast rain objects (black numerical labels) and observed rain objects (white numerical labels) with the corresponding interest matrix at right. Orange-shaded objects are matched whereas blue shading denotes no match. Hypothetical total interest values >0.7 are shown in red numbers in the matrix.

Not shown in Fig. 1 is an additional factor used for potentially merging multiple objects in a given field into a single cluster. We consider the contiguous area within objects defined with the same R, but a threshold of $T/4$. The lower threshold results in larger rain areas and will sometimes connect features that do not touch when using the threshold T. In these cases, the objects defined with the threshold T are merged. The motivation is to allow nearby features to be considered as part of a larger-scale entity. The use of this "double thresholding" technique only affects the merging of objects within a given field and therefore affects what is included in object clusters. However, the total interest for pairs of simple objects is still computed based on objects derived from the threshold T.

In the present paper we will derive verification scores from total interest values for forecast and observed pairs of simple objects rather than for clusters. Clusters are perhaps more intuitive groupings of features, and will appear in the figures in the present paper as a convenient

means of visualizing corresponding features in forecasts and observations. However, statistics will be computed based entirely on simple objects.

A useful metric of forecast quality is the maximum interest value attained for each forecast or observed object, corresponding to the maximum values across rows or down columns, respectively, in the matrix in Fig. 1. By aggregating values of maximum interest over all objects in a sample, one obtains a metric that can be used to compare forecasts from different models. This metric is the median of the distribution of maximum interest values, referred to as the median of maximum interest (MMI). The MMI can be computed with respect to the forecast objects (MMIF) or the observed objects (MMIO). The MMIF finds the maximum interest across each row of the interest matrix. In this case, those values are 0.9, 0.8, and 0.55, with a median value of 0.8. The median rather than the mean is used to reduce the impact of outliers and because there is no basis to assume any particular underlying distribution of the data. The MMIO is computed from the maximum interest values in each column, 0.9 and 0.8, yielding 0.85 (assuming linear interpolation between ranks 1 and 2). Note that MMIO and MMIF will not generally be equal. False alarms will lower the MMIF whereas misses will lower the MMIO. In the simple example here, there were no missed events. For a single statistic, hereafter simply MMI, the distributions of maximum interest values from the forecasts and observations are joined, yielding a rank ordered set {0.55, 0.8, 0.8, 0.9, 0.9} whose median is 0.8. The full distribution of maximum interest values is typically aggregated over several tens of objects at a single time, and over several tens of days. An advantage of using the MMI as a metric of quality is that it does not depend on imposing any matching threshold and takes into account all of the attributes used in the calculation of total interest.

As a relatively simple example of quantitative results from MODE, consider the synthetic forecasts and observations presented in Fig. 2. These particular examples also appeared in D06a and A09. The Gilbert skill score [GSS, also known as the equitable threat score; Schaefer (1990)], which measures the amount of overlap between two fields, is slightly negative for each of the first four forecast examples (Figs. 2–e). Herein, we simply use the letter representing the panel of Fig. 2 to denote each case. The negative score arises because there is no overlap between the forecast and observed fields, and this is less overlap than would be expected by chance. Although the use of a particular forecast ultimately determines the utility, for most applications it is probably true that the forecast in b is a better forecast than, say, the forecast in d because it has a smaller displacement error and zero intensity bias. Only f reveals a positive GSS owing to the overlap between the forecast and observed ''precipitation'' areas.

The ordering of the forecast quality is substantially different with MODE. The forecast with the greatest total interest is b in which the position offset is small and the structural match is exact. The forecast with the smallest interest is c in which the position offset is large. Example f is not the poorest forecast (as some might assume) because the algorithm as applied here is weighted toward position errors and also gives some ''credit'' for overlap. Giving substantially less weight to the intersection area and more weight to the area ratio would have allowed the total interest in d to exceed that in f.

A potentially counterintuitive result is that the forecast in e has a lower total interest than that in d. Much of the distinction between d and e comes from the confidence assigned to attributes of centroid separation and angle difference. Recall from (1) that the denominator of the expression for total interest is the sum of the product of weight and confidence. For a round object, as in d, we have no confidence in the angle assigned. The angle-difference term therefore drops out of both the numerator and denominator. In case e, the angles are well defined. The 90° difference yields a partial interest value of zero, but the denominator still retains the weight of the angle-difference term. This fact tends to lower the total interest in e compared with d. Centroid displacement affects the results in a similar way because the confidence value is equal to the area ratio (appendix A). For objects differing greatly in size, the contribution of the weight of the centroid separation to the denominator of (1) will be reduced. The fact that the area ratio is poorly predicted in d limits the numerator of (1), but not enough to make the total interest smaller in d than in e.

There is no ''correct'' answer in these geometric cases, but the results discussed here indicate some of the possible sensitivities of the MODE algorithm. The influences of different attributes on the total interest should be selected to represent the aspects of the forecast that are most important to the user of the forecast. The preceding analysis also illustrates that MODE produces results in these idealized cases that are probably more consistent with subjective evaluation than does the ETS metric.

3. Evaluation of numerical forecasts

a. Evaluation of forecasts for nine cases

The focus of the remainder of the present paper is the application of MODE to a comparison of numerical forecasts produced by the NMM and ARW models. The ARW model (Skamarock et al. 2005) is a nonhydrostatic, terrain-following, mass-coordinate model that

