1.1 EVALUATING METHODS FOR DOWN-SELECTING NWP MULTIPHYSICS ENSEMBLES FOR WIND PREDICTION

Jared A. Lee[1,2,*], Sue Ellen Haupt[1,2], George S. Young[2],
Walter C. Kolczynski[2,3], and Tyler C. McCandless[2]

[1]Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO
[2]Department of Meteorology, The Pennsylvania State University, University Park, PA
[3](Current affiliation: Department of Meteorology, Naval Postgraduate School, Monterey, CA)

1. INTRODUCTION

Ensembles of numerical weather prediction (NWP) models are used to predict the range of possible future atmospheric states, and the corresponding forecast uncertainty. How best to configure NWP ensembles is an area of active research in the community. First, there are many practical issues that need to be considered: How many members should be in the ensemble? What NWP model(s) should be used? What horizontal and vertical resolution should be employed? What area should the domain(s) cover? What should be the forecast duration? Sacrifices often have to be made in one or more of those considerations because of computational limitations, whether in research or operations.

When configuring or evaluating an ensemble system, effort should be made to ensure that the ensemble forecasts are calibrated. If an ensemble is perfectly calibrated, then the ensemble variance and the ensemble-mean error variance will match (Grimit and Mass 2007; Kolczynski et al. 2009, 2011). Even when attempting to account for various sources of error, however, most ensembles are still under-dispersive and thus require calibration (Raftery et al. 2005). This is true even for very large ensembles (Kolczynski et al. 2011). NWP ensembles must therefore be "dressed" with statistical estimates of the true error distribution via post-processing (Roulston and Smith 2003).

When configuring an NWP ensemble, there are also many possible approaches for representing forecast uncertainty and variability. While initial condition (IC) and lateral boundary condition (LBC) uncertainty certainly play a large role in NWP forecast errors, model error is an important source of error in NWP ensembles, particularly for short-range forecasts (e.g., Stensrud et al. 2000; Fujita et al. 2007; Clark et al. 2008). There are two main types of model error: one stems from lack of knowledge about the processes that are being modeled, and the other stems from uncertainty in the values of the model parameters themselves. Thus there are several approaches to representing model error, including multimodel, multiphysics, and stochastic perturbation ensembles, or combinations thereof (e.g., Eckel and Mass 2005; Hacker et al. 2011).

When constructing a multiphysics ensemble, it is not clear *a priori* what sets of physics schemes are the best to choose. For instance, in the Weather Research and Forecasting (WRF) NWP model, there are hundreds of possible combinations of physics schemes to choose from. Previous research

*Corresponding author address: Jared A. Lee, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307, USA. Email: jaredlee@ucar.edu.

proposes an objective method to choose, or "down-select," a subset of 14 ensemble members that represent the forecast probability density function (pdf) nearly as well as the full ensemble of 24 WRF members (Lee et al. 2012). The post-processing method used in that study is principal component analysis (PCA). PCA is not the only possible technique for down-selection, however. In this study we use a larger WRF multiphysics ensemble of 42 members that has variability in more types of physics schemes than Lee et al. (2012), and we also compare PCA with another candidate down-selection method, k-means clustering. In both Lee et al. (2012) and this study, Bayesian model averaging (BMA; Raftery et al. 2003, 2005) is used to calibrate the forecasts for the full ensemble and to dress the down-selected subset ensembles.

We discuss our ensemble configuration and verification procedures in section 2. In section 3 we demonstrate our down-selection techniques, PCA and k-means clustering. We describe calibration with BMA in section 4. We present verification results in section 5. Section 6 summarizes the study and suggests extensions.

2. DATA

2.1 Ensemble configuration

We choose to configure a multiphysics ensemble that uses the same ICs/LBCs to isolate the effects model error for two reasons. First, it is not clear how any down-selection approach would be physically meaningful if it were applied to an ensemble with only equally likely IC/LBC random perturbations, because members would then be exchangeable and statistically indistinguishable (Fraley et al. 2010). A second reason for this choice is that one of the goals of this study is to define a small set of physics members that could be used in a later ensemble that would also have IC/LBC variability included.

Our 42-member physics ensemble is created with version 3.3 of the Advanced Research WRF (WRF-ARW) NWP model (Skamarock et al. 2008). At least three different options are used for each type of physics scheme in the ensemble, as detailed in Table 1. Skamarock et al. (2008) contains details and references for all the parameterization schemes we use. We use a slightly modified version of the Mellor-Yamada-Janjic (MYJ) ABL scheme, as in Lee et al. (2012).

The coarse domain uses a horizontal grid spacing of 36 km, while the one-way nested fine domain uses 12-km grid spacing. The geographic area spanned by the domains can be seen in Fig. 1. In this study we use time steps of 90 s and 30 s. The vertical resolution is identical to that of Lee et al. (2012). There are 45 full vertical levels in each simulation, with high vertical resolution in the lowest 2 km so that we can resolve processes in the atmospheric boundary layer (ABL) well.

We initialize the 48-h forecasts every fifth day at 0000 UTC starting on 1 Dec 2009, and continuing through Feb 2010, for a total of 18 forecast periods during this winter evaluation period. Table 2 lists all the initialization dates for these forecasts. No data assimilation is used in this study because we desire to simulate a forecasting system.

The LBCs for all 42 members in this study come from the 0.5˚x0.5˚-resolution Global Forecast System (GFS) forecast cycles initialized at each of the simulation times. We use sea surface temperature (SST) analyses from the National Centers for Environmental Prediction (NCEP) real-time global 0.083˚ dataset. We use daily snow analyses from the National Environmental

Satellite, Data, and Information Service (NESDIS).

The ICs use the 0-h GFS forecast and are blended with standard WMO observations to produce a more accurate initial state. We use the Obsgrid objective analysis software to perform this blending. Obsgrid is part of the WRF modeling system and developed by the National Center for Atmospheric Research (NCAR), and uses multiple passes of the objective analysis scheme to modify the first-guess field (NCAR Mesoscale & Miscroscale Meteorology (MMM) Division 2011, chap. 7). In Obsgrid we use the Cressman objective analysis scheme, assigning each observation a distance-weighted flow-dependent radius of influence (Cressman 1959).

2.2 Verification and quality control

All post-processing, verification, and analysis is only performed on the inner 12-km domain.

We use standard WMO surface and upper-air observations to verify our WRF ensemble forecasts. We quality control these observations against the GFS analysis fields that are interpolated by the WRF Pre-processing System (WPS), using Obsgrid as described above and in Lee et al. (2012).

Of the 18 forecasts we create for the winter evaluation period, the first 12 are used as training data for both the down-selection process and for calibration. The remaining six forecasts are set aside for verification (Table 2). We recognize that training on only 12 forecast periods is less than ideal, as Raftery et al. (2005) use 25 forecast periods to train their Bayesian model averaging (BMA) calibration technique. In the future we plan to have a longer WRF ensemble dataset available to use.

The observations used in this study are temperature and wind at the surface and the mandatory levels of 925 hPa, 850 hPa, and 700 hPa. Model predictions are horizontally and vertically interpolated to the observation locations. We perform verification on wind direction, wind speed, vector wind difference, and the zonal (u) and meridional (v) components of the wind. We also only examine the forecasts at four lead times: 12 h, 24 h, 36 h, and 48 h. We do this because these are the only times for which standard radiosonde observations are available.

We use root-mean squared error (RMSE) as our deterministic verification metric and continuous ranked probability score (CRPS) as our probabilistic verification metric. RMSE assesses accuracy and is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(f_i - o_i\right)^2} \tag{1}$$

where o is the value of the observation i, f is the forecast value at the time and location of observation i, and N is the total number of observations. The CRPS assesses both accuracy and sharpness and is defined as (Wilks 2006):

$$CRPS = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\left(p_i^f(x) - p_i^o(x)\right)^2$$
$$p_i^o(x) = \begin{cases} 0 & x < o_i \\ 1 & x \geq o_i \end{cases} \tag{2}$$

where $p_i^f(x)$ is the forecast cumulative probability of the forecast variable being $\leq x$ at the space-time location of observation i, and all other variables are as before. Both RMSE and CRPS are negatively oriented, with zero representing a perfect score for both.

3. ENSEMBLE DOWN-SELECTION TECHNIQUES

The goal of our ensemble down-selection techniques is to retain the subset of ensemble members that span the uncertainty

space of the forecast but to eliminate those that are most redundant. We do this because ensembles are most useful when members each sample a different portion of the atmospheric probability density function (pdf). In this study we compare two techniques for down-selection, principal component analysis and k-means cluster analysis.

3.1 Principal component analysis

PCA and its application as a down-selection technique is discussed more completely in Lee et al. (2012), but we briefly describe it here for clarity. PCA is a mathematical technique that reduces the dimensionality of a dataset from K variables to N variables (N < K) (Jolliffe 2002; Wilks 2006; Witten and Frank 2005). The N new variables are the principal components (PCs), and are mutually orthogonal. Additionally, each PC is a linear combination of all K old variables; in this study, the 42 ensemble members are themselves the K variables. PCs are also ordered so that each PC (PC_1, PC_2, PC_3, etc.) accounts for the greatest amount of remaining variance in the original dataset, subject to orthogonality with all previously defined PCs. Therefore, PCA allows for a given amount of variance to be explained by the minimum number of vectors.

For this study we use MATLAB to perform PCA on the forecast errors during the 12-forecast training period (Table 2). We perform PCA separately for each forecast lead time (12 h, 24 h, 36 h, 48 h), and separately for temperature errors and vector wind difference (VWD). For both temperature errors and VWD we combine all levels (surface, 925 hPa, 850 hPa, 700 hPa) into a single bin at each lead time.

For each lead time/variable combination we tally each member that is the top contributor to each PC that individually explains at least 2.0% of the error variance of the ensemble. Thus we define two candidate down-selection subsets from PCA: subset P07W, a seven-member subset determined by PCA on wind vector errors; and subset P07T, a seven-member subset determined by PCA on temperature errors. Lists of members that are included in the down-selection subsets appear in Table 3.

The members in subset P07W have diversity in most types of physics schemes, except for the land surface scheme, where six of the seven members selected use the Noah land surface scheme. Six of the seven members in subset P07T also use the Noah land surface scheme, and all seven members use the Kain-Fritsch cumulus scheme. This seems to imply that the Noah land surface scheme contributes substantially to the variance of the temperature errors and VWD, and that the Kain-Fritsch cumulus scheme also contributes substantially to the variance of temperature errors.

It should also be noted that all 42 members contributed nearly equally to the first PC for VWD. Moreover, the first PC for VWD at all four lead times explained between 87-89% of the total variance (for temperature errors, the first PC contains 74-81% of the variance). Therefore for VWD, it does not appear that any one member contains the bulk of the error variance. This may indicate a weakness of PCA for identifying sources of variability in this ensemble, or may indicate that there is very little variability in VWD across the entire ensemble over the 12-forecast training period.

3.2 K-means cluster analysis

Cluster analysis is a post-processing technique by which similar data are grouped together. The grouped data are called

clusters. Several studies use various cluster analysis techniques to group similar members in an NWP ensemble forecast (e.g., Legg et al. 2002; Alhamed et al. 2002; Yussouf et al. 2004; Johnson et al. 2011). There are several types of cluster analysis, including *K*-means clustering. *K*-means cluster analysis is a non-hierarchical clustering technique, which means that data can be reassigned to different clusters on successive iterations through the algorithm. *K*-means is also the most commonly used non-hierarchical clustering method (Wilks 2006).

We perform clustering separately at each of the four lead times for both temperature errors and VWD at all four levels during the 12-forecast training period (Table 2), as we do for PCA. The MATLAB algorithm used for performing *K*-means clustering is as follows:

1) Specify the number *K* clusters that are in the data; therefore, a range of values for *K* must often be tested.
2) Split the *N* data vectors (temperature errors or VWD for each ensemble member at all *O* observation locations) into an initial guess at membership in *K* clusters. In our study we determine these by randomly selecting *K* data vectors to serve as seeds, or initial cluster centroid positions.
3) Calculate the absolute distances between each data vector and cluster centroid, and assign all data vectors at once to the nearest cluster.
4) Re-calculate all cluster centroids.
5) Repeat steps 3 and 4 until convergence is achieved.
6) Complete a pass through the data vectors, and individually reassign a data vector to a different cluster if doing so results in a smaller sum of distances.
7) Re-calculate cluster centroids after each reassignment.
8) Repeat steps 6 and 7 until there are no more reassignments.

Each replicate of this algorithm will yield a local minimum, but is not guaranteed to find a global minimum of the total sum of data-centroid distance in all *O* dimensions. Therefore a large number of replicates are generally necessary to increase the probability that the global minimum will be found. In this study we use 5,000 replicates of the clustering algorithm, and report results from the case with the smallest total sum of data-centroid distances.

In this study $O > 3000$ for all four lead time-variable combinations. While the clusters cannot be visualized in the full 3000+ dimensions, they can be visualized in two dimensions, by comparing the errors for two of the observations, and the components of the centroid positions in those two dimensions. One such visualization, for VWD clusters at a lead time of 48 h, is shown in Fig. 2. While the ensemble members (closed and open circles) may not be closer to other centroids (rotated and upright crosses) in these two dimensions, one should keep in mind that the clusters are determined according to minimizing data-centroid distance in all *O* dimensions, not just these two dimensions.

While clusters may technically be populated by only one data vector, for this study we require that a cluster must contain at least two ensemble members to be considered valid. For both variables we test a range of numbers of clusters, to find the largest number of clusters that does not have any single-member clusters across all four lead times. In this way we determine there to be ten clusters for VWD and eight clusters for temperature errors. When an individual

member is assigned to different clusters at different lead times, we consider it to belong to the cluster to which it was assigned for the majority of the lead times. Tables 4 and 5 contain lists of the ensemble members that belong to the various clusters for VWD and temperature errors, respectively.

One of the strengths of *K*-means cluster analysis is that the clusters are generally amenable to straightforward physical interpretation, as members within each cluster share certain characteristics. For both sets of clusters the common thread among all members within each cluster is a common land surface scheme, ABL scheme, or both. Additionally, some of the clusters for temperature errors and VWD are identical. In some clusters the members also share the same microphysics, radiation, or cumulus schemes, but that is not true universally. These observations indicate that the choice of land surface and ABL scheme have the greatest influence on predictions of low-level temperature and wind. This implies that an ensemble should at least contain diversity in land surface and ABL schemes.

To configure candidate down-selection subset ensembles from these clusters, we randomly choose one member from each cluster. In future work we will compare this random selection with other selection methods, such as choosing the ensemble member that is closest to the cluster centroid. The VWD *K*-means subset is subset K10W, and the temperature errors *K*-means subset is subset K08T. The members that comprise both subsets K10W and K10T are listed in Table 3.

4. ENSEMBLE CALIBRATION

We use Bayesian model averaging (BMA; Raftery et al. 2003) to "dress" the full ensemble and all down-selected ensembles to better approximate the pdf of the forecast distribution (Roulston and Smith 2003). BMA estimates the weights and parameters for each ensemble member, and then during a training period (12 forecast periods in this study), these weights and parameters are trained to best match the observations. The BMA weights and standard deviations are then applied to forecasts in a verification period to create a better ensemble forecast pdf.

We perform BMA on the temperature and on the zonal (u) and meridional (v) wind component forecasts at each forecast lead time (12, 24, 36, and 48 h) and for each level (surface, 925 hPa, 850 hPa, and 700 hPa). As in Lee et al. (2012) and Raftery et al. (2005) we assume a normal distribution for the temperature. Lee et al. (2012) assume a normal distribution for both wind components separately, but here we assume a bivariate normal distribution for the wind components, and perform BMA on the u and v together at each level and lead time. As in Lee et al. (2012) we also perform a single domain-wide bias correction and calibration for each variable at each lead time and level.

A benefit of using BMA to dress the pdf is that the relative values of the weights provide a rough indication of the relative importance of the various schemes. The optimal BMA weights for 2-m temperature forecasts are displayed as a donut chart in Figure 3 as an example. Interestingly, in contrast to the finding in Lee et al. (2012) over a summer training period that the members with the Noah land surface scheme had the highest BMA weights for 2-m temperature forecasts, in this study over a winter training period the members with Noah all have the smallest BMA weights, while the members with the thermal diffusion land surface scheme have the largest weights. This seems to imply, at least in a relative sense, that the Noah land

surface scheme performs more poorly in winter than in summer among the selected schemes, with the converse true for the thermal diffusion scheme. There is also a discernible trend in weights for 2-m temperature according to the choice of ABL scheme, but less pronounced than the trend involving land surface schemes. Members with the YSU boundary layer scheme had higher weights than members with the ACM2 scheme. Members with the MYNN-2.5 scheme had the lowest weights, except when paired with the RUC land surface scheme; among those members, the members with the MYJ ABL scheme had the lowest weights. There are no discernible trends in the weights with other physics schemes.

The BMA weights for 10-m wind also exhibit a signal among the various ABL and land surface schemes, with the highest weights occuring for the members using the MYNN-2.5 ABL scheme, followed by ACM2, as shown in Figure 4. The land surface schemes yield a secondary signal, with RUC having larger weights, and Noah having smaller weights.

At 925, 850, and 700 hPa, for both temperature and wind components, the weights tend to become more similar to each other, although there are still some weak differences in weights between members with different land surface and ABL schemes. Above the surface, the RUC land surface and YSU ABL schemes tend to yield the highest weights, both for temperature and wind. The stronger signals apparent at the surface as well as the weaker signals aloft indicate that, consistent with Lee et al. (2012), varying the choice of ABL scheme and land surface scheme appears to add to variability in the ensemble for low-level wind and temperature forecasts.

5. RESULTS

We calculate RMSE and CRPS for both the full 42-member ensemble and each of the four ensemble subsets defined above, K10W, K08T, P07W, and P07T (Table 3), over the six-forecast verification period. First we examine the RMSE and CRPS for the wind and temperature for the impact of calibration. For the wind components, there is little difference in RMSE between the equal-weighted and BMA-weighted ensembles, at both the surface (Fig. 5) and 850 hPa (Fig. 6). For the v-wind components, the BMA-weighted ensembles generally have a slightly lower (i.e., better) RMSE than the equal-weighted ensembles, but there is no clear tendency for the u-wind components between the BMA-weighted and equal-weighted ensembles. Calibrating the ensemble via BMA substantially improves the CRPS for both wind components, however, both at the surface (Fig. 7) and 850 hPa (Fig. 8), for all ensembles and lead times. These findings are consistent with Lee et al. (2012), where calibration also did not substantially improve RMSE values, but did substantially improve CRPS values. This suggests that for the multi-physics ensembles in this study and Lee et al. (2012), BMA primarily improves the sharpness of the ensemble pdf, rather than the accuracy. Even though accuracy does not appear to be improved much, the improvement in sharpness still illustrates the benefit of calibrating an ensemble forecast.

The story is similar for RMSE and CRPS for 850-hPa temperature (Figs. 10 and 12). The RMSE and CRPS values for 2-m temperature (Figs. 9 and 11) are much larger than those for the 850-hPa temperature, however, and much larger than those for 2-m temperature over summer 2009 in Lee et al. (2012) as well. We speculate that this is due to poor modeling of the surface layer and land

surface in winter, or due to errors in the snow cover analysis. For instance, if snow cover is analyzed to be present in the model but there is no snow cover actually on the ground (or vice versa), there will be a large RMSE for 2-m temperature for those locations. We have not proven that this is the cause of the large errors in 2-m temperature because it is outside the scope of this study, however. Additionally, calibration substantially improves the performance of the ensembles here, using both RMSE and CRPS. We suspect that the large RMSE for 2-m temperature is what allowed for calibration to have a larger, more noticeable positive impact.

As for comparison of the candidate ensemble subsets, the RMSE and CRPS values show the same trends generally (i.e., when one subset has a higher RMSE it also has a higher CRPS), so we focus our analysis on the CRPS plots. At the surface, the CRPS for both *K*-means subsets (K10W and K08T) is approximately the same as the CRPS for the full ensemble. This is true for both 10-m wind components (Fig. 7) and 2-m temperature (Fig. 11). At the surface, both *K*-means subsets perform better than either PCA subset (P07W and P07T). Therefore, for surface variables, it appears that using *K*-means cluster analysis as an ensemble down-selection technique yields a forecast pdf nearly equivalent to the full 42-member ensemble, while subsets from PCA perform more poorly.

Above the surface for both wind components and temperature, subsets K10W and K08T still perform nearly as well as the full ensemble, while subsets P07W and P07T have only slightly higher CRPS values than the full ensemble at 925 hPa (not shown), and at 850 hPa the CRPS for all the subsets are approximately the same as the full ensemble (Figs. 8 and 12). The reason for this improvement in performance of the PCA subsets above the surface is difficult to discern.

Both *K*-means and PCA down-selection methods perform well in this study, but the improved performance at the surface gives *K*-means an advantage. It is also clearer conceptually that selecting one member from each of several clusters results in better sampling of the ensemble pdf than PCA, especially when the PCA technique results (in this study) in selecting members that, for the most part, had less diversity in physics schemes than did *K*-means clustering. The ensemble subsets chosen by PCA in Lee et al. (2012) have greater physics diversity than does either PCA subset here, and the reason for this difference is an area of continuing research. It does indicate that PCA may be a less reliable method for determining useful down-selected ensemble subsets when compared to *K*-means cluster analysis, however.

6. SUMMARY

We have created a 42-member WRF multi-physics ensemble for December 2009-February 2010 over the northeastern third of the U.S. We examine and compare two statistical post-processing techniques, principal component analysis and *K*-means cluster analysis, to use for down-selecting subsets of members from the full ensemble. We also calibrate the full ensemble and subset ensembles with Bayesian model averaging to dress the ensemble pdf. We train these post-processing techniques over twelve consecutive forecasts. We then verify the wind component and temperature forecasts at four different levels over six consecutive forecasts using root-mean squared error and continuous ranked probability score as metrics.

Ensemble calibration is extremely beneficial. Ensemble calibration with BMA does not result in values of RMSE that differ substantially from the equal-weighted ensembles, except for 2-m temperature, where BMA substantially improved the RMSE, although that RMSE is quite large to begin with. The calibrated full ensemble and subset ensembles have values of CRPS about 15-20% lower (better) than the equal-weighted ensembles, however. Therefore, for this ensemble, the primary benefit of calibration appears to be improved sharpness of the ensemble pdf, though for 2-m temperature, calibration substantially improved the accuracy as well.

The CRPS values for both down-selection subsets determined by *K*-means cluster analysis (K10W and K10T) are similar to the full 42-member ensemble for both wind components and temperature, at the surface, 925 hPa, 850 hPa, and 700 hPa. Thus, choosing one member from each cluster appears to represent the ensemble pdf well, with little or no degradation in forecast quality. The CRPS values for both down-selection subsets determined by PCA (P07W and P07T) are somewhat degraded for both wind components and temperature at the surface, but differ little from the full ensemble or other subsets above the surface. The reason for this difference in behavior is unclear.

The subsets determined by *K*-means clustering have a great deal of diversity in several types of physics schemes. This physics diversity arises because, in this study, members that cluster together share a common boundary layer scheme, land surface scheme, or both. Therefore, the members that are chosen as representative of each cluster use several boundary layer-land surface combinations, which appears to account for most of the variability in low-level wind and temperature forecasts. The PCA subsets, on the other hand, tend to have much less diversity in physics. This is one possible explanation for the PCA subsets performing more poorly at the surface. While both down-selection techniques perform fairly well, the *K*-means subsets are more clearly sampling from different parts of the ensemble pdf.

Because of these factors, it appears at the current time that *K*-means clustering is a more promising and consistent down-selection technique. More testing needs to be done in other seasons, however. Testing of the length of the training period required for robust results also needs to be performed. The sensitivity of the down-selection results to incorporating each above-surface level also ought to be studied.

ACKNOWLEDGMENTS

We gratefully acknowledge grants provided by both the Extreme Science and Engineering Discovery Environment (XSEDE) and the Computational Information Systems Laboratory (CISL) at the National Center for Atmospheric Research (NCAR) for allowing the computation and storage of the WRF ensemble data. XSEDE and NCAR are both supported by the National Science Foundation (NSF). We thank Aijun Deng of Penn State for providing the NESDIS snow analyses for the WRF initial conditions, and we also thank David Stauffer of Penn State for helpful discussions during the course of this project. Authors Lee and Haupt are also grateful to Xcel Energy for funding that partially supported this project.

REFERENCES

Alhamed, A., S. Lakshmivarahan, and D.J. Stensrud, 2002: Cluster analysis of

multimodel ensemble data from SAMEX. *Mon. Wea. Rev.*, **130,** 226-256.

Clark, A.J., W.A. Gallus Jr., 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.

Cressman, G.P., 1959: An operational objective analysis system. *Mon. Wea. Rev.*, **87,** 367-374.

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

Fraley, C., A.E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138,** 190-202.

Fujita, T., D.J. Stensrud, and D.C. Dowell, 2007: Surface data assimilation using an ensemble Kalman filter approach with initial condition and model physics uncertainties. *Mon. Wea. Rev.*, **135,** 1846-1868.

Grimit, E.P., and C.F. Mass, 2007: Measuring the ensemble spread-error relationship with a probabilistic approach: Stochastic ensemble results. *Mon. Wea. Rev.*, **101,** 968-979.

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,** 625-641.

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: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139,** 3694-3710.

Jolliffe, I.T., 2002: Principal component analysis. 2nd ed., Springer, 487 pp.

Jones, M.S., B.A. Colle, and J.S. Tongue, 2007: Evaluation of a mesoscale short-range ensemble forecast system over the Northeast United States. *Wea. Forecasting*, **22,** 2305-2319.

Kolczynski, W.C., D.R. Stauffer, S.E. Haupt, and A. Deng, 2009: Ensemble variance calibration for representing meteorological uncertainty for atmospheric transport and dispersion modeling. *J. Appl. Meteor. Climat.*, **48,** 2001-2021.

Kolczynski, W.C., D.R. Stauffer, S.E. Haupt, N.S. Altman, and A. Deng, 2011: Investigation of ensemble variance as a measure of true forecast variance. *Mon. Wea. Rev.*, **139,** 3954-3963.

Lee, J.A., W.C. Kolczynski, T.C. McCandless, and S.E. Haupt, 2012: An objective methodology for configuring and down-selecting an NWP ensemble for low-level wind prediction. *Mon. Wea. Rev.*, in press.

Legg, T.P., K.R. Mylne, and C. Woolcock, 2002: Use of medium-range ensembles at the Met Office I: PREVIN – a system for the production of probabilistic forecