Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of object-oriented cluster analysis method for precipitation fields

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

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

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

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

Submitted to *Monthly Weather Review*
Jan. 14, 2011
Revised May 7, 2011

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

Abstract

Convection-allowing ensemble forecasts with perturbations to model physics, dynamics, and initial (IC) and lateral boundary conditions (LBC) generated by the Center for the Analysis and Prediction of Storms for the NOAA Hazardous Weather Testbed (HWT) Spring Experiments provide a unique opportunity to understand the relative impact of different sources of perturbation on convection-allowing ensemble diversity. Such impacts are explored in this two-part study through an object-oriented Hierarchical Cluster Analysis (HCA) technique.

In part I, an object-oriented HCA algorithm, where the dissimilarity of precipitation forecasts is quantified with a non-traditional Object-based Threat Score (OTS), is developed. The advantages of OTS-based HCA relative to HCA using traditional Euclidean distance and Neighborhood probability-based Euclidean Distance (NED) as dissimilarity measures are illustrated by hourly accumulated precipitation ensemble forecasts during a representative severe weather event.

Clusters based on OTS and NED are more consistent with subjective evaluation than clusters based on traditional Euclidean distance because of the sensitivity of Euclidean distance to small spatial displacements. OTS improves the clustering further compared to NED. Only OTS accounts for important features of precipitation areas, such as shape, size and orientation, and OTS is less sensitive than NED to precise spatial location and precipitation amount. OTS is further improved by using a fuzzy matching method. Application of OTS-based HCA for regional sub-domains is also introduced. Part II uses the HCA method developed in Part I to explore systematic clustering of the convection-allowing ensemble during the full 2009 HWT Spring Experiment period.

1. Introduction

Since ensemble forecasting was recognized as a practical way to provide probabilistic forecasts (Leith 1974), global-scale medium range ensemble forecasting has undergone dramatic advancement (e.g., Toth and Kalnay 1993; Molteni et al. 1996; Houtekamer et al. 1996; Hamill et al. 2000; Wang and Bishop 2003, 2005; Wang et al. 2004, 2007; Wei et al. 2008). Meso/regional-scale short-range ensemble forecasting has also been studied for over a decade (e.g., Du et al. 1997; Stensrud et al. 2000; Hou et al. 2001; Stensrud and Yussouf 2003; Eckel and Mass 2005; Clark et al. 2008, 2009; Bowler and Mylne 2009; Berner et al. 2011; Hacker et al. 2011). The extent to which results based on mesoscale ensembles are applicable when convective motions are explicitly included is not known. For example, cumulus parameterization in mesoscale ensembles has been shown to dominate precipitation forecast uncertainty resulting from model physics (Jankov et al. 2005). Additionally, growth rates of convective-scale perturbations that may not be resolved at mesoscale resolution can be highly non-linear (Hohenegger and Schar 2007).

Since 2007, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma has run convection-allowing[1], or Storm-Scale, Ensemble Forecasts (SSEF) over a near-CONUS (CONtinental United States) domain during the National Oceanic and Atmospheric Administration Hazardous Weather Testbed (NOAA HWT) Spring Experiments (Xue et al. 2007, 2008, 2009, 2010; Kong et al. 2007, 2008, 2009, 2011). The CAPS Spring Experiment data sets provide a unique opportunity to study many scientific issues for convection-allowing forecasts, as listed in Xue et al. (2009), and have helped answer many

[1] Convection-allowing resolution refers to grid spacing less than or equal to 4 km which allows vertical redistribution of heat and moisture to be effectively represented by grid-scale convection (Weisman et al. 1997), making cumulus parameterization unnecessary. The term convection-resolving is avoided because the convective scale details are not necessarily adequately resolved (Bryan et al. 2003; Petch 2006).

questions related to SSEF (Kong et al. 2007, 2008, 2009, 2011; Clark et al. 2009, 2010a, 2010b; Coniglio et al. 2010; Kain et al. 2010; Schwartz et al. 2010; Xue et al. 2010).

The above studies have examined the impacts of convection-allowing resolution, model physics, and Initial and Lateral Boundary Condition (IC/LBC) perturbations on spread, skill and statistical consistency of non-precipitation variables as well as precipitation forecast bias and skill. New post-processing methods for SSEFs have also been shown to improve skill over traditional methods (Clark et al. 2009; Schwartz et al. 2010). Yet, many research questions on SSEFs still remain to be answered by the data sets.

This two-part study uses the SSEFs produced during the 2009 Spring Experiment to study how the ensemble member forecasts are clustered and to relate the clusters to how the ensemble members were generated. This is done with a Hierarchical Cluster Analysis technique (HCA, Anderberg 1973; Alhamed et al. 2002). Such study can help to understand the impacts and importance of the sources of uncertainty in model physics, model dynamics, and IC/LBCs on ensemble diversity for a convection-allowing ensemble, which will be discussed in detail in part II.

A requirement for HCA is a suitable measure of the dissimilarity or "distance" between forecasts. For the high resolution precipitation forecasts emphasized in this study traditional metrics of measuring the distance between forecasts based on a point-wise comparison, such as Equitable Threat Score or Mean Square (or Absolute) Error, are inappropriate. Traditional metrics are inappropriate because of the small horizontal scale of features compared to the horizontal scale of acceptable spatial errors (Baldwin et al. 2001). This limitation of point-wise metrics is further exaggerated by a double penalty whereby high-amplitude small-scale features with small spatial errors are penalized once for missing the correct location and again for

forecasting in the incorrect location (Baldwin et al. 2001). As a result, traditional metrics can disagree with subjective evaluations (Davis et al. 2006).

In Part I, an object-oriented HCA method is developed. In this new HCA method, the distance between precipitation forecasts is quantified using an object-oriented measure based on the Method for Object-based Diagnostic Evaluation (MODE, Davis et al. 2006). The new distance measure allows for improved automated clustering of precipitation forecasts over traditional distance measures because the object-oriented distance is not based on a point-wise comparison of the forecasts. Instead, distance is based on features of discrete objects within the forecasts, which is more appropriate for comparing precipitation fields at high resolution (Baldwin et al. 2001; Davis et al. 2006; Gilleland et al. 2009).

Part I and Part II are organized as follows. This paper (Part I) develops the object-oriented HCA method and illustrates it with a representative case from 13 May 2009, during the 2009 NOAA HWT Spring Experiment. Part II (Johnson et al. 2011) uses the new HCA method, developed in Part I, to explore systematic clustering of the ensemble members over the entire 2009 NOAA HWT Spring Experiment. Section 2 of the present paper introduces the forecast and observation data, followed by a brief overview of the severe weather case examined in Section 3. The HCA algorithm is described in Section 4, followed by a discussion of bias adjustment in Section 5. HCA results using different distance measures are compared in Section 6. Section 7 shows how the results change when focused on a smaller region and Section 8 presents a summary and discussion.

2. Convection-allowing ensemble and verification data

Recent advances in computational resources have allowed the CAPS to produce experimental real-time SSEF for several weeks for the NOAA HWT Spring Experiment over a near-CONUS domain at a convection-allowing resolution. During the spring of 2009, the ensemble consisted of 20 members, with 10 members from the Weather Research and Forecast (WRF) Advanced Research WRF (ARW; Skamarock et al. 2005), 8 members from the WRF Non-hydrostatic Mesoscale Model (NMM; Janjic 2003), and 2 members from the CAPS Advanced Regional Prediction System (ARPS; Xue et al. 2000, 2001, 2003). The grid spacing is 4 km and none of the forecasts use cumulus parameterization. A more detailed description of the ensemble configuration can be found in Xue et al. (2009). The data set consists of 28 sets of forecasts run out to 30 hours, initialized at 00 UTC of weekdays from 30 April 2009 to 5 June 2009 after discarding 2 days due to incomplete data. Each member is labeled according to its model core and IC perturbation (e.g., ARW N1, NMM N2, etc.). The details of how ensemble members were generated are listed in Table 1.

Quantitative Precipitation Estimates (QPE) from the National Severe Storm Laboratory (NSSL) Q2 product are used for verification of precipitation forecasts and referred to as the observations. The NSSL QPE is interpolated from a 1 km grid to the same 4 km grid as the model forecasts for direct comparison. The QPE is obtained from radar estimates as described in Zhang et al. (2005). Data are only examined over a sub-domain within the full forecast grid (Fig. 1) to reduce the impact of lateral boundary conditions.

3. Description of Case

We selected a case from 13 May 2009 to introduce a new HCA framework because a significant severe weather outbreak occurred over a large area. Intense convection developed in the afternoon along a cold front extending from western Oklahoma to northwest Wisconsin where surface dewpoints were in the middle and upper 60's F and strong winds aloft were indicated by a strong 500 hPa height gradient (Fig. 2). Several tornado and severe hail reports between 23 UTC 13 May and 01 UTC 14 May are found in the Storm Prediction Center (SPC) storm log (http://www.spc.noaa.gov). Forecasts initialized at 00 UTC 13 May 2009, valid at 00 UTC 14 May 2009, are the focus of much of this paper and are shown in Fig. 3 for reference throughout the paper.

The 13 May 2009 case is used, together with expected scales and features of interest for forecasting intense precipitation, to tune the configuration of some MODE parameters. Results are also compared to several other cases with minimal additional tuning to verify that the parameters perform well on other cases with diverse forecast scenarios (e.g., 2 May 2009 and 2 June 2009). These other cases are not shown because the discussion of the 13 May 2009 case is representative of the other cases as well.

4. Method of Clustering

This section first describes the HCA algorithm and the traditional measures of distance that are used to cluster ensemble forecasts. A newly defined HCA using a non-traditional object oriented distance measure is then introduced.

a. Hierarchical Clustering Analysis (HCA) algorithm

HCA is a method of identifying potentially important relationships in a complex data set in order to facilitate hypothesis development (Jain and Dubes 1988; Gong and Richman 1995). HCA consists of initially identifying each forecast as a single-element cluster then iteratively merging two clusters into one until all forecasts are in the same cluster (e.g., Alhamed et al. 2002). HCA is selected for the present study because it requires no *a priori* assumptions about how many clusters exist (Jain and Dubes 1988), efficient and widely used algorithms (e.g., Ward 1963) are available, and primary clusters as well as secondary sub-clusters can be simultaneously identified (Fovell and Fovell 1993).

Ward's method (Ward 1963; Jain and Dubes 1988) is selected as the specific objective clustering algorithm because initial results showed better agreement with a manual clustering of forecasts based on our subjective evaluations (hereafter referred to as subjective clustering) compared to other potential methods. In Ward's algorithm, the distance between (i.e., dissimilarity of) single-forecast clusters is quantified with the squared Euclidean distance. The distance between multiple-forecast clusters is quantified as the increase of the Error Sum of Squares (ESS; Ward 1963) that would result from merging them into a single cluster. The two clusters with the smallest distance between them are merged at each step. For convenience, we define a new quantity, called variability, in this paper, in place of the ESS. The variability is defined as

$$ESS \propto \frac{2}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}, (i \neq j) \equiv variability \tag{1}$$

where N is the number of forecasts in the cluster, i and j are the index of each forecast in the cluster in turn, and d_{ij} is the distance between forecasts i and j. It can be shown that the

variability is proportional to the ESS when the distance between individual forecasts in the cluster, d_{ij}, is the squared Euclidean distance.

Hierarchical clustering is visualized graphically as a dendrogram (e.g., Fig. 4) with the ensemble of forecasts along the bottom horizontal axis. The merging of forecasts and clusters is depicted as two solid lines joining into one as the clustering proceeds from the bottom to the top of the dendrogram. The vertical axis is a cumulative measure of variability, summed over all clusters at that level. The distance between merged clusters is the increase of variability resulting from the merge. Therefore, the difference in the vertical axis values, $y_i - y_{i-1}$, is the distance between the clusters merged at the i^{th} iteration. In the dendrogram, lower level clusters contain more similar forecasts than higher level clusters.

b. Traditional and neighborhood probability Euclidean distance measure for HCA

Traditional distance measures are commonly defined in terms of a point-wise comparison of two fields. The standard measure for Ward's algorithm is squared Euclidean Distance (ED) which is defined between two forecasts, *i* and *j*, of a variable, *x*, at *K* grid points, where the index *k* refers to each grid point in turn:

$$ED_{ij} = \sum_{k=1}^{K} (x_i^k - x_j^k)^2 \qquad (2)$$

Thus the traditional implementation of Ward's algorithm uses $d_{ij} = ED_{ij}$ in Eq. 1.

A neighborhood method (Ebert 2008) is applied to the forecasts before computing the ED with the goal of reducing the impact of small spatial differences, and the corresponding double penalty, on ED. This provides a baseline for comparison to an object-oriented distance defined in section 3c.The neighborhood probability method used in the present study follows that of Schwartz et al. (2010) (see also Theis et al. 2005). To apply the neighborhood probability

method, each field of 1 hour accumulated precipitation is converted to a probability field. The resulting value at each point is defined as the percentage of grid points within a radius of 30 km that have hourly accumulated precipitation greater than 10 mm. The ED between these neighborhood probability fields (Neighborhood squared Euclidean Distance, NED) is used as a distance measure between forecasts. A threshold of 10 mm is chosen to emphasize heavy rainfall events. The threshold is applied over a radius of 30 km which is equal to 7.5 times the model grid spacing. Such parameter settings provide good balance between smoothing of features on unpredictable scales and retaining the larger scale structures, most consistent with a subjective interpretation of the forecasts (e.g., Fig. 5).

c. Object-Oriented distance measure based on MODE

MODE identifies objects in a gridded field by first smoothing the raw forecast into a convolved field. A threshold is then applied so that each contiguous area in the convolved field that exceeds a user-specified threshold defines the area of an object (Davis et al. 2006). User-specified attributes describing each object, such as shape, size or other properties of interest, are then calculated. In the new HCA framework, instead of using ED, the distance between two precipitation forecasts is determined by comparing the attributes of objects in the two fields. Thus the forecasts are no longer a set of spatial locations with a forecast value associated with each grid point, but are a smaller set of objects with several attributes associated with each object. Advantages of MODE include its easy adaptability to specific applications and the fact that it is maintained and made freely available by the National Center for Atmospheric Research as part of their Model Evaluation Tools package[2].

[2] Available for download at http://www.dtcenter.org/met/users/

As in the matching between a verification field and a forecast field that MODE is originally applied to, tunable parameters must be predefined. In our application of MODE for HCA, those tunable parameters are selected based on features and scales of interest, including the location, structure and organization of intense precipitation on meso- and storm-scales. Subjective evaluation of the quality of the HCA results also played a role in parameter selection. The parameters were tuned to give subjectively reasonable matching of objects on several independent cases with a variety of weather scenarios in addition to the 13 May 2009 case emphasized in this paper. For a detailed description of the parameters involved and how they were chosen in this study, please refer to Davis et al. (2009) and appendix A of this paper.

The object-oriented distance measure used to quantify distance between forecasts for the HCA, referred to here as the Object-based Threat Score (OTS), is a modification of the traditional Threat Score for use with the MODE algorithm. The OTS is defined as a weighted sum of the area of corresponding objects in both fields divided by the total area of all objects in both fields (see Davis et al. 2006 and appendix A for matching algorithm and interest functions):

$$OTS_{ij} = \frac{1}{A_i + A_j} \left\{ \sum_{p=1}^{P} w^p (a_i^p + a_j^p) \right\}, \qquad (3)$$

where A_i is the total area of objects in field i, A_j is the total area of objects in field j, P is the number of pairs of corresponding objects which have area of a_i^p and a_j^p, and w^p is the weight applied to object pair p. As defined in Appendix A, Eq. A1, the degree of similarity between a pair of objects is defined by a quantity called "total interest" which has a value between 0 and 1. Given an object in one field, the corresponding object in the opposing field is defined as the object with the highest total interest value. In practice, the corresponding object pairs are assigned as follows. First, the total interests between all possible pairs of objects from the

opposing fields, *i* and *j*, are calculated and sorted from highest to lowest. Then, the objects of the first pair (i.e., with highest total interest) are considered to correspond to each other and all other pairs containing one of those two objects are removed from the list. The process is then repeated with the next pair remaining in the sorted list until the list is empty. This process ensures that each object can correspond to at most one object in the opposing field. Thus $P = \min(M_i, M_j)$, where M_i and M_j are the number of objects in field *i* and *j*, respectively.

OTS can be considered in a binary or a fuzzy context. In a binary context $w^p = 1$ if the total interest between corresponding objects is greater than a matching threshold and $w^p = 0$ otherwise. The matching threshold in the binary context is defined as 0.6 based on good agreement of the resulting clusters with the subjective clustering. The effectiveness of the matching threshold depends on the choice of attributes and interest functions comprising the total interest. Several thresholds were tried (including 0.7 used in Davis et al. 2009) and we found a threshold of 0.6 provided better clustering results in our study. In a fuzzy context, w^p is equal to the total interest for that pair of corresponding objects, and thus varies continuously between 0 and 1. We call it "fuzzy" because unlike the binary case, there is not a clear distinction between similar and dissimilar. Binary OTS equal to 1 occurs when all objects in both fields are sufficiently similar to a unique object in the opposing field to be considered a match and both fields contain the same number of objects. Conversely, binary OTS equal to 0 occurs when none of the objects in either field are sufficiently similar to be considered a match to an object in the opposing field. In contrast, fuzzy OTS is only equal to 1 when the two fields are identical and approaches 0 as the interest between every possible pair of objects approaches zero. When used as a distance measure for HCA, OTS is first subtracted from 1.

The binary OTS (i.e., w^p is either 1 or 0) has been referred to previously as the Area Weighted Critical Success Index (AWCSI; Weiss et al. 2009) and the "fraction of rain area within matched objects" (Davis et al. 2009, their table 4). To the author's knowledge, it has not previously been applied in a fuzzy context. Davis et al. (2009) note the limitations of using a binary decision to determine matched objects and define a Median of Maximum Interest (MMI) to measure the distance between forecasts and observations based on the distribution of (fuzzy) total interest values. Our initial results suggest that the MMI is less suitable than the fuzzy OTS for the present application (not shown). The OTS terminology is used here for brevity and because "area weighted" is implied by analogy to the traditional Threat Score which can be interpreted as the intersection area divided by the union area.

d. Applicability of distance defined by the OTS in HCA

Ward's algorithm for HCA merges the two clusters at each step that result in the smallest increase of variability as defined in Eq. 1, with $d_{ij} = ED_{ij}$. In the object-oriented framework the forecasts are not represented as a gridded field of values so ED_{ij} is undefined. We therefore define an object-oriented measure of variability by replacing the ED with the OTS so that now $d_{ij} = OTS_{ij}$ in Eq. 1. This modification of Ward's algorithm is referred to as object-oriented HCA. Appendix B demonstrates that object-oriented variability is a reasonable measure of within cluster variability in sub-section Ba. Sub-section Bb demonstrates that the traditional algorithm for implementing Ward's algorithm applies to object-oriented variability.

5. Bias adjustment for HCA

A commonly occurring characteristic of precipitation forecasts at convection-allowing resolution is a large positive bias that can depend on the physics configuration (Schwartz et al. 2010). HCA of ensemble forecasts of precipitation amount, and therefore the amplitude bias, are of interest to many users such as hydrological prediction centers. However, in this study we focus on the location, structure and organization of the precipitation forecasts from the perspective of operational forecasters at the SPC (as described in appendix A). In order to minimize the impact of amplitude bias and focus on other aspects of the forecasts, the forecasts are adjusted for known biases before they are clustered. Object attributes related to the intensity of rain rate and intensity distribution within objects are also not included in the determination of the object-oriented distance to be consistent with this focus.

For the NED-based HCA, the neighborhood probability forecasts are adjusted to account for bias by tuning a different precipitation threshold for each member. This threshold is determined based on the total area within the verification domain (Fig. 1) that the neighborhood probability exceeds 0.25, averaged over all days at the same 24 hour forecast range. The bias-adjusted threshold for each member is tuned so this average area is within 5% of that of the observations using a 10 mm threshold for the observations. Results are not sensitive to a range of the neighborhood probability chosen (0.25-0.35, not shown).

For the OTS-based HCA, the forecasts are adjusted to account for bias in a similar manner, using a method based on the determination of thresholds in Skok et al. (2009). Skok et al. (2009) used MODE thresholds to ensure that the total area of objects is consistent with the total area of rainfall exceeding a threshold of interest on average. In contrast to the method in Skok et al. (2009), the goal here is to ensure that the total area of MODE objects from forecasts

in each ensemble member is consistent with the total area of MODE objects from the observations on average. A MODE threshold of 6.5 mm results in a total area of objects, averaged over 26 days, nearly twice as large in the NMM control forecasts as in the ARW control forecasts, and nearly three times as large in the NMM control forecasts as in the observed fields (Table 2). Even among members with the same model there are differences as large as a factor of 2 between the average total area of MODE objects for different members (Table 2). Qualitatively similar results were found in Davis et al. (2009) for a threshold of 3 mm hr^{-1} using a different set of forecasts. The thresholds are therefore adjusted for each member until the average area for each member is within 5% of the observation average (Table 2). An observation threshold for MODE of 6.5 mm is chosen to be lower than the 10 mm threshold used for the NED. While MODE objects subjectively appear more reasonable on many cases (not shown) with the lower 6.5 mm threshold,10 mm creates NED fields that look more similar to the raw fields than with 6.5 mm (Fig. 5). This is consistent with the fact that the MODE thresholds are applied to a convolved field while the NED thresholds are applied to raw fields. Compared to clusters without bias adjustment (not shown), bias-adjusted clustering is more consistent with subjective clustering. The bias adjustment methods adopted are intended for a diagnostic understanding of ensemble clustering. Further work is needed for real-time applications of bias-adjusted clustering where the bias can be estimated from the latest months preceding the current forecast.

6. Understanding differences in HCA with ED, NED, and OTS from a case study

Clusters of 24 hour forecasts of 1 hour accumulated precipitation, initialized at 00 UTC 13 May 2009, are created using ED, NED, binary OTS, and fuzzy OTS as distance measures and

subjectively evaluated in this section. The ED and NED distance measures are sensitive to effects of small spatial differences but not the structure of forecast features. In contrast, the object-oriented measures are able to appropriately cluster forecasts that are spatially close but do not quite coincide in location, particularly when features have similar structure. The OTS is also found to create more reasonable clusters when considered in a fuzzy, rather than binary, context. The results in this section are representative of other independent cases that were examined (not shown) and demonstrate the effectiveness of object-oriented HCA for clustering high resolution ensemble precipitation forecasts.

a. Comparison of ED to NED for the HCA

Clustering based on ED (i.e., ED HCA) is first compared to clustering based on NED (i.e., NED HCA). Figure 4 shows the dendrogram for the ED HCA, valid at 00 UTC 14 May 2009 (see Fig. 3 for the corresponding ensemble forecasts). The only clusters that subjectively make sense occur where large precipitation maxima are precisely co-located in those members only. For example, NMM N2, NMM P4 and ARW N1 are clustered together which makes sense because only those forecasts show an east to west oriented rainfall maximum in northern Illinois and a thin east-northeast to west-southwest oriented band of weaker precipitation across northern Missouri.

The ED HCA can be sensitive to small placement differences of otherwise similar features. For example, NMM CN and NMM C0 are not clustered together even though they look very similar subjectively. Both have a maximum along the northern Missouri/Illinois border with a thin band extending to the Oklahoma/Kansas border, along with smaller isolated maxima in western Oklahoma and along the Illinois/Indiana border. However, the ED HCA did not cluster

them together because of the dominating influence of small spatial differences in the high amplitude maxima near the Missouri/Illinois border, in spite of their similar structure (Fig. 3).

The ED HCA can also be sensitive to the amount of precipitation which can result in subjectively unrealistic clusters. High-amplitude, small-scale, features are rarely located at precisely the same grid point. When using the ED HCA, such features are thus typically compared to grid points without precipitation in the opposing field, rather than being compared to a corresponding feature. Thus, the ED is largely determined by the amplitude of such features. This effect is further magnified by the double penalty. Therefore two forecasts with a lot of precipitation tend to have a larger ED than two forecasts with little precipitation. The smaller ED between forecasts with low precipitation causes the cluster of forecasts that have dissimilar structure of forecast features but less precipitation than the other forecasts (ARW N2, NMMP1, and ARW P1). The sensitivity to amplitude can create unrealistic clusters even for co-located features. For example, for our application, we emphasize interpreting the forecasts in terms of convective mode and organization. Although amplitude differences between these co-located features are relatively unimportant in our application, they are still emphasized by the ED HCA.

In contrast to ED, NED makes use of nearby grid points and acts as a type of smoothing which relaxes the strict spatial sensitivity of ED (Ebert 2008). The NED HCA (Fig. 6) therefore results in improvement over the ED HCA. For example, the ED HCA clusters ARW P1 with ARW N2 and NMMP1 due to the relatively small amount of forecast precipitation shared by these members (Fig. 4). In contrast, the NED HCA clusters ARW P1 with ARW CN which is subjectively more reasonable because ARW P1 and ARW CN both show weaker disorganized showers over a broad area in Illinois and Missouri. Furthermore, unlike the ED HCA which clusters ARW P2 with NMM N4 and OBS that subjectively look different, the NED HCA

clusters ARW P2 with NMM N2 and NMM P4. This is another example of subjectively more reasonable clustering by the NED HCA than the ED HCA because these three members forecast the heavy precipitation to be focused mainly in north-central Illinois.