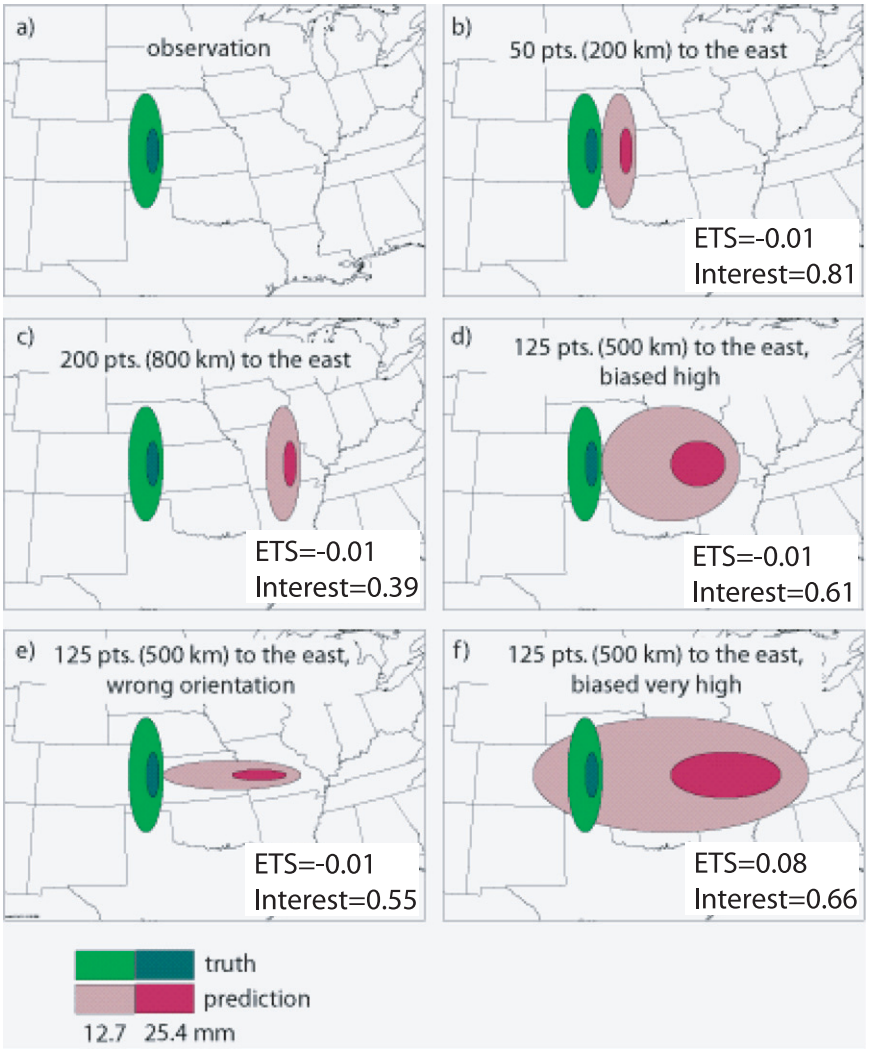


FIG. 2. Results of MODE applied to idealized rain areas that are elliptical with two intensity values. The total interest and equitable threat score (GSS) are annotated on each example. Results were calculated on a grid increment of 4 km. Parameters for the computations of interest are described in the text and are the same as those used for numerical forecasts. No thresholding was applied in these idealized cases.

was integrated with a grid increment of 4 km. The NMM (Janjić 2004), integrated in a mass-based, terrain-following coordinate, utilized a grid increment of 4.5 km. The physical parameterizations in each model configuration are summarized in Table 1 (adapted from Baldwin and Elmore 2005). Additional information about the setup of the models may be found in Baldwin and Elmore (2005) and Kain et al. (2008).

Both models were integrated daily beginning at 0000 UTC during the 2005 NSSL/SPC Spring Program from 18 April to 3 June (Baldwin and Elmore 2005). Here, we examine only 1-h precipitation accumulations from 24-h forecasts valid at 0000 UTC (i.e., precipitation accumulated from 23 to 24 h of the forecast). The time of 0000 UTC (early evening) is typically when convection is intensifying and organizing across much of the central United States. Forecasts were not available from both models on all days during the Spring Program. A total of 32 forecasts were evaluated during the period from 23 April to 4 June. Nine of these are discussed in detail by A09.

Precipitation forecasts and observations valid 0000 UTC 1 June 2005, one of the nine cases discussed in A09, are summarized in Fig. 3. The stage II precipitation analyses were used as observations (Lin and Mitchell 2005). The precipitation from the NMM model (hereafter the NMM4 model) and ARW model (hereafter the ARW4 model) was interpolated to the stage II grid using bilinear

TABLE 1. Configuration of numerical models: NAM, North American Model of NCEP; MYJ, Mellor–Yamada–Janjić (Janjić 2002); and YSU, Yonsei University (Noh et al. 2003). The WRF single-moment six-class (WSM6) scheme is derived from Hong et al. 2004) and the Ferrier scheme from Ferrier et al. (2005).

	WRF-NMM4	WRF-ARW4
Grid spacing	4.5 km	4.0 km
Vertical levels	35	35
Planetary boundary layer scheme	MYJ	YSU
Microphysics	Ferrier	WSM6
Initial conditions	NAM	NAM

interpolation. The interpolated precipitation fields suggest an elongated synoptic-scale precipitation system extending from the Texas Panhandle into the upper Midwest forming a general arc shape. Within this broad area, the ARW4 indicates two concentrations of rainfall whereas the NMM4 produces one larger contiguous rain area with a smaller area over western Texas and eastern New Mexico. Some of the smaller rain areas remain unmatched, especially those in the NMM4 model observations.

The differences in matching that occur with each of the two models can be understood in terms of the interest functions for each forecast–object pair. The large rain area in the NMM4 forecast (green) matches only the area of convection centered over the intersection of Iowa, Nebraska, Kansas, and Missouri. However, the arc-shaped system observed over Oklahoma is not matched in the NMM4 forecast. The reasons are that the centroid of the NMM4 rain area is in central Kansas, rather far from the centroid of the observed feature, and that there is a significant size mismatch between the two features. Both of these detract from the total interest, which turns out to be 0.64 for this pair of objects. If the total interest had been above 0.7, the observed area over Oklahoma would have matched the large area in the NMM4 forecast. In this case, the large area in the NMM4 forecast would match the two regions: one over Oklahoma, the other near the intersection of Nebraska, Iowa, Kansas and Missouri. These two observed regions would then be considered part of the same cluster. But the fact that they are not results from the NMM4 model predicting an area that is too large and not sufficiently similar in structure to the observed features. The ARW4 model predicts a clear break between the rain areas over northeastern Kansas and eastern Nebraska and a rain area over central and western Oklahoma and the Texas Panhandle. This break, being more like the observations, results in a better overall match in this case. On the other hand, the ARW4 model produces only a weak convective cell over eastern New Mexico that does not survive the convolution and thresholding processes, and the storm area therefore appears as a missed event. In this region, the NMM4 model produces cells with areas and intensities closer to those observed than does the ARW4 model.

The results shown in Fig. 3 are obtained with a particular choice of convolution radius (R) and threshold (T). It is instructive to examine the performance of each model across a range of these parameters, thereby considering the performance of models for different spatial scales and different intensities. The results are summarized with MMI values plotted as a function of R and T in Fig. 4. Near the origin (bottom-left corner in Fig. 4), there is essentially no filtering of the original data. The number of objects is greatest in this region of the plot. Few if any objects are found in the upper-right corner of Fig. 4, where the convolution radius is large (strong smoothing) and the threshold is high (only large, heavy rain areas can pass through the filter). With increasing convolution radius along the abscissa, more smoothing is applied so there tend to be fewer, larger rain areas. With increasing threshold near the ordinate, there are fewer areas from which only the most intense portions are retained.

For the case of 1 June, it can be seen in Fig. 4 that there is a broad swath of higher MMI values for the ARW4 model extending horizontally for thresholds near 1 mm. There is a suggestion that the relatively large values of MMI are found at higher thresholds for convolution radii between 5 and about 15 grid lengths (20–60 km). In D06b, it was suggested that a convolution radius of around five grid lengths (20 km) was reasonable for highlighting mesoscale convection features. Relatively larger values of MMI are found in the NMM4 model only for the smallest thresholds. Both models tend toward low MMI values in the upper-right portion of Fig. 4 because there are essentially no objects. For modest threshold values, the MMI increases with convolution radius. This is consistent with enhanced forecast quality when predicting features on larger scales.

It should be emphasized that the types of errors seen here may vary from case to case. Some of the apparently sensitive dependence of matching on object intensity or contiguousness tends to average out unless one of the models has a systematic bias to make contiguous rain areas that are too large, for instance. Thus, it is important to view statistics over a reasonably large sample of cases before any conclusions can be drawn. The nine-case sample discussed in A09 offers an extension of the number of cases, but for which a more detailed analysis of individual cases is still possible.

