Cluster Analysis for Object-Oriented Verification of Fields: A Variation

CAREN MARZBAN

Applied Physics Laboratory, and Department of Statistics, University of Washington, Seattle, Washington, and Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

SCOTT SANDGATHE

Applied Physics Laboratory, University of Washington, Seattle, Washington

(Manuscript received 16 August 2006, in final form 20 June 2007)

ABSTRACT

In a recent paper, a statistical method referred to as cluster analysis was employed to identify clusters in forecast and observed fields. Further criteria were also proposed for matching the identified clusters in one field with those in the other. As such, the proposed methodology was designed to perform an automated form of what has been called object-oriented verification. Herein, a variation of that methodology is proposed that effectively avoids (or simplifies) the criteria for matching the objects. The basic idea is to perform cluster analysis on the *combined set* of observations and forecasts, rather than on the individual fields separately. This method will be referred to as combinative cluster analysis (CCA). CCA naturally lends itself to the computation of false alarms, hits, and misses, and therefore, to the critical success index (CSI). A desirable feature of the previous method—the ability to assess performance on different spatial scales—is maintained. The method is demonstrated on reflectivity data and corresponding forecasts for three dates using three mesoscale numerical weather prediction model formulations—the NCEP/NWS Nonhydrostatic Mesoscale Model (NMM) at 4-km resolution (nmm4), the University of Oklahoma's Center for Analysis and Prediction of Storms (CAPS) Weather Research and Forecasting Model (WRF) at 2-km resolution (arw2), and the NCAR WRF at 4-km resolution (arw4). In the small demonstration sample herein, model forecast quality is efficiently differentiated when performance is assessed in terms of the CSI. In this sample, arw2 appears to outperform the other two model formulations across all scales when the cluster analysis is performed in the space of spatial coordinates and reflectivity. However, when the analysis is performed only on spatial data (i.e., when only the spatial placement of the reflectivity is assessed), the difference is not significant. This result has been verified both visually and using a standard gridpoint verification, and seems to provide a reasonable assessment of model performance. This demonstration of CCA indicates promise in quickly evaluating mesoscale model performance while avoiding the subjectivity and labor intensiveness of human evaluation or the pitfalls of non-object-oriented automated verification.

1. Introduction

It has become evident that the performance of numerical weather prediction (NWP) models must be assessed within a framework that acknowledges the existence of "objects" in the spatial field of observations and forecasts. Standard verification techniques ignore the spatial structure of forecast and observation fields and treat errors inappropriately. For example, a misplaced forecast's contribution to the mean squared error is independent of the magnitude of the displacement. Or, if there is partial overlap between an observed and a forecast object, then the forecast is penalized twice (both as a false alarm and a miss). These issues are discussed by Brown et al. (2004).

The aforementioned objects either can be discontinuous parameters, such as precipitation or cloud area, or they can be more conceptual entities that are defined by a large aggregate of features, and, therefore, are more difficult to assess (e.g., tropical or extratropical cyclones). Considerable progress has been made in this direction (Baldwin et al. 2001, 2002; Brown et al. 2002, 2004; Bullock et al. 2004; Chapman et al. 2004; Davis et al. 2006a,b; Du and Mullen 2000; Ebert and McBride 2000; Venugopal et al. 2005). The main thrust of these works is to identify and delineate objects in the two

Corresponding author address: Caren Marzban, Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, OK 73019.
E-mail: marzban@caps.ou.edu

DOI: 10.1175/2007MWR1994.1

fields in a meteorologically meaningful fashion in order to quantify model forecast skill. This injection of meteorology or mental models into the analysis is one of the strengths of these approaches. Other approaches have also been pursued (Casati et al. 2004; Nachamkin 2004).

In a recent article (Marzban and Sandgathe 2006, hereafter MS06), an alternative methodology was proposed that does not require such quantification of the objects. Specifically, a statistical procedure was employed to perform cluster analysis (CA) (Everitt 1980) on the observed and forecast field, separately, thereby automatically identifying objects in the two fields. In contrast to the aforementioned papers, here the identified objects are not parameterized at all.[1] Although this allows for the possibility that a given cluster may not be physically meaningful, it does have the advantage of allowing for automatic, objective verification.[2]

An important contribution of the CA-based approach is that one can assess the performance of an NWP model as a function of spatial scale. The specific CA algorithm that is employed is iterative (hierarchical), wherein the number of clusters is varied from N, the total number of grid points in the data, down to 1. As such, the number of clusters constitutes a measure of scale; a large number of clusters corresponds to verification on a small scale, and a small number of clusters is associated with large-scale verification. The two quantities N_o and N_f (the number of clusters in the observed field and that in the forecast field, respectively) span a 2D space, over which one can compute some scalar measure of performance. The resulting "error surface" captures the quality of the forecasts on different scales.

The error surfaces computed in MS06 plot a measure of distance between the observed and forecast fields. Several such measures of distance were computed, but they are all based on some metric of distance between matched clusters. As such, false-alarm and missed clusters did not contribute to the error surface; their numbers were reported separately. However, because their numbers can be computed, one can compute a scalar measure of performance, such as the critical success index (CSI), in which case the error surface would be a plot of CSI as a function of N_o and N_f. One such error surface is shown in the next section.

If one decides to assess performance in terms of false alarms, misses, CSI, or some other categorical measure (as opposed to a distance measure), then CA may be employed more efficiently.[3] Instead of performing CA on the two fields separately, and then matching the clusters between the two fields (which is effectively a third application of CA), one can perform one CA on the *combined set* of observations and forecasts. Each cluster will have some number of points belonging to the observation field n_o and some number of points originating from the forecast field n_f. Then, a comparison of these numbers can indicate whether the cluster should be counted as a hit, a false alarm, or a miss. As such, CSI can be computed upon a single application of CA. This method will be called combinative cluster analysis (CCA), and is further described in section 3.

The number of clusters in the combined dataset may still be interpreted as a measure of scale, in the same sense that the number of clusters in the separate fields is related to scale (as in MS06). A small (large) number of clusters in the combined set corresponds to the situation wherein the verification is done at low (high) resolution. Although there is no simple relationship between the number of clusters and the more traditional notion of spatial resolution, in the discussion section it is argued that the former is a more natural notion of scale within an object-oriented framework.

The aim of the current work is to introduce CCA. Although some general results are reported (see appendix B), this work is mostly an application of the methodology to forecasts of reflectivity from the University of Oklahoma's Center for Analysis and Prediction of Storms (CAPS) Weather Research and Forecasting Model (WRF) at 2-km resolution (arw2), the National Center for Atmospheric Research (NCAR) WRF at 4-km resolution (arw4), and the National Centers for Environmental Prediction (NCEP)/National Weather Service (NWS) Nonhydrostatic Mesoscale Model (NMM) at 4-km resolution (nmm4). The next section will review the basic idea of CA-based verification. It is followed by a section that presents the details of the more direct CCA-based approach. The paper ends with the conclusions and a deeper discussion of the results.