The NED HCA clusters subjectively appear more reasonable than the ED HCA clusters but they are far from perfect. For example, NMM CN and NMM CO are subjectively similar in terms of structure. However, like the ED HCA, this is not reflected in the NED HCA (Fig. 6). The NED HCA also suffers the same problems as the ED HCA due to sensitivity to precipitation amount rather than storm structure. For example, neither NMM N4 nor NMM P1 is subjectively similar to ARW N2. However, they cluster at a low level in the NED HCA (Fig. 6) because these three members have low precipitation relative to the other forecasts. The relatively low precipitation in these forecasts reduces the double penalty induced by small spatial errors. Although the sensitivity to the overall precipitation amount by the ED HCA and the NED HCA can be ameliorated by using a normalization such as standardized anomalies before clustering (Alhamed et al. 2002), the object-oriented distance measure as shown in the next sub-section, is more flexible and effective in clustering the forecasts in terms of the structure and mode of the features. The difference is that the object-oriented distance is based on a comparison of object attributes rather than a point-wise comparison. Normalization only accounts for domain total precipitation amount and not small spatial errors that still dominate high-resolution precipitation forecasts.

b. Comparison of the NED HCA to the OTS HCA

The OTS-based HCA (i.e., OTS HCA) further improved the clustering. This is a result of two main advantages of the OTS HCA over the NED HCA. First, the OTS HCA is sensitive to

the structure of features (i.e., size, shape, and orientation). Second, the OTS HCA is less sensitive than the NED HCA to precise spatial location and precipitation amount.

One advantage of the OTS HCA, relative to the NED HCA, is the ability to take into account structural similarity of features, regardless of their spatial location. For example, the ED and NED HCA (Fig. 4 and Fig. 6, respectively) show cophenetic proximity[3] between NMM CN and NMM C0 of 0.89 and 0.75, respectively. This implies a lack of relative similarity between these forecasts which is not consistent with the subjective analysis of these members in the previous subsection. In contrast, the OTS HCA (Fig. 7) shows cophenetic proximity between NMM CN and NMM C0 of 0.15 which is more consistent with their structural similarity.

A second advantage of the OTS HCA, relative to the NED HCA, is a reduced sensitivity of the OTS HCA to precise spatial location and precipitation amount. This reduced sensitivity is due to the OTS being based on a comparison of object attributes rather than a point-wise comparison of precipitation values. Furthermore, since the OTS is defined by user-selected tunable parameters, amplitude differences between particular features can be ignored or limited through the choice of object attributes and interest functions. This improves the resulting HCA because subjective clustering for this application is concerned with storm structure and approximate location but not necessarily precipitation amount. For example, neither NMM N4 nor NMM P1 is subjectively similar to ARW N2. However, they cluster at a low level in the NED HCA because of the low overall precipitation amount (Fig. 6). In comparison, ARW N2 has a relatively large OTS distance to NMM P1 and NMM N4 and therefore they cluster at relatively high level in the OTS HCA (Fig. 7) which is subjectively more reasonable. The result

[3] Cophenetic proximity (Jain and Dubes 1988) is the height on the dendrogram where two members first merge into the same cluster, as a fraction of total dendrogram height. It is an indication of the dissimilarity of the members relative to the dissimilarity to other members in the hierarchical clustering.

is a cophenetic proximity between ARW N2 and NMMP1 (NMM N4) of 0.73 (1.0) in the OTS HCA (Fig. 7) instead of 0.08 (0.13) in the NED HCA (Fig. 6).

c.) Comparison of binary and fuzzy OTS for HCA

Both the binary OTS HCA and the fuzzy OTS HCA have the advantages over the NED HCA that are discussed in the previous subsection. However the fuzzy OTS HCA has two additional advantages over the binary OTS HCA because the fuzzy OTS does not require a matching threshold.

The first advantage of the fuzzy OTS HCA, relative to the binary OTS HCA, is that it avoids a discontinuity in the distance calculation among a large group of forecasts. This is illustrated in Fig. 8 showing forecasts from ARW N3, NMM N3 and NMM P4. Both the binary and fuzzy distances between NMM N3 and NMM P4 are quite low, indicating similar forecasts, largely because the total interest between the large object in northern Illinois in both forecasts is 0.81. NMM N3 and ARW N3 have similarly small distances between them so it is reasonable to expect that NMM P4 and ARW N3 are similar. However, NMM P4 has the maximum possible binary OTS distance to ARW N3 of 1.0. In contrast, the fuzzy OTS distance between NMM P4 and ARW N3 is only 0.234 (0.219) larger than the fuzzy OTS distance between NMM P4 and NMM N3 (ARW N3). The difference in the binary OTS distance is due to the total interest between the large objects in northern Illinois being 0.57 between ARW N3 and NMM P4. Since this is just below the matching threshold of 0.6 there is a large and discontinuous difference in the binary OTS while the difference in the fuzzy OTS is gradual and continuous. In general, there is sometimes a large subjective difference between two forecasts that has little impact on the binary OTS distance to a third forecast. Other times a small subjective difference between

two forecasts has a large impact on the binary OTS distance to a third forecast. This does not occur for the fuzzy OTS because there is no matching threshold.

The second advantage of the fuzzy OTS is that it is conceptually more robust than the binary OTS since it can discriminate marginal matches and non-matches from very good matches and completely spurious objects, respectively. In contrast, the binary OTS will give 2 forecasts (A and B) an equal distance of 0.0 to a third forecast (C) if all objects in A and B match all objects in C, even if the objects in A are subjectively much more similar to the objects in C than are the objects in B. In such a case the fuzzy distance between A and C would be smaller than the fuzzy distance between B and C while the binary distance for both would be the same. This limitation of binary OTS HCA cannot be avoided by raising the matching threshold because then the limitation would be that all unmatched objects are treated equally. Figure 8 also illustrates this second advantage since the binary OTS distance between ARW N3 and NMM P4 gives no weight to the large object in northern Illinois, even though the total interest is close to the matching threshold. In contrast, the fuzzy OTS distance gives partial weight to this object for almost being a match.

7. Regional OTS HCA

Although the OTS HCA can be applied to the full verification domain (Fig. 1) to mimic subjective impressions of overall similarity among ensemble members, for certain practical applications a local or regional OTS HCA may be more appropriate. For example, on the 13 May 2009 case the fact that the forecasts from ARW N2 and NMM P2 are entirely different in the Mississippi and Ohio River Valley region is irrelevant when evaluating the potential for, and/or

organization of, convection in the Southern Plains. Therefore this section demonstrates a method to apply the object-oriented HCA to a particular geographic region.

The 13 May 2009 case can be divided into two regional forecasting problems (Fig. 9). The first problem is forecasting the mode and organization of convection in the Midwest. The second problem is forecasting the potential for convection in the Southern Plains and the southward extent of convection along the cold front. For this example we choose two center points so the regions within 600 km of the center point encompass most of the precipitation forecast by all members in the Midwest and Southern Plains with minimal overlap. Objects with a centroid more than 600 km from the center point are excluded from the calculation of the OTS. A sharp distance threshold would make the results very sensitive to small spatial differences in objects with a centroid close to the edge of the region. Therefore full weight is only given to the area of objects with a centroid less than 300 km from the center point. Otherwise, a factor, decreasing linearly from 1.0 at 300 km distance to 0.0 at 600 km distance, is multiplied by the area of objects before performing calculations of OTS[4].

The fuzzy OTS dendrogram for the Southern Plains region (Fig. 10) demonstrates the effectiveness of regional OTS HCA. The majority of clusters in the northern region are unchanged from the full domain OTS HCA in this case, and therefore not shown. The majority of precipitation on this case was forecast in the northern region so the results in section 7 are already dominated by this region. In the southern region (Fig. 10), members NMM N4 and NMM P4 are clustered with close cophenetic proximity of less than 0.05 while the same members have the most distant possible cophenetic proximity of 1.0 in the full domain OTS HCA (Fig. 7). This change makes sense when focusing on the southern region because both

[4] A localization based on decreasing weight with increasing distance from a point of interest is conceptually similar to localization used in data assimilation (e.g., Janjic et al. 2011)

members have a thin line of convection in southeast Kansas and a small isolated cell in western Oklahoma. Other clusters that form at a low dendrogram height in Fig.10 are also more representative of subjective impressions over the Southern Plains than the clusters from the full domain OTS HCA (Fig. 7) where the same members do not merge until much higher on the dendrogram (e.g., cluster of ARWP1, ARW P2, and NMM P1, or cluster of ARW CN and ARW P3). Similarly, members with little subjective similarity in the Southern Plains (e.g., ARW C0 vs. ARPS CN) that have close cophenetic proximity in the full domain OTS HCA (Fig. 7) have much more distant cophenetic proximity in the regional OTS HCA (Fig. 10).

8. Summary and Discussion

This paper is the first of a two-part study which seeks a systematic understanding of the impacts and relative importance of different sources of uncertainty within the 2009 CAPS Spring Experiment convection-allowing ensemble through an automated Hierarchical Clustering Analysis (HCA). Instead of using the traditional squared Euclidian distance (ED), an Object-based Threat Score (OTS) is defined in a fuzzy context and used to quantify dissimilarity of precipitation forecasts in the HCA. The fuzzy OTS is defined as the sum of the area of all paired objects from two fields, weighted by a fuzzy value between 0 and 1 representing their degree of similarity, divided by the total area of all objects in the two fields. The objects are identified using MODE where each member is tuned to use a different convolved threshold for object identification in order to account for different forecast biases in each member. The fuzzy OTS is then used to quantify the dissimilarity among ensemble members to conduct an HCA on convection-allowing hourly accumulated precipitation forecasts over a large verification domain as well as smaller regional domains. The effectiveness of the fuzzy OTS HCA is illustrated by

comparison to the ED HCA, the NED (Neighborhood Euclidian Distance) HCA, and the binary OTS HCA during a severe weather event on 13 May 2009.

The Fuzzy OTS HCA results in clusters that are more consistent with subjective clustering than the ED HCA, the NED HCA, and the binary OTS HCA. The ED HCA is the least effective on the representative case of 13 May 2009. Only features with similarity at a precise grid-point are clustered with the ED HCA while the similarity is otherwise determined by the precipitation amount, as expected from previous studies noting the impact of the double penalty (e.g., Baldwin et al. 2001). The NED HCA shows some improvement by relaxing the strict grid-point precision required of the ED HCA. The bias adjusted NED HCA shows even further improvement by removing the impact of systematic differences in precipitation amount. However, the bias adjusted NED HCA is still sensitive primarily to the location of precipitation features as well as the precipitation amount. The Binary OTS HCA improves the clustering further to be more consistent with the subjective clustering due to its capability to explicitly account for the size, shape, and orientation of precipitation areas. The Fuzzy OTS HCA is the most effective clustering method because it retains the positive qualities of the binary OTS HCA without suffering from the discontinuity issue in the clustering that was caused by the use of a pre-specified matching threshold in the binary OTS.

Compared to the binary OTS HCA, there are at least two advantages of the fuzzy OTS HCA arising from the absence of a matching threshold. One advantage is that large, discontinuous changes in the fuzzy OTS do not occur for small changes in the forecast. Another advantage is that all matched (and unmatched) objects are not treated equally allowing better and worse matches to be discriminated with the fuzzy OTS.

By demonstrating a relatively effective method of clustering convection-allowing precipitation forecasts, this paper provides a framework for a more systematic examination of the ensemble clustering tendencies. This is undertaken in Part II with a goal of understanding the impacts and importance of different ensemble perturbations in the 2009 CAPS Spring Experiment ensemble to inform future researchers and designers of convection-allowing ensembles.

Acknowledgments.

The authors are grateful to NSSL for the QPE verification data. We thank Nusrat Yussouf for providing tested and documented software for performing hierarchical cluster analyses, Dave Stensrud for helpful discussions of the results, NCAR for making MODE source code available, and two anonymous reviewers whose comments and suggestions greatly improved the manuscript. This research was supported by Science Applications International Corporation (SAIC) as a sponsor of the AMS Graduate Fellowship program and University of Oklahoma faculty start-up award 122-792100 and NSF AGS-1046081. The CAPS real-time forecasts were produced at the Pittsburgh Supercomputing Center (PSC), and the National Institute of Computational Science (NICS) at the University of Tennessee, and were mainly supported by the NOAA CSTAR program (NA17RJ1227). Some of the computing for this project was also performed at the OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma (OU). Kevin Thomas, Yunheng Wang, Keith Brewster and Jidong Gao of CAPS are also thanked for the production of the ensemble forecasts.

Appendix A. MODE configuration

While the MODE algorithm closely follows the description in Davis et al. (2009), there are numerous tunable parameters to be specified. The minimum object area is specified at 16 grid points ($4\Delta x \times 4\Delta y$). A smoothing radius of 4 grid-points is also applied to the raw forecasts to smooth features on unresolved scales. The attributes used to describe objects and their associated weights and confidence values are shown in Table A1. The degree of similarity between attributes of different objects (i.e., interest value) is quantified using interest functions shown in Figure A1.

The ensemble in this particular study was developed and used in the context of real time forecasting of severe weather so the forecasts are subjectively interpreted from the perspective of operational forecasters at the Storm Prediction Center (SPC). The SPC forecasters typically use convection-allowing model guidance for predicting the location and timing of convective initiation, storm modes, and the potential for evolution into a larger scale system (Weiss et al. 2004; Coniglio et al. 2010). Thus the precise amplitude and location of forecast features are of only secondary relevance to this study relative to the structure, organization and approximate location of intense convection. The flexibility of the object-oriented approach allows us to focus on this specific application although different users might emphasize different features.