We applied MODE to the nine cases (26 April; 13, 14, 18, 19 and 25 May; 1, 3, and 4 June) and examined the

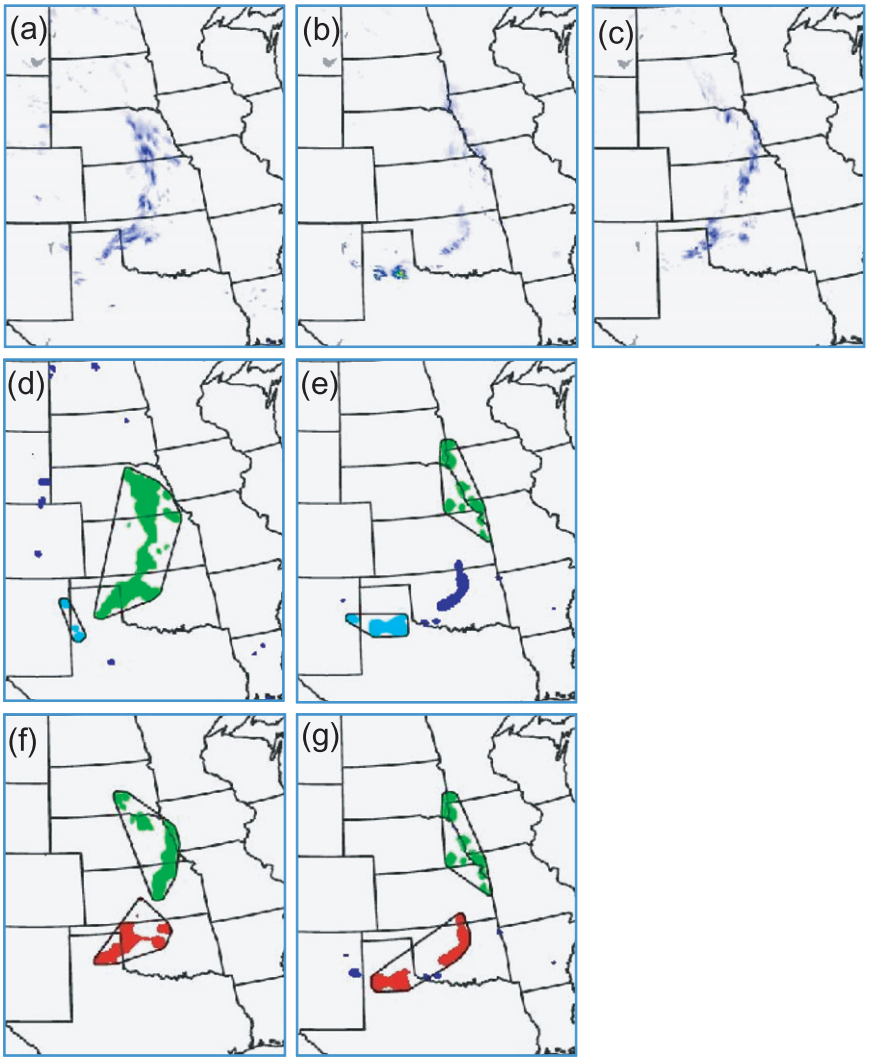


FIG. 3. One-hour rainfall accumulations from 24-h forecasts valid at 0000 UTC 1 June from (a) NMM4 and (c) ARW4, with the stage II observations (OBS) in (b). (d)–(g) MODE objects are colored with $R = 5$ and $T = 3$ mm. Like colors in (d) (NMM4) and (e) (OBS) indicate matching objects (dark blue indicates no match). The same is true for like colors in (f) (ARW4) and (g) (OBS). Bounding convex hulls (black lines) delineate object clusters, but are not used for computations.

MMI metric (Fig. 5). For most cases, the MMI was between 0.6 and 0.7. However, the forecasts valid at 0000 UTC on 13 May and 4 June are both notably poorer in the NMM4 model. From the summary of the objects for each case (Figs. 6 and 7), it is apparent that the NMM4 model produced regions of convective cells or small convective systems that were not present in the observations. For instance, at 0000 UTC 13 May numerous rain areas predicted by NMM4 were grouped into a cluster (green) that matched one tiny cell in the observations over the eastern part of the domain. This occurred because one cell in the forecast was close enough to the observed cell that a match was found, and the proximity of the forecast cells to that one matching cell resulted in a cluster of merged areas in the forecast. However, the interest values for most of the simple objects in the cluster paired with the tiny observed cell were small, mainly due to large centroid separation, boundary separation, and discrepancies in object sizes. Such an overprediction of the number of objects adds numerous rows to the interest matrix (e.g., Fig. 1) and the physical discrepancies between the forecast and observed objects mean that the maximum interest values in those rows are small. The distribution of maximum interest values, keyed on the forecasts, thus gains a large number of small values, pulling the MMIF (and hence

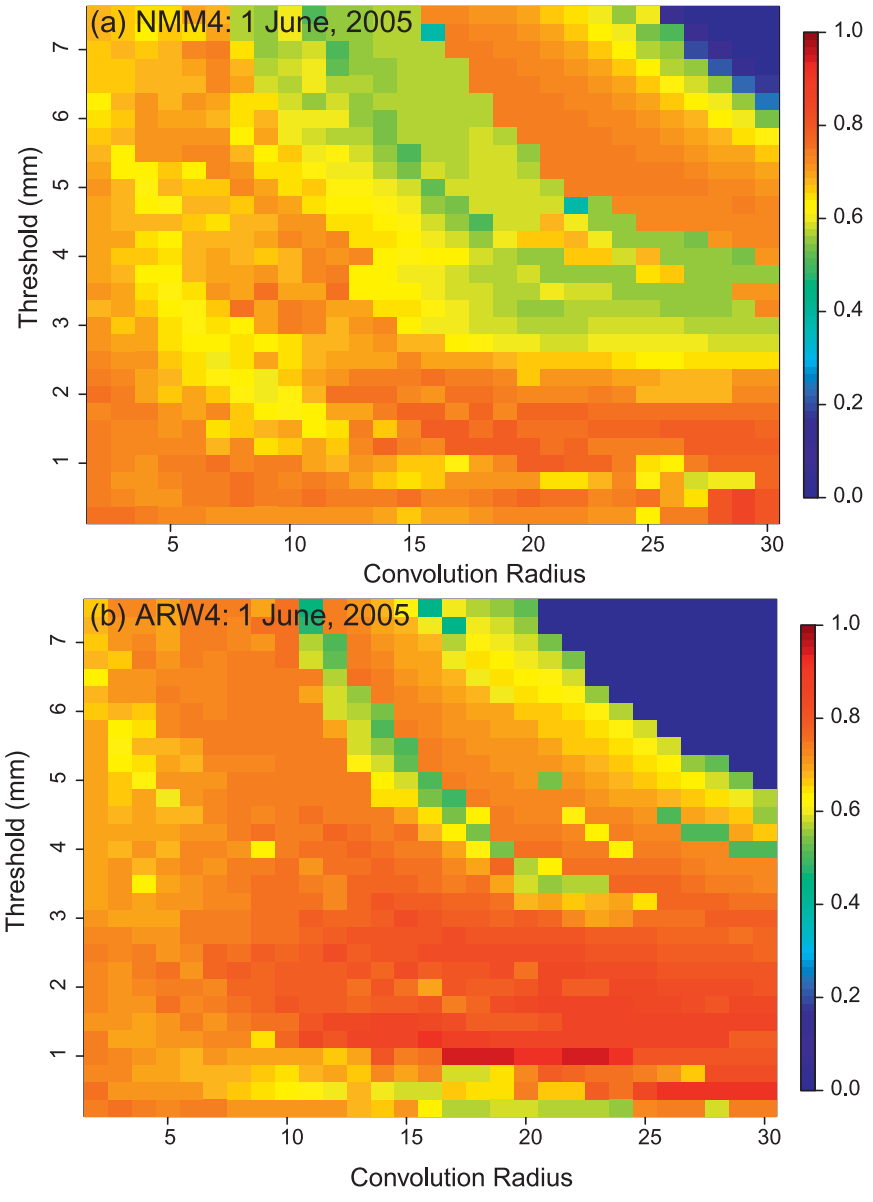


FIG. 4. Plots of the MMI as a function of convolution radius (grid squares) and threshold (mm) for the 24-h forecast of 1-h rainfall accumulation valid at 0000 UTC 1 Jun 2005.

the MMI) toward lower values. A similar situation occurred with the NMM4 forecast on 4 June where a large number of small rain areas was erroneously predicted over the northern and southeastern parts of the domain.

