9A.3 TOWARD OPTIMAL CONVECTION-ALLOWING ENSEMBLE DESIGN: OBJECT-ORIENTED CLUSTER ANALYSIS, PROBABILISTIC VERIFICATION, AND CALIBRATION OF CAPS ENSEMBLE FOR 2009 HWT SPRING EXPERIMENT

Aaron Johnson*[1], Xuguang Wang[1,2], Fanyou Kong[2], and Ming Xue[1,2]

[1] University of Oklahoma, School of Meteorology, Norman, OK
[2] Center for Analysis and Prediction of Storms (CAPS), Norman, OK

1. INTRODUCTION

This study is a first step toward understanding the impacts and importance of the sources of uncertainties in model physics, model dynamics, and initial and lateral boundary conditions (IC/LBC) for convection-allowing ensemble forecasts. Some of the key issues for future study of optimal ensemble design and post-processing are inferred through a Hierarchical Cluster Analysis (HCA) of a 20-member convection allowing ensemble from the 2009 Hazardous Weather Testbed (HWT) Spring Experiment (Xue et al 2009). Non-precipitation variables (10m Wind Speed and 500hPa Temperature) are clustered using Ward's minimum variance algorithm (Ward 1963) and hourly accumulated precipitation is clustered using a new object-oriented form of Ward's algorithm.

Ward's algorithm is traditionally based on Euclidean distance which often does not agree with subjective evaluation of convection-allowing precipitation forecasts (Baldwin et al 2001). Precipitation forecasts are therefore evaluated using the object-oriented Method for Object-based Diagnostic Evaluation (MODE; Davis et al 2006). MODE is used to compute an Object-based Threat Score (OTS) that is defined, discussed, and compared to a Neighborhood-based Euclidean Distance (NED) in the present study. It is found that OTS is a more effective distance measure for the HCA than NED and that OTS is more effective when forecast objects have a fuzzy degree of similarity rather than a binary classification as matching or not matching. Therefore "Fuzzy" OTS is used to create dendrograms composited over multiple forecasts in order to better understand the systematic clustering of explicit forecasts of convection.

The design of the CAPS ensemble that is analyzed in this study is summarized in Table 1. The control members used initial conditions from the operational NCEP NAM analysis with additional radar observations, along with mesoscale wind and temperature observations, assimilated using ARPS 3DVAR and cloud analysis (Hu and Xue 2007). One member from each of the three models (ARW C0, NMM C0, and ARPS C0) used identical configuration to the control member with the same model (ARW CN, NMMCN and ARPS CN respectively) except without radar and mesoscale data assimilation. Initial condition perturbations were generated by taking the control analysis as a base state and adding only the perturbations from the NCEP SREF members indicated in Table 1. Perturbed LBCs were taken directly from the SREF member forecasts, while control member LBCs were taken directly from NCEP NAM forecasts. A more thorough description of the Spring Experiment and the CAPS ensemble design is found in Xue et al (2009).

The goal of this study is to infer the issues related to ensemble design that require further research in order to optimally design useful ensembles for the explicit prediction of convective-scale phenomena such as severe storms. It is found that object-oriented precipitation forecasts cluster primarily by model dynamics at all forecast times, with secondary sub-clusters corresponding to microphysics scheme at 3hr forecast time (valid 03UTC) and, for NMM members, according to Planetary Boundary Layer (PBL) scheme at 24hr forecast time (valid 00UTC).

Post-processing is needed to effectively communicate probabilistic information based on the large amounts of data generated by a convection-allowing ensemble. Initial probabilistic verification results are also presented to help understand the potential advantages of different types of post-processing.

* *Corresponding Author Address*: Aaron Johnson, School of Meteorology, University of Oklahoma, 120 David L. Boren Blvd., Suite 5900, Norman OK 73072; email: ajohns14@ou.edu

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

Table 1: Details of ensemble configuration, modified from Xue et al (2009), showing the IC/LBC source, whether radar data is assimilated (R), and which microphysics scheme (MP), planetary boundary layer scheme (PBL), shortwave radiation scheme (SW), and land surface model (LSM) was used with each member. Symbols identifying MP and PBL schemes in other figures are also included. Perturbations added to CN members and LBC conditions are from NCEP SREF (Du et al 2006).

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

Table 2: 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)

2. OBJECT-ORIENTED CLUSTER ANALYSIS

HCA iteratively merges *N* clusters of 1 forecast each into 1 cluster of *N* forecasts, where *N* is the number of forecasts being clustered. This study uses Ward's algorithm to determine which two clusters to merge next. Ward's algorithm merges the two clusters which result in the smallest increase of total within cluster Error Sum of Squares (ESS) (Ward 1963). Ward's algorithm is modified for use with convection-allowing precipitation forecasts by replacing squared Euclidean distance with an object-oriented measure of distance and replacing ESS with an object-oriented measure of variability as the objective function to be minimized at each step. HCA results are illustrated with dendrograms (Alhamed et al 2002) showing the entire sequence of cluster merging.

2.1 Object-Oriented Distance

In this study, the distance between forecasts is calculated using a new measure, OTS, which is based on total interest, I, between forecast objects. Total interest is a weighted sum of the interest values for each of *M* object attributes (Davis et al 2009):

$$I_j = \frac{\sum_{i=1}^{M} c_i * w_i * F_{ij}}{\sum_{i=1}^{M} c_i * w_i} \qquad (1)$$

Where *c* is the confidence in an attribute, *w* is the weight assigned to an attribute, and *F* is the interest value of the i^{th} attribute for the j^{th} pair of objects. The user of MODE must choose several parameters and those most relevant to the present study are illustrated in Table 2, and Figure 1. It should also be noted that a different convolved threshold is used for each ensemble member so the average total area of all objects forecast by a given member is within 5% of the average total area of observed objects. These thresholds are intended to minimize the impact of systematic forecast bias.

The attributes in Table 2 are selected to quantify location (centroid distance), organization (area), and structure (aspect ratio and orientation angle) of intense rainfall. Confidence for angle difference follows Davis et al (2009) to give less weight to angle difference of circular objects, while angle difference and aspect ratio confidence is 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 for objects of similar size in similar locations. This is 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 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.

Figure 1 maps differences in object attributes to a fuzzy interest value. Approximate, rather than precise, location is emphasized by assigning objects with up to 40 km centroid distance an interest value of 1.0. A linear form of all interest functions is chosen for simplicity in lieu of established guidelines otherwise. The x-intercepts in Figures 1c and 1d are selected to be consistent with subjective evaluations of how well the total interest described the degree of similarity over a large number of different object pairs.

OTS is then calculated between forecast *i* and forecast j as:

$$OTS_{ij} = \frac{1}{A_i + A_j}\left\{\left[\sum_{k=1}^{N_i} I_k a_k\right] + \left[\sum_{k=1}^{Nj} I_k a_k\right]\right\} \qquad (2)$$

Where *A* is the total area of all objects in the forecast, *N* is the number of objects in the forecast, *a* is the area of the k^{th} object in the forecast, and *I* is the fuzzy value of total interest between the k^{th} object and its corresponding object in the other forecast. The corresponding object is the object with highest total interest that doesn't already correspond to a different object with higher total interest. Thus each object in one forecast corresponds to exactly one object (at most) in another forecast and the correspondence is the same in reverse.

If *w* in eqn (2) were replaced with a binary 1 or 0 depending on whether its corresponding object has total interest above a specified matching threshold, then OTS becomes the Area Weighted Critical Success Index (AWCSI; Weiss et al 2009, also fraction of area in matched objects; Davis et al 2009). Some differences between (fuzzy) OTS and AWCSI (i.e. binary OTS) are discussed further below. When used as a distance measure OTS is first subtracted from 1.

2.2 Object-Oriented Variability

Ward's algorithm is modified by defining the distance between clusters as the increase in a new measure of cluster variability, rather than ESS, that would result from a merge of those clusters. Variability is defined as the average distance, *d*, between all pairs of forecasts in the cluster, multiplied by *N*-1 where *N* is the number of forecasts in the cluster:

$$\begin{aligned} variability &= (N-1) * \frac{2}{N*(N-1)} \sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} \\ &= \frac{2}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} \end{aligned} \qquad (3)$$

When squared Euclidean distance is the distance measure, *d*, variability is equal to ESS and the modified Ward's algorithm is equivalent to the traditional Ward's algorithm.

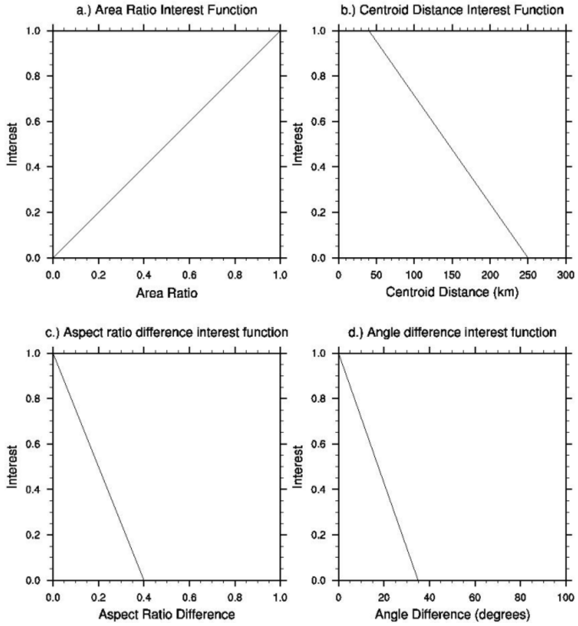

Figure 1: Functions mapping attribute value to interest value for (a) area ratio, (b) centroid distance, (c) aspect ratio difference, and (d) angle difference.

Variability, as defined here, 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 2009). For example, consider three clusters of three members in Figure 2 from a case study of forecasts valid 00 UTC 14 May 2009. 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 IL while the forecasts in MO range from a single linear object, to several small objects, to nothing at all. The cluster in column (b) has less spread because all the forecasts have a large rain area although they have large differences in placement. The cluster in column (c) has the least spread because they all forecast a large rain area in northern IL and have similar placement and structure of objects in MO. This subjective comparison is also reflected in the variability for columns (a), (b) and (c) of 1.36, 1.11, and 0.66 respectively. Most other cases that were subjectively examined exhibited the same correspondence between variability and subjective impressions of spread.

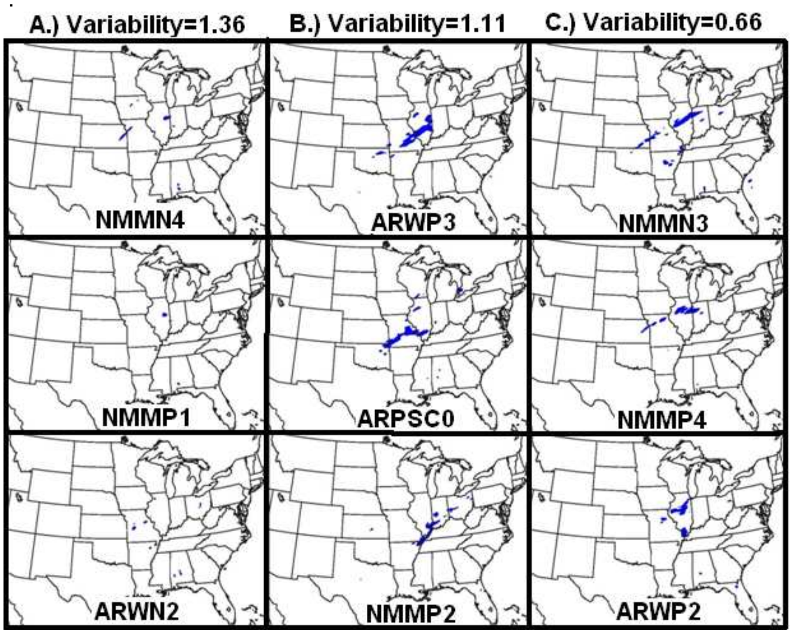

Figure 2: Object-oriented variability for clusters of forecasts valid 00 UTC 14 May 2009 including (a) members NMM N4, NMM P1, and ARW N2, (b) members ARW P3, ARPS C0, and NMM P2, and (c) NMM N3, NMM P4, and ARW P2.

2.3 OTS vs. NED

Clusters created using fuzzy OTS as the distance measure (i.e. 1-OTS) with the modified Ward's algorithm are compared to clusters created using Euclidean Distance of Neighborhood probability forecasts with traditional Ward's algorithm (NED). Neighborhood probability is defined as the percentage of grid points within a search radius that exceed a threshold of interest (Theis et al 2005). Clusters created using fuzzy OTS agree with subjective analysis more than clusters created using NED on several case study days (not shown) for two main reasons: OTS accounts for forecast features that are more closely related to convective mode and organization and OTS is not as sensitive as NED to overall forecast precipitation amount because OTS does not suffer from the "double penalty". Even though the neighborhood ED relaxes the strict spatial accuracy required of traditional ED, NED is still unable to properly account for similar forecast features at different locations. This is demonstrated with a brief case study of a severe weather event on 13 May 2009 (Figs. 3 through 5).

In terms of convective mode, organization, and coverage the NMM CN and NMM C0 forecasts subjectively appear more similar to each other than to the ARW P1 forecast. This is because NMM CN and NMM C0 both show a cluster of cells of intense precipitation near the MO/IL border, with a line of smaller and generally weaker cells extending southwestward to the OK/KS border. In contrast, ARW P1 shows just one strong cell in central IL with much weaker showers elsewhere (Fig. 3).

The NED dendrogram (Fig. 3) indicates that NMM CN is more similar to ARW P1 than to NMM C0. A relative lack of intense precipitation in ARW P1, combined with the largest maximum in ARW P1 being precisely co-located with a maximum in NMM CN, decreases NED compared to other members. At the same time, the NED from NMM CN to NMM C0 is penalized once because NMM C0 forecast maxima are at grid points without maxima in NMM CN and is penalized again because NMM C0 has no maxima at the grid points where NMM CN does have maxima. This is the essence of the double penalty.

The OTS dendrogram (Fig. 5) indicates NMM CN and NMM C0 forecasts are particularly similar relative to the other forecasts. This clustering is caused by the similarity of the main forecast features in terms of approximate location, total area, aspect ratio, and orientation angle. These attributes are also more likely to influence the subjective interpretation of severe weather forecasters interested in convective mode and organization than a Euclidean-based distance.

Another reason that OTS is preferred over NED as a distance measure for this cluster analysis is that NED is very sensitive to the overall precipitation amount. For example, Figure 3 indicates that ARW N2 and NMM P1 are the two most similar forecasts on this case. However, these two forecasts actually have different looking storms in completely different locations (Fig. 4). These members simply have in common an overall lower amount of precipitation than the other forecasts which results in a small Euclidean distance between them. This is also related to a reduced impact of the double penalty.

2.4 Binary vs. Fuzzy OTS

Fuzzy OTS has two main advantages over binary OTS, both of which result from the lack of a matching threshold in the fuzzy context.

The first advantage of fuzzy OTS, relative to binary OTS, is an increase in self-consistency of the distances among a large group of forecasts. Binary OTS does not change as large objects get incrementally less similar until the threshold is reached and a sudden large change in distance occurs. The result is that sometimes a large subjective difference between forecasts has little impact on binary OTS while other times a small subjective difference between forecasts has a very large impact on binary OTS. In contrast, fuzzy OTS changes continuously as forecasts get incrementally less similar.

The second advantage of fuzzy OTS is that it is conceptually more robust since it can discriminate matches that are very good from matches that are not as good. In contrast, binary OTS will give 2 forecasts (A and B) an equal distance to a third forecast (C) if the same objects in A and B match the same objects in C. This is true even if the objects in A are subjectively much more similar to the objects in C than are the objects in B. This limitation of binary OTS cannot be avoided by raising the matching threshold because then the limitation would be that all unmatched objects are treated equally.

3. CLUSTER ANALYSIS OF COMPOSITE DENDROGRAMS

Fuzzy OTS distance is used to examine systematic clustering of ensemble member forecasts of precipitation at forecast times of 3 and 24 hours, valid 03UTC and 00UTC respectively.

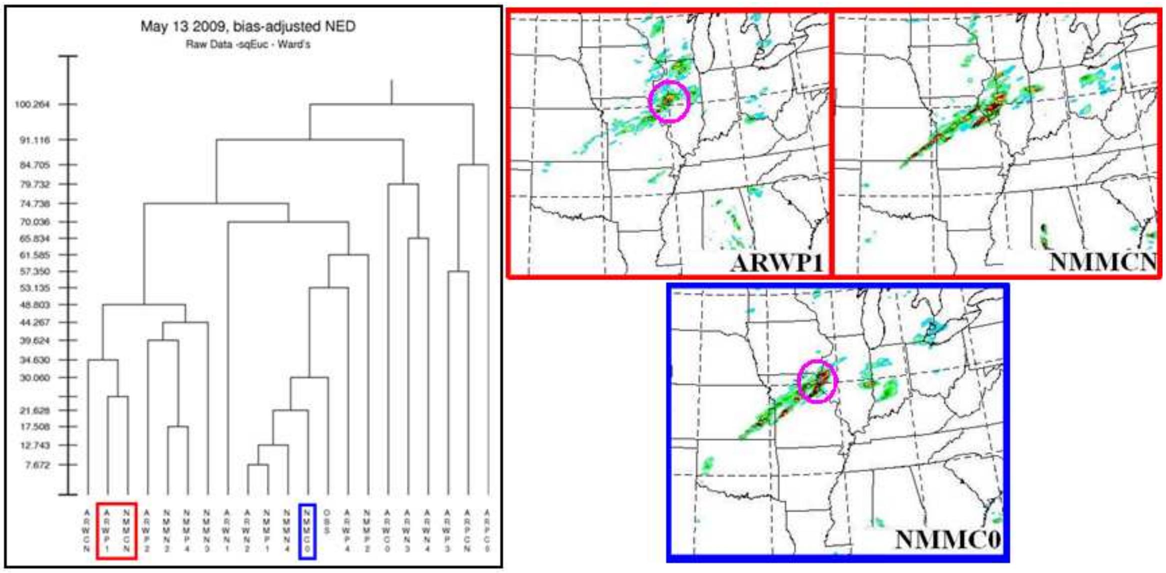

Figure 3: Dendrogram resulting from clustering the forecasts of hourly accumulated precipitation valid at 00UTC 14 May 2009, using NED as the distance measure. Also shown are the raw forecasts from NMM CN, NMM C0 and ARW P1 members.