The emphasis of this study on the structure, organization and approximate location of intense convection not only motivates our use of object-oriented distance measures, but also guides our choice of the object attributes (Table A1) of area, centroid location, orientation angle, and aspect ratio. Area is selected to represent the amount of upscale organization of convective systems. Orientation angle and aspect ratio are selected to represent convective mode (e.g. linear or cellular). Centroid location is selected because approximate location is also important.

Approximate, rather than precise, location is emphasized by assigning objects with up to 40 km centroid distance an interest value of 1.0 (Fig. A1). A linear form of all interest functions is chosen for simplicity in lieu of established guidelines otherwise. The x-intercept in Figures A1c and A1d was selected to be consistent with subjective impressions of how well the total interest (Eqn. A1) described the degree of similarity over a large number of different object pairs.

A total interest, I, is defined for the r^{th} object pair is a weighted sum of the interest values of each of the S object attributes, denoted by s index (Davis et al. 2009):

$$I_r = \frac{\sum_{s=1}^{S} c_s w_s F_{sr}}{\sum_{s=1}^{S} c_s w_s} \qquad \text{(A1)}$$

In Eqn. A1, c is the confidence in an attribute, w is the weight assigned to an attribute, and F is the interest value of the attribute for the object pair (e.g., Fig. A1). Since the interest values of each attribute are defined between 0 and 1 and the effective weight applied to each interest value summed over all attributes is equal to 1, the only constraint on c and w is that they are non-negative. The total interest, I, is a value between 0 and 1.

In Table A1, confidence for angle difference follows Davis et al. (2009) to give less weight to angle difference when objects are not linear, while confidences for angle difference and aspect ratio difference are also multiplied by the product of area ratio (AR) and centroid distance interest (CDI). Thus the effective weights become half location and half size for objects that are far apart or very different in area and become one third location, one third size, and one third structure (aspect ratio and angle) for objects of similar size in similar locations. This was done because as size or location becomes less similar there is less confidence that the objects

represent the same feature so it is less relevant whether they have similar structure. The default confidence for centroid distance, equal to the area ratio, is used resulting in small weight to centroid distance interest when the area ratio is very small. The confidence value for area ratio is a function of centroid distance (CD) so that objects that are extremely far apart (i.e., CDI of 0.0) but happen to have similar size (i.e., AR about 1) have a near zero interest (rather than 0.5) since those objects do not correspond to each other.

Appendix B. Non-Euclidean distance measure in Ward's algorithm

a. Correspondence between object-oriented variability and ensemble spread

Object-oriented variability, as defined in Eq. 1 with $d_{ij}=OTS_{ij}$, is intended to provide an automated comparison of spread in different groups of forecasts in a way that mimics how a subjective analyst would compare them manually. In this way it is consistent with the intended use of MODE as a way to mimic a subjective analysis (Davis et al. 2009). For example, consider the three clusters of three members in Figure B1 from the 13 May 2009 North region which have object-oriented variability for columns (a), (b) and (c) of 1.36, 1.11, and 0.66 respectively[5]. The cluster in column (a) subjectively appears to have a lot of spread since it includes forecasts both with and without an object in east-central Illinois while the forecasts in Missouri range from a single linear object, to several small objects, to nothing at all. The cluster in column (b) has less spread subjectively because all the forecasts have a large rain area although they have large differences in placement. The forecasts in column (c) have the least spread subjectively because

[5] Note that $d_{ij}=OTS_{ij}$ in Eq. 1 is here based on a regional subdomain centered over western Illinois. The regional emphasis is achieved when calculating the OTS (Eq. 3) by giving full weight to objects within 300 km of the region's center and linearly decreasing the weight given to the area of each object between 300 and 600 km from the center. This is reflected in Fig. B1 by not showing objects located more than 600 km from the center of the region (Fig. 9b) and using lighter shading for partially-weighted objects between 300 and 600 km from the center.

they all have a large object in northern Illinois and have similar placement and structure of objects in Missouri. In other words, higher variability corresponds to larger subjective impressions of spread. Most other cases that were subjectively examined exhibited the same correspondence between object-oriented variability and subjective impressions of spread.

b.) Implementation of object-oriented HCA

Object-oriented HCA can be implemented with the same algorithm commonly used for the traditional Ward's algorithm. Ward's algorithm is commonly implemented by defining the distance, D_{ij}, between clusters *i* and *j*, where *j* is the new cluster resulting from merging clusters *k* and *l* from the previous step, as follows (Anderberg 1973; Jain and Dubes, 1988):

$$D_{ij} = \left(\frac{N_k + N_i}{N_k + N_l + N_i} \right) D_{ik} + \left(\frac{N_l + N_i}{N_k + N_l + N_i} \right) D_{il} - \left(\frac{N_i}{N_k + N_l + N_i} \right) D_{kl} \quad , \quad \text{(B1)}$$

where N_i, N_j, N_k are the number of elements in clusters *i*, *j*, and *k* respectively. Note that the distance between clusters of multiple forecasts, D_{ij}, and the distance between individual forecasts, d_{ij}, are only equal for clusters of size N=1. The advantage of Eq. B1 is that it is efficient for large data sets because the variability (Eq. 1) does not have to be calculated for each possible merge of 2 clusters. Anderberg (1973) shows that, for the special case where $d_{ij}=ED_{ij}$ in Eq. 1[6], merging the two clusters with the smallest D_{ij} is equivalent to merging the two clusters associated with the smallest increase of variability.

[6] Anderberg (1973) uses Error Sum of Squares (ESS) instead of variability. The two can be shown to be proportional making the clustering result equivalent.

We now show that the above equivalence is true for *any* distance measure, d_{ij}, between the individual forecasts. In other words we will show that D_{ij} in Eq. B1 can also be defined as the variability of the new cluster minus the variability of each of the old clusters:

$$D_{ij} = \frac{2}{N_{ij}} \sum_{m=1}^{N_{ij}} \sum_{n=1}^{N_{ij}} d_{mn} - \frac{2}{N_i} \sum_{m=1}^{N_i} \sum_{n=1}^{N_i} d_{mn} - \frac{2}{N_j} \sum_{m=1}^{N_j} \sum_{n=1}^{N_j} d_{mn} \qquad \text{(B2)}$$

where d_{mn} is the distance between the individual forecasts m and n.

The equivalence of (B1) and (B2) can be proved if the following is true:

$$D_{ij} - \left(\frac{N_k + N_i}{N_k + N_l + N_i} \right) D_{ik} - \left(\frac{N_l + N_i}{N_k + N_l + N_i} \right) D_{il} + \left(\frac{N_i}{N_k + N_l + N_i} \right) D_{kl} = 0 \qquad \text{(B3)}$$

with each of D_{ij}, D_{ik}, D_{il}, and D_{kl} in (B3) being defined by (B2). Noting that $N_j = N_k + N_l$ and $N_{ij} = N_i + N_k + N_l$, and substituting (B2) into the left hand side of (B3), the left hand side becomes

$$\frac{2}{N_i + N_j} \sum_{m=1}^{N_i+N_j} \sum_{n=1}^{N_i+N_j} d_{mn} - \frac{2}{N_i} \sum_{m=1}^{N_i} \sum_{n=1}^{N_i} d_{mn} - \frac{2}{N_j} \sum_{m=1}^{N_j} \sum_{n=1}^{N_j} d_{mn}$$

$$- \left(\frac{N_k + N_i}{N_i + N_j} \right) \left(\frac{2}{N_i + N_k} \sum_{m=1}^{N_i+N_k} \sum_{n=1}^{N_i+N_k} d_{mn} - \frac{2}{N_i} \sum_{m=1}^{N_i} \sum_{n=1}^{N_i} d_{mn} - \frac{2}{N_k} \sum_{m=1}^{N_k} \sum_{n=1}^{N_k} d_{mn} \right)$$

$$- \left(\frac{N_l + N_i}{N_i + N_j} \right) \left(\frac{2}{N_i + N_l} \sum_{m=1}^{N_i+N_l} \sum_{n=1}^{N_i+N_l} d_{mn} - \frac{2}{N_i} \sum_{m=1}^{N_i} \sum_{n=1}^{N_i} d_{mn} - \frac{2}{N_l} \sum_{m=1}^{N_l} \sum_{n=1}^{N_l} d_{mn} \right)$$

$$+ \left(\frac{N_i}{N_i + N_j} \right) \left(\frac{2}{N_j} \sum_{m=1}^{N_j} \sum_{n=1}^{N_j} d_{mn} - \frac{2}{N_k} \sum_{m=1}^{N_k} \sum_{n=1}^{N_k} d_{mn} - \frac{2}{N_l} \sum_{m=1}^{N_l} \sum_{n=1}^{N_l} d_{mn} \right)$$

If we then expand the 4th and 7th terms using the summation identity,

$$\sum_{m=1}^{N_a+N_b}\sum_{n=1}^{N_a+N_b} d_{mn} = \sum_{m=1}^{N_a}\sum_{n=1}^{N_a} d_{mn} + \sum_{m=1}^{N_b}\sum_{n=1}^{N_b} d_{mn} + 2\sum_{m=1}^{N_a}\sum_{n=1}^{N_b} d_{mn} ,$$

We can collect like terms, many of which cancel after using $N_{ij} = N_i + N_j$ and $N_j = N_k + N_l$.

After dividing both sides of A7 by $2/N_{ij}$, the left hand side of (B3) becomes:

$$\sum_{m=1}^{N_{ij}}\sum_{n=1}^{N_{ij}} d_{mn} - \sum_{m=1}^{N_i}\sum_{n=1}^{N_i} d_{mn} - \sum_{m=1}^{N_j}\sum_{n=1}^{N_j} d_{mn} - 2\sum_{m=1}^{N_i}\sum_{n=1}^{N_k} d_{mn} - 2\sum_{m=1}^{N_i}\sum_{n=1}^{N_l} d_{mn}$$

$$= \sum_{m=1}^{N_{ij}}\sum_{n=1}^{N_{ij}} d_{mn} - \sum_{m=1}^{N_i}\sum_{n=1}^{N_i} d_{mn} - \sum_{m=1}^{N_j}\sum_{n=1}^{N_j} d_{mn} - 2\sum_{m=1}^{N_i}\sum_{n=1}^{N_j} d_{mn}$$

$$= 0$$

Therefore equations (B1) and (B2) are equivalent.

References

Alhamed, A., S. Lakshmivarahan, D. J. Stensrud, 2002: Cluster analysis of multimodel ensemble data from SAMEX. *Mon. Wea. Rev.* **130.** 226-256.

Anderberg, M.R., 1973: *Cluster Analysis for Applications.* Academic Press, 359 pp.

Atger, F., 1999: Tubing: An alternative to clustering for the classification of ensemble forecasts. *Wea. Forecasting,* **14,** 741–757.

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2001: Verification of mesoscale features in NWP models. Preprints, *9th Conf. on Mesoscale Processes*, Ft. Lauderdale, FL, Amer. Meteor. Soc., 255-258.

Benjamin, S. G., G. A. Grell, J. M. Brown, T. G. Smirnova, and R. Bleck, 2004: Mesoscale weather prediction with the RUC hybrid isentropic-terrain-following coordinate model. *Mon. Wea. Rev.,* **132**, 473–494.

Berner, J,, S.-Y. Ha, J. P. Hacker, A. Fournier, C. and Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multi-physics representations. *Mon. Wea. Rev*., In Press.

Bowler N. E., K. R. Mylne, 2009: Ensemble transform Kalman filter perturbations for a regional ensemble prediction system. *Quart. J. Roy. Meteor. Soc* .**135**, 757-766.

Clark, A.J., W.A. Gallus, and T.C. Chen, 2008: Contributions of mixed physics versus perturbed initial/lateral boundary conditions to ensemble-based precipitation forecast skill. *Mon. Wea. Rev.*, **136**, 2140–2156.

Clark, A. J., W. A. Gallus, M. Xue, F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and -parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140.

Clark, A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2010a: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612.

Clark, A. J., W. A. Gallus Jr., M. Xue, F. Kong, 2010b: Convection-allowing and convection-parameterizing ensemble forecasts of a mesoscale convective vortex and associated severe weather environment. *Wea. Forecasting* **25**:4, 1052-1081.

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

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

Davis, C.A., B.G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The Method for Object-Based Diagnostic Evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. *Wea. Forecasting*, **24**, 1252–1267.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.,* **125,** 2427–2459.

Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107.

Ebert, E.E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteorol. Appl.*, **15**, 51-64.

Eckel, F. A., and C.F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

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

Ferrier, B. S., 1994: A double-moment multiple-phase four-class bulk ice scheme. Part I: Description. *J. Atmos. Sci*., **51**, 249–280.