However, from the perspective of the observed objects, the one small green-colored object in Fig. 6b matches well with one object in the forecast, so the distribution of maximum interest values keyed on the observations gains a high value and MMIO is actually increased. But being only a single object, this cannot raise the MMI significantly. On 4 June, there are no observed objects in these regions, so the MMIO and MMI are unaffected. Thus, the difference in forecast quality, as summarized by the MMI metric in Fig. 5, can be isolated to the MMIF (e.g., maximum interest across rows of the interest matrix) in these two cases. For the 13 May case, the ARW4 and NMM4 values of MMIF were 0.59 and 0.49, respectively. For the 4 June case, the values were 0.67 and 0.42. It turns out that the MMIO values (median of maximum interest down columns of the interest matrix) were larger than the MMIF values for both models on 13 May (0.75 and 0.77, respectively), and for the NMM4 model on 4 June (0.70). Both models erred more by false alarms than by misses on these days.

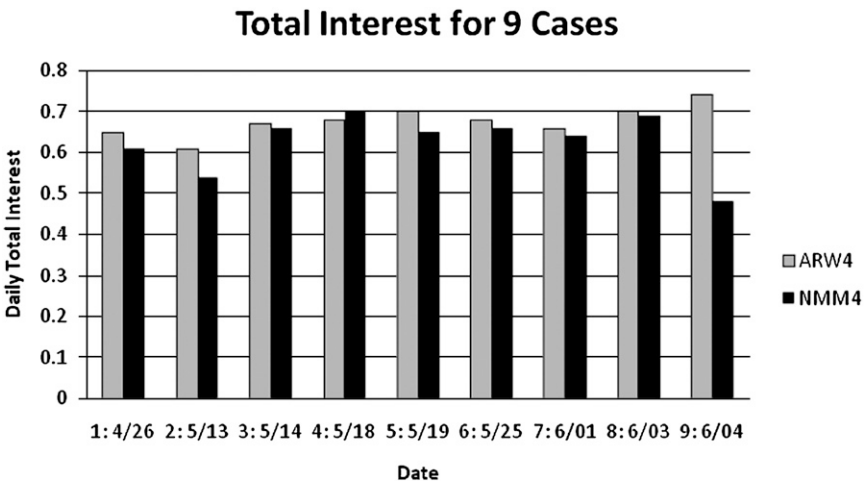


FIG. 5. Bar graph showing values of MMI computed separately for each of the nine cases from the 2005 NSSL/SPC Spring Program dataset highlighted in A09. Black bars indicate the NMM4 model and gray bars indicate the ARW4 model. Dates are indicated as M/DD, where M is for month and DD for day.

In the study of A09, ''traditional'' skill scores were applied to each of the nine cases noted above. The Gilbert skill score (Schaefer 1990) applied to the NMM4 forecasts was lowest for 18 May, and 1 and 4 June (Fig. 6 from A09), whereas the MMI metric from MODE was notably lower than other values only on 4 June. One reason for this behavior becomes clear upon examination of Fig. 8, which shows the rain areas derived from the NMM4 and stage IV data using MODE for 18 May. While the forecast rain area over the Dakotas is spatially close to what is observed, there is essentially no overlap between the forecast and observed rain areas. Thus, the traditional verification approach indicates nearly zero skill. However, the NMM4 forecast quality derived from MODE is comparable to other cases considered to be ''good forecasts'' by subjective and traditional evaluation metrics (A09). The 18 May case appears to be a real-atmosphere counterpart to the example in Fig. 2b and it accentuates the difference in interpretation between a traditional method and a spatial verification method such as MODE.

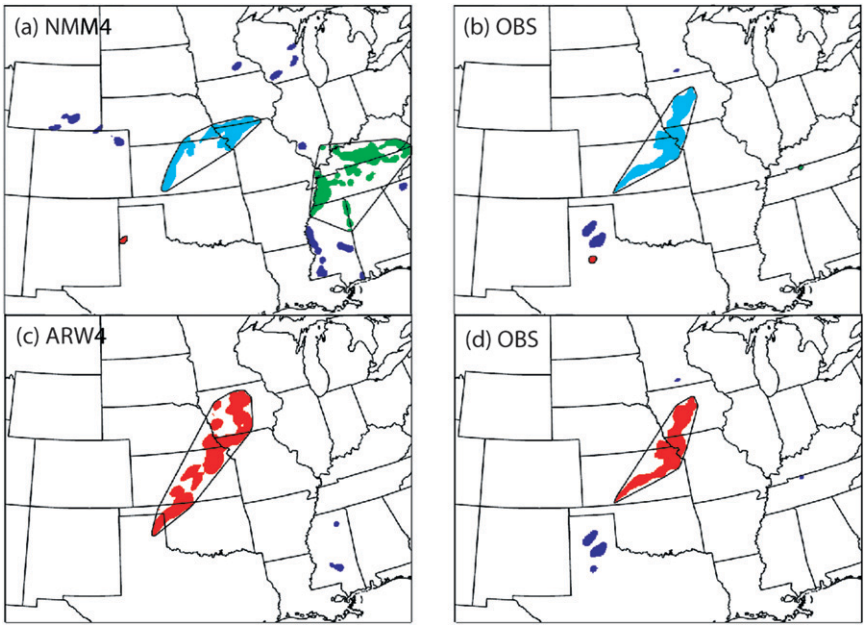


FIG. 6. Results of MODE for 0000 UTC 13 May 2005. Fields are based on a 1-h accumulated precipitation from a 24-h forecast. Format follows Figs. 3d–g.

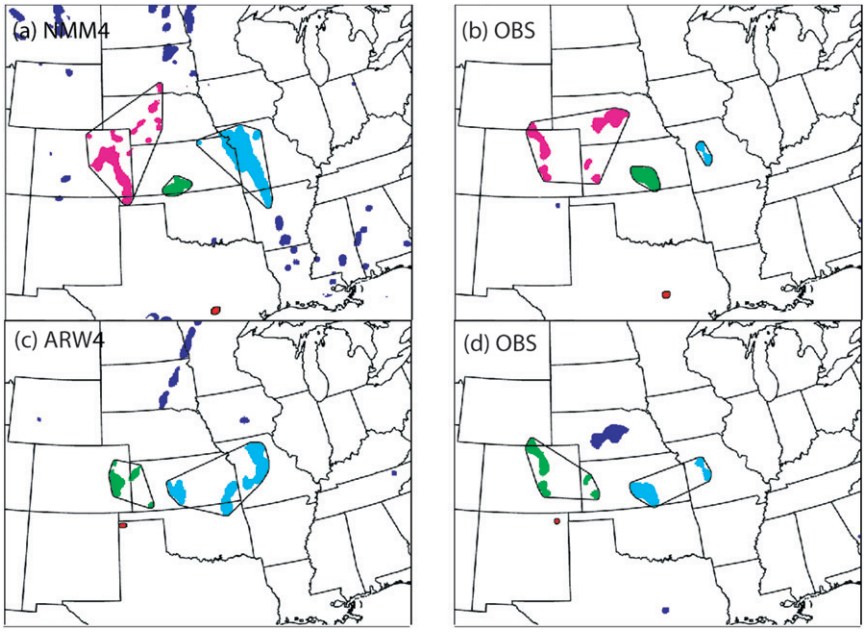


FIG. 7. As in Fig. 6, but for 0000 UTC 4 Jun 2005.