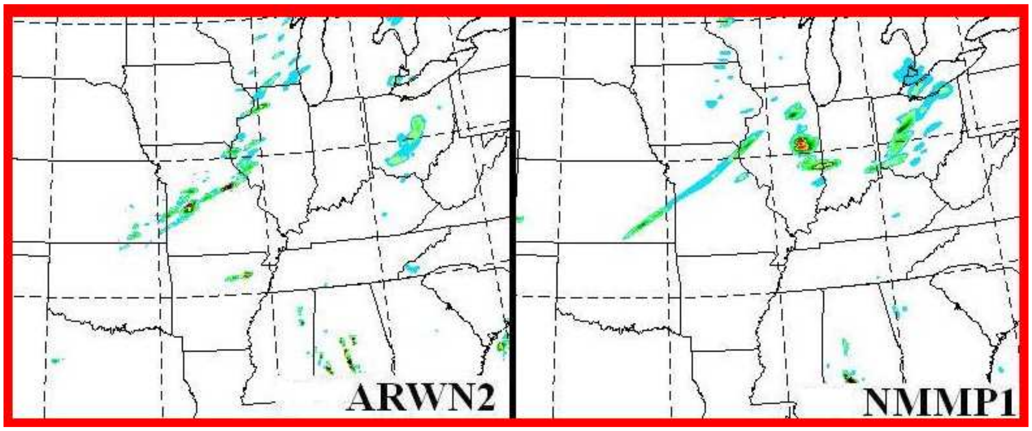

Figure 4: Raw forecasts valid 00UTC 14 May 2009 from ARW N2 and NMM P1 members for comparison to Figure 3.

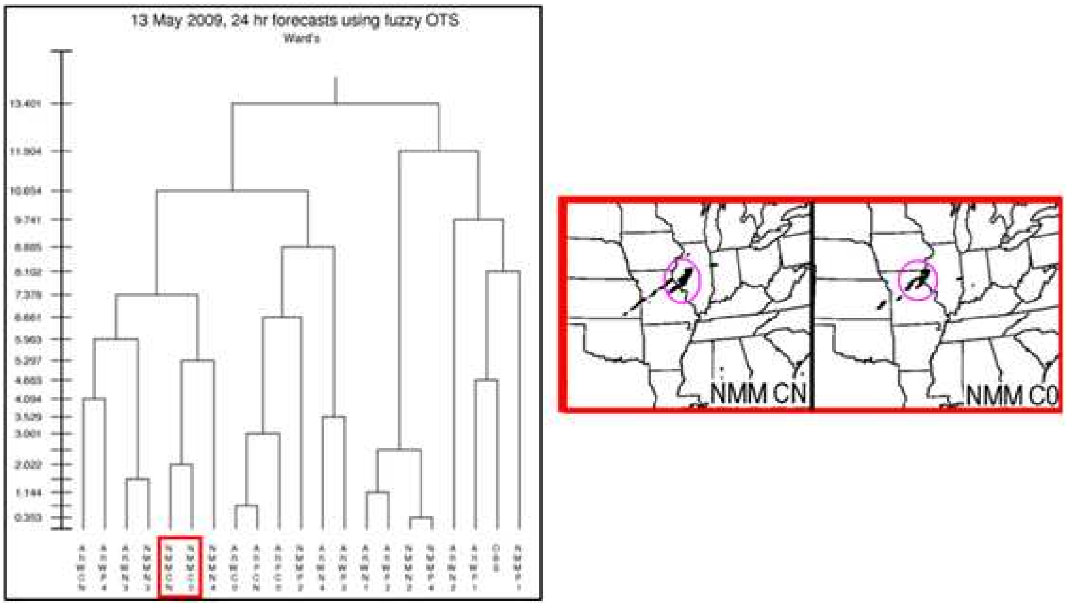

Figure 5: Dendrogram resulting from clustering the forecasts of hourly accumulated precipitation valid at 00UTC 14 May 2009, using fuzzy OTS as the distance measure. Also shown are the MODE objects from NMM CN and NMM C0 members

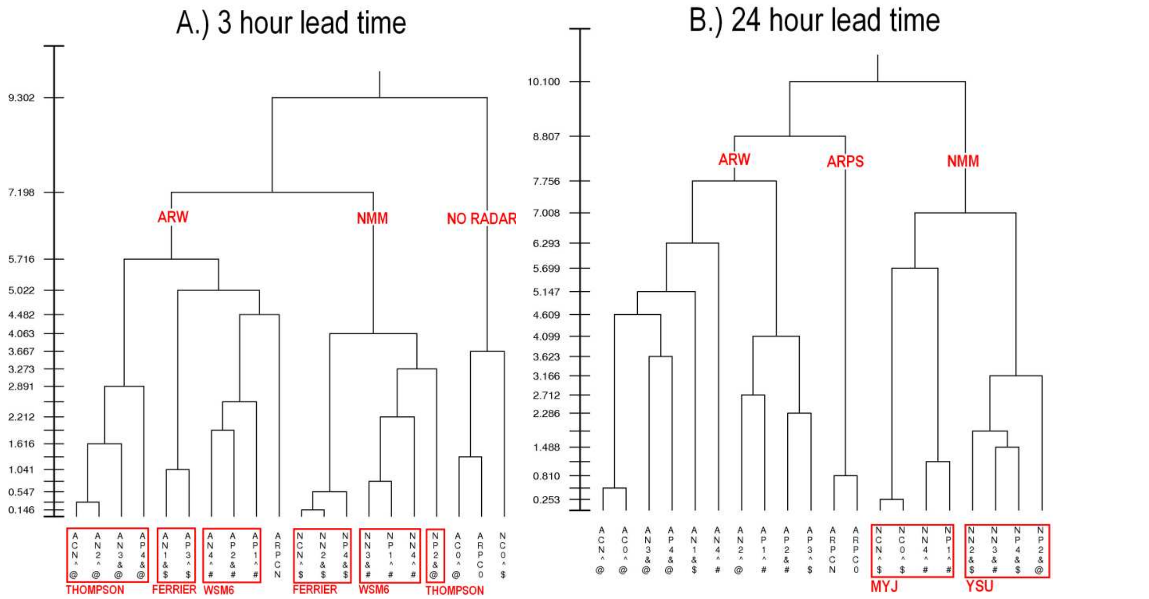

Figure 6: Dendrograms composited over 26 days using fuzzy OTS as distance measure between hourly accumulated precipitation forecasts at (a) 3 hour lead time valid 03UTC and (b) 24 hour lead time valid 00UTC. Labels are defined in Table 1.

The *systematic* clustering is examined by defining a composite distance between members as the average normalized distance between those members over all cases. The normalized distance is defined as the distance minus the largest distance between any pair in each case, divided by the range of distances on that case.

The composite dendrogram at 3hr lead time (valid 03UTC; Fig. 6a) shows that the primary distinction among members is based on the assimilation of radar and mesoscale data. The remaining members form two clusters according to WRF model dynamics while ARPS CN is included in the ARW cluster. These primary clusters of model dynamics contain sub-clusters that are entirely determined by the microphysics scheme for both ARW and NMM.

The composite dendrogram at 24hr lead time (valid 00UTC; Fig 6b) also contains three primary clusters of members with common model dynamics (ARW, NMM and ARPS). The NMM cluster has two sub-clusters, one containing all the NMM members with MYJ PBL scheme and another containing all the NMM members with YSU PBL scheme. Unlike the NMM cluster, the ARW cluster does not have sub-clusters with a common PBL scheme.

4. POST-PROCESSING METHODS

Two methods of post-processing to provide calibrated probabilities for hourly accumulation exceeding 2.54 mm/hr (.1 in/hr) are described in this section and compared in the following section.

4.1 NEIGHBORHOOD-BASED CALIBRATION

The first post-processing method obtains forecast probability from the percentage of grid points exceeding the accumulation threshold within a neighborhood (radius of 48km) of a point, averaged over all members. Over-forecasting is evident in such forecasts in a typical reliabilitiy diagram (Fig. 7a) showing the observed frequency corresponding to each forecast probability.

The neighborhood method is calibrated by excluding the day of the forecast from the reliability diagram, applying a Gaussian, and using the result to convert forecast probability to a calibrated probability (Fig 7b). No calibration is applied to very high forecast probabilities that rarely, if ever, occurred during the training period.

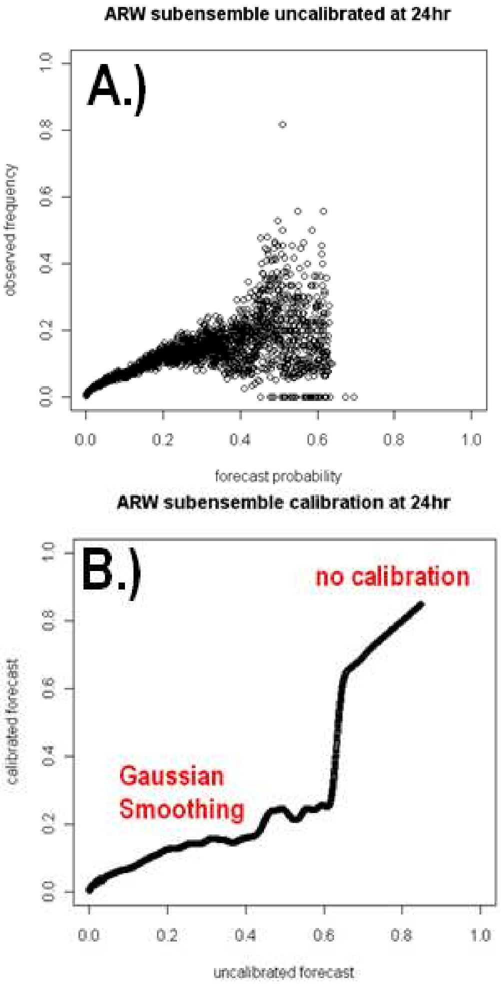

Figure 7: Reliability Diagram of a representative ensemble (a subensemble of 8 ARW members) (a) before and (b) after applying gaussian smoothing.

4.2 LOGISTIC REGRESSION

The second post-processing method applies logistic regression to the ensemble mean hourly accumulated precipitation. Following Hamill et al (2004), forecasts at each grid point from each day (except the day of the forecast) are used to fit values of β_0 and β_1 to the following equation:

$$P = 1 - \frac{1}{1 + \exp(\beta_0 + \beta_1 * x)}$$

where P is the probability of exceeding the threshold and x is the predictor variable (i.e., ensemble mean hourly accumulated precipitation). As suggested by Hamill et al (2004) we actually use x' = $x^{0.25}$ in this equation which we also find to slightly improve the performance. A typical curve of P vs. x' is shown in Fig. 8a and the same curve in terms of P vs. x is shown in Fig. 8b for direct comparison to Fig. 7. Additional predictors had little impact on our initial results so only the one predictor variable is used and discussed here.

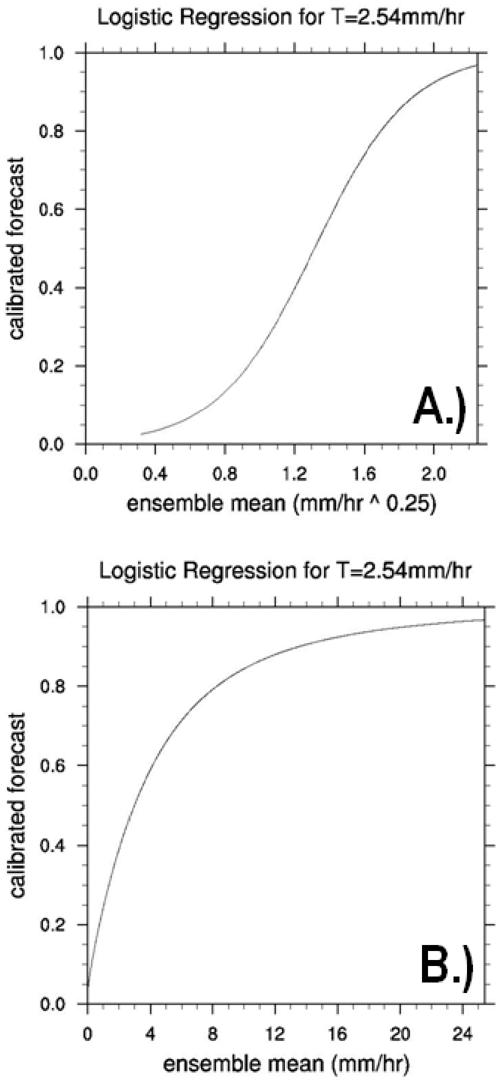

Figure 8: Representative example of fitted Logistic Regression Function at 24 hour lead time. Forecast probability is on vertical axis and horizontal axis is (a) ensemble mean accumulation raised to ¼ power and (b) ensemble mean accumulation.

5. PROBABILISTIC VERIFICATION

This section presents a verification of probabilistic forecasts for hourly accumulated precipitation exceeding 2.54 mm/hr (.1 in/hr). The first subsection uses 8 member sub-ensembles to compare the skill of these perturbations. The second subsection compares the two methods of post-processing, described in section 4. Both subsections use the Brier Score which allows the contributions to skill from reliability and resolution to be decomposed (Stephenson et al 2008, their eqn. 7). For calculation of skill score a reference forecast is defined separately for each grid point as the observed frequency of occurrence, over 3 spring seasons, on all days except the day of the forecast, at the same lead time.

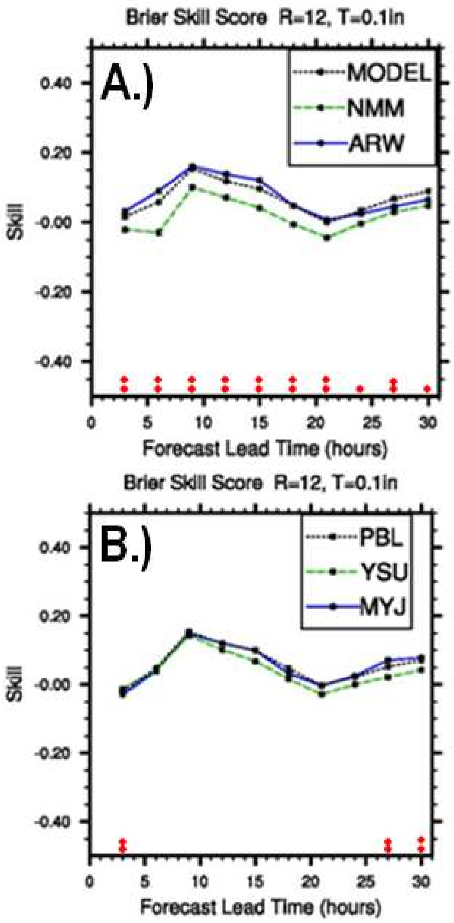

Figure 9: Brier Skill Score as a function of forecast lead time for uncalibrated neighborhood ensemble probability forecasts for (a) model-based sub-ensembles and (b) PBL scheme-based sub-ensembles.

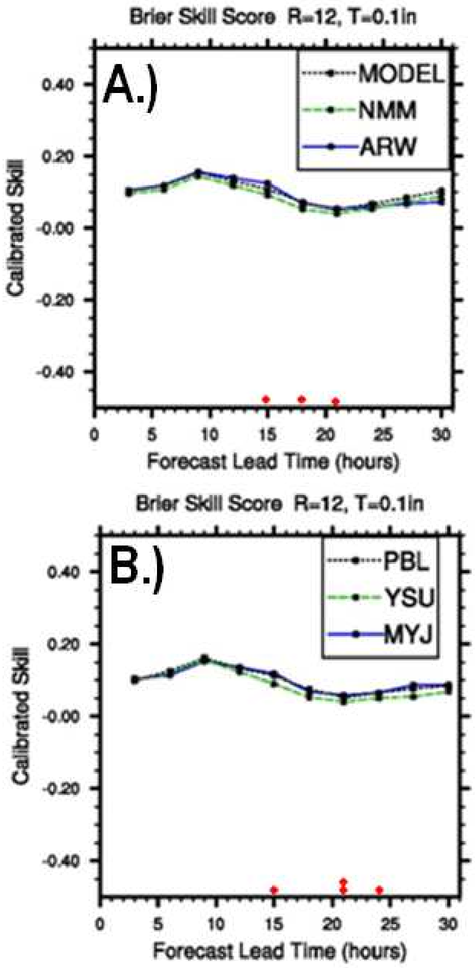

Figure 10: Brier Skill Score as a function of forecast lead time for calibrated neighborhood ensemble probability forecasts for (a) model-based sub-ensembles and (b) PBL scheme-based sub-ensembles.

5.1 IMPACT OF CLUSTERS ON SKILL

The cluster analysis showed model and PBL scheme perturbations to be most effective at creating forecast diversity so 8 member sub-ensembles are defined as follows. ARW, NMM, and MODEL sub-ensembles have 8 ARW members, 8 NMM members, and 4 of each chosen randomly each day, respectively. Likewise, MYJ, YSU and PBL sub-ensembles have 8 MYJ members, 8 YSU, and 4 of each chosen randomly each day, respectively.

The Brier Skill Score for each of these uncalibrated ensembles is shown in Fig. 9. Red dots indicate a statistically significant difference between ARW and NMM subensembles or YSU and MYJ subensembles at the 95% (two dots) and 90% (one dot) confidence level. Statistical Significance is assessed using the paired Wilcoxan Signed Rank Test on the daily Brier Scores (Hamill 1999). Fig. 9 shows that ARW tends to be significantly more skillful than NMM and MYJ tends to be more skillful than YSU (although the latter difference is generally not significant).

After calibration the skill of ARW and NMM sub-ensembles becomes nearly indistinguishable, statistically (Fig. 10). This is because the main difference between the uncalibrated sub-ensembles was due to much poorer reliability component for NMM than ARW (not shown) caused by larger positive bias in NMM members than ARW members. The calibration makes both sub-ensembles very reliable. In contrast, MYJ still tends to be more skillfull than YSU after the calibration although the difference is smaller than before calibration (Fig. 10). This is because the uncalibrated YSU sub-ensemble is worse than MYJ in terms of both reliability and resolution components of the brier score (not shown). Since calibration has negligible impact on the resolution component, the poorer resolution of YSU is reflected in the lower skill of the calibrated forecasts.

5.2 IMPACT OF CALIBRATION ON SKILL

In order to compare the two calibration methods, the Brier Skill Score and Brier Score components are examined for the full 20 member ensemble (Fig. 11), using the Neighborhood probability method, both before and after calibration, Logistic Regression probabilities, and the traditional percentage of ensemble members exceeding the threshold at a point.

Fig. 11 shows that skill scores are highest for the neighborhood-based calibration. The reliability component is better (i.e., lower) for both calibrated methods than the uncalibrated methods (Fig. 11b). The resolution component is best (i.e., highest) for the neighborhood-based methods, regardless of calibration (Fig. 11c).

In summary, both types of calibration result in increased skill due to better reliability while the neighborhood-based post-processing also shows a further improvement in resolution.

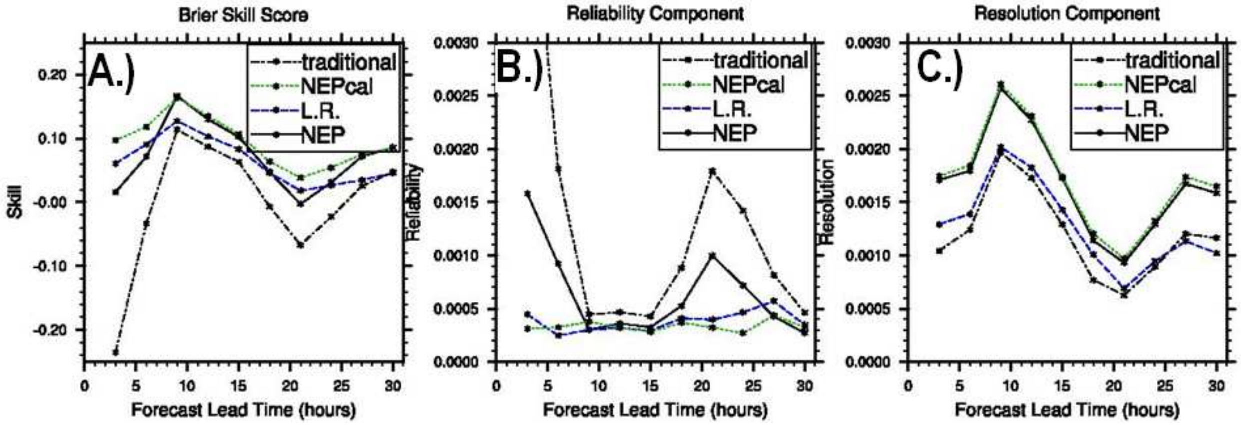

Figure 11: Verification of probabilistic forecasts using percentage of ensemble members exceeding .1in/hr at a point (traditional; dashed black), uncalibrated Neighborhood Ensemble Probability (NEP; solid black), Logistic Regression (LR; blue) and calibrated NEP (green) using (a) Brier Skill Score, (b) Reliability component of Brier Score and (c) Resolution component of Brier Score.

6. DISCUSSION OF ENSEMBLE DESIGN AND VERIFICATION

This study demonstrates an object-oriented distance measure can be applied to Hierarchical Cluster Analysis to obtain subjectively more reasonable clusters than traditional methods. A subsequent cluster analysis revealed that model and PBL scheme perturbations are particularly important to emphasize when designing an ensemble forecast system because of their large impact on clustering.

Probabilistic forecast verification revealed an ensemble based on NMM model is less reliable than an ensemble based on ARW model but the difference is reduced through a simple calibration procedure. Calibration using Neighborhood Method and Logistic Regression both improved reliability while only the Neighborhood Method was also able to improve resolution.

In order to determine if one convection-allowing ensemble design is *better* than another, appropriate methods of verification are needed. There are two concerns that cast doubt on the appropriateness of verification metrics such as those in the previous section for convection-allowing forecasts. First, point-wise metrics were shown to be inadequate for cluster analysis so they also may not be ideal for verification purposes. Second, the storm scale details that are one of the justifications of the expense of using a convection-allowing resolution are often lost in the process of creating probabilistic forecasts.

Figure 12 shows the energy decomposition of deterministic forecasts, traditional ensemble probability, and observations using a Haar Wavelet Filter (see Casati and Wilson 2007). While deterministic forecasts from individual members have somewhat realistic energy spectra the probabilistic forecasts (all methods considered in this study show a similar spectrum) have most of the energy at much larger scales than what is observed (Fig. 12). Since increased realism is one of the benefits of convection allowing ensembles more work is clearly needed to determine the best ways to obtain both skillfull and realistic calibrated probabilities.

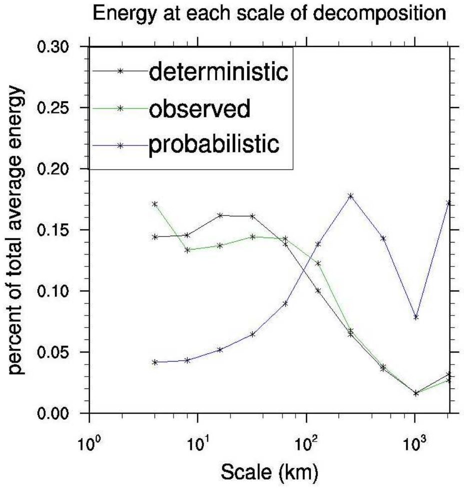

Figure 12: Haar Wavelet decomposition of energy by spatial scale averaged over 2009 season for deterministic forecasts of all ensemble members (black line), observed precipitation (green line) and traditional ensemble probability forecasts (blue line)

A potential alternative verification method was also examined using the MODE algorithm, and is now discussed. For this method the ARW CN member is arbitrarily selected as the “best guess” or control forecast. The remaining members are used to assign a probability (of being matched by an observed object) to each object in the control forecast according to the percentage of members that have a matching forecast object. The verification “event” is the occurrence of an observed object matching each control forecast object. Unlike all previous methods in this study, 6-hourly accumulations are used here as a demonstration of the method.

The reliability diagram and ROC curve corresponding to this verification method are shown in Fig. 13. A calibration was also approximated from Fig. 13 as a constant factor of 0.7 multiplied by the uncalibrated forecast. Verification statistics are shown in Table 3. Even before calibration there is positive skill that is improved even further by calibration which shows that the ensemble can skillfully forecast the probability of an object begin “matched” by the observations.

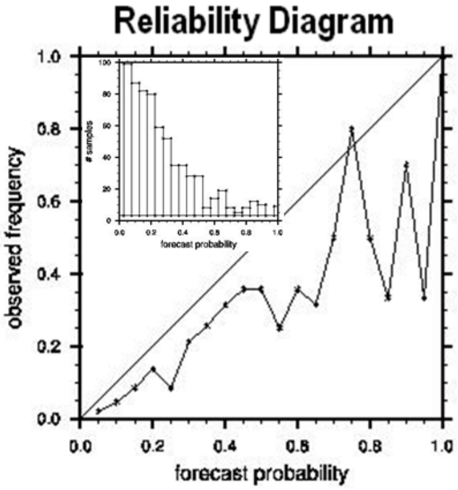

Figure 13: Reliability Diagram for probabilistic forecasts of an observed object “matching” control member forecast objects.

	BSS	Rel.	Res.	samples	ROCarea
Uncal.	.077	.02	.03	681	.786
Calibr.	.176	.005	.03	681	.786

Table 3: Brier Skill Score, Reliability Component of Brier Score, Resolution Component of Brier Score, number of objects (samples) forecast by control member, and ROC area for both calibrated and uncalibrated probabilistic forecasts of an observed object matching control member forecast objects

It is important to note that a different interest map than that used for the cluster analysis was found to subjectively give the most reasonable results. However, the statistics in Table 3 are very sensitive to the choice of an interest map used to determine if objects are “matching” or not. Even if a suitable interest map is identified, a “matching” object is difficult to translate into an operationally useful forecast, although forecasters may still find such information useful.

In order to allow for verification that is consistent with commonly used methods we also tried using the above approach only to assign probabilities but a point-wise verification. Here the verification event is an observed 6.5mm accumulation (Table 4). The advantage of this approach is that it allows storm-scale detail to be retained while providing probabilistic forecasts for the same event as more traditional methods. The disadvantages include the lack of resolution with this method and the fact that, like the previous method, scenarios only forecast by members other than the control member are not reflected in the control forecast.

	BSS	Rel.	Res.	Unc.
6.5mm/hr	-.086	.0008	.00003	.0081
6.5mm/6hr	-.216	.0146	.00265	.04779

Table 4: Brier Skill Score and Brier Score components for hourly and 6hourly probabilistic forecasts of exceeding 6.5mm accumulation at a grid point, using objects to determine probabilities.

These alternative verification methods are not proposed as replacements for traditional methods at convection-allowing resolution. Instead, they are meant to illustrate the challenges when trying to objectively evaluate such forecasts.

ACKNOWLEDGEMENTS

The authors are grateful to CAPS and NSSL for providing the ensemble forecast and QPE verification data that made this study possible. The authors also wish to thank Nusrat Yussouf and Dave Stensrud from NSSL for providing tested and documented software for performing hierarchical cluster analyses and NCAR for making MODE source code available. The real-time forecasts were produced at the Pittsburgh Supercomputing Center (PSC), and at the National Institute of Computational Science (NICS) at the University of Tennessee. 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). This research was partially 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.

5. REFERENCES

Alhamed, A., S. Lakshmivarahan, D. J. Stensrud, 2002: Cluster Analysis of Multimodel Ensemble Data from SAMEX. *Monthly Weather Review.* **130.** Pp 226-256.

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.

Casati, B., L. J. Wilson, 2007: A New Spatial-Scale Decomposition of the Brier Score: Application to the Verification of Lightning Probability Forecasts. *Mon. Wea. Rev.*, **135**, 3052–3069.

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.

Ebert, Elizabeth E., 2009: Neighborhood Verification: A Strategy for Rewarding Close Forecasts. *Wea. Forecasting*, **24**, 1498–1510

Hamill, Thomas M., 1999: Hypothesis Tests for Evaluating Numerical Precipitation Forecasts. *Wea. Forecasting*, **14**, 155–167.

Hamill, Thomas M., Jeffrey S. Whitaker, Xue Wei, 2004: Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447.

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.

Stephenson, D. B., C. A. S. Coelho, I. T. Jolliffe, 2008: Two Extra Components in the Brier Score Decomposition. *Wea. Forecasting*, **23**, 752–757.

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

Ward J. 1963: Hierarchical Grouping to minimize an objective function. *Journal of the American Statistical Association.* Vol. **58**, 236-244

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., 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.,* Omaha, Nebraska, June 1-5, 2009