Cluster Analysis for Verification of Precipitation Fields

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 20 July 2004, in final form 21 November 2005)

ABSTRACT

A statistical method referred to as cluster analysis is employed to identify features in forecast and observation fields. These features qualify as natural candidates for events or objects in terms of which verification can be performed. The methodology is introduced and illustrated on synthetic and real quantitative precipitation data. First, it is shown that the method correctly identifies clusters that are in agreement with what most experts might interpret as features or objects in the field. Then, it is shown that the verification of the forecasts can be performed within an event-based framework, with the events identified as the clusters. The number of clusters in a field is interpreted as a measure of scale, and the final "product" of the methodology is an "error surface" representing the error in the forecasts as a function of the number of clusters in the forecast and observation fields. This allows for the examination of forecast error as a function of scale.

1. Introduction

Given the rapid advances of numerical weather prediction (NWP) systems, it is imperative to devise a framework within which the performance of these systems can be assessed objectively and possibly without human intervention. Most NWP systems produce outputs (forecasts) that are spatial fields, with quantities of interest taking values at every grid point. These fields, in turn, have features that most experts (though not all) might identify as events or objects, within which the grid values are highly correlated. As such, it is difficult to interpret verification results based on a simple comparison of the forecasts and observations on a grid-by-grid basis without accounting for the existence of these features. Such features are not only peculiar to the forecast fields produced by NWP, but also to the observation fields. It is then natural to perform the verification of the NWP forecasts within a framework that acknowledges the existence of such features in the two fields.[1] To that end, event-based or object-oriented verification techniques have been put forth (Ebert and McBride 2000; Du and Mullen 2000) for precipitation verification, and central pressure tracking techniques are common for tropical cyclone and midlatitude cyclone prediction verification. Numerous extensions and generalizations have been developed by Baldwin et al. (2001, 2002), Brown et al. (2002), Bullock et al. (2004), and Chapman et al. (2004). Alternative approaches aimed at addressing such scale-related issues in verification have also been proposed by Nachamkin (2004) and Casati et al. (2004).

In this paper, a similar approach is developed; it is based on a statistical method generally referred to as cluster analysis. In image processing circles, cluster analysis is referred to as image segmentation, and it has been employed for storm and cloud identification (Lakshmanan et al. 2003; Peak and Tag 1994). The primary aim of cluster analysis is the identification of clusters whose members are in some sense more similar to one

Corresponding author address: Caren Marzban, Dept. of Statistics, University of Washington, Box 354322, Seattle, WA 98195. E-mail: marzban@caps.ou.edu

[1] Hereafter, any reference to "the two fields" or "both fields" refers to the observation and forecast fields.

another than to members of other clusters. If the data consist of only spatial coordinates, say x and y, then cluster analysis can produce a scatterplot of the data where dissimilar regions are identified (e.g., by different colors). These dissimilar regions can be identified as objects or events in the aforementioned sense. However, cluster analysis is not restricted to two spatial coordinates; it can be performed on any number of variables. The output of the analysis in higher dimensions is not as visually appealing as in two dimensions, but the primary aim of the analysis is still the same: the identification of clusters that have similar features, with a feature quantified by all the variables.

Given the interpretation of a cluster in cluster analysis as an event or object in a gridded field, it is natural to perform the analysis on a forecast and a verifying observation field. Having identified objects in both fields, it is further natural to compare the two fields in terms of the clusters found within them. This is the main theme of the current paper: the verification of forecast precipitation fields in terms of objects identified in forecast and observed precipitation fields, where the objects are defined objectively via cluster analysis.[2]

As in any statistical tool, whether cluster analysis produces physically meaningful results is a matter that depends on the choice of several parameters that are not universally determined. In other words, different variants of cluster analysis produce different results, some of which may be entirely meaningless. Here, it is shown that one of the common variants of cluster analysis does produce clusters that, although not perfectly sensible, do generally agree with the human visual interpretation of a meteorological event.

Having motivated for the use of cluster analysis for object identification, several additional criteria are introduced in order to allow for verification, that is, an objective comparison of the two fields. Although these criteria produce further ambiguity in the outcome of the analysis, it is shown that even simplistic criteria allow for reasonably meaningful verification results.

In most applications of cluster analysis an important quantity is the number of clusters in the field. In one type of cluster analysis that quantity is prespecified (e.g., k-means cluster analysis), while in another type it is inferred according to some criterion (e.g., Bayes's information criterion; see the next section.) Here, however, it is neither prespecified nor inferred. Instead, it is treated as a variable that allows one to explore a field on different scales. In fact, even the comparison of two fields (i.e., verification) is performed in a framework where the number of clusters in both fields is treated as variable. As such, the final outcome of the verification is not a single value for error, but an array of error values for different numbers of clusters in the two fields. It is suggested that this array of numbers be viewed as an "error surface" in a three-dimensional space whose x and y coordinates are the number of clusters in the observation and forecast field, NC_o, and NC_f, respectively. The "height" of the surface at a given point expresses the error of the forecasts at the corresponding scale.

The structure of the paper is as follows. In the next section cluster analysis is reviewed, followed by a section that discusses the proposed methodology in further detail. After discussing the data, the paper proceeds to provide illustrations of the application of cluster analysis, followed by a demonstration of the proposed verification methodology. The paper concludes with a summary of the conclusions and a discussion of further details and of future directions for research. An appendix provides further details of cluster analysis.

2. Cluster analysis: Review

Cluster analysis refers to a large class of techniques designed to classify a multivariate dataset into some number of clusters whose members are more similar to one another than to members of other clusters. These techniques are divided into many classes based on different notions of similarity, different emphasis placed on merging versus splitting clusters, and so on. Here, only one subclass of these techniques is discussed; the class is sufficiently large to allow for a demonstration of the abilities and restrictions of cluster analysis. Details can be found in Everitt (1980). The specific subclass examined here is called agglomerative hierarchical cluster analysis, hereafter referred to as CA.

The algorithm begins by assigning every data point to a cluster (of size 1). The distance between every pair of clusters is computed, and the two closest clusters are merged into a single cluster. The procedure is then repeated with the new set of clusters. The number of clusters, therefore, begins with N the sample size, and is systematically reduced to 1, that is, the entire dataset. Each step is referred to as an iteration. This iterative approach is desirable for verification of gridded data, because it explores the fields at different scales, but still within an object-oriented framework. At one extreme, it addresses individual grid points (i.e., at the first iteration), and it ends with the entire field treated as a single event.

[2] It is also suggested that performing cluster analysis on the *joint* set of the two fields can provide an objective method for estimating false alarms and misses. See section 7.

The choice of the measure of intercluster distance (a.k.a. cluster similarity) gives rise to different variations of CA. One common measure is the group average distance, which is computed as the average of the distances between every pair of data in two clusters. In the so-called SLINK (for shortest link) version of CA, the intercluster distance is taken to be the shortest distance between the elements of the clusters. Adopting the largest distance between the elements to gauge intercluster distance gives rise to the CLINK (for complete link) variant. These three distance measures yield clusters with different characteristics (Everitt 1980). For example, CLINK results in tightly packed, small clusters.