[1] As described in the data section, the analysis here is performed only on grid points with reflectivity values exceeding some prespecified threshold; that threshold does not parameterize *identified* objects, however.

[2] A type of CA has been utilized by Lakshmanan et al. (2003) and Peak and Tag (1994) for both storm and cloud identification.

[3] In an object-oriented verification of two fields, one can unambiguously compute the number of false alarms, misses, and hits. The remaining element of the contingency table (i.e., the number of correctly forecast nonevents) is not readily computable. As such, only scalar performance measures that are independent (at least, explicitly) of this element are desirable. One such measure is the CSI, defined as the number of hits divided by the total number of hits, false alarms, and misses.

2. Prior work

MS06 employed agglomerative hierarchical CA for the purpose of identifying clusters in a forecast and an observation field. Several measures of intercluster distance were examined, for example, the group average, shortest distance, and longest distance. Additionally, the distance between two points (in one or two clusters) was computed with L2 and L1 norm.[4] It was found that CA based on the group-averaged L2 (i.e., Euclidean) distances yields clusters that are physically reasonable.

A particularly desirable feature of the hierarchical approach is that the number of clusters is not a fixed quantity. The specific type of CA adopted in both MS06 and here begins with a number of clusters equal to the number of data points (exceeding some value), and iteratively merges them into larger clusters. At the end of the procedure there is only one cluster, containing all of the data. This is desirable for verification, because the number of clusters is effectively a measure of spatial scale. The first iteration of CA corresponds to the very fine scale, where every grid point is considered as an independent object, while the last iteration of CA corresponds to a very course scale, where the entire field is considered to be a single homogeneous object. In short, CA-based verification conveys information about the quality of the forecasts at any given scale, and therefore, across all scales. This issue is further elaborated in the discussion section.

In MS06, at every iteration of CA, with N_o clusters in the observed field and N_f clusters in the forecast field, all $N_o \times N_f$ distances between interfield clusters were computed, and the clusters with the smallest distance were considered as a candidate match. This distance was recorded. Upon excluding these (nearest) clusters, the next smallest distance was recorded and the corresponding clusters were considered as another candidate match. This process was repeated for all clusters. The distribution of all of the recorded distances was then considered and only those within one standard deviation, for example, of the median, were accepted as matches (hits); clusters with distances outside that range were defined as either a false alarm or miss, depending on the field to which they belong.

In short, CA-based verification was performed by first performing CA in the two fields, separately, and then utilizing certain matching criteria for identifying clusters between the two fields. Given the matching, it was possible to compute the interfield distance as a simple average of the distances between the matched clusters. This distance was plotted as a function of N_o and N_f, resulting in the aforementioned error surface.

Moreover, CA was performed not only in the two-dimensional space of (x, y) values, labeling grid coordinates, but also in the 3D space of (x, y, p) values, with p representing the amount of precipitation accumulated at coordinates (x, y) over a specified period of time. A CA-based verification in (x, y) captures performance only in terms of the spatial placement of the clusters, and their size, whereas CA-based verification in (x, y, p) includes precipitation amount as well. MS06 may be consulted for further detail.

Because that methodology is capable of identifying hits (i.e., matches), false alarms, and misses, it is possible to also compute a categorical measure, such as CSI, as a function of N_o and N_f. (This surface may be more aptly called a performance surface, because CSI is a positive measure of performance; that is, higher CSI implies better performance.) An example of such a surface for precipitation forecasts from an NCAR WRF forecast is shown in Fig. 1. It can be seen that CSI is generally higher when an equal number of clusters exists in both fields. This is not surprising; however, what is surprising is that the height of the surface along the diagonal undulates. The highest CSI values appear on the scale of 5–10 clusters, followed by another peak (albeit lower) at 40–50 clusters; even on the scale of 80–100 clusters there is another low and broad peak. One explanation of this pattern is that the observation and forecast fields have inherent scales, and they better match one another on those scales. Away from the inherent scales, the agreement between the two fields is smaller. Another application of such performance surfaces can be in model comparison. In other words, one can compare two (or more) NWP models by comparing their performance surfaces. The graphical challenges are formidable and so will not be considered here. The main point is as follows: performance depends on scale, and so should be assessed accordingly.

As mentioned in the previous section, if a categorical performance measure is deemed as being sufficient, then one can compute it more efficiently using CCA; as a side effect, it also avoids the aforementioned graphical challenges. The details are provided next.

3. Combinative cluster analysis methodology

The basic idea examined here is to perform CA on the combined set of forecasts and observations. But first, to properly understand this methodology, one must distinguish (conceptually) between a cluster in the combined data, as identified by CA, and "clusters" in

[4] The L2 and L1 distances between points x and y refer to $(x - y)^2$ and $|x - y|$, respectively.

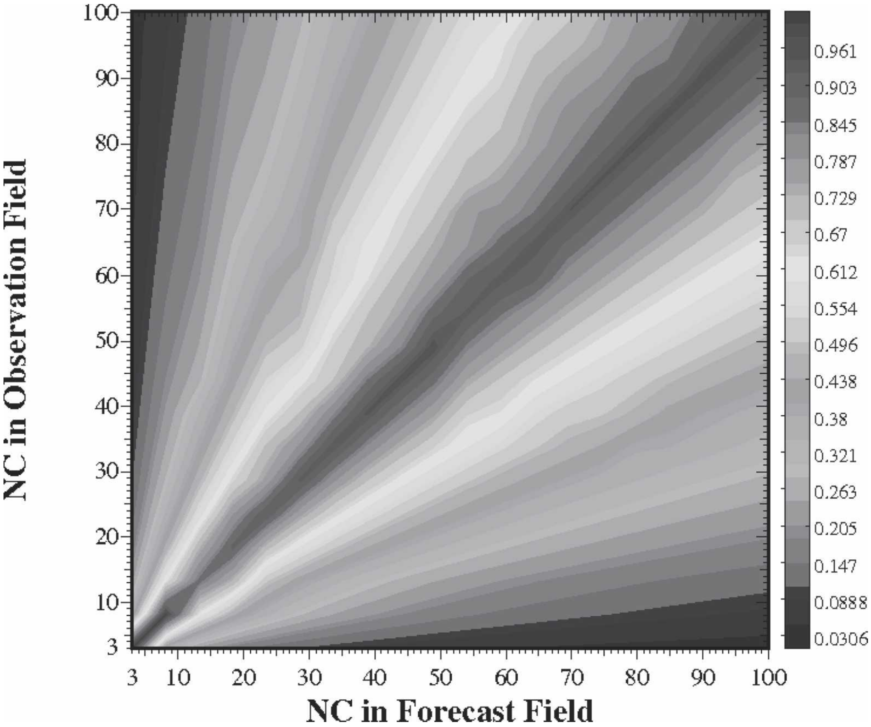

FIG. 1. An example of a CSI performance surface is shown for arw2 24-h forecasts, according to the methodology of MS06, with the CA performed in (x, y, Z).

the separate observation and forecast fields. Here, we will refer to the former as simply *cluster* and to the latter as *underlying cluster*. It is important to point out that a cluster is an entity determined by CA, while underlying clusters are what a human analyst, for example, might perceive as objects in the two fields.