Fovell, R. G., and M. Y. C. Fovell, 1993: Climate zones of the conterminous United States defined using cluster analysis. *J. Climate*, **6**, 2103–2135.

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

Gong, X. and Richman, M.B., 1995: On the application of cluster analysis to growing season precipitation data in North America east of the Rockies, *J. Climate*, **8**, 897-931.

Hacker, J. P., S.-Y. Ha, C. Snyder, J. Berner, F. A. Eckel, E. Kuchera, M. Pocernich, S. Rugg, J. Schramm, and X. Wang, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus* **63A**, 1-17.

Hamill, T. M., C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular vector, and perturbed observation ensembles. *Mon. Wea. Rev.,* **128,** 1835–1851.

Hohenegger, Cathy, Christoph Schär, 2007: Predictability and error growth dynamics in cloud-resolving models. *J. Atmos. Sci.*, **64**, 4467–4478.

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

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.,* **129,** 73–91.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev*., **124**, 1225-1242.

Jain, A. J., and R. C. Dubes, 1988: *Algorithms For Clustering Data.* Prentice Hall. Pp 72-80.

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

Janjic´, Z. I., 2003: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys.,* **82,** 271–285.

Janjic, T., L. Nerger, A. Albertella, J. Schroter, S. Skachko, 2011: On domain localization in ensemble based Kalman filter algorithms. *Mon. Wea. Rev*. In Press.

Jankov, I., W.A. Gallus, M. Segal, B. Shaw, and S.E. Koch, 2005: The impact of different WRF model physical parameterizations and their interactions on warm season MCS rainfall. *Wea. Forecasting*, **20**, 1048–1060.

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

Kain J., and coauthors, 2010: Assessing advances in the assimilation of radar data and other mesoscale observations within a collaborative forecasting-research environment. *Wea. Forecasting*. **25**, 1510–1521.

Kong, F., and co-authors, 2007: Preliminary analysis on the real-time storm-scale ensemble forecasts produced as a part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. *Preprints, 22th Conf. on Weather Analysis and Forecasting and 18th Conf. on Numerical Weather Prediction* Amer. Meteor. Soc., Park City, UT, 3B.2.

Kong, F., M. Xue, M. Xue, K. K. Droegemeier, K. W. Thomas, Y. Wang, J. S. Kain, S. J. Weiss, D. Bright, and J. Du, 2008: Real-time storm-scale ensemble forecasting during the 2008 Spring Experiment. *24th Conf. Several Local Storms*, Savannah, GA, Ameri. Meteor. Soc., Paper 12.3.

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

Kong, F., M. Xue, K. W. Thomas, Y. Wang, K. Brewster, X. Wang, J. Gao, S. J. Weiss, A. Clark, J. S. Kain, M. C. Coniglio, and J. Du, 2011: Evaluation of CAPS multi-model storm-scale ensemble forecast for the NOAA HWT 2010 Spring Experiment. *24th Conf. Wea. Forecasting/20th Conf. Num. Wea. Pred.*, Amer. Meteor. Soc., P452.

Lacis, A. A., and J. E. Hansen, 1974: A parameterization for the absorption of solar radiation in the earth's atmosphere. *J. Atmos. Sci*., **31**, 118–133.

Leith, C., 1974: Theoretical skill of monte carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Lin, Yuh-Lang, Richard D. Farley, Harold D. Orville, 1983: Bulk parameterization of the snow gield in a cloud model. *J. Climate Appl. Meteor.*, **22**, 1065–1092.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, 122, 73-119.

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

Rutledge, G. K., J. Alpert, and W. Ebisuzaki, 2006: NOMADS: A climate and weather model archive at the National Oceanic and Atmospheric Administration. *Bull. Amer. Meteor. Soc.,* **87,** 327–341.

Schwartz, C.S., J.S. Kain, S.J. Weiss, M. Xue, D.R. Bright, F. Kong, K.W. Thomas, J.J. Levit, M.C. Coniglio, and M.S. Wandishin, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280.

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

Skok, G., J. Tribbia, J. Rakovec, and B. Brown, 2009: Object-based analysis of satellite-derived precipitation systems over the low- and midlatitude Pacific Ocean. *Mon. Wea. Rev.*, **137**, 3196–3218.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev*., **128**, 2077–2107.

Stensrud, D.J., and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.,* **131**, 2510-2524.

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

Theis S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev*., **136**, 5095–5115.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.*,**74**, 2317-2330.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.,* **60,** 1140–1158.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Q. J. R. Meteo. Soc.*, **131**, 965-986.

Wang, X., C. H. Bishop and S. J. Julier, 2004: Which is better, an ensemble of positive–negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.,* **132,** 1590–1605.

Wang, X., T. M. Hamill, J. S. Whitaker and C. H. Bishop, 2007: A comparison of hybrid ensemble transform Kalman filter-OI and ensemble square-root filter analysis schemes. *Mon. Wea. Rev.*, **135**, 1055-1076.

Ward J. 1963: Hierarchical grouping to minimize an objective function. *J. Amer. Statistical Association.* Vol. **58**, 236-244.

Wei, M., Z. Toth, R.Wobus, and Y.Zhu, 2008: Initial perturbations based on the Ensemble Transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62-79.

Weiss, S. J., J. S. Kain, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2004: Examination of several different versions of the Weather Research and Forecasting (WRF) model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. Preprints, *22nd Conf. on Severe Local Storms,* Hyannis, MA, Amer. Meteor. Soc., 17.1.

Weiss S., J. Kain, M. Coniglio, D. Bright, J. Levit, G. Carbin, R. Sobash, J. Hart, R. Schneider, 2009. NOAA Hazardous Weather Testbed Experimental Forecast Program Spring Experiment 2009: Program Overview and Operations Plan. pg. 49. http://hwt.nssl.noaa.gov/Spring_2009/

Xue, M., K. K. Droegemeier, and V. Wong, 2000: The Advanced Regional Prediction System (ARPS)—a multiscale nonhydrostatic atmospheric simulation and prediction tool. Part I: Model dynamics and verification. *Meteor. Atmos. Phys.,* **75,** 161–193.

Xue, M., and Coauthors, 2001: The Advanced Regional Prediction System (ARPS)—a multi-scale nonhydrostatic atmospheric simulation and prediction tool. Part II: Model physics and applications. *Meteor. Atmos. Phys.,* **76,** 143–166.

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

Xue, M., F. Kong, D. Weber, K. W. Thomas, Y. Wang, K. Brewster, K. K. Droegemeier, J. S. K. S. J. Weiss, D. R. Bright, M. S. Wandishin, M. C. Coniglio, and J. Du, 2007: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA hazardous weather testbed 2007 spring experiment. *22nd Conf. Wea. Anal. Forecasting/18th Conf. Num. Wea. Pred.,* Salt Lake City, Utah, Amer. Meteor. Soc., CDROM 3B.1

Xue, M., F. Kong, K. W. Thomas, J. Gao, Y. Wang, K. Brewster, K. K. Droegemeier, J. Kain, S. Weiss, D. Bright, M. Coniglio, and J. Du, 2008: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2008 Spring Experiment. *24th Conf. Several Local Storms*, Savannah, GA, Ameri. Meteor. Soc., Paper 12.2.

Xue, M., F. Kong, K. W. Thomas, J. Gao, Y. Wang, K. Brewster, K. K. Droegemeier, X. Wang, J. Kain, S. Weiss, D. Bright, M. Coniglio, and J. Du, 2009: CAPS realtime 4-km multi-model convection-allowing ensemble and 1-km convection-resolving forecasts from the NOAA Hazardous Weather Testbed 2009 Spring Experiment. *Extended Abstract, 23rd Conf. Wea. Anal. Forecasting/19th Conf. Num. Wea. Pred. Amer. Meteor. Soc.,* Paper 16A.2.

Xue, M., F. Kong, K. W. Thomas, Y. Wang, K. Brewster, J. Gao, X. Wang, S. Weiss, A. Clark, J. Kain, M. Coniglio, J. Du, T. Jensen, and Y.-H. Kuo, 2010: CAPS realtime storm scale ensemble and high resolution forecasts for the NOAA Hazardous Weather Testbed 2010 Spring Experiment. *25th Conf. Severe Local Storms*, Amer. Meteor. Soc., Paper 7B.3.

Yussouf N., D. J. Stensrud, S. Lakshmivarahan, 2004: Cluster analysis of multimodel ensemble data over New England. *Mon. Wea. Rev.* **132.** 2452-2462.

Zhang, J., K. Howard, and J. J. Gourley, 2005: Constructing three-dimensional multiple-radar reflectivity mosaics: Examples of convective storms and stratiform rain echoes. *J. Atmos. Ocean. Tech.*, **22,** 30-42.

List of Figures

FIG. 8. MODE objects and OTS distances for 1 hour accumulated precipitation forecasts valid 00 UTC 14 May 2009 for (a) ARW N3, (b) NMM N3, and (c) NMM P4.

FIG. 9. Regions selected for clustering of forecasts valid 00 UTC 14 May 2009. Center of region is white dot and shaded area is the region within 600 km of the center. (a) north region and (b) south region.

FIG. 10. Dendrogram of 1 hour accumulated precipitation forecasts valid 00 UTC 14 May 2009 using fuzzy OTS as distance measure and focusing on southern region.

FIG. A1. Functions mapping attribute value to interest value for (a) area ratio, (b) centroid distance, (c) aspect ratio difference, and (d) angle difference.

FIG. B1: MODE objects in forecasts valid at 00 UTC 14 May 2009 for (a) NMM N4, NMM P1 and ARW N2 (top to bottom), (b) ARW P3, ARPS C0, and NMM P2 (top to bottom), and (c) NMM N3, NMM P4, and ARW P2 (top to bottom). The variability of each column, defined by Eq. 1 with $dij=OTSij$, is given at the top of the column. Objects within 300 km of center of North Region (defined in fig 9a) are shaded black and objects centered between 300km and 600km of center of North Region, making a partial contribution to OTS, are shaded gray.

TABLE 1. Details of ensemble configuration with columns showing the members, Initial Conditions (ICs), Lateral Boundary Conditions (LBCs), whether radar data is assimilated (R), and which Microphysics scheme (MP; Thompson (Thom., Thompson et al. 2008), Ferrier (Ferr., 1994),WRF Single Moment 6-class (WSM6, Hong et al. 2004), or Lin (Lin et al. 1983) microphysics), Planetary Boundary Layer scheme (PBL; Mellor-Yamada-Janjic (MYJ, Janjic´ 1994), Yonsei University (YSU, Noh et al. 2003) or Turbulent Kinetic Energy(TKE)-based (Xue et al. 2000) scheme), Shortwave radiation scheme (SW; Goddard (Tao et al. 2003), Dudhia (1989) or Geophysical Fluid Dynamics Laboratory (GFDL, Lacis and Hansen 1974) scheme), and Land Surface Model (LSM; Rapid Update Cycle (RUC, Benjamin et al. 2004) or NOAH ((NCEP-Oregon State University-Air Force-NWS Office of Hydrology, Ek et al. 2003))) was used with each member. NAMa and NAMf are the direct NCEP-NAM analysis and forecast, respectively, while the CN IC has additional radar and mesoscale observations assimilated into the NAMa. Perturbations added to CN members to generate the ensemble of ICs, and LBCs for the SSEF forecasts are from NCEP SREF (Du et al 2006). SREF members are labeled according to model dynamics: nmm members use WRF-NMM, em members use WRF-ARW (i.e., Eulerian Mass core), etaKF members use Eta model with Kain-Fritsch cumulus parameterization, and etaBMJ use Eta model with Betts-Miller-Janjic cumulus parameterization.

Member	IC	LBC	R	MP	PBL	SW	LSM
ARW CN	CN	NAMf	Y	Thom.	MYJ	Goddard	Noah
ARW C0	NAMa	NAMf	N	Thom.	MYJ	Goddard	Noah
ARW N1	CN – em	em N1	Y	Ferr.	YSU	Goddard	Noah
ARW N2	CN – nmm	nmm N1	Y	Thom.	MYJ	Dudhia	RUC
ARW N3	CN - etaKF	etaKF N1	Y	Thom.	YSU	Dudhia	Noah
ARW N4	CN - etaBMJ	etaBMJ N1	Y	WSM6	MYJ	Goddard	Noah
ARW P1	CN + em	em N1	Y	WSM6	MYJ	Dudhia	Noah
ARW P2	CN + nmm	nmm N1	Y	WSM6	YSU	Dudhia	Noah
ARW P3	CN + etaKF	etaKF N1	Y	Ferr.	MYJ	Dudhia	Noah
ARW P4	CN + etaBMJ	etaBMJ N1	Y	Thom.	YSU	Goddard	RUC
NMM CN	CN	NAMf	Y	Ferr.	MYJ	GFDL	Noah
NMM C0	NAMa	NAMf	N	Ferr.	MYJ	GFDL	Noah
NMM N2	CN - nmm	nmm N1	Y	Ferr.	YSU	Dudhia	Noah
NMM N3	CN - etaKF	etaKF N1	Y	WSM6	YSU	Dudhia	Noah
NMM N4	CN - etaBMJ	etaBMJ N1	Y	WSM6	MYJ	Dudhia	RUC
NMM P1	CN + em	em N1	Y	WSM6	MYJ	GFDL	RUC
NMM P2	CN + nmm	nmm N1	Y	Thom.	YSU	GFDL	RUC
NMM P4	CN + etaBMJ	etaBMJ N1	Y	Ferr.	YSU	Dudhia	RUC
ARPS CN	CN	NAMf	Y	Lin	TKE	2-layer	Noah
ARPS C0	NAMa	NAMf	N	Lin	TKE	2-layer	Noah

TABLE 2. Total area (number of gridpoints) of all objects in verification domain, averaged over 26 days, for each ensemble member. First column is for using 6.5 mm threshold for all members. Second column is for using different thresholds as shown in the third column for the purpose of bias adjustment.