b. Evaluation of 32 cases

1) MAXIMUM OF MEDIAN INTEREST

To examine a still larger sample, we aggregate the MMIF and MMIO metrics over 32 cases (see appendix B) during the 2005 Spring Program. Each of these 32 cases represents a 24-h forecast of hourly rainfall accumulation valid at 0000 UTC. These MMIF and MMIO values appear in Table 2 along with the number of forecast objects in the sample. Because results will vary with the choice of R and T, we examine a realistic range of these parameters. The MMIF metric (Table 2) is generally higher for the ARW4 forecasts than for the NMM4 forecasts. The MMIO values were similar overall between the two models, and larger than the MMIF values, especially the NMM4 model. A Wilcoxon rank sum test was applied to the distributions underlying the MMIF values. We found that the null hypothesis that the maximum

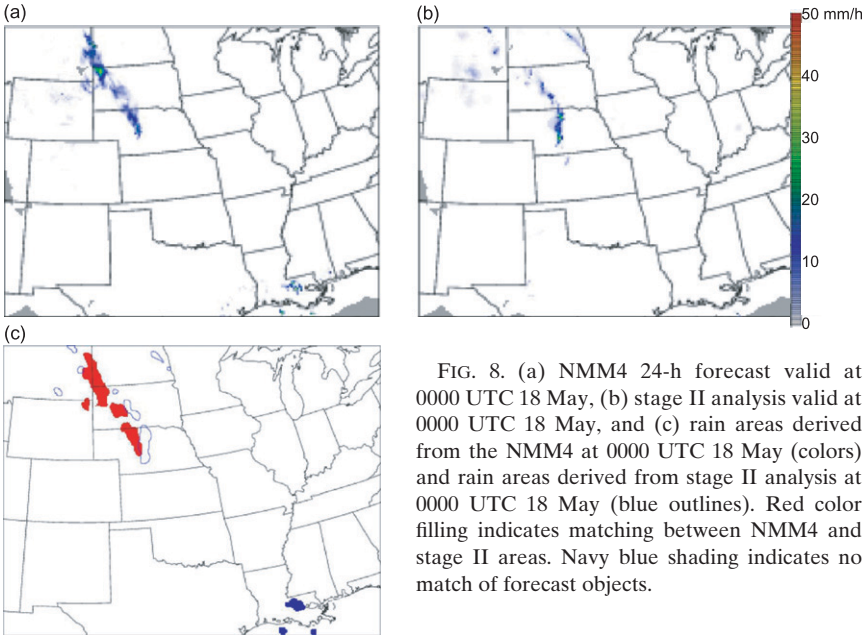


FIG. 8. (a) NMM4 24-h forecast valid at 0000 UTC 18 May, (b) stage II analysis valid at 0000 UTC 18 May, and (c) rain areas derived from the NMM4 at 0000 UTC 18 May (colors) and rain areas derived from stage II analysis at 0000 UTC 18 May (blue outlines). Red color filling indicates matching between NMM4 and stage II areas. Navy blue shading indicates no match of forecast objects.

TABLE 2. Median of maximum interest based on forecast (MMIF) and observed objects (MMIO), displayed as MMIF/MMIO for different combinations of R and T. The number of forecast objects is in parentheses.

	ARW4				NMM4		
	$T = 1.5$	$T = 3$	$T = 6$		$T = 1.5$	$T = 3$	$T = 6$
$R = 10$	0.63/0.70 (292)	0.67/0.68 (173)	0.59/0.62 (95)	$R = 10$	0.58/0.73 (459)	0.55/0.69 (315)	0.52/0.63 (155)
$R = 5$	0.66/0.70 (690)	0.65/0.68 (435)	0.66/0.67 (263)	$R = 5$	0.62/0.73 (1027)	0.60/0.72 (752)	0.56/0.68 (414)

interest values from each model were drawn from the same distribution could be rejected with greater than 99% confidence. This was true for each value of R and T. From this perspective, the forecast quality of the ARW4 model was greater.

Also shown in Table 2 is the number of forecast objects for each choice of R and T. It is clear that the NMM4 model predicts far more discrete rain areas than the ARW4 model. It turns out that both models produce too many objects, but the NMM4 model errs in this regard by roughly a factor of 2 whereas the number of objects in the ARW4 is about 30% greater than is observed.

2) BIASES IN AREA AND INTENSITY

The number of rain areas, by itself, is not necessarily indicative of a problem with the forecast because a high-resolution forecast might be expected to produce more areas than a coarse-resolution forecast. However, in the present case, both observations and models are projected onto the same grid (preserving area-mean rainfall), with the same convolution and threshold parameters. This suggests that the character of convection, particularly in the NMM4 model, is rather different than is observed. This finding is further supported by examining the error in the total area of the objects, expressed as the ratio of the total forecast rain area within objects to the total observed rain area within objects. From Table 3, it is apparent that the NMM4 model has a bias as large as 3 by this measure, whereas the ARW4 model is also biased in the same sense, but the departure of the bias from unity is typically about 1/3 of the bias of the NMM4 model. The bias is generally largest for the large, intense rain areas. In the NMM4 model, the bias in the number of objects is roughly a factor of 2, but the bias in area is generally larger than a factor of 2. This suggests that not only are there too many forecast areas, but the forecast areas are too large. A similar conclusion pertains to the ARW4 model, but the biases therein are considerably smaller.

From Table 4 it is also apparent that the fraction of the rain area within matched objects decreases as the threshold increases, particularly for large, intense rain areas. Not only is there a smaller total area that survives the convolution and thresholding procedures, but the fraction of that area that resides within matching forecast or observed objects also decreases with increasing rainfall threshold. The errors exemplified in Table 4, together with the results of Tables 2 and 3, imply that the more intense rainfall features tend to be overpredicted by the models, particularly the NMM4 model, and the more intense areas often have no counterpart in the observations.

We can also compare the distribution of rainfall intensity within the rain areas. We consider percentiles of the rainfall distribution for each object and focus on the 90th percentile to indicate whether forecasts have a realistic fraction of heavier rainfall amounts within rain areas. There was a positive bias in both models, with the 90th percentiles in the NMM4 and ARW4 models being 11.9 and 11.3 mm, respectively, whereas the observed 90th percentile was 9.7 mm. This result is only for simple objects that match using a total interest threshold of 0.7. It should be noted that intensity was not used as a

TABLE 3. Area bias (the total area of forecast objects divided by the total area of observed objects). The total observed area in 10^6 km^2 appears in parentheses.

	ARW4				NMM4		
	$T = 1.5$	$T = 3$	$T = 6$		$T = 1.5$	$T = 3$	$T = 6$
$R = 10$	1.3 (3.4)	1.4 (1.5)	1.9 (0.4)	$R = 10$	2.1 (3.4)	2.3 (1.5)	3.1 (0.4)
$R = 5$	1.3 (3.4)	1.4 (1.7)	1.7 (0.6)	$R = 5$	2.1 (3.4)	2.3 (1.7)	2.9 (0.6)

TABLE 4. The total area of matched objects (forecast + observed) divided by the total area of all objects (forecast + observed).