With this terminology, at any given iteration of CA, a cluster will consist of a subset of underlying clusters. If forecasts are perfect in every respect, then a cluster will be exactly equal to the union of two underlying clusters—one from each field. Also, a cluster is designed to consist of similar points. It follows, then, that for perfect forecasts, a cluster will consist of identical underlying clusters, with one from each field. For imperfect forecasts, a cluster will consist of similar underlying clusters, with one from each field as well. If CA is performed in (x, y) space, then the underlying clusters are similar in terms of their spatial placement. Similarly, CA in (x, y, Z) will yield clusters whose underlying clusters are similar in terms of their spatial placement and reflectivity (Z). Note, however, that the aforementioned similarity is not quantifiable, because the underlying clusters are not quantifiable. The latter would require performing CA in the separate observation and forecast fields, that is, the methodology of MS06, which is not the approach followed here.

The question then arises as to *how much* of a cluster is from an underlying observed cluster, or, equivalently, *how much* of it is from a forecast underlying cluster? This question is important in deciding whether a cluster should be treated as a miss, hit, or false alarm. One criterion for assessing this quantity involves the *size* of the underlying clusters. (Recall that the *placement* and *intensity* of the clusters have already been quantified within CA itself.) If, for example, less than 10% of a cluster is composed of grid points in the observation field, then that cluster may be declared to be a false alarm. Similarly, if more than 90% of a cluster is composed of grid points in the observation field, then the cluster may be marked as a miss. Finally, if a cluster is composed of comparable contributions from the two fields, then it may be considered a hit.

The size of a cluster is determined by the number of grid points defining it. Each cluster is then composed of some number of points (n_o) from the observed field and some number of points (n_f) from the forecast field. Then, the proportion $n_o/(n_o + n_f)$ can be employed to define hits, false alarms, and misses. For example, if that proportion for some cluster is near zero, then that cluster may be defined as a false alarm. At the other extreme, a cluster with a near-1 proportion could be considered a miss, while clusters with intermediate proportions would qualify as hits. Therefore, an unambiguous definition of the three quantities (false alarm, hit, and miss) follows from the specification of two thresholds. To simplify even further, in this work a single threshold is considered. A cluster with $n_o/(n_o + n_f)$ less than the threshold is considered a false alarm, and one with a proportion $1 - n_o/(n_o + n_f)$ exceeding the threshold is considered a miss; a cluster with any other ratio is defined to be a hit. This operational definition of false alarm, hit, and miss is shown in Fig. 2. In summary, if the threshold is low (e.g., 0.01), then the procedure is more generous in allowing a cluster to be classified as a hit. As the threshold approaches 0.5, fewer clusters will be classified as hits.[5]

In principle, one may consider all threshold values, but given that the emphasis of this work is an exploration of the new combinative cluster analysis method, only two values are considered: 0.01 and 0.1. The former threshold means that if less than 1% of a cluster is composed of observed points, then it will be classified as a false alarm. Also, if less than 1% of a cluster consists of forecast points, then it is classified as a miss. A threshold of 0.1 is similar, except the relevant percentage is 10%.

One may question the sensibility of this threshold criterion for the identification of false alarms, hits, and misses. After all, even if a cluster consists of equal amounts of two underlying clusters (one from each field), it makes little sense to consider that cluster a hit if the underlying clusters are unreasonably distant from each other, that is, if an expert might consider the two underlying clusters as two separate objects, unrelated in any physical way.

This is a valid objection; however, one must then recall the iterative nature of the CA-based methodology. Specifically, although it is true that at some iteration of CA two distant underlying clusters may be identified as a single cluster, it is also true that at some earlier iteration they would have been considered as two separate clusters. As such, in deciding whether a cluster should be classified as a false alarm, hit, or a miss, the iterative nature of the CA-based approach obviates any need to address the "closeness" between the two underlying clusters, because different scales are scanned as CA proceeds from one iteration to another.

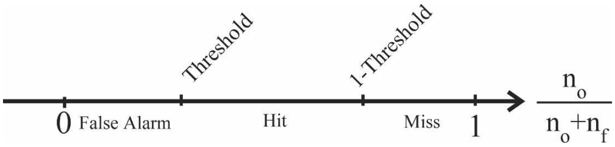

FIG. 2. The definition of false alarm, hit, and miss, as a function of a threshold placed on $n_o/(n_o + n_f)$ is shown, that is, the fraction of the grid points in a cluster that belong from the observation field.

CA does have a number of "parameters" that must be specified (see MS06). But here, given the exploratory nature of the work, the role of these parameters (e.g., choice of distance measure and norm) is not examined; only group-averaged distances with an L2 norm are considered.

The analysis is performed in both (x, y) and (x, y, Z) space, where Z refers to reflectivity. In both cases, all of the coordinates are standardized (i.e., for every coordinate, the mean is subtracted out and the result is divided by the standard deviation). In other words, each coordinate has a mean of 0 and a standard deviation of 1.

Finally, there are two components in CA that are computationally intensive: 1) the computation of a group-averaged distance between two clusters of size N_1 and N_2, respectively, requiring $\sim N_1 \times N_2$ distance computations; and 2) the identification of the two nearest clusters, which requires the computation of the N^2 group-averaged distance, where N stands for the number of clusters. To expedite these components of the analysis, group-averaged distances are computed only for 50% (randomly sampled) of the points in each cluster. Also, not all of the intercluster distances are computed; on the first iteration of CA, only 50% of the clusters are randomly selected, and their distances are computed. That percentage is increased according to an exponential rate, guaranteeing that 100% of the clusters are considered at the last iteration.[6]

4. Data

An extensive dataset dealing with reflectivity forecasts and observations is currently being compiled by M. Baldwin of Purdue University (2007, personal com-

[5] It may appear that decreasing the threshold can improve performance on all accounts, because the number of hits increases and the number of false alarms and misses decreases. Of course, this occurs only because one component of performance is neglected here, namely, the correct forecast of nonevents. Unfortunately, that component of performance is not uniquely defined for the verification of fields, and so one cannot employ measures that explicitly account for all four components of performance, for example, Heidke's skill score. As such, CSI can be utilized only for the purpose of comparing models, and even then it is important to keep in mind all of the defects associated with CSI (Marzban 1998).

[6] Computing the group-averaged distance using only 50% of the points in each cluster is similar to estimating a population parameter by a sample statistic. As such, it is a relatively standard procedure. However, it is more cavalier to match clusters by starting with only 50% of the intercluster distances. However, experimentation (not shown) suggests that although the specific clustering of the data is affected for the first few iterations of CA, the final configuration of the clusters is relatively robust. Again, here, the main purpose of this step is to expedite the analysis.