Member	**6.5 mm threshold area**	**Bias-adjusted threshold area**	**Bias-adjusted threshold**
ARWCN	3178	2064	8.5
ARWC0	3014	1983	8.5
ARWN1	3770	2110	9.0
ARWN2	2070	2070	6.5
ARWN3	2175	2050	6.75
ARWN4	3972	2011	10.0
ARWP1	2538	2033	7.5
ARWP2	2403	2143	7.0
ARWP3	3549	2053	9.0
ARWP4	2964	2000	8.5
NMMCN	5859	2006	14.0
NMMC0	5711	1985	14.0
NMMN2	3862	2013	10.5
NMMN3	3747	2045	10.5
NMMN4	5041	2012	13.0
NMMP1	5453	2027	14.0
NMMP2	3471	2143	9.5
NMMP4	3739	2049	10.25
ARPSCN	3289	2044	9.0
ARPSC0	3135	1991	8.8
OBS	2055	2055	6.5

TABLE A1. Attributes and parameter values used for MODE fuzzy matching algorithm (CD denotes Centoid Distance, CDI denotes Centroid Distance Interest, AR denotes Area Ratio, T denotes aspect ratio)

Attribute	**Weight**	**Confidence**
Centroid Distance	2.0	AR
Area Ratio	2.0	1.0 if CD ≤ 160 km 1 – [(CD – 160) / 640] if 160 km < CD < 800 km 0.0 if CD ≥ 800 km
Aspect Ratio Difference	1.0	CDI * AR
Orientation Angle Difference	1.0	CDI * AR * $\sqrt{a^2 + b^2}$ Where a,b are $(\frac{(T-1)^2}{T^2-1})^{0.3}$ for the two objects being compared

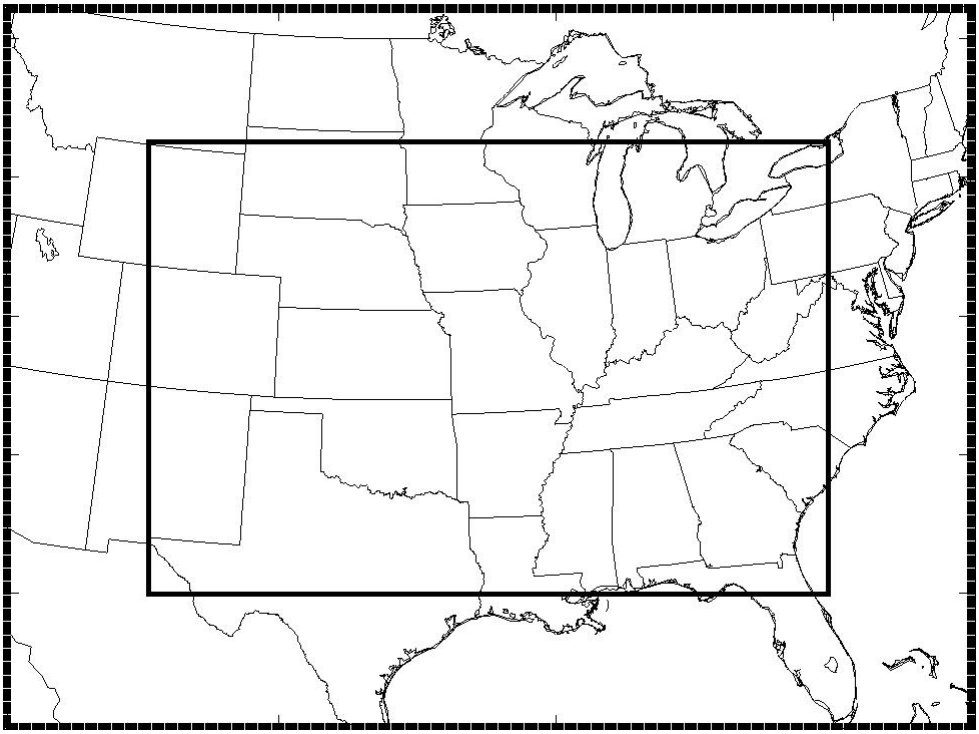

FIG. 1. Outer box is the model domain and inner box is the analysis domain used in the present study.

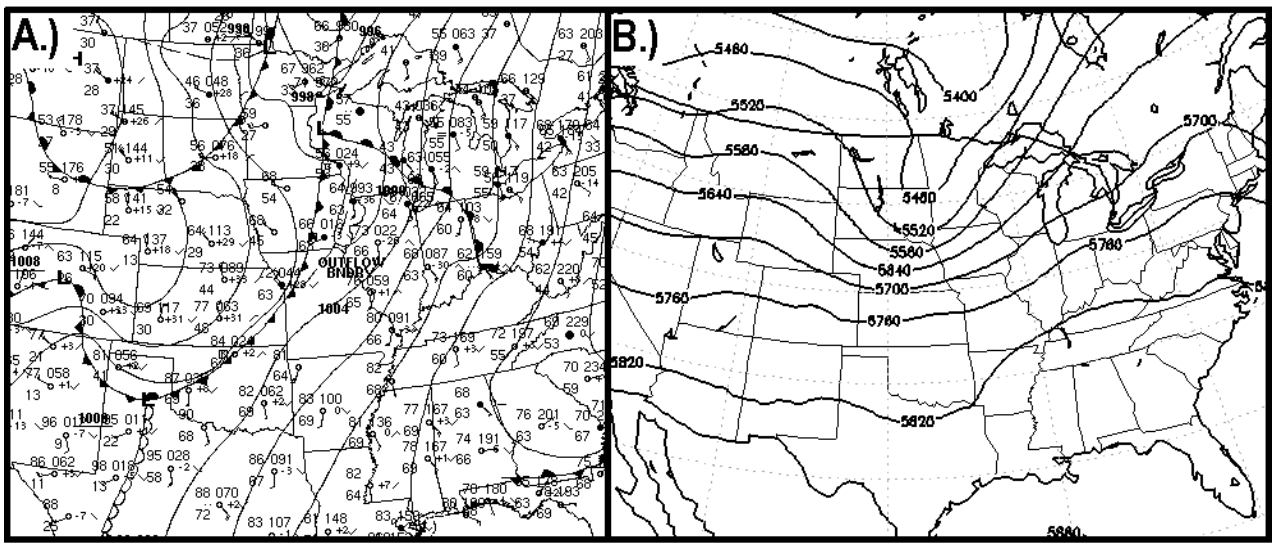

FIG. 2. (a) Surface analysis valid on 00 UTC 14 May 2009 from Hydrometeorological Prediction Center (HPC) (http://www.hpc.ncep.noaa.gov/html/sfc_archive.shtml) and (b) North American Regional Reanalysis of 500 hPa geopotential height at 00 UTC 14 May 2009 (obtained from NOMADS online archive, Rutledge et al 2006).

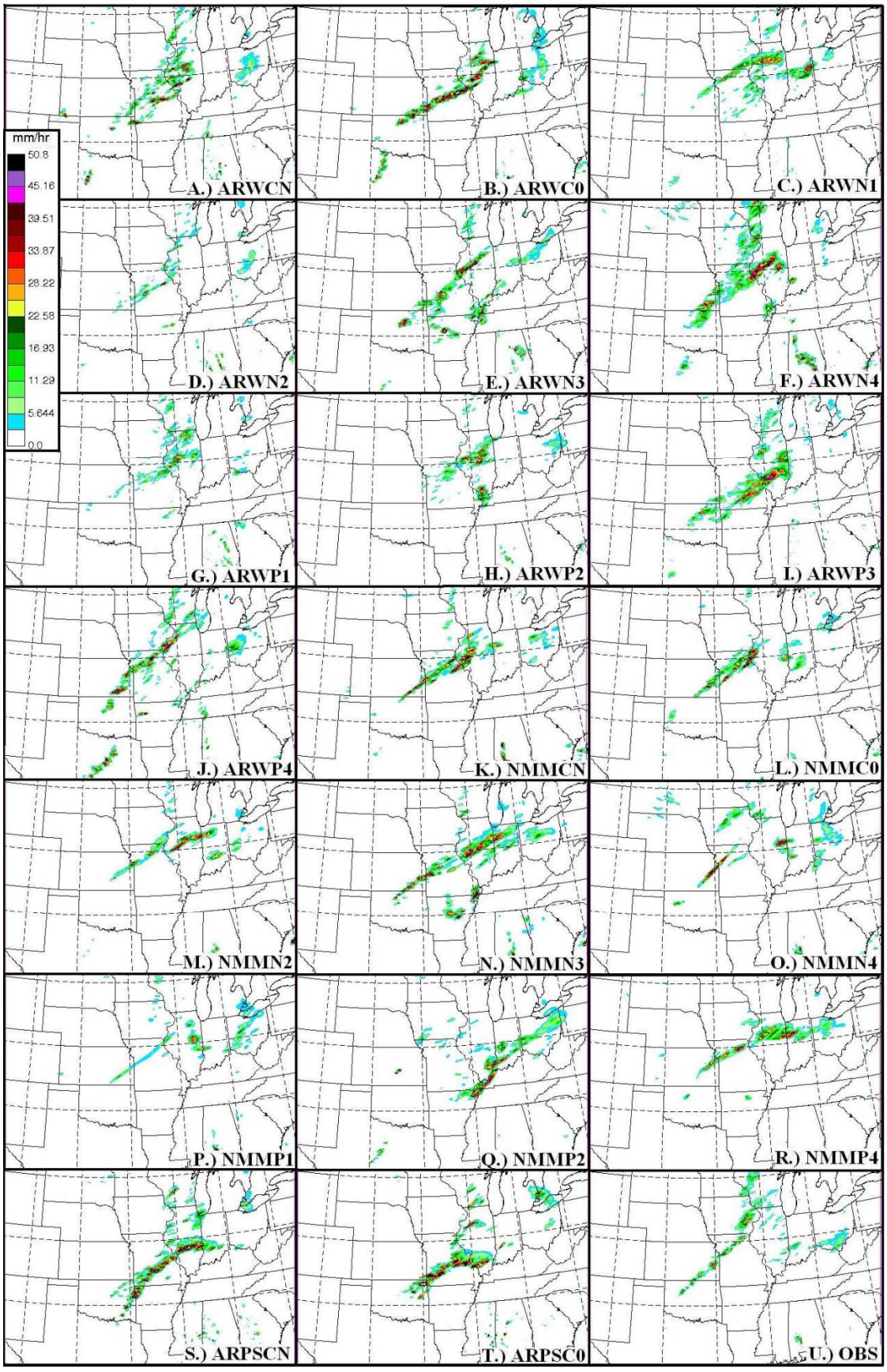
mm/hr
50.8
45.16
39.51
33.87
28.22
22.58
16.93
11.29
5.644
0.0
A.) ARWCN
B.) ARWC0
C.) ARWN1
D.) ARWN2
E.) ARWN3
F.) ARWN4
G.) ARWP1
H.) ARWP2
I.) ARWP3
J.) ARWP4
K.) NMMCN
L.) NMMC0
M.) NMMN2
N.) NMMN3
O.) NMMN4
P.) NMMP1
Q.) NMMP2
R.) NMMP4
S.) ARPSCN
T.) ARPSC0
U.) OBS

FIG. 3. 24 hour forecasts of 1 hour accumulated precipitation (mm) valid 00 UTC 14 May 2009 for ensemble members (a) ARWCN, (b) ARWC0, (c) ARWN1, (d) ARWN2, (e) ARWN3, (f) ARWN4, (g) ARWP1, (h) ARWP2, (i) ARWP3, (j) ARW P4, (k) NMMCN, (l) NMMCO, (m) NMMN2, (n) NMMN3, (o) NMMN4, (p) NMMP1, (q) NMMP2, (r) NMMP4, (s) ARPSCN, (t) ARPSC0 and (u) observations (OBS).