	ARW4				NMM4		
	$T = 1.5$	$T = 3$	$T = 6$		$T = 1.5$	$T = 3$	$T = 6$
$R = 10$	0.83	0.73	0.58	$R = 10$	0.79	0.65	0.44
$R = 5$	0.81	0.78	0.70	$R = 5$	0.78	0.77	0.63

matching criterion. The bias of 1.16 in the ARW4 model is somewhat smaller than noted in D06b for forecasts integrated on a similar 4-km grid with an earlier version of the ARW model.

3) OBJECT-BASED GILBERT SKILL SCORE

An alternative, object-based, summary metric of model performance is the Gilbert skill score, which in this case is derived from the number of matched simple objects (hits) N_m, total forecast objects N_f, and unmatched observed objects (misses) M:

$$\mathrm{GSS} = \frac{N_m - \varepsilon}{N_f + M - \varepsilon}, \tag{2}$$

where ε is the number of hits expected due to random guessing. This GSS application differs from the traditional application of the GSS in precipitation verification because the evaluation here is performed for objects rather than individual grid cells. We present the object-based GSS because it has a statistical link with traditional verification metrics but retains the advantage of the object-based perspective.

While ε is trivial to compute in the traditional, grid-point-based application of the GSS, its meaning is not so obvious in terms of objects. Perhaps the best way to estimate ε for the object-based GSS is to compare a large sample of forecast and observed fields that have been randomly chosen but drawn from the same model and observed climatologies as the forecasts and observations we wish to evaluate. Wernli et al. (2008) pursued a similar strategy in their object-based verification study. Given a sufficient statistical sample, which may require running MODE through hundreds of cases, we can develop statistics of randomly matched objects.

A more expedient approach is to use the statistics of observed rain areas directly. It turns out that observed rain areas occupy about 1%–1.5% of the total geographic area in our dataset at any time. However, matching does not require overlap, so practically speaking, objects that nearly touch but have no overlap will be matched. A relevant area for considering matching is roughly 4 times the object area.[1] Therefore, the effective fraction of area occupied for the purpose of matching, f_A, is roughly 5% of the total area. Using this definition, the "null" fractional area is 95% of the domain. This means that, on average, 95% of the domain is empty; thus, a forecast object placed in this area would be unmatched. If we accept that 5% of the domain is the effective "size unit" of an object, then there would be 19 such size units in the 95% of the domain that is empty. That is, there would be 19 null objects for every observed object. In terms of the fraction area of an observed object, f_A, and N_o (the number of observed objects), the number of null objects is $(1 - f_A)N_o/f_A = 19N_o$.

The above definition of a null object allows us to compute the number of matches that will occur due to chance. This number is generally expressed as

$$\varepsilon = \frac{N_f N_o}{N_f + M + D}, \tag{3}$$

where D is the number of correct null forecasts (i.e., correctly predicting that no object will occur). We represent D as the number of null observations minus the number of forecast objects (because forecast objects cannot be correct null forecasts):

$$D = \frac{(1 - f_A)N_o}{f_A} - N_f. \tag{4}$$

Using the above expressions, we can compute the GSS for object matching. The chance of a correct random forecast with the above assumptions works out to be about 10% of the hits from the NMM4 forecasts and about 5% of the hits from the ARW4 forecasts. Including this definition of matches due to chance, both models attained a GSS of 0.42 when aggregated over the 32-day period. Here, it was assumed that a match occurred when the total interest for a forecast–observed object pair was at least 0.7. As indicated by A09, the traditional GSS values were typically around 0.1. One reason for the difference in the two sets of values is that spatial overlap of forecast and observed objects is not

[1] Consider an observed object represented by a circle of radius r centered at the origin. A forecast object will just touch the observed object if it is centered at a radius $2r$. Therefore, a forecast object placed anywhere within a radius of $2r$ from the center will probably match. If A is the area of the observed object, then a forecast object placed anywhere within an area of $4A$ will result in a match. For matching not to occur, the forecast object would need to be centered outside an area $4A$ centered on the observed object. Thus, the effective "cross section" of the observed object is roughly $4A$, not A. In practice, because matching can occur even if forecast and observed objects do not touch, the effective region where matching would occur is somewhat larger than $4A$.

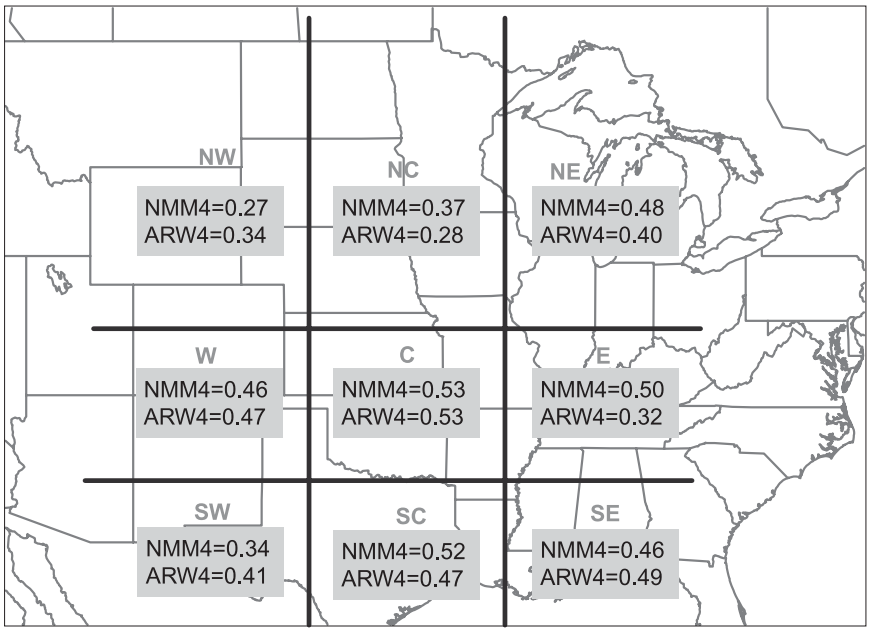


FIG. 9. Regional dependence of ETSs defined based on numbers of objects matching, false alarms, and misses. Results are valid for 32 cases. Regions are NW, northwest; W, west; SW, southwest; NC, north central; C, central; SC, south central; NE, northeast; E, east; and SE, southeast. Regions were defined based on terciles of the distributions of latitude and longitude of observed object centroids.

required to achieve a positive score in the object-based GSS, but such overlap is essential in the traditional application of the GSS.

Regional performance of both models was also examined using the object-based GSS. We divided the forecast domain into nine regions (Fig. 9) based on tertiles of the distributions of the latitude and longitude of observed objects. This was roughly the smallest division that retained a few tens of hits per region and hence could provide meaningful statistics. The GSS values (Fig. 9) show that the ARW4 model had somewhat higher scores in the western part of the domain, whereas the NMM4 model had higher scores in the eastern part of the domain.