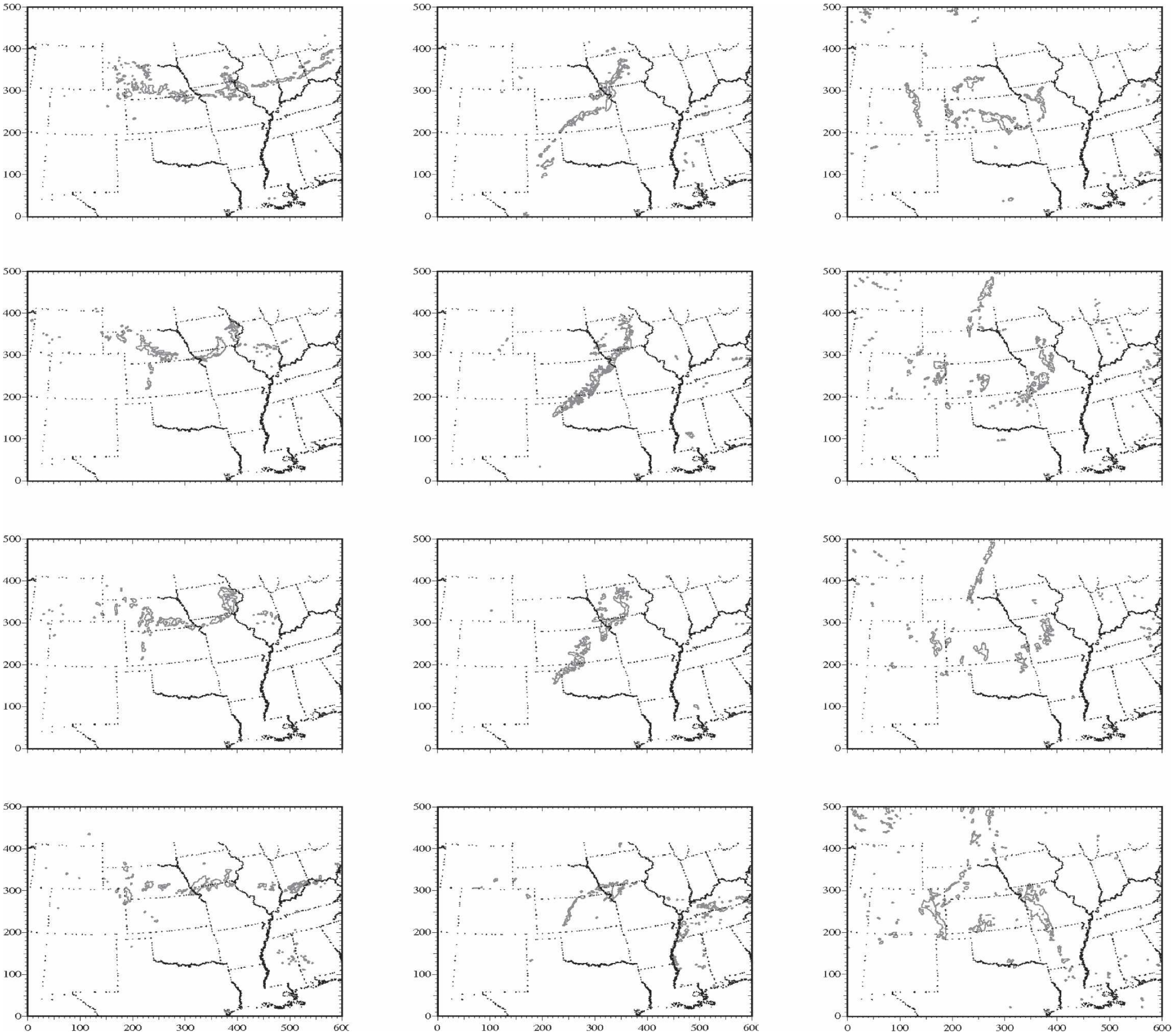

FIG. 3. (top row) Reflectivity observations and (other rows, top to bottom) the corresponding 24-h forecasts according to arw2, arw4, and nmm4. The columns refer to the three dates examined: 12 and 13 May and 4 Jun 2005. The coordinates of the region are 30°N, 70°W; 27°N, 93°W; 48°N, 67°W; 44°N, 101°W, which covers the United States east of the Mississippi. Only grid points whose reflectivity exceeds some value are displayed (see text).

munication). Here, however, only three specific days—12 and 13 May, and 4 June 2005—from that dataset are employed. The grid spacing is 4.7625 km. Figure 3 displays the observations and the 24-h forecasts according to arw2, arw4, and nmm4 (but only for reflectivity exceeding some value; see the next paragraph). The coordinates of the four corners of the region are 30°N, 70°W; 27°N, 93°W; 48°N, 67°W; and 44°N, 101°W, covering the United States, east of the Mississippi.[7] As an aside, it is telling to note that, even at a very high resolution and short forecast interval, the model predictions are closer to each other than they are to the observations.

The analysis is performed on only data whose reflectivity exceeds 40 dBZ for the 12 and 13 May data, and 35 dBZ for the 4 June data. There are two reasons for "thresholding" the reflectivity: first, it reduces the amount of data, and therefore expedites the CA phase of the analysis; and second, it isolates the more significant events, rendering the data more "lumpy" (i.e., more clusters in the two fields), and therefore provides for a better environment for testing the CCA approach.

[7] We are grateful to Mike Baldwin for providing the data for this analysis.

The specific thresholds of 35 and 40 dBZ are selected to ensure that all three dates yield approximately the same sample size for the analysis.

5. Results

CCA begins by assigning every point in the data (exceeding the reflectivity threshold) to a single cluster of size 1. It then identifies the two nearest clusters and merges them into a new cluster. The procedure is iterated until all of the data are merged into a single cluster. It is possible to examine the clusters at each iteration. For CA-based verification, it is more useful to view the clusters in such a way as to convey some information regarding the quality of the forecasts. To that end, in this work, the clusters are displayed as follows: if a cluster passes the test of being a hit (see Fig. 2), then its n_o observation points are plotted in one panel, and the n_f forecast points are plotted in an adjacent panel. Similarly, if a cluster is declared to be a false alarm, then its n_o observation points and n_f forecast points are plotted in adjacent panels. Missed clusters are split between two adjacent panels in the same fashion. Each of the three pairs of panels is displayed on a single figure. In the terminology of section 3, every cluster is decomposed into its underlying clusters and displayed separately.

Examples of this layout are shown in Fig. 4, for 13 May (top) and 4 June (bottom), respectively, with 13 clusters in the joint set of observations and arw2 forecasts. The CA is performed in (x, y, Z) space, and the threshold is 0.1. This higher threshold is employed for the presentation, because it allows for more false alarms and misses to appear in the figures; the lower threshold of 0.01 produces too few false alarms and misses, and results in noninformative figures (not shown). The top two panels display the hits, with the observation points plotted in the left panel and the forecast points plotted in the right panel. Clearly, it is desirable that forecasts should populate these two panels much more so than the remaining panels, which, in these examples, they do. Moreover, the matched colors between the two panels indicate a reasonable matching of the clusters between the two fields. Note that the southern extension of the line of reflectivity located in the left (western) third of the display is forecast too far to the east, and the northern extension of the reflectivity in the right (eastern) third of the display is forecast to extend too far to the north, yet CCA identifies these areas as "hits." In other words, these features are forecast with only an error in location (or timing). This is a good example of the strength of object-oriented methods such as CCA.

The false alarms are shown in the two panels in the second row of Fig. 4, and the misses are displayed in the two panels in the third row. The arw2 forecast for this particular date misses a significant portion of the eastern extension of the reflectivity band. This is easily determined by CCA, along with a few other misses. The low number of false alarms and misses is the result of the forgiving threshold choice of 0.1. Analogous results (not shown) for smaller thresholds, for example, 0.01, display even fewer clusters in the middle-left and bottom-right panels. In short, the rarity of clusters in these two panels is a direct consequence of the smallness of the threshold.

The analogous results for 4 June are shown in the bottom half of Fig. 4. The arw2 overforecasts for this day, including a significant reflectivity band in the upper-middle portion that does not occur in the observations. CCA identifies this feature as a false alarm and identifies a number of additional false alarms. CCA also correctly identifies a significant number of isolated reflectivity areas as misses.

The discussion of the previous few paragraphs is intended to reveal the inner workings of CCA in a verification setting. Independent of that discussion, one can compute CSI at each iteration, and in an objective and automatic fashion. Figure 5 displays the results for arw2, arw4, and nmm4 with the analysis performed in (x, y, Z) space, with a threshold of 0.01.

The manner in which the error bars for CSI are computed is described in appendix A. They are intended to convey a rough estimate of the sampling variation. It should be noted that the error bars are only standard errors, and not true confidence intervals. The latter would require some distributional assumptions. Confidence intervals would be larger than the error bars shown in this work. As such, an overlap between two error bars would suggest "statistical equivalence," but an absence of an overlap may not imply that the curves are statistically distinct. Given the crude manner in which they are computed, these error bars should be interpreted with care; their appearance in these graphs is intended to be only a reminder that the CSI values are not free of sampling error. For the purpose of visual clarity, however, the error bars are not shown in all figures.

Referring to Fig. 5, the following observations can be made:[8]

[8] To provide a reference for comparing these CSI values, a simple gridpoint verification is performed. For all dates and models the CSI values are in the range of 0.01–0.06. Such low CSI values are entirely expected, according to the "double penalty" that occurs in non-object-oriented verification (Brown et al. 2002).

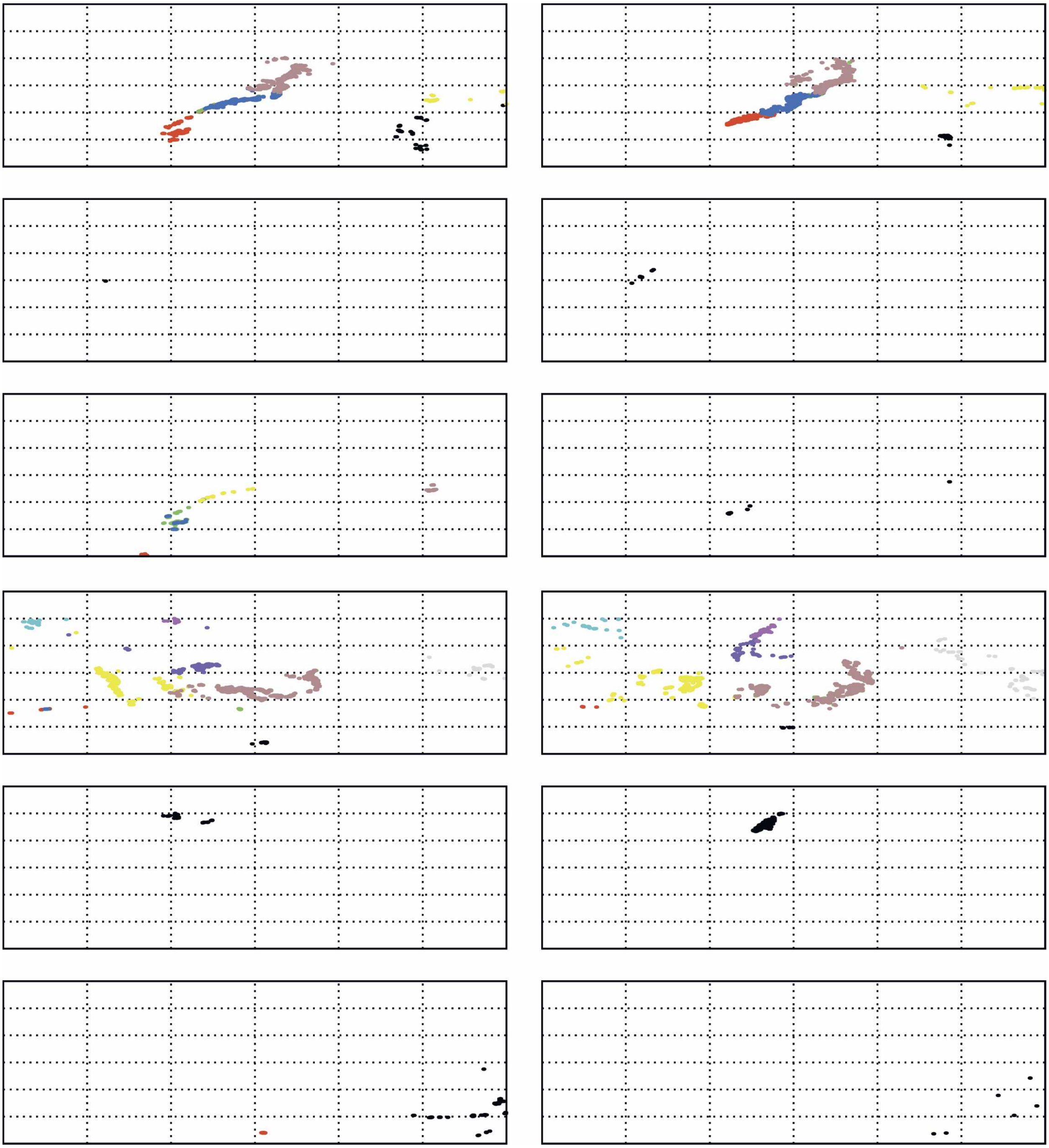

FIG. 4. The (left) observed and (right) arw2 forecast fields on (top half) 13 May and (bottom half) 4 Jun. The top, middle, and bottom rows in each half represent hits, false alarms, and misses, respectively; see text for details. The colors represent different clusters according to CA, and similarly colored clusters in adjacent panels indicate matched pairs. These images are extracted from CA at the iteration corresponding to 13 clusters. The threshold is set at 0.1 in order to allow for more false alarms and misses to appear in these panels.