The choice of the distance measure is further multiplied by the ambiguity in how the distance between members of clusters is computed. For example, one may compute distances in a Euclidean sense (i.e., with L2 norm), or as a city-block distance (i.e., L1 norm). In short, one has at least six measures of intercluster distance:

$$\begin{aligned} \text{group average distance} &= \frac{1}{n_1 n_2} \sum_{i}^{n_1} \sum_{j}^{n_2} D_{i,j}, \\ \text{SLINK distance} &= \min(D_{ij}), \\ \text{CLINK distance} &= \max(D_{ij}), \end{aligned} \tag{1}$$

where $D_{i,j}$ stands for the distance between the ith element of the first cluster (size n_1) and the jth element of the second cluster (size n_2), and it can be computed with an L2 or L1 norm,

$$\sqrt{\sum_{i}^{D} (x_i - y_i)^2}, \sum_{i}^{D} |x_i - y_i|, \tag{2}$$

respectively, where x and y are D-dimensional vectors representing the coordinates of the two members. The characteristics of these distances are illustrated in section 5, below, and a few additional features are addressed in the appendix.[3]

The number of clusters in CA is not specified a priori. An alternative technique, not considered here, where the number of clusters is fixed, is called k-means clustering. There exist a number of methods for inferring the optimal number of clusters from the data itself. A simple, visual method relies on dendrograms (Everitt 1980). These diagrams often reveal a natural number of clusters underlying the data. There exist more quantitative criteria for inferring the optimal number of clusters, based on Bayes's or Akaike's information criteria (BIC and AIC, respectively). Here, however, the ultimate task is to perform verification at different scales, and so the "optimal" number of clusters in the two fields is not of concern.

Another reason for treating the number of clusters in the two fields as an issue that should be addressed jointly (involving both fields) is that observations occasionally suffer from errors, as well. In other words, an observation field is occasionally no more accurate than a forecast field. In this sense, it is natural to treat them on the same footing.

3. Verification method

The verification procedure can be performed either in the two-dimensional space spanned by the spatial coordinates only, or in the three-dimensional space that also includes the precipitation amount. In what follows, these two cases will be referred to as x–y, and x–y–p, respectively. (See section 7 on the possibility of adding other variables to the analysis.) Performing the verification in x–y space assesses the agreement between the two fields in terms of the size, shape, and displacement errors, while the distances in x–y–p space assess the sum of all four errors: size, shape, displacement, and precipitation amount.

To assure that the variables (x, y, and/or p) are treated on the same footing, they are normalized to vary over the same range. This is done by converting each variable to a z score by subtracting from each variable its mean and dividing by its standard deviation. This gives all variables a mean of 0 and a standard deviation of 1. The normalization is performed over the joint set of observation and forecasts. Moreover, because of the highly skewed distribution of precipitation, the natural log of precipitation is employed instead. At the end of the procedure, all variables are transformed back to their original units. There are situations where one does not desire for the various variables to be treated on the same footing; for example, when precipitation amount is more important than the placement of some cluster. That issue is addressed in section 7.

At every iteration of CA performed on both fields, a criterion is required for matching together the NC_o observed clusters and the NC_f forecast clusters. A simple approach is to compute the intercluster distance between all $\mathrm{NC}_o \times \mathrm{NC}_f$ pair of clusters, and to identify the smallest element in that set; the corresponding clusters are defined as "matched" and then excluded from the set of $\mathrm{NC}_o \times \mathrm{NC}_f$ elements. This is a reasonable approach, except that it leads to all NC_o observed clusters

[3] The distance between the centroids of the clusters would be an alternative measure. However, it is not employed in the current work because it is not well suited to elongated or irregularly shaped objects.

being matched with all NC_f forecast clusters; Any false alarms or misses will be a direct consequence of $NC_0 \neq NC_f$. In other words, the equality of NC_o and NC_f precludes false alarms or misses, even if the clusters are not "adequately" matched. Although, the difference between NC_o and NC_f can be thought of as a performance measure in itself, it is unreasonable to assume that a forecasting model that produces the same number of forecast clusters as that of the observation field will also adequately place them.

This raises the question of what is adequate placement? A related question is the following: When an expert views a single pair of forecast and observation fields, what feature of the two fields is most important in aiding identification of anomalously placed clusters? The term "anomalous" suggests the answer: outliers. In other words, the expert's mind analyzes the two fields and creates an estimate of the typical, or average, displacement between forecast and observed clusters. It then identifies any cluster in one field whose distance to other clusters in the second field is uncharacteristically large, and labels it as a false alarm or a miss.

To simulate this behavior, the above cluster-matching criterion is revised. At every iteration of CA performed on both fields, $NC_o \times NC_f$ distances are computed, and the minimum distance is identified. A list of the minimum distances is recorded for all iterations, and a histogram is computed.[4] It is found (not shown) that the histograms are generally bell shaped, though skewed to the left. As such, one can argue that any distance that resides in the upper tail of the histogram can be considered an outlier. Here, matched clusters whose distances fall outside of the (median + $z\sigma$) of the histogram are defined as "unmatched."[5] An unmatched cluster in the forecast field is then counted as a false alarm, and an unmatched observed cluster is counted as a miss. These two scalar quantities are two of the verification measures reported here. An alternative, more objective, technique for identifying false alarms and misses is put forth in section 7.

Another measure of the agreement between the two fields is naturally provided by the average of the distances between matched clusters. Of course, the measure of these distances is subject to ambiguity, for it can be computed in any of the six ways defined in Eqs. (1) and (2). Note that, although the distance measures in Eqs. (1) and (2) arise for measuring the distance between clusters *within* a given field, they can also be utilized for measuring the distance between clusters in *different* fields. Here, no attempt is made to promote one measure over another, because ultimately the choice of the measure is problem dependent. As shown below, for the problem at hand, the group average distance with an L2 norm appears to be adequate for both the CA and the verification tasks.

Therefore, at every iteration of CA (i.e., at every scale) several measures can be computed to assess the quality of the agreement between the observed and forecast fields: 1) $|NC_o - NC_f|$, 2) number of false alarms, 3) number of misses, and 4) the average distance between the matched clusters. Although, all four measures are reported here, the last measure is treated more heavily as an overall measure of forecast error. Of course, one may combine the four measures into a single measure; however, that requires an assessment of the relative importance of the four measures—a task that is again problem dependent.

One may be concerned that ignoring some of these facets of performance can lead to an overproduction of false alarms or misses, but as is shown below, this is not a serious concern. Specifically, if the intercluster distance measure is the group average distance [Eq. (1)], with the point-to-point distances computed as a Euclidean (L2) distance [Eq. (2)], then that distance measure has the unusual property that it increases with decreasing cluster size. As such, false alarms or misses are discouraged. The appendix presents further details. Nevertheless a consequence of neglecting the contribution of false alarms and misses to the total error is that the skill of the forecasts as assessed in this approach should be considered a measure of potential skill.[6] The issue of assigning error to false alarms and misses has also been addressed by Marshall et al. (2004), although in a somewhat different context.

Another concern may arise in addressing the matching of multiple clusters in one field to a single cluster in the other field. Here, it is unnecessary to consider an explicit match of multiple forecast clusters to a single observed cluster, or vice versa, because it is implicitly incorporated into the iterative nature of CA. In other words, if it turns out that multiple forecast clusters should be matched with a single observed cluster (in the sense of producing better agreement between the two

[4] This is not a histogram of all $NC_o \times NC_f$ distances, but a histogram only of the minimum distances. As such, the procedure is somewhat slow, because it requires computing the distances twice, but it is necessary because it is the distribution of the minimum distances that sets the scale for what can be considered anomalous placement.

[5] The median is utilized, because it is closer to the mode of the distribution. The value of z is set to 1; see section 7 for the effect of varying z.

[6] The term "potential skill" is employed in a sense similar to that of Murphy (1995), where a skill score is said to measure potential skill if it ignores some facet of performance such as bias.

fields), then the distance/error between the fields will be lower at some iteration where the multiple forecast clusters are considered a single cluster.

4. Data

Three datasets are utilized. The first consists of 3-h precipitation forecasts from the University of Washington Mesoscale Ensemble with the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5) on a 12-km grid (Grimit and Mass 2002).[7] Precipitation forecasts for 9 September 2003 over the Pacific Northwest (centered on Seattle, Washington) are employed to illustrate CA and the effect of distance measures on the clusters. No verification is performed on this dataset.

The second dataset is synthetic in both observations and forecasts. It is employed for illustrating the clustering and verification procedure. It consists of five unambiguously distinct events in the observation field (Fig. 3a), and five events in the forecast field (Fig. 3b); both fields are on a 100 × 100 grid. Apart from an unmatched forecast event (false alarm) and an unmatched observed event (miss), the remaining events differ from the observed ones in size, the amount of spatial displacement, and precipitation intensity. Their shape is the same, namely a circular disc.

The forecasts for the third and final dataset are collected from The Naval Research Laboratory Coupled Ocean–Atmospheric Mesoscale Prediction System (COAMPS; Hodur 1997), and the verifying observations are derived from River Forecast Center (RFC) 24-h precipitation reports, interpolated from their original 4-km grid to the COAMPS 27-km grid. These data are employed to illustrate CA and the verification procedure on realistic data.

The COAMPS forecast fields are given in terms of three types of precipitation: 1) total accumulated resolved liquid, 2) total accumulated snow, and 3) total accumulated convective precipitation. In this study, only the latter is compared with the observations. The specific forecasts are for 25 July 2003, over the continental United States. The comparison is made between observations at 1200 UTC of the precipitation accumulated over the previous 24 h, and the forecasts initialized at 1200 UTC on the previous day.

5. Cluster analysis illustration

In this section CA is performed on MM5 data. The main purpose of this exercise is to demonstrate the effect of different distance measures on the type of clusters that emerge from CA. Figure 1 displays a contour plot of the MM5 data.

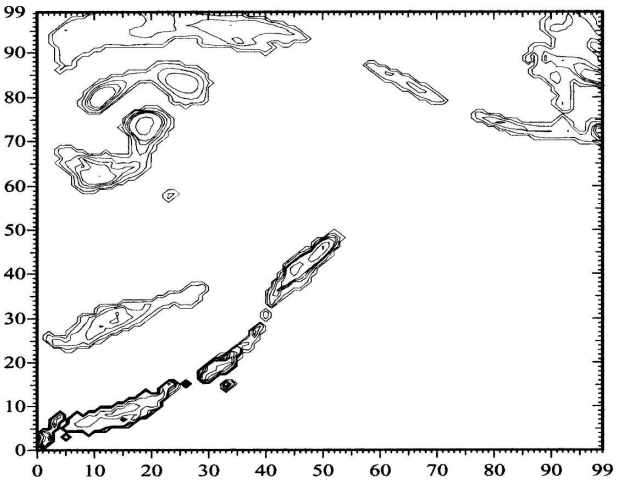

FIG. 1. Contour plot and *x*–*y* coordinates of precipitation forecasts from MM5.

Figure 2 displays the results of clustering performed in *x*–*y*–*p* space, at iterations corresponding to 10 (top) and 5 (bottom) clusters. Quite generally, it can be shown that for the specific type of cluster analysis adopted here embedded clusters are not possible in *x*–*y* space. But, as can be seen in these figures, such clusters do emerge in *x*–*y*–*p* space, and they generally represent regions of heavy precipitation in Fig. 1. The rows in Fig. 2 correspond to the three distance measures in Eq. (1), and the columns refer to the two norms given in Eq. (2). For this particular dataset, it is evident that the L2 norm generally produces smaller clusters than the L1 norm. Similarly, the group average distance and CLINK generate smaller and tighter clusters than SLINK.

The MM5 precipitation data are centered on the NE Pacific and indicate a major low pressure system in the NW corner, a pair of rainbands associated with a frontal system in the SW quadrant, and another low system to the NE. A key aspect of these synoptic rain patterns is the very intense embedded precipitation regions. It is a difficult verification issue to decide whether these areas of intense precipitation should be treated separately from the surrounding lighter-precipitation regions. From many perspectives, these intense rainfall areas are significantly more critical than the overall region of rainfall. The method of CA chosen—group average, SLINK, CLINK, and so on—allows these regions to be highlighted (group average, CLINK) or suppressed (SLINK). SLINK and the lower number of clusters (five) clearly reproduce the synoptic aspects of the precipitation without unduly singling out intense precipi-

[7] The operational initialization employed here is based on the Aviation Model.

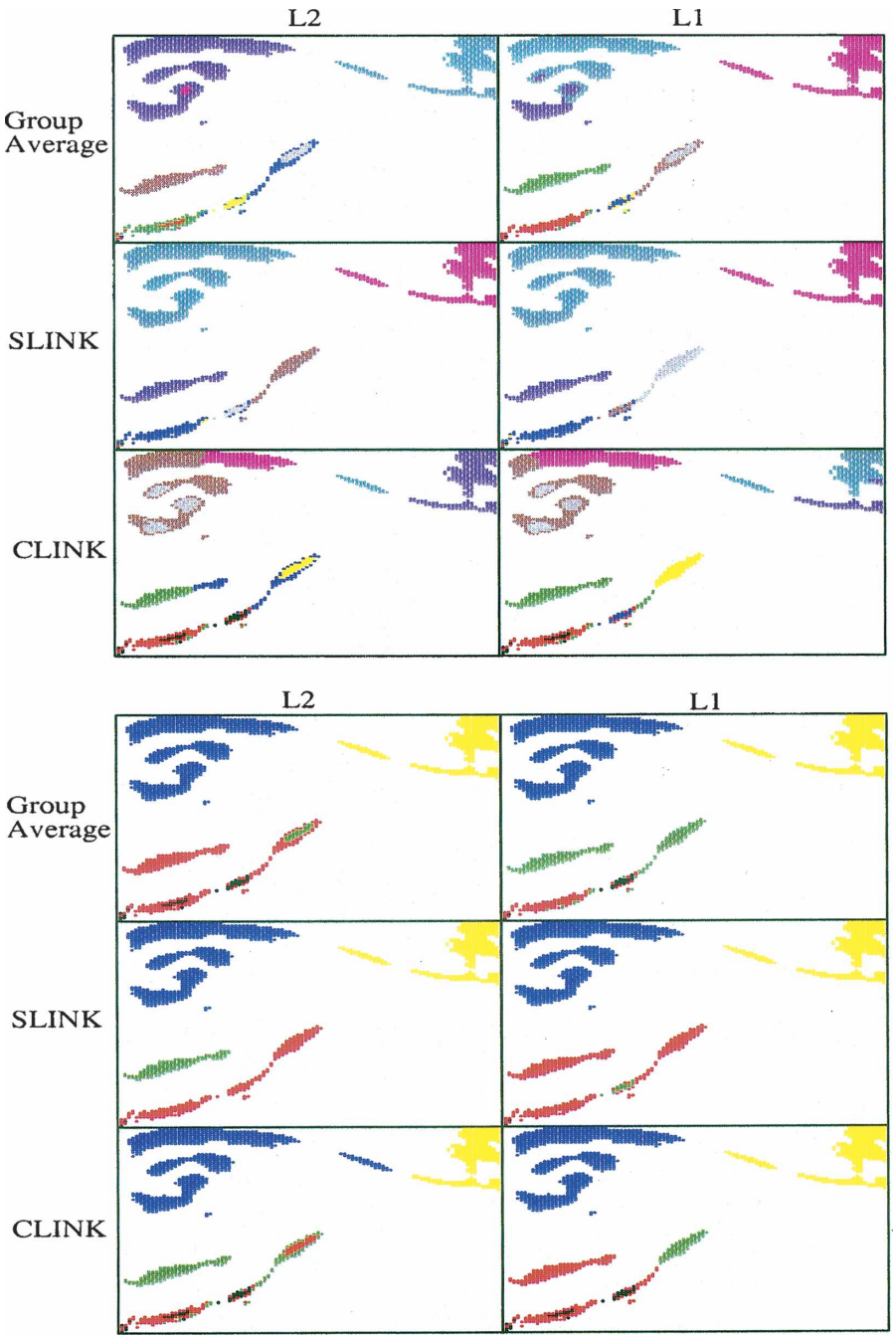

FIG. 2. The CA results for the six different distance measures [Eqs. (1) and (2)] on MM5 data.

tation. Verification of the intense precipitation regions could also be accomplished by successive thresholding; however, CA in x–y–p space may prove more useful.