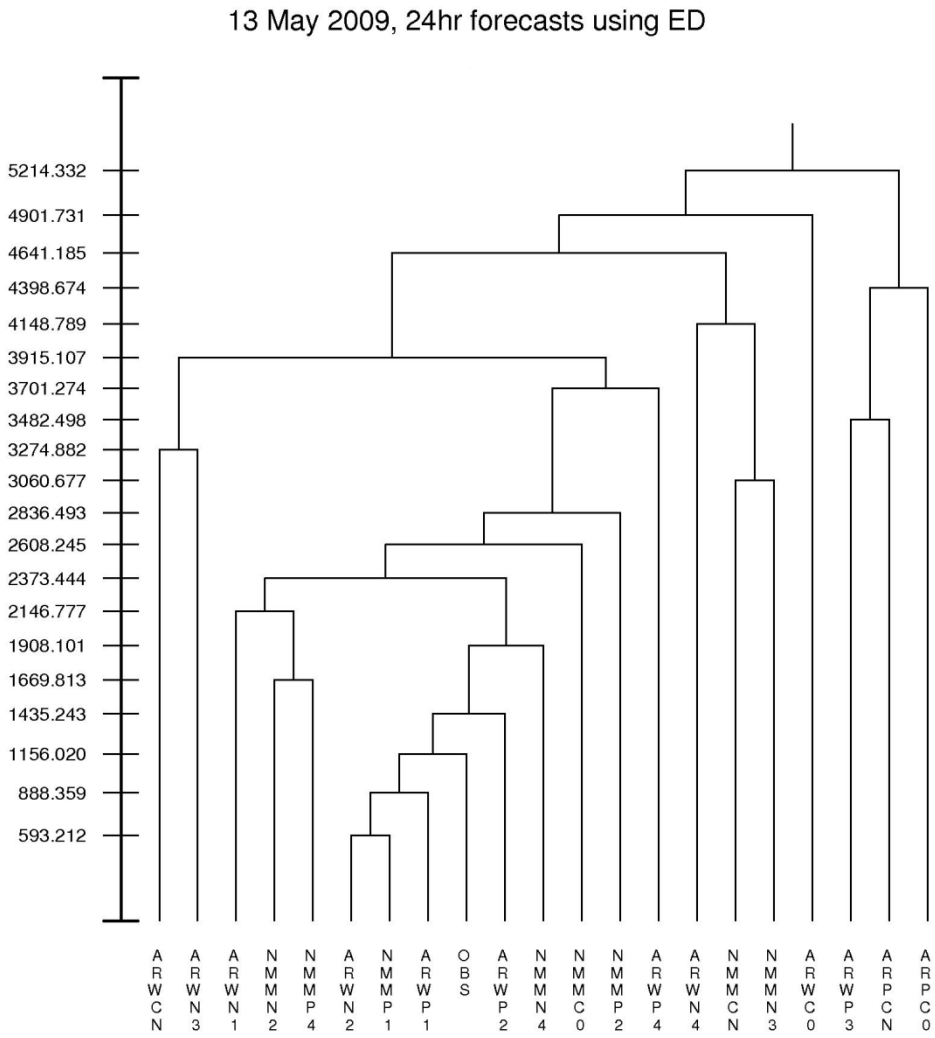

FIG. 4. Dendrogram of raw forecasts of 1 hour accumulated precipitation valid 00 UTC 14 May 2009, using ED as distance measure.

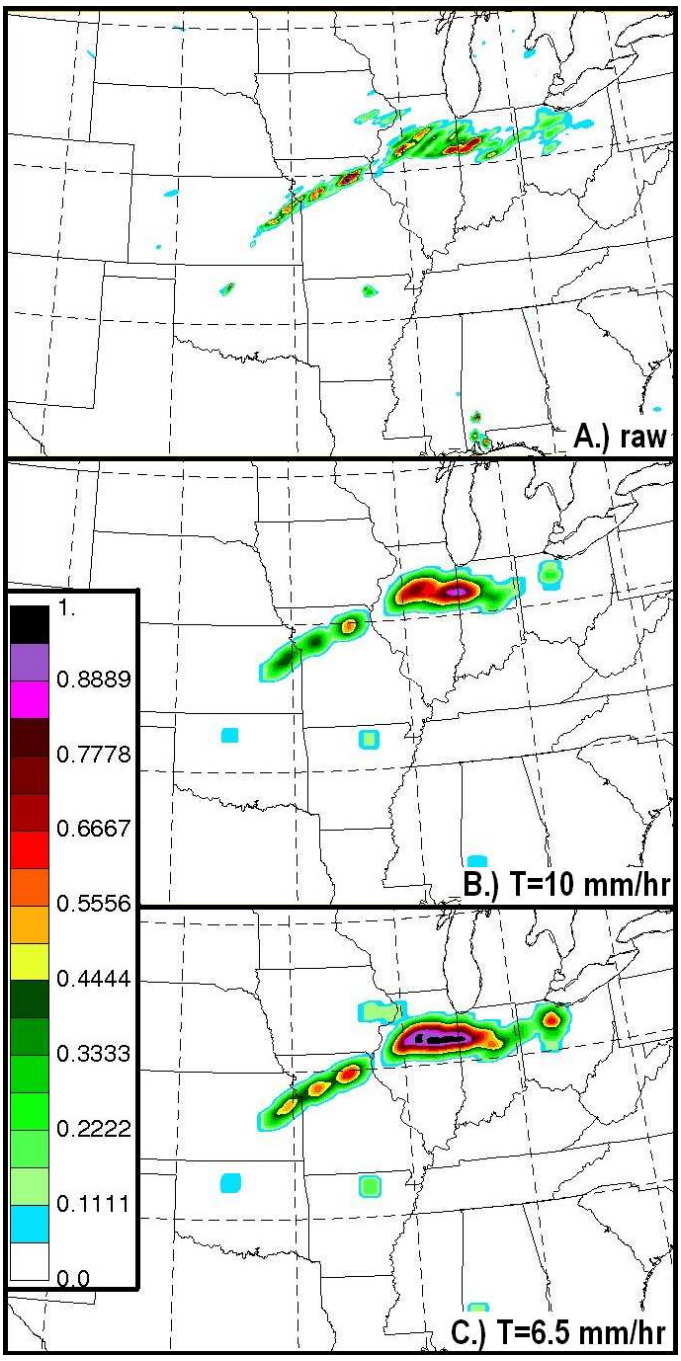

FIG. 5. Forecast from NMM P4 member, valid at 00 UTC 14 May 2009 showing (a) raw forecast with same color scale as Figure 3, (b) Neighborhood probability field with radius of 30 km and threshold of 10 mm, and (c) Neighborhood probability field with radius of 30 km and threshold of 6.5 mm.

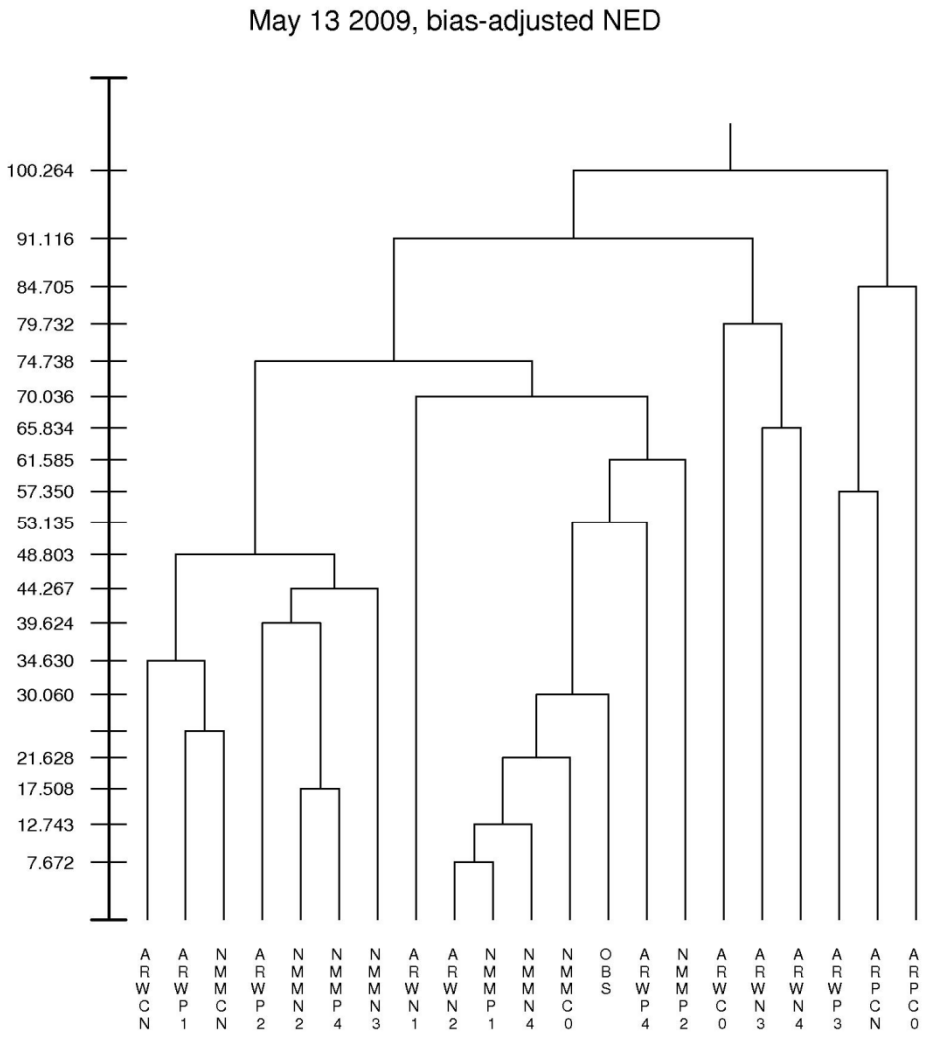

FIG. 6. Dendrogram of forecasts of 1 hour accumulated precipitation valid 00 UTC 14 May 2009, using bias-adjusted NED as distance measure.

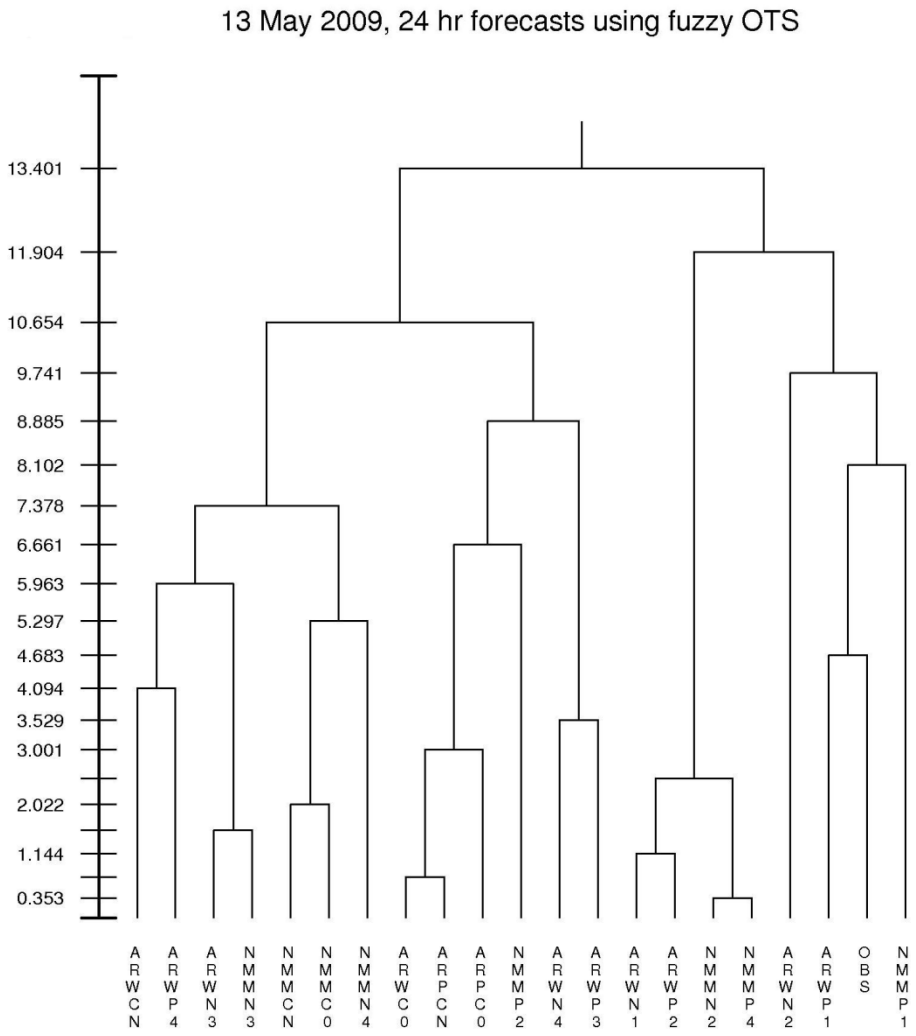

FIG. 7. Dendrogram of forecasts of 1 hour accumulated precipitation valid 00 UTC 14 May 2009, using bias-adjusted fuzzy OTS as distance measure.

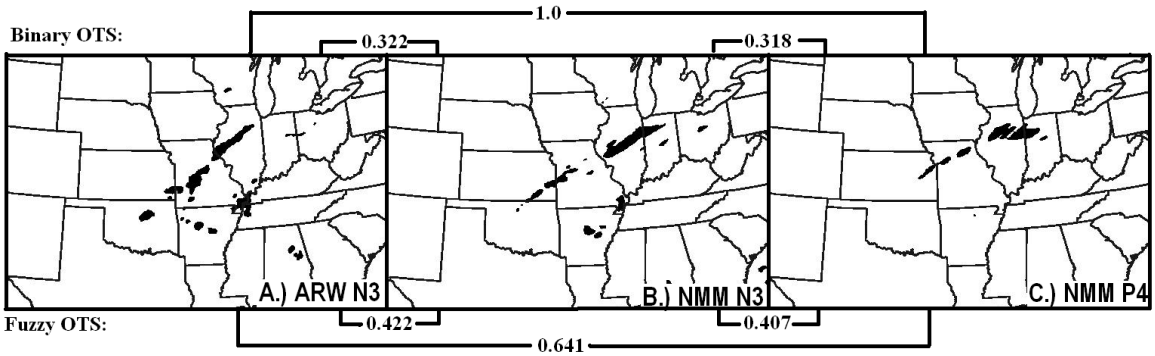

FIG. 8. MODE objects and OTS distances for 1 hour accumulated precipitation forecasts valid 00 UTC 14 May 2009 for (a) ARW N3, (b) NMM N3, and (c) NMM P4.

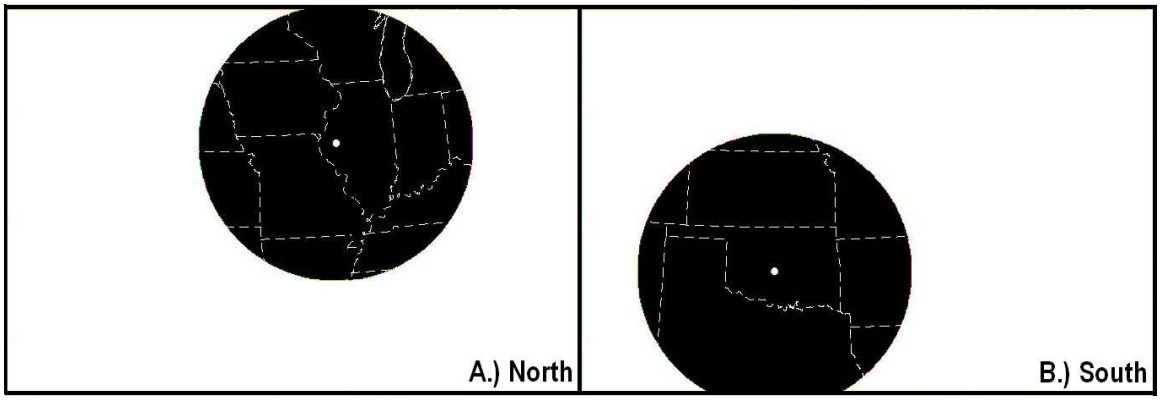

FIG. 9. Regions selected for clustering of forecasts valid 00 UTC 14 May 2009. Center of region is white dot and shaded area is the region within 600 km of the center. (a) north region and (b) south region.

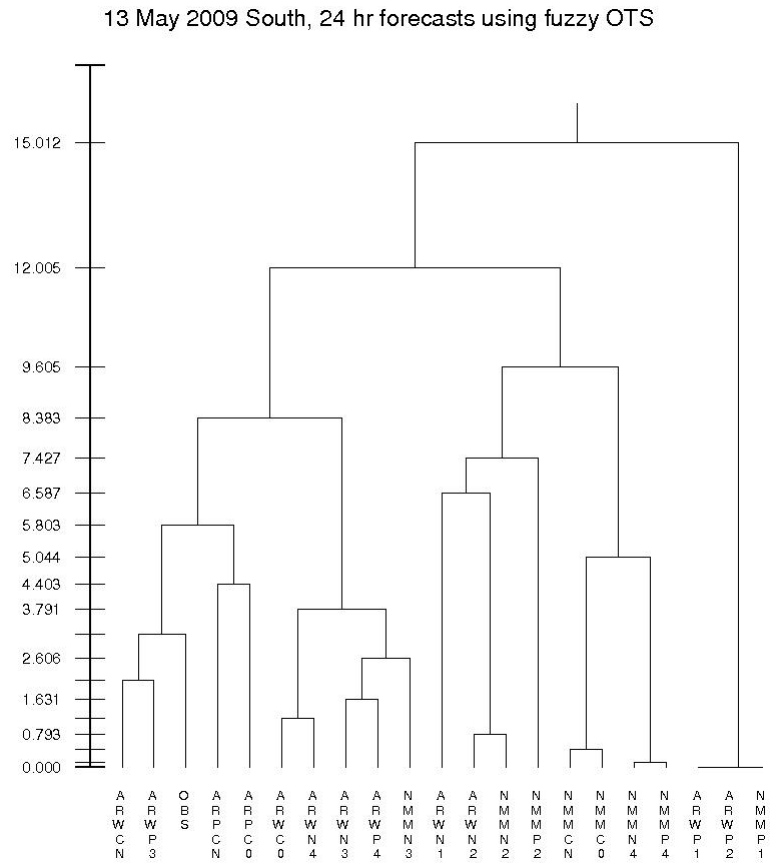

FIG. 10. Dendrogram of 1 hour accumulated precipitation forecasts valid 00 UTC 14 May 2009 using fuzzy OTS as distance measure and focusing on southern region.

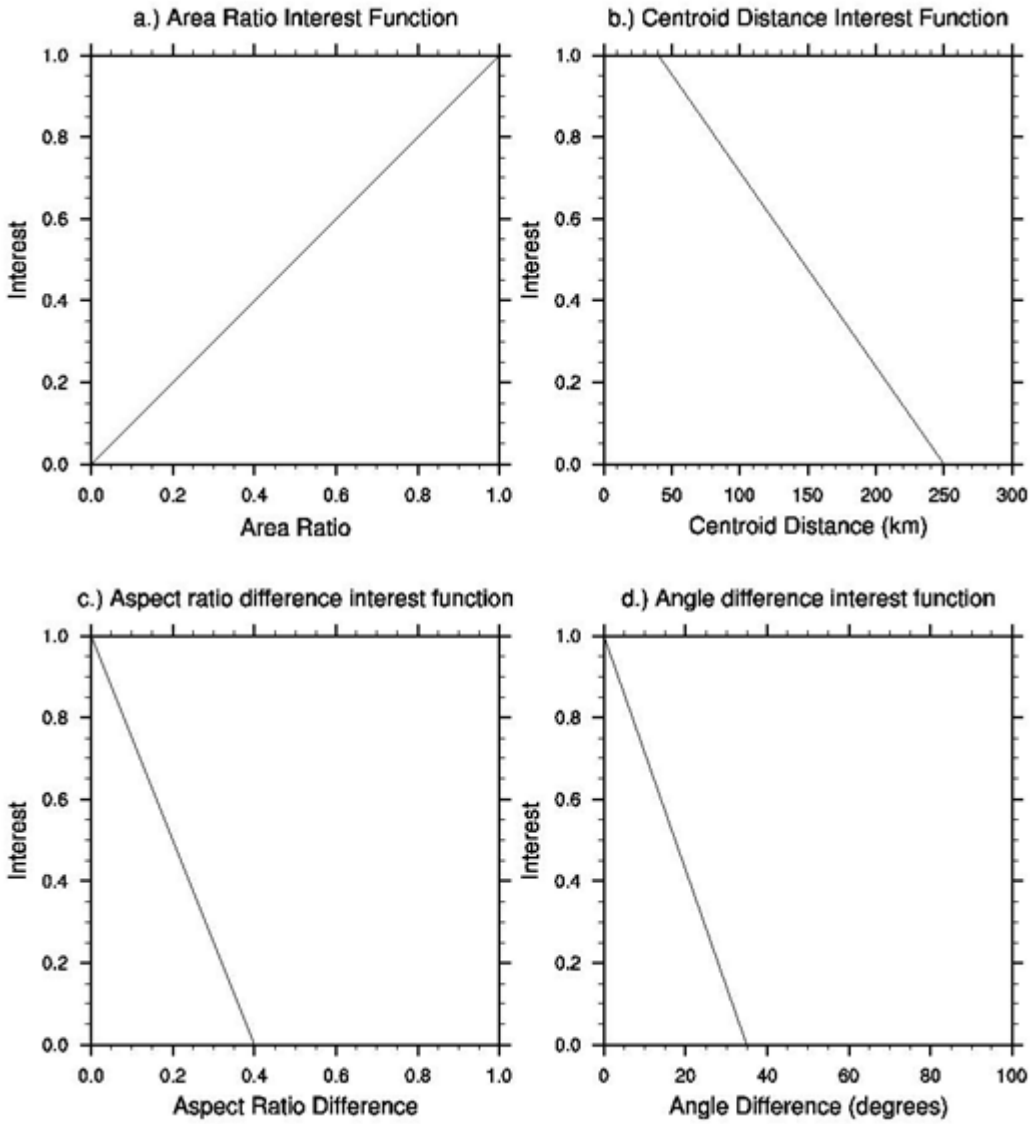

FIG. A1. Functions mapping attribute value to interest value for (a) area ratio, (b) centroid distance, (c) aspect ratio difference, and (d) angle difference.

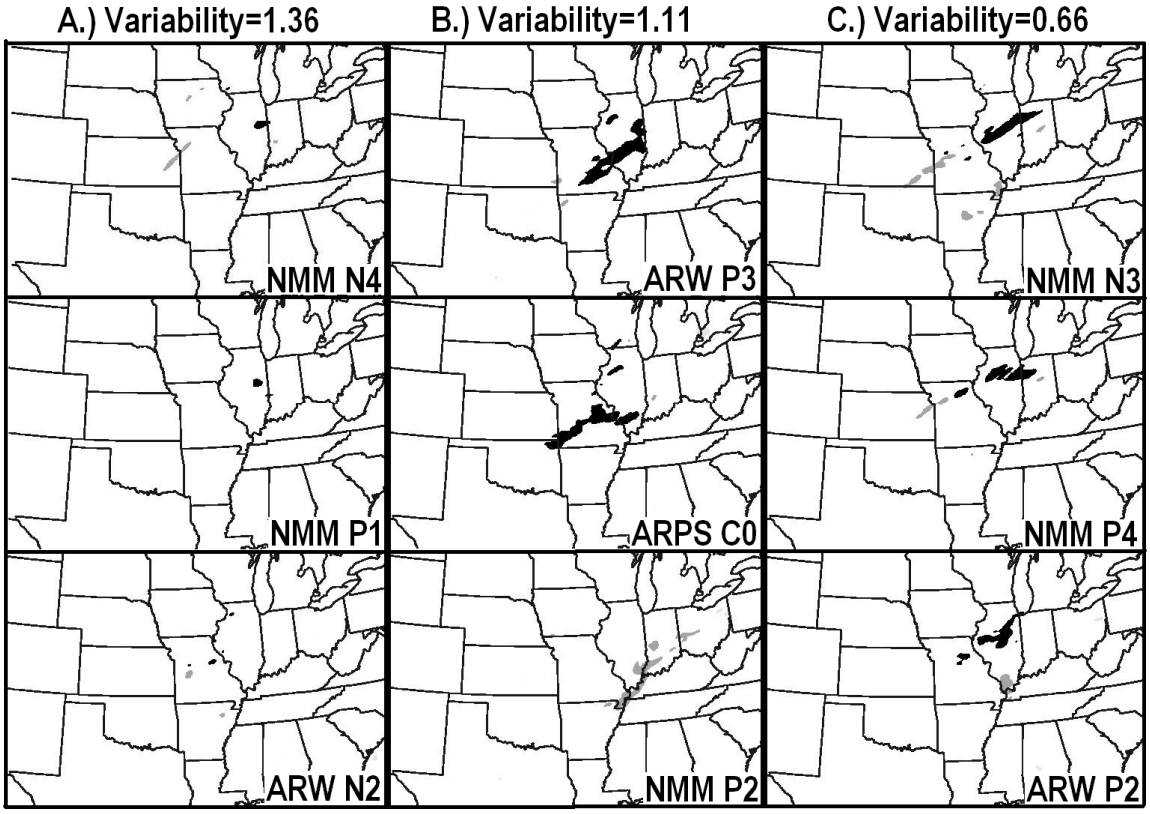

FIG. B1: MODE objects in forecasts valid at 00 UTC 14 May 2009 for (a) NMM N4, NMM P1 and ARW N2 (top to bottom), (b) ARW P3, ARPS C0, and NMM P2 (top to bottom), and (c) NMM N3, NMM P4, and ARW P2 (top to bottom). The variability of each column, defined by Eq. 1 with *dij=OTSij*, is given at the top of the column. Objects within 300 km of center of North Region (defined in fig 9a) are shaded black and objects centered between 300km and 600km of center of North Region, making a partial contribution to OTS, are shaded gray.