1) On all three dates, arw2 outperforms arw4 and nmm4 to some extent. This should be expected, considering its ability to better resolve smaller-scale features and convective precipitation. (The difference between arw2 and arw4 is least on 13 May).
2) This outperformance is generally true across most scales examined, except on the extremes (see the following items).
3) On larger scales (small cluster numbers), there are very large variations in CSI. This reflects the merg-

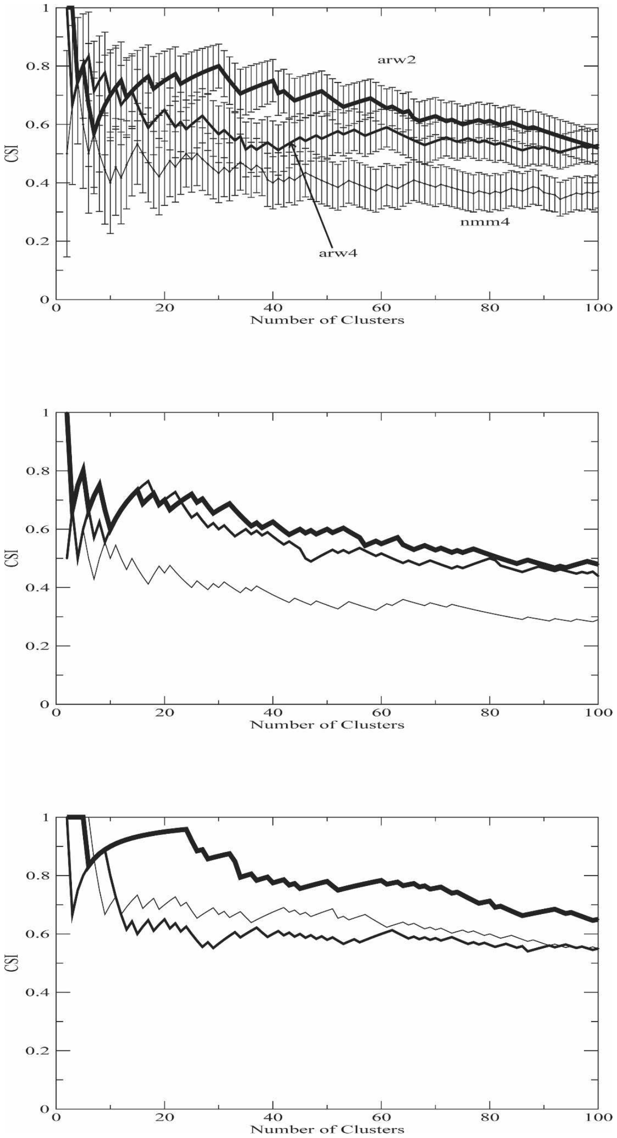

FIG. 5. The CSI values for the three models on (top) 12 and (middle) 13 May, and (bottom) 4 Jun. The analysis is done in (x, y, Z) with a threshold at 0.01. The error bars are no more than a rough estimate of sampling variation.

ing of generally unrelated areas of reflectivity together into single clusters, causing erratic verification results. Visually scanning the forecasts for the three dates (Fig. 3) reveals that no fewer than 10–15 clusters should be considered for such a large region.

4) On smaller scales (larger number of clusters), the differences between the models appear to diminish.
5) The performance of all three models falls off with increasing cluster number.

Appendix B provides some theoretical explanation of these and other behaviors. It is not surprising that all three models have lower performance when forecasting at smaller scales than on larger scales.

To examine the effect of the threshold, the analysis is repeated with the threshold raised to 0.1. Figure 6 (middle column) displays the results, while the left column is Fig. 5, reproduced here for comparison. Recall that a larger threshold is expected to be accompanied with fewer hits, but more false alarms and misses. This expectation is borne out in these results. The behavior of the CSI curves is generally the same for the different threshold values. One noteworthy difference is that the differences between the three models are diminished at the higher threshold, albeit to different degrees for the three dates. Another way of stating the effect of the threshold is that when the verification procedure is allowed to be more generous in allowing false alarms and misses, then the three models have more similar performances in terms of CSI, across all scales.

To assess the effect of the reflectivity (Z) on the analysis, Fig. 6 (right column) shows the analog of the left column, but with the analysis performed in (x, y) space. Recall that an (x, y) analysis assesses performance only in terms of the spatial characteristics of the clusters, while an (x, y, Z) analysis includes reflectivity as well. A comparison of the left and the right columns in this figure suggests that the CSI curves for the (x, y) analysis start higher but fall off much faster with increasing cluster number than those for the (x, y, Z) analysis. In other words, with reflectivity included in the analysis, performance is generally higher and more stable across different scales. Moreover, the differences between the three models are less pronounced when only spatial characteristics of the forecasts are taken into account; that is, the models are more similar in terms of the spatial characteristics of their forecasts, and the inclusion of reflectivity in the analysis sets the models apart. It is likely that arw2, with higher resolution, is better able to predict the extremes in reflectivity, which give it an advantage in the (x, y, Z) analysis.

6. Summary, conclusions, and discussion

Reflectivity forecasts from arw2, arw4, and nmm4 are verified against observations in a framework that acknowledges the existence of clusters (i.e., objects) in the two fields. The methodology is based on a class of statistical methods called the agglomerative hierarchical cluster analysis (CA). It is found that when the quality of the forecasts is assessed in terms of the critical success index, the three models appear to have similar performance, at least when only the spatial placement of the clusters is concerned. However, when reflectivity