The fact that both models performed similarly according to the object-based GSS, but not for the MMI metric (Table 2), may be reconciled in two ways. The first way involves hedging the GSS by overforecasting the number of objects. We can represent the effects of overforecasting the number of objects by considering every forecast object to be two adjacent objects that occupy exactly the same grid squares as the original object. For convenience, if we further assume that the numbers of hits, misses, and false alarms are all equal (which is not far from the real situation), and further that ε represents 10% of the hits, it turns out that the GSS for the model that overforecasts the number of objects (by a factor of 2) will be greater by a factor of 1.16 (0.31 versus 0.36). Thus, it is possible that the NMM4 was gaining some advantage by overforecasting the number of objects. Despite these drawbacks to using an object-based GSS, it may still have a useful application as an object-based verification metric that has a clear analogy with a traditional verification metric.

Another interpretation is that while total interest penalizes extreme outlier forecast objects more than ''near–false alarms,'' the GSS for object matching does not distinguish these. In other words, if we adopt a threshold for matching, as long as the maximum interest value among all object pairs containing a given forecast object remains less than this threshold, the GSS is insensitive to variations in the closeness of a match. Based on examples such as 4 June (Fig. 6), it could be that there are more outliers in the NMM4 model such that the interest distribution suffers more than the GSS. We view such penalization as a desirable outcome of using MMI (as MMIF or MMIO) as a metric of forecast quality.

4. Conclusions

We have used the Method for Object-based Diagnostic Evaluation (MODE) to compare forecasts made from the ARW and NMM models during the 2005 NSSL/SPC Spring Program. Both models were run without a traditional cumulus parameterization scheme on horizontal

grid lengths of 4 km (ARW) and 4.5 km (NMM). MODE was used to evaluate 1-h rainfall accumulations from 24-h forecasts valid at 0000 UTC on 32 days during the period from 23 April to 4 June 2005. A nine-case subset was also evaluated in more detail.

The primary variable used for evaluation was a ''total interest'' derived from a fuzzy-logic algorithm that compared several attributes of forecast and observed rain features. Total interest comprised such factors as the separation of the object centroids, minimum edge separation, orientation angle relative to the grid axis, the ratio of the areas of the two objects, and the fraction of area common to both objects. Total interest was defined for each pair of forecast and observed objects identified at a particular time. A matrix (called the interest matrix) was constructed in which the rows were forecast objects, columns were observed objects, and the matrix elements were total interest values. The maximum value of the total interest values along a row defined the best match for a given forecast object while the maximum interest down a column represented the best match for a given observed object. The median of the distribution of maximum interest values for forecast or observed objects (denoted MMIF or MMIO, respectively) defined a metric of forecast quality. This metric could be aggregated over a single forecast or any collection of forecasts to obtain a summary measure.

Results from the 32 cases suggest that, overall, the ARW4 model performed better than the NMM4 model based on the MMIF metric, whereas models performed nearly identically based on the MMIO metric. We demonstrated the statistical significance of the results for the MMIF metric using a Wilcoxon rank sum test. The primary reason for the poorer performance of the NMM4 model was the larger number of forecast rain areas with no observed counterpart (false alarms). Broadly similar conclusions regarding model performance were obtained by Marzban et al. (2008) and Marzban and Sandgathe (2009) using completely different verification methods.

However, it was noted that the performance varied considerably from day to day with most days featuring indistinguishable performance, and a small number of poor NMM4 forecasts likely producing the overall result. In particular, 0000 UTC 13 May and 0000 UTC 4 June were two times that revealed significant convective activity and an excessive number and size of rain areas in the NMM4.

We also compared the regional quality of forecasts by examining a Gilbert skill score for object matching. Within the construct of a 2×2 contingency table, matches were considered hits, objects forecast but not matched were false alarms, and objects observed but having no forecast counterpart were misses. We used the climatological fractional area covered by observed objects (5%) to estimate the number of hits due to chance. We found that the ARW4 performed somewhat better in the western part of the domain (the high plains) and the NMM4 was better over the Ohio Valley and southeast. Both models achieved GSS values near 0.42 overall. This result conflicted somewhat with the results based on total interest. However, the bias of the NMM4 model to forecast too many objects was found to artificially inflate the GSS and could have offset the somewhat lower forecast quality of that model relative to the ARW4 that was inferred from the maximum interest metric.

There are many possibilities for further study. One is a deeper investigation into the causes of the NMM overforecast of the number and extent of rain areas. The study by Skamarock (2004) suggests that excessive near-grid-scale variability can occur if numerical dissipation is not properly tuned. Skamarock and Dempsey (2005) showed that the NMM, as configured during the Developmental Testbed Center Winter Forecasting Experiment (see Bernardet et al. 2008 for configuration details), produced too much near-grid-scale energy in the lower troposphere. We speculate that given moist, unstable thermodynamic conditions, this energy could lead to excessive small-scale convection. Once initiated, this convection could then grow upscale. Alternatively, many nearby small-scale convective features could be grouped together as a large area by MODE. Either way, this could result in the occasionally poor NMM forecasts we noted, and could contribute to the overall biases in rain object area and intensity we have reported.

While the fuzzy-logic approach to matching behaves sensibly in most cases and is tunable by the user, it may be desirable to have a matching algorithm that is less empirical, with fewer parameters. Gilleland et al. (2008) demonstrated a computationally efficient implementation of the Baddeley delta image metric (Baddeley 1992), which encompasses a comparison of many object attributes into a generalized distance metric. This metric could augment or replace the fuzzy-logic method for estimating the quality of a match between object pairs.

Another avenue of further research is introducing dynamical variables into MODE so that the covariance of precipitation features with other aspects of the flow could be considered (a) to more intelligently perform matching and (b) to better link results back to the model dynamics in order to understand precipitation errors. For instance, one would be better able to categorize objects phenomenologically based on conditions near rain areas such as vertical wind shear and convective available potential energy. One could also make more intelligent choices about matching and merging rain areas given dynamical information such as the presence

of frontal boundaries, or physiographic information such as the presence of mountains or coastlines. The aim of these methodological augmentations is to more closely mimic the reasoning process an expert would use in assessing the quality of correspondence between forecast and observed features.

Acknowledgments. The authors wish to thank the Air Force Weather Agency and NCAR for supporting the development of MODE, and we thank David Ahijevych of NCAR for his helpful comments on an earlier draft of the manuscript. We also thank Mike Baldwin for providing the data from the NSSL/SPC Spring Program. NCAR is sponsored by the National Science Foundation.

APPENDIX A

Interest Maps

The attributes whose interest functions are nonzero are (a) centroid distance separation, (b) minimum separation distance of object boundaries, (c) orientation angle difference, (d) area ratio, and (e) intersection area. These attributes are weighted by (a) 24%, (b) 35%, (c) 12%, (d) 17%, and (e) 12%, respectively, in the computation of total interest. These are subjective choices based on empiricism. The minimum boundary separation was assigned the largest value because it was felt that this was the most robust factor describing forecast quality. The proximity of forecast and observed objects is represented by the centroid displacement, boundary displacement, and intersection area. Collectively, these account for nearly 70% of the total interest and treat objects having a wide variety of shapes and sizes. The other parameters (angle difference and area ratio) were deemed slightly less indicative of forecast quality, but still important enough to include.

Figure A1 shows the interest functions for each attribute. Only the absolute value of the orientation angle difference is considered. Note that this attribute is restricted to values of 90° or less. The area ratio is defined to fall between zero and unity. Therefore, this function does not distinguish whether the forecast or observed object is larger.

Confidence functions, which multiply the interest derived from a particular attribute, are less than unity if a particular attribute becomes overly sensitive to small changes in the data or otherwise fails to provide useful information about the forecast error in certain situations. For instance, if the aspect ratio of an object is nearly unity, the orientation angle becomes too sensitive to small changes in the shape of the object. We may then obtain large differences in orientation between forecast and observed objects for subtle changes in either shape. Hence, the confidence that we can meaningfully assign an error in the orientation angle approaches zero as the aspect ratio of either object approaches unity. The confidence function is expressed in terms of the aspect ratio (r) as

$$c = \left[\frac{(r-1)^2}{r^2+1}\right]^{0.3}. \quad \text{(A1)}$$

From (A1), it can be seen that the confidence is symmetric about an aspect ratio of unity. To derive the confidence for a pair of objects (forecast and observed), we compute the square root of the product of the confidence values obtained for each object separately. When this confidence value is nearly zero, we effectively remove the angle difference contribution to the total interest.

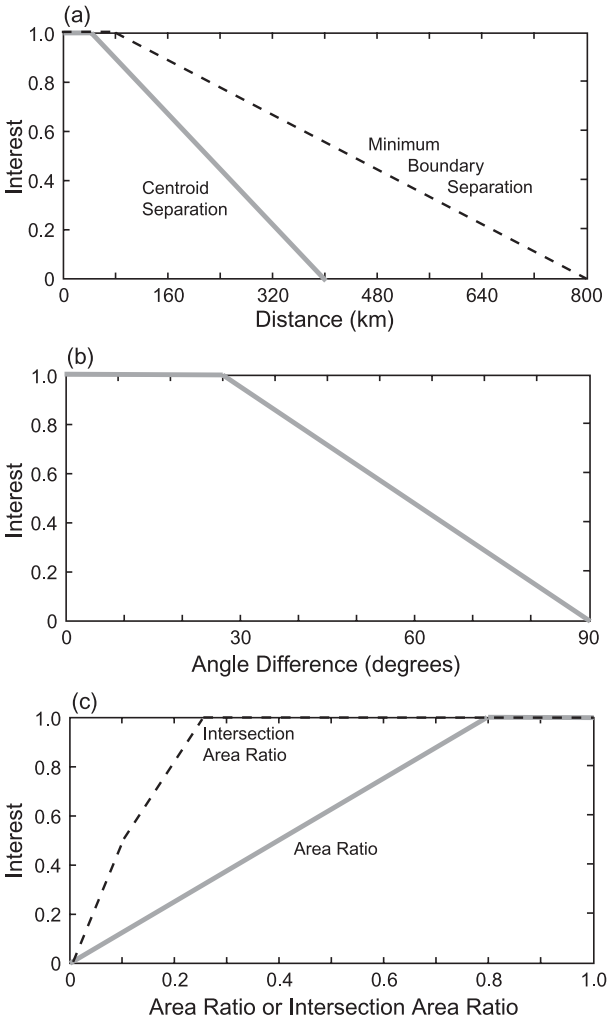


FIG. A1. Interest functions for (a) centroid separation (solid, gray) and minimum boundary separation (dashed), both in km; (b) angle difference (°); and (c) area ratio (solid, gray) and area intersection ratio (dashed).

The other nontrivial confidence function pertains to the distance separation of the two centroids of an object pair. If the two objects have a greatly different area (i.e., a small area ratio), centroid separation is a less meaningful measure of forecast error than if the forecast and observed objects have a similar size. We define the confidence for centroid separation to be equal to the area ratio so that when the area ratio is zero, the centroid separation is not considered in the computation of total interest.

APPENDIX B

Dates Examined

The following is the list of dates examined, for which forecasts from both the NMM4 and ARW4 models were available. The time evaluated was 0000 UTC in all cases, representing the 1-h accumulation of rain from a forecast initialized 24 h prior to the time shown. Dates highlighted in the nine-case sample appear in italics.

1)	23 April	17)	15 May
2)	*26 April*	18)	*18 May*
3)	27 April	19)	*19 May*
4)	29 April	20)	20 May
5)	30 April	21)	21 May
6)	1 May	22)	24 May
7)	3 May	23)	*25 May*
8)	5 May	24)	26 May
9)	6 May	25)	27 May
10)	7 May	26)	28 May
11)	8 May	27)	29 May
12)	9 May	28)	30 May
13)	10 May	29)	*1 June*
14)	11 May	30)	2 June
15)	*13 May*	31)	*3 June*
16)	*14 May*	32)	*4 June*

REFERENCES

Ahijevych, D., E. Gilleland, B. Brown, and E. Ebert, 2009: Application of spatial verification methods to idealized and NWP gridded precipitation forecasts. *Wea. Forecasting,* in press.

Baddeley, A. J., 1992: Errors in binary images and an L^p version of the Hausdorff metric. *Nieuw. Arch. Wiskunde,* **10,** 157–183.

Baldwin, M. E., and K. L. Elmore, 2005: Objective verification of high-resolution WRF forecasts during 2005 NSSL/SPC Spring Program. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction.* Washington, DC, Amer. Meteor. Soc., 11B.4. [Available online at http://www.amGSSoc.org/ams/pdfpapers/95172.pdf.]

Bernardet, L., and Coauthors, 2008: The Developmental Testbed Center and its winter forecasting experiment. *Bull. Amer. Meteor. Soc.,* **89,** 611–627.

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

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

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

Gilleland, E., T. C. M. Lee, J. Halley Gotway, R. G. Bullock, and B. G. Brown, 2008: Computationally efficient spatial forecast verification using Baddeley's delta image metric. *Mon. Wea. Rev.,* **136,** 1747–1757.

——, D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification. *Wea. Forecasting,* **24,** 1416–1430.

Hoffman, R., Z. Liu, J.-F. Louis, and C. Grassoti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.,* **123,** 2758–2770.

Hong, S.-Y., J. Dudhia, and S.-H. Chen, 2004: A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Mon. Wea. Rev.,* **132,** 103–120.

Janjić, Z. I., 2002: Nonsingular Implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso Model. NCEP Office Note 437, 61 pp.

——, 2004: The NCEP WRF core. Preprints, *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., 12.7. [Available online at http://ams.confex.com/ams/pdfpapers/70036.pdf.]

Kain, J. S., S. J. Weiss, D. R. Bright, M. E. Baldwin, and J. J. Levit, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting,* **23,** 931–952.

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

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. Preprints, *19th Conf. on Hydrology,* San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at http://ams.confex.com/ams/pdfpapers/83847.pdf.]

Marzban, C., S. Sandgathe, and H. Lyons, 2008: An object-oriented verification of three NWP model formulations via cluster analysis: An objective and a subjective analysis. *Mon. Wea. Rev.,* **136,** 3392–3407.

——, and ——, 2009: Verification with variograms. *Wea. Forecasting,* **24,** 1102–1120.

Noh, Y., W. G. Cheon, S. Y. Hong, and S. Raasch, 2003: Improvement of the K-profile model for the planetary boundary layer based on large eddy simulation data. *Bound.-Layer Meteor.,* **107,** 421–427.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

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

——, and D. Dempsey, 2005: High-resolution winter season NWP: Preliminary evaluation of the WRF ARW and NMM models in the DWFE forecast experiment. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction.* Washington, DC, Amer. Meteor. Soc., 16A.3. [Available online at http://ams.confex.com/ams/pdfpapers/94988.pdf.]

——, 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 TN-468+STR. 88 pp.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.,* **136,** 4470–4487.