An important point is the appearance of apparently nonsensible clusters when the analysis is performed in x–y–p space. For example, one notes spatially distant entities appearing as a single cluster (i.e., with the same color). This simply implies that the agreement between the amount of precipitation in the two clusters is to such a high degree so as to justify the merging of the two into a single cluster in spite of the spatial distance between them. Conversely, spatially close clusters may emerge as being distinct clusters. The implication of this occurrence would be that the clusters are dissimilar in terms

of their precipitation amount. In short, apparently nonsensible clusters may in fact be perfectly sensible when viewed in the larger *x–y–p* space. As such, *x–y–p* clustering results may appear as misleading. The virtue of CA in *x–y–p* space is, therefore, not in producing visually appealing clusters, but rather in providing a more realistic and more accurate assessment of the quality of the forecasts.

The occurrence of small dots (i.e., a one-member clusters) scattered throughout some of the graphs is disconcerting. It is clear, however, that the extent of this problem depends on the choice of the distance and norm. It is likely that the inclusion of other variables in CA will further alleviate this problem.

In summary, it would appear that CA does a reasonable job of identifying physically meaningful clusters if the distance measure is suitably selected. It would appear that the group average distance and the CLINK distance (i.e., the longest distance) with either an L2 or L1 norm provide comparable clusters. Although group average distance with an L2 norm is adopted for presentation in the remainder of the analysis, the other measures have been examined, and confirmed to produce similar results.

6. Verification

The methodology will be tested on a synthetic dataset with known characteristics (number of clusters, size, location, and amount of precipitation). After the procedure has been tested (next section), it will be applied to realistic data.

a. Verification of synthetic data

This dataset is described in section 4 and is shown in Figs. 3a and 3b. The results of CA at each iteration are not shown. Instead, Table 1 shows the distances between the observed and forecast fields at iterations three through seven, that is, with three through seven clusters in each field. These distances are measured as a group average distance with an L2 norm in *x–y–p* space.

This table is a representation of the error surface mentioned in the introduction. The *x* and *y* coordinates of the three-dimensional space (in which the surface is embedded) are the rows and columns of the table, and the height of the surface (i.e., error at that scale) corresponds to the elements of the table.

Evidently, the error surface has a parabolic shape, in that there is a single absolute minimum at (5, 5). It is worth emphasizing that although the error between the fields is minimum when there are five clusters in the observed field and five in the forecast field—precisely the true numbers—it is the table *as a whole* that represents the error of the forecasts. The emergence of the point (5, 5) as a clear minimum is a consequence of the explicit scale injected into the structure of the two fields (Fig. 3). At this scale, the cluster-matching procedure identifies cluster A in the forecast field with cluster A in the observed field and, similarly, with clusters B, C, and D. For the E clusters, see below.

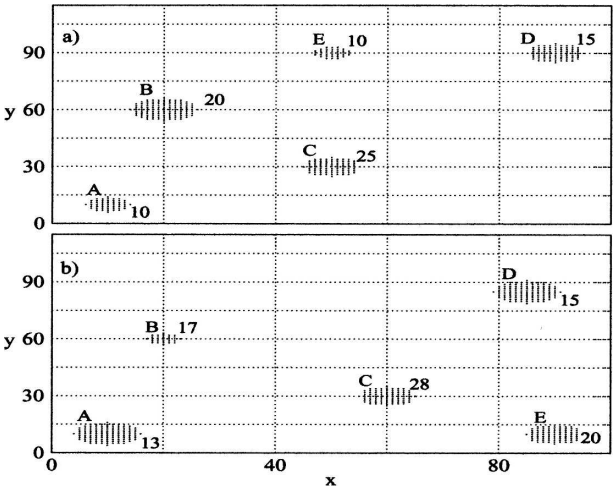

FIG. 3. Synthetic (a) observed and (b) forecast fields. The numbers indicate the amount of precipitation at each grid point in each cluster. The letters label the clusters.

The existence of a natural scale (i.e., five clusters in each field) allows one to examine the individual errors between the clusters at that scale. These are shown in Table 2. When the clustering is done in *x–y* space, the results indicate that the A and B clusters are in fact the best forecast clusters, while the C cluster represents the worse forecast.

By contrast, in the *x–y–p* case the best forecast is the C cluster, and the worst is the A cluster. The reversal of the A cluster from the best forecast in the *x–y* analysis

TABLE 1. The distance between observation and forecast fields for different NC in each. The boldface entry is the smallest distance (error), corresponding to five clusters in the observed field and five clusters in the forecast field—precisely the true values. The analysis is done in *x–y–p* space.

	NC_f				
NC_o	3	4	5	6	7
3	1.23	0.89	0.89	0.89	0.89
4	1.07	0.74	0.74	0.98	0.98
5	0.83	0.68	**0.55**	0.78	0.78
6	0.83	0.68	0.99	1.11	1.02
7	0.83	0.68	0.98	1.11	1.25

TABLE 2. The matched observed and forecast clusters in Fig. 3, and the distance/error between them, assuming five clusters in each of the observed and forecast fields, respectively.

Cluster	*x–y*	*x–y–p*
A	0.15	0.87
B	0.14	0.55
C	0.33	0.51
D	0.26	0.26
Avg	0.22	0.55

to the worst forecast in the *x–y–p* analysis suggests that the forecasting model has misestimated the size and/or amount of precipitation in the A cluster, even though the placement is correct.

As mentioned in section 3 the error values in Table 1 are computed only from matched clusters, and the unmatched ones (false alarms or misses) are identified as outliers in the distribution of all distances in the field. This criterion correctly identifies the E clusters as a false alarm and a miss, respectively.

b. Verification of real data

In this section CA-based verification is applied to COAMPS and RFC data (see section 4). Figure 4 displays the observation field (top) and the forecast field (bottom), for precipitation $\geq$5 mm.[8]

The *x–y* analysis is presented first. The complexity of the two fields precludes a representation of the error surface as a table (as in Table 1). For presentation purposes, the surface is plotted as a contour plot in Fig. 5a. Several dominant features appear, some of which are easily explained; other features are more enigmatic, calling for more research. For example, it can be seen that the forecast errors are generally high (red through yellow) when the number of clusters in either field is small. This feature is likely to be a universal feature of the error surface, because it simply reflects the fact that the distance/error between two fields is relatively large when there are a few large clusters in one field and a large number of smaller clusters in the other field.

The errors are lowest when the number of clusters in either field (but not both) is large. This, too, is a consequence of the methodology; specifically, because false alarms and misses are not allowed to contribute to the overall error. As a result, the procedure rewards the production of large numbers of false alarms and misses. As mentioned previously, it may seem inappropriate to not penalize the procedure for producing false alarms and misses, but this is a necessary simplification at the current developmental stage of the methodology. It does not, however, preclude utilizing the error surface for assessing forecast errors; for instance, in comparing the forecasts from two different forecast systems, it is the relative heights of the two error surfaces that contain the relevant information.

Another feature of this error surface is the ridge along the diagonal. This occurs because "forcing" an equal number of clusters in the two fields leads to large errors, especially when the "true" numbers are unequal. As such, the existence of the ridge is a consequence of different cluster numbers in the two fields. Better forecasts imply a smaller ridge. Note that the ridge does not extend all across the diagonal; its strength decreases for larger cluster numbers. The extent of its reach is further addressed in section 7.

Note the mostly symmetric structure about the diagonal. This symmetry implies that the forecasts and observations are more agreeable when they are verified on the same scale. This in itself can be viewed as a measure of (good) performance, for if the symmetry were absent, then one could conclude that the forecasts generally have the wrong scale. Indeed, the error surface in Fig. 5a is not entirely symmetric along the diagonal. The error surface falls to lower values (dark blue) when the number of clusters in the forecast field is generally larger than that in the observed field. This finding is consistent with the fields shown in Fig. 4; there appears to be more clusters in the forecast field than in the observed field.

In the *x–y–p* case, the contour plot of the error surface is shown in Fig. 5b. Although many of the features are similar to that of the *x–y* analysis in Fig. 5a, there is an additional asymmetry of the ridge (green), which suggests that when precipitation amount is included in the analysis, then the forecasts tend to agree with the observations even when the number of clusters in the forecasts is larger than that in the observations. Said differently, Fig. 5b shows that in *x–y–p*, the forecast model does not faithfully reproduce the more intense precipitation over Florida and in the Northeast (Fig. 4); hence, the overall "score" is worse, and there is a tendency for more false alarms and misses based on precipitation amount. However, in the West (Fig. 4), there are many more forecast clusters than observed clusters with reasonably similar precipitation amounts. The CA rightly associates these forecast and observed clusters, matching multiple forecast clusters with fewer observed clusters.

One perplexing feature of the *x–y–p* error surface is in its lowest value; it occurs when the number of clusters in the observation field is larger than that of the

[8] This threshold is introduced only to reduce the size of the data and to expedite CA.

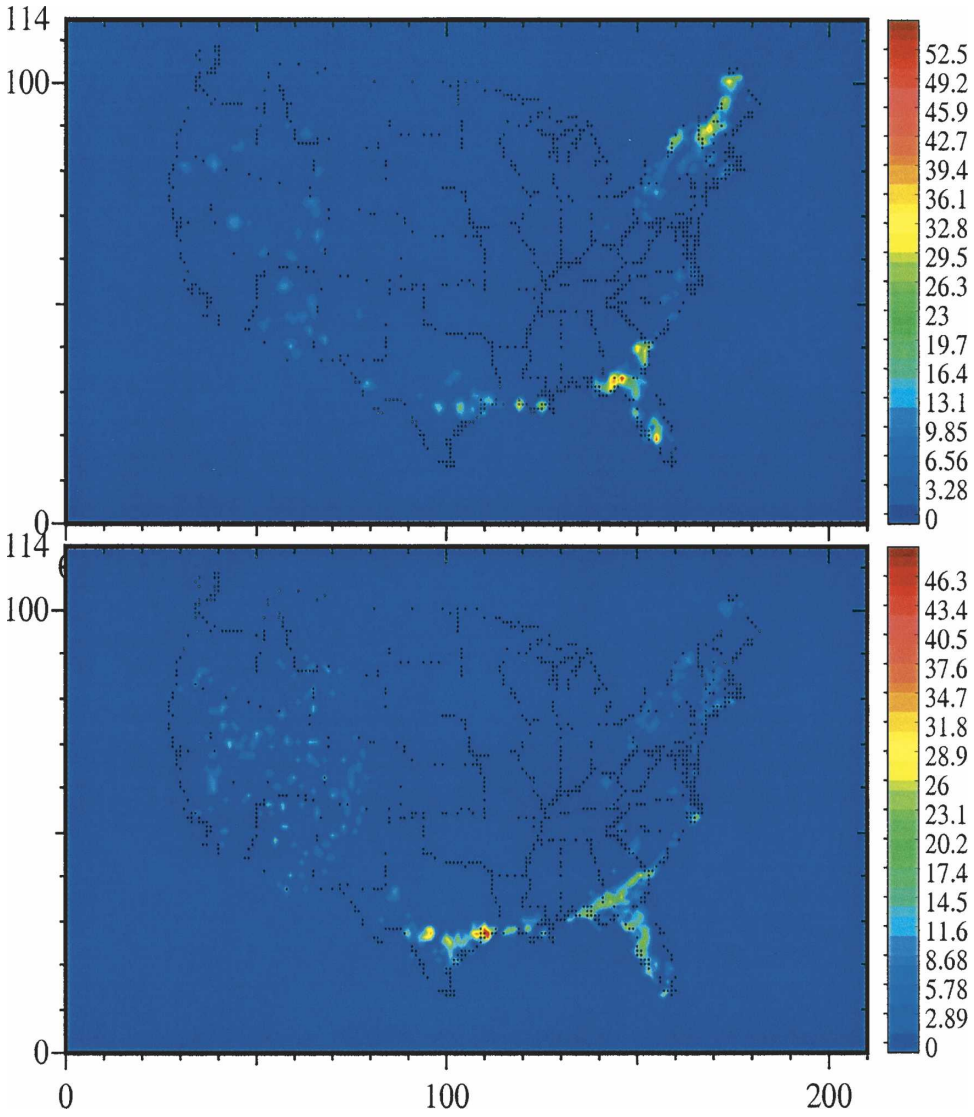

FIG. 4. (top) Observation and (bottom) forecast fields in COAMPS data.

forecast field. This is the opposite of the situation in the *x–y* analysis. This is likely due to the greater variability of precipitation amount in the observation field. While visually there are more clusters in the forecast field than the observation field, especially in the West, when precipitation amount is taken into account, the more variable observation field naturally decomposes into more clusters than the smoother forecast field.

Moving away from the analysis of the error surface, one may compare the forecast field with the observation field at a specific scale. For illustration purposes (and economy of color), the point ($NC_o = 5$, $NC_f = 5$) is selected. Figures 6 and 7 show the clusters, and the distances are shown in Table 3. The matching of the clusters is visually consistent. For the *x–y* analysis (Fig. 6) according to Table 3, the black, red, and blue clusters have the lowest errors. The scattered clusters in the western region have the highest error. The orange-colored cross-shaped clusters are the false alarms and misses.