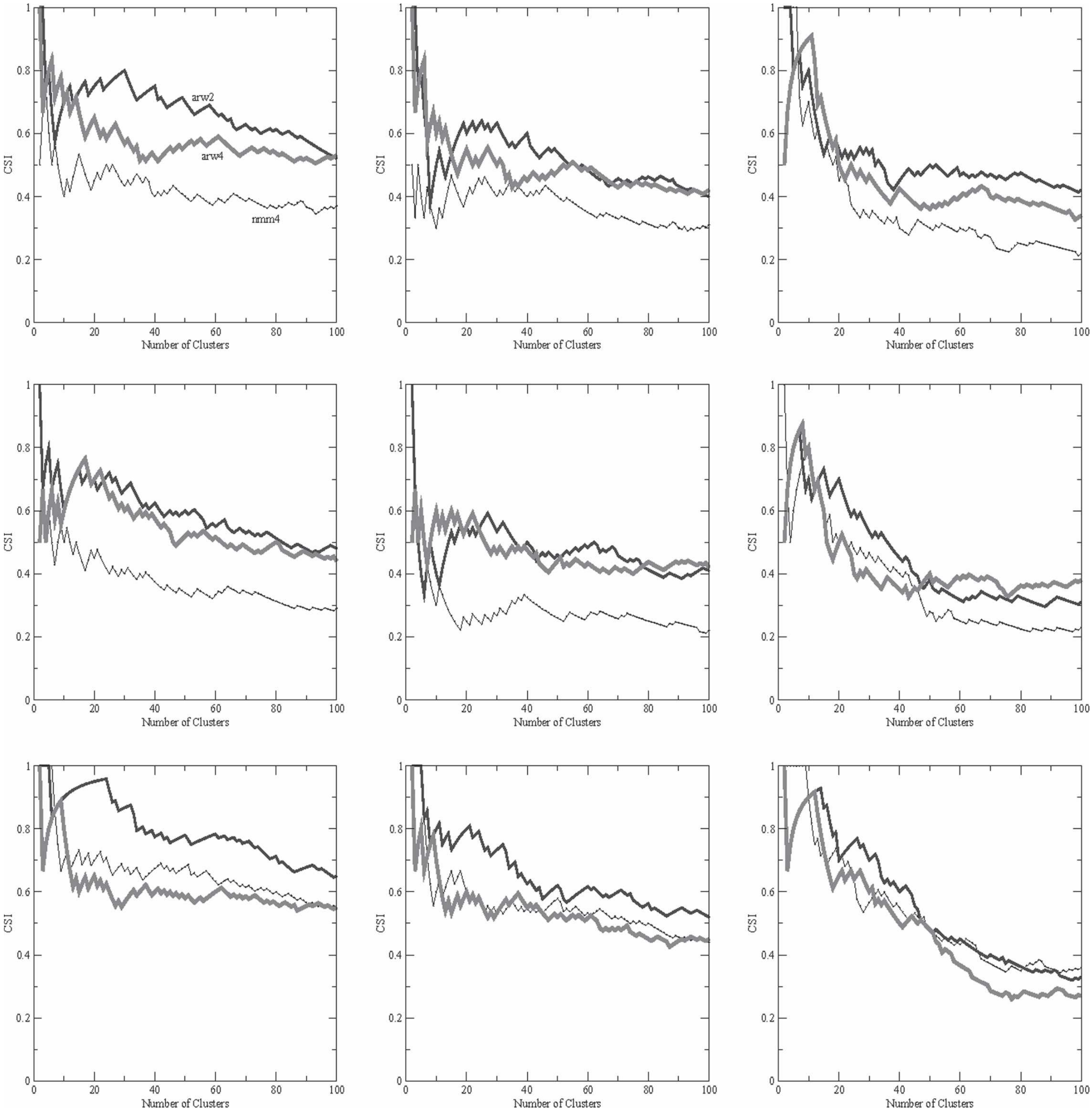

FIG. 6. Similar to Fig. 5, for all three dates: (top) 12 May and (middle) 13 May, and (bottom) 4 Jun. The left column is a reproduction of Fig. 5, but with the error bars suppressed for visual purposes. The middle column is the same but with the threshold at 0.1. The right column is with the threshold at 0.01 (same as the left column), but with the analysis performed in (x, y) space.

is also included in the analysis, then arw2 appears to have a slight edge over the other contenders, likely reflecting a better capability to predict reflectivity extremes or gradients. The slightly higher-quality forecasts of arw2 appear to persist across the full range of scales examined herein.

In short, the main conclusions of this study are that the proposed methodology for automatic object-oriented verification appears to be sound, producing reasonably meaningful results. Furthermore, based on this methodology, arw2 forecasts emerge as marginally superior to those of arw4 and nmm4 across a wide range of spatial scales for three dates. A sample size of 3 is clearly inadequate to establish model performance; however, visual inspection and gridpoint verification bear out the success of the methodology.

The issue of scale requires further discussion. In a CA-based approach to verification, the number of clusters is treated as a measure of spatial scale. A natural question is how is this notion of scale related to the more traditional notion of scale, namely ordinary/spatial distance? There cannot be a general answer to this question, because the answer depends on the meteorological phenomenon under examination. For example, if one is assessing the quality of forecasts of an organized system (e.g., a frontal system) occupying a major portion of the forecast field, then examining the problem on a fine scale will not affect the number of clusters. On the other hand, if the weather phenomenon is highly "lumpy," such as an outbreak of airmass thunderstorms, then the number of clusters will depend on the spatial scale being analyzed; this is true for both a CA-based approach as well as for a human forecaster. In short, there is no simple relationship between the two notions of spatial scale: number of clusters and resolution. Indeed, one may argue that in an object-oriented approach to verification the former provides a more natural notion of scale, because it is based on the number of objects themselves.

A comment about the general behavior of the CSI curves seen here is in order. A common feature of these curves is that they appear to have "phases" during which they increase, only to be followed by a decreasing phase. The right panels in Fig. 6 display this clearly. Appendix B presents some theoretical arguments to explain this behavior. Briefly, the transition between the two phases occurs when the decrease in CSI, resulting from a split in a cluster, is compensated by an increase in the number of hits.

There are several directions in which this analysis can be generalized. As mentioned in section 3, in the current work the notion of a false alarm, miss, and hit depends on a single threshold (see Fig. 2). It may be worth exploring two different thresholds delineating false alarms, hits, and misses. It may also be worthwhile exploring a wider range of the two thresholds, beyond the values 0.01 and 0.1 considered here. In this way, one will again have a performance *surface* over a 2D space spanned by different values of the threshold. It will also be interesting to compare the resulting performance surface with that obtained by the method described in MS06. Moreover, the analysis should be extended to other dates and other types of weather phenomena.

Acknowledgments. The authors would like to acknowledge Mike Baldwin, Barbara Brown, Chris Davis, and Randy Bullock for contributing to all levels of this project. Partial support for this project was provided by the Weather Research and Forecasting Model Developmental Testbed Center (WRF/DTC), and by the National Science Foundation Grant 0513871.

APPENDIX A

Error Bar for CSI

There are numerous methods for computing error bars for a test statistic such as CSI. Perhaps the most accurate ones are confidence intervals derived from a resampling approach (e.g., bootstrap), wherein the distribution of CSI is approximated by the histogram of some number of CSI values, each obtained from a sample taken from the data. This approach, however, is computationally intensive, and not entirely necessary for the task at hand. The main purpose for introducing error bars in the current analysis is simply to emphasize that CSI is a random variable. As such, it is sufficient to adopt a simple (and crude) probability model for the errors in the various elements of the contingency table, and then simply "propagate" those errors to CSI.

In statistics, this is often referred to as the delta method.

Consider the contingency table

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{A1}$$

where b stands for the number of false alarms, c for the number of misses, and d for the number of hits. The marginals $N_0 = a + b$ and $N_1 = c + d$ are the total number of observed nonevents and events, respectively. Similarly, $F_0 = a + c$ and $F_1 = b + d$ are the total number of forecast nonevents and events, respectively. A reasonable assumption is that b is a binomial random variable with parameters N_0 and a/N_1.[A1] However, as mentioned previously, a is ambiguous (or extremely large; at best, equal to the number of grid points associated with no forecasts and no events) in the verification problem at hand.

Instead, and in the spirit of CSI, it is possible to assume that b is a binomial random variable with parameters F_1 and b/F_1. Similarly, one may assume that c is a binomial random variable with parameters N_1 and c/N_1. Then, the expected value and variance of b and c are given by

[A1] A binomial random variable x, with parameters N and π, is defined by the following mass function: $p(x; N, \pi) = \binom{N}{x} \pi^x (1 - \pi)^{N-x}$.

$$E[b] = F_1 \frac{b}{F_1} = b, \tag{A2}$$

$$\mathrm{Var}[b] = F_1 \frac{b}{F_1}\left(1 - \frac{b}{F_1}\right) = \frac{bd}{F_1}, \tag{A3}$$

$$E[c] = N_1 \frac{c}{N_1} = c, \tag{A4}$$

$$\mathrm{Var}[c] = N_1 \frac{c}{N_1}\left(1 - \frac{c}{N_1}\right) = \frac{cd}{N_1}. \tag{A5}$$

Writing CSI as

$$\mathrm{CSI} = 1 - \frac{b + c}{b + c + d}, \tag{A6}$$

and assuming $b + c + d$ to be a constant, it follows that the standard deviation of CSI is given by

$$\sigma_{\mathrm{CSI}} = \mathrm{CSI}\sqrt{\frac{1}{d}\left(\frac{b}{b+d} + \frac{c}{c+d}\right)}. \tag{A7}$$

Note that the two terms in the parentheses are the false-alarm ratio and the miss rate. The error bars reported in this work are given by this σ_{CSI}.

It is worth pointing out that this notion of an error bar is based on a number of simplifying assumptions whose validity is not self-evident. For example, it is assumed that b and c are random variables, but the sum $b + c + d$ (i.e., the denominator of CSI) is not. Alternatively, one may assume that d is a random variable, as well. The above formula for σ_{CSI} changes in that case; however, that possibility poses an inherent ambiguity, which is undesirable. On the one hand, one may assume that d has a binomial distribution with parameters N_1 and d/N_1; on the other hand, d may be a binomial with parameters F_1 and d/F_1. Interestingly, the resulting formulas for σ_{CSI} in the two cases are the first and second terms in Eq. (A7). As such, the error bars given in Eq. (A7) are more conservative, and that is one of the reasons why they were adopted here. Another reason why the latter error bars are not employed in this paper is that they do not display the $b \leftrightarrow c$ symmetry displayed by CSI itself.

APPENDIX B

CSI and the Number of Clusters

The general behavior of CSI as the number of clusters varies can be anticipated to some extent. Consider CA at an iteration corresponding to n clusters; let the value of CSI at this iteration be CSI_n. A natural question is "How much does CSI change every time a cluster splits into two other clusters?" In other words, in going from one iteration to the next, how much does CSI change? A partial answer follows when one notes that an increase of 1 in the number of clusters will increment only one of the three components of CSI.

Specifically, in the notion of appendix A, and writing $\mathrm{CSI}_n = d/(b + c + d)$, incrementing n by 1 will result in one of the following:

$$\mathrm{CSI}_{n+1} = \frac{d}{b + c + d + 1} = \left(\frac{n}{n+1}\right)\mathrm{CSI}_n \quad \text{if } b \text{ or } c \text{ are incremented,} \tag{B1}$$

$$\mathrm{CSI}_{n+1} = \frac{d+1}{b + c + d + 1} = \left(\frac{n}{n+1}\right)\left(1 + \frac{1}{d}\right)\mathrm{CSI}_n \quad \text{if } d \text{ is incremented.} \tag{B2}$$

In the first situation, when b or c is incremented, one can solve the iterative equation for CSI_n, exactly:

$$\mathrm{CSI}_n = \frac{1}{n}\mathrm{CSI}_1.$$

In other words, in the first situation, CSI falls off as $1/n$.

In the second situation, where a split in a cluster increases the number of hits, the appearance of $1/d$ in the equation makes the solution nontrivial. However, one can still estimate CSI for limiting values of d. For large values of d, the $1/d$ term can be ignored, and the solution reduces to the one found in the first situation, that is, $\mathrm{CSI}_n \sim 1/n$. For small values of d, the equation becomes

$$\mathrm{CSI}_{n+1} = \left(\frac{n}{n+1}\right)\left(\frac{1}{d}\right)\mathrm{CSI}_n,$$

which implies that CSI_{n+1} may actually be larger than CSI_n. In other words, if the number of hits is small, then CSI may increase with n, while CSI is constrained to decrease with n monotonically if the number of hits is large. This explains the behavior of the CSI curves in Figs. 6 and 7; there are regions in n where CSI increases, while the overall behavior is a $1/n$ fall off.

REFERENCES

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

——, ——, and ——, 2002: Development of an "events-oriented" approach to forecast verification. Preprints, *19th Conf. on*

Weather Analysis and Forecasting and 15th Conf. on Numerical Weather Prediction, San Antonio, TX, Amer. Meteor. Soc., 255–258.

Brown, B. G., J. L. Mahoney, C. A. Davis, R. Bullock, and C. K. Mueller, 2002: Improved approaches for measuring the quality of convective weather forecasts. Preprints, *16th Conf. on Probability and Statistics in the Atmospheric Sciences,* Orlando, FL, Amer. Meteor. Soc., 20–25.

——, and Coauthors, 2004: New verification approaches for convective weather forecasts. Preprints, *11th Conf. on Aviation, Range, and Aerospace,* Hyannis, MA, Amer. Meteor. Soc., 9.4.

Bullock, R., B. G. Brown, C. A. Davis, K. W. Manning, and M. Chapman, 2004: An object-oriented approach to quantitative precipitation forecasts. Preprints, *17th Conf. on Probability and Statistics in the Atmospheric Sciences,* Seattle, WA, Amer. Meteor. Soc., J12.4.

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.,* **11,** 141–154.

Chapman, M., R. Bullock, B. G. Brown, C. A. Davis, K. W. Manning, R. Morss, and A. Takacs, 2004: An object oriented approach to the verification of quantitative precipitation forecasts: Part II—Examples. Preprints, *17th Conf. on Probability and Statistics in the Atmospheric Sciences,* Seattle, WA, Amer. Meteor. Soc., J12.5.

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

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

Du, J., and S. L. Mullen, 2000: Removal of distortion error from an ensemble forecast. *Mon. Wea. Rev.,* **128,** 3347–3351.

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

Everitt, B. S., 1980: *Cluster Analysis.* 2nd ed. Heinemann Educational Books, 136 pp.

Lakshmanan, V., R. Rabin, and V. DeBrunner, 2003: Multiscale storm identification and forecast. *J. Atmos. Res.,* **67–68,** 367–380.

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting,* **13,** 753–763.

——, and S. Sandgathe, 2006: Cluster analysis for verification of precipitation fields. *Wea. Forecasting,* **21,** 824–838.

Nachamkin, J. E., 2004: Mesoscale verification using meteorological composites. *Mon. Wea. Rev.,* **132,** 941–955.

Peak, J. E., and P. M. Tag, 1994: Segmentation of satellite imagery using hierarchical thresholding and neural networks. *J. Appl. Meteor.,* **33,** 605–616.

Venugopal, V., S. Basu, and E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res.,* **110,** D08111, doi:10.1029/2004JD005395.