To assess the effect of precipitation amount, the *x–y–p* results are shown at the same scale (i.e., $NC_o = 5$, $NC_f = 5$). Figure 7 shows the structure of the clusters in the two fields, and Table 3 shows the distances between the various matched clusters. The pattern is similar to the *x–y* pattern, with one exception: the black and red clusters now have some of the highest errors. The reason these clusters emerge as poor forecasts in the *x–y–p* analysis is, therefore, the wrong amount of precipitation forecast in these clusters, a fact that is clear in

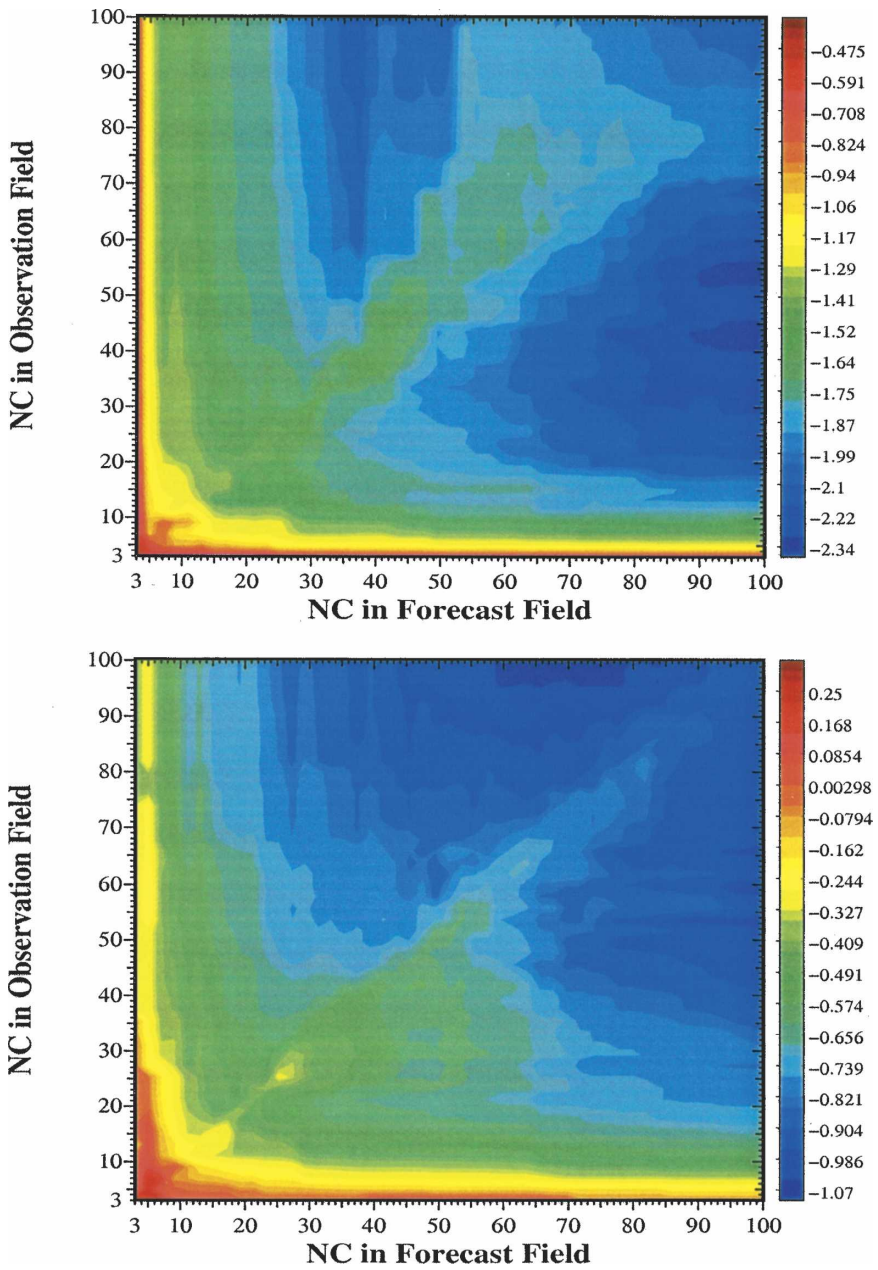

FIG. 5. A contour plot of the error surface based on (top) *x–y* and (bottom) *x–y–p* analysis. The sidebar displays the natural log of the errors.

Fig. 4. The scattered clusters in the western region are now matched, with the corresponding forecast clusters without any misses. The false alarms and misses are now in the eastern regions and are again labeled with orange-colored cross-shaped symbols. Recall that these conclusions are based on *x–y–p* results, and so "good" and "bad" forecasts refer to the placement of a cluster as well as the amount of precipitation forecast within it.

7. Conclusions and discussion

It is shown that agglomerative hierarchical CA can identify sensible clusters in an observation and a forecast field. The introduction of a few other concepts (e.g., the distance between a cluster in an observation field and one in the forecast field) sets the stage for a methodology whereby verification can be performed objectively, and in an object-oriented fashion. The final

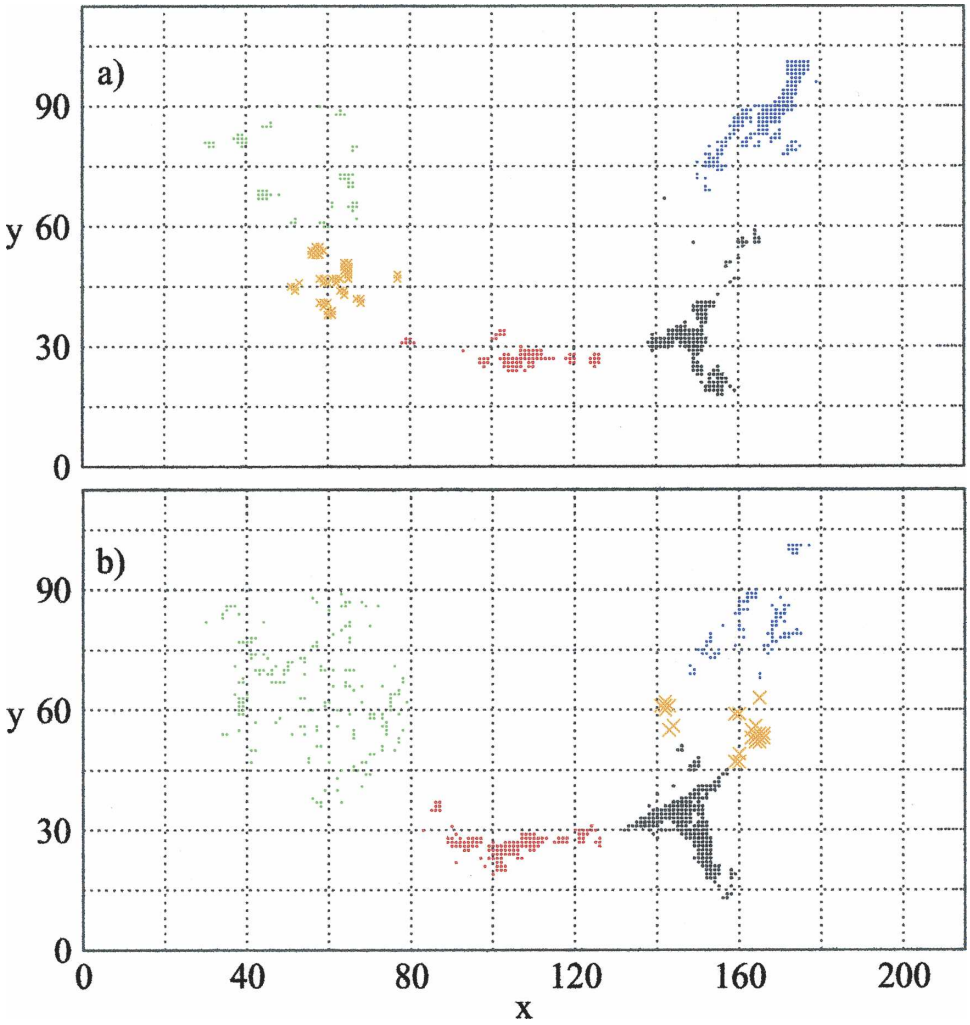

FIG. 6. Clusters identified in the (a) observation and (b) forecast fields, with the colors indicating the match. The analysis is done in *x*–*y* space. The orange crosses mark false alarms and misses.

"product" of the methodology is an "error surface" that represents forecast error values as a function of two quantities: the number of clusters in the observation field and that in the forecast field. It is argued that these two quantities span different scales, and so the error surface assesses the error of the forecasts at different scales.

The practical utility of the error surface is in its ability to represent forecast errors over all scales in an event-based or object-oriented sense. Ideally, of two forecast systems, the one that produces a lower error surface would be considered the better one. Of course, in practice, the error surfaces of the two forecast systems may cross one other, in which case one can then state the scales over which one system outperforms the other.

In its current form, the proposed methodology cannot be fully automated without some loss of information. Some amount of automation is possible, as in the CA portion of the analysis and the computation of the distances between the clusters. However, the outcome of the procedure is a multidimensional entity (e.g., the contour plot assessing the error of the forecasts at different scales) that calls for some interpretation. It is, of course, possible to specify the scale of interest a priori, thereby reducing the dimensionality of the problem. One may even distill the information contained in the contour plot to a scalar quantity, such as an average over all scales. However, it remains true that the contour plot representing the error surface carries more information and is more useful in identifying the scales over which the forecasts are poor.

It would be desirable to include in the procedure variables that allow the separation of strong convective

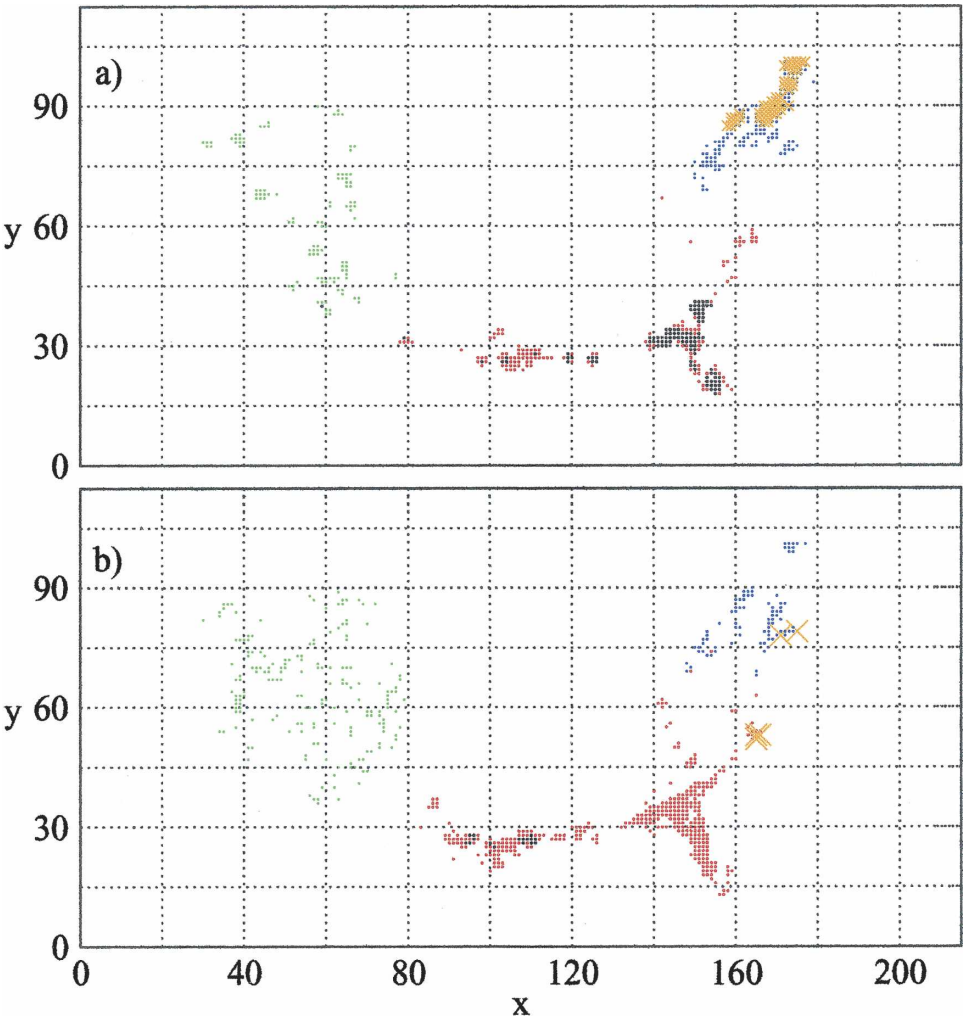

FIG. 7. Same as in Fig. 6 but in x–y–p space.

events from stratiform precipitation, such as cloud base and height, precipitation rate, horizontal extent of cloud, and so on. This idea will be further developed in the future.

Another desirable revision to the methodology would replace the proposed cluster-matching criterion with a more "fuzzy" criterion that would allow clusters to be matched even if their distance does not strictly meet some criterion. Such an approach would not only be more consistent with the statistical nature of the methodology, but it could potentially reduce the excessive appearance of single-member clusters in a field. The works of Brown et al. (2004), Bullock et al. (2004), and Chapman et al. (2004) are important in this connection.

As previously mentioned, although CA is incapable of identifying embedded clusters if the clusters are defined in terms of only spatial coordinates, it does allow for embedded clusters if the clusters are defined in terms of spatial coordinates and the amount of precipitation. There exist other types of cluster analysis (model based) techniques that do allow for a larger variety of clusters (even in x–y space). It will be worthwhile to explore such methods.

Given that the methodology matches an observed

TABLE 3. The closest observed and forecast clusters, and the distance between them for $NC_o = 10$ and $NC_f = 10$. The "X" here (orange crosses in Fig. 6) corresponds to a false alarm cluster.

Observation	x–y	x–y–p
Black	0.45	1.56
Red	0.34	1.37
Green	0.74	1.11
Blue	0.47	0.95
Orange	X	X
Avg	0.50	1.25

cluster with a forecast cluster, it is possible to assess the quality of the match in terms of the amount of precipitation in the two. Specifically, it is natural to compare the *distribution* of precipitation in the two clusters. Preliminary tests, based on a Student's t test, have been performed. On the particular datasets examined here, the tests suggest that the difference between the means of the distributions is not statistically significant. In the future, more robust tests (e.g., Kolmogorov–Smirnov) will be explored.

The criterion for identifying false alarm and missed clusters developed here relies on the idea that the notion of an anomalous cluster is one that is perceived from the placement of all other clusters in a field. After all, if a forecasting model places all clusters precisely in observed locations (and with the right amount of precipitation), then even the smallest deviation will be considered an anomalous cluster. As such, the criterion is adaptive in the sense that it varies from forecast to forecast. And yet, it does require specifying a distance threshold above which clusters are labeled as false alarm or missed; the parameter that controls that distance in this analysis is the multiplier z in (median + $z\sigma$), here set to 1. An alternative, more objective, approach that is being tested is to perform CA on the *combined set* of forecasts and observations. In principal, CA should cluster observed and forecast clusters that are relatively close to one another. Clusters will then be composed of both observed and forecast clusters. As such, cluster matching can be done by CA itself. Thus, any cluster that does not contain a comparable amount of observed and forecast clusters can then be called a false alarm or a miss. Preliminary work suggests that this is a promising method in terms of rendering the cluster-matching step here more objective.

As mentioned previously, experimentation shows that z also controls the extent of the reach (toward the upper-right corner in Fig. 5) of the ridge appearing in the error surface (e.g., Fig. 5a). Recall that the initial motivation for introducing this multiplier was to allow for an objective identification of outliers, that is, false alarms and misses. Experiments suggest that the ridge strengthens with smaller values of z, which in turn yields more false alarms and misses. As such, any error associated with the false alarms and misses is apt to affect the ridge. This issue will be addressed in the future when an objective method for assigning these errors is developed.

As also mentioned previously, one virtue of cluster analysis for performing verification is that the clustering may be performed on multiple variables. Here, the addition of precipitation to the set of spatial coordinates is shown to allow for more complex clusters. It is likely that the inclusion of variables that represent different physical processes, such as cloud base and height (to distinguish cloud type) or even surface temperature (to distinguish air mass), will lead to more physically meaningful clusters.

In a related issue, the normalization of the variables assures that the spatial variables' contribution to distance measures is comparable to that of the precipitation variable. However, it is conceivable that certain applications would require more weight being placed on the spatial variables, or on the precipitation variable. It is, therefore, natural to introduce a metric on the space of variables that defines the contribution each variable is expected to make to the distances. Such a metric can be useful even in the current analysis. For example, the rather distant association of the black clusters in Fig. 7 is clearly and undesirably dominated by the amount of precipitation in those clusters. A metric that downweighs the contribution of precipitation will alleviate such problems.

The proposed methodology is demonstrated on precipitation fields. Other fields that are more continuous may not benefit from this approach, for physically sensible clusters may not naturally emerge in the data. It will be interesting to test the idea of cluster-based verification with sea level pressure data, such as in tracking low pressure systems and identifying missed systems for both NWP and climate prediction verification.

It must be acknowledged that issues related to scale are treated somewhat cavalierly in this paper, for the primary aim of the study is the introduction and illustration of an object-oriented verification method. It is easy to show (although not done here) that the variance of the data is highly dependent on the scale of the problem.[9] As such, the points raised by Tustison et al. (2001, 2003) and Harris et al. (2001) are highly relevant and must be taken into account for more practical (less pedagogical) applications.

Acknowledgments. The authors are grateful to Mike Baldwin, Valliappa Lakshmanan, and Jason Nachamkin for invaluable contributions. Leah Heiss and Timothy R. Whitcomb of the Applied Physics Laboratory are acknowledged for providing some of the datasets employed here. This project was funded by the ONR Marine Meteorology Program under ONR Grant N00014-01-G-0460 for mesoscale verification. Addi-

[9] In other words, the variance of the data depends on whether it is computed for the "raw" data on a grid of size 1, or box-averaged data corresponding to boxes of size 2, 3, and so on. In fact, with precipitation data, it is common to find the variance monotonically decreasing with the size of the box.

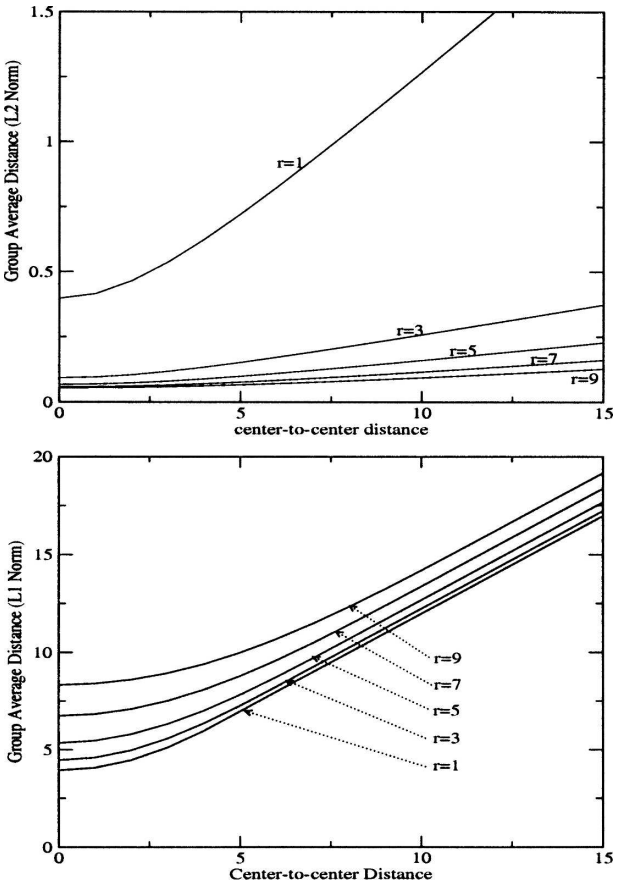

FIG. 8. Group average distance as a function of the distance between cluster centers, for a range of cluster sizes (radius = 1, 3, 5, 7, and 9) of one cluster; the other cluster has radius = 5. The distance between members of clusters is computed with an (top) L2 and (bottom) L1 norm.

tional funding was provided under a cooperative agreement between the National Oceanic and Atmospheric Administration (NOAA) and the University Corporation for Atmospheric Research (UCAR). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA, its subagencies, or UCAR.

APPENDIX

Intercluster Distance as a Function of Cluster Size

As mentioned above, the group average distance (and CLINK) with an L2 norm has the unusual property that it increases with decreasing cluster size. In this appendix, all six distance measures are examined. Some lend themselves to exact/analytic results, while others are examined numerically. Specifically, consider two circular clusters of radius r and R, with their centers separated by a distance d, residing on a grid of cell size 1. We ask: What is the intercluster distance as a function of the size of the clusters and the distance between their centers?

From the definitions in Eq. (1), the answer is evident for the SLINK and CLINK methods: $(d - r - R)$ and $(d + r + R)$, respectively. Note that while they both increase with d, the former decreases with increasing cluster size. For group average distance, however, the sums are difficult to perform analytically.[A1] The top panel in Fig. 8 shows group average distance with the L2 norm as a function of d, for five different values of $r = 1$, 3, 5, 7, and 9, while R has been fixed at 5. The bottom panel in Fig. 8 shows the same quantity but with an L1 norm. First, note that both increase with center-to-center distance d. However, the former decreases as a function of cluster size r, while the latter increases with r.

In short, the group average distance and CLINK have the property that they increase with smaller cluster size. This property helps in preventing a proliferation of false alarms and misses in the analysis.

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 J. S. Kain, 2002: Development of an "events-oriented" approach to forecast verification. Preprints, *15th Conf. on Numerical Weather Prediction,* San Antonio, TX, Amer. Meteor. Soc., 210–213.

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., CD-ROM, 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., CD-ROM, 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

[A1] Even in the continuous limit, where each of the sums is replaced by two integrals, the resulting quadruple integral does not yield a closed expression.

forecasts: Part II—Examples. Preprints, *17th Conf. on Probability and Statistics in the Atmospheric Sciences,* Amer. Meteor. Soc., CD-ROM, J12.5.

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.* 2d ed. Heinemann, 252 pp.

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting,* **17,** 192–205.

Harris, D., E. Foufoula-Georgiou, K. K. Droegemeier, and J. J. Levit, 2001: Multiscale statistical properties of a high-resolution precipitation forecast. *J. Hydrometeor.,* **2,** 406–418.

Hodur, R. M., 1997: The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Mon. Wea. Rev.,* **125,** 1414–1430.

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

Marshall, S. F., P. J. Sousounis, and T. A. Hutchinson, 2004: Verifying mesoscale model precipitation forecasts using an acuity-fidelity approach. Preprints, *20th Conf. on Weather Analysis and Forecasting,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, J13.3.

Murphy, A. H., 1995: The coefficients of correlation and determination as measures of performance in forecast verification. *Wea. Forecasting,* **10,** 681–688.

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

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

Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.,* **106,** 11 775–11 784.

——, E. Foufoula-Georgiou, and D. Harris, 2003: Scale-recursive estimation for multisensor quantitative precipitation forecast verification: A preliminary assessment. *J. Geophys. Res.,* **108,** 8377, doi:10.1029/2001JD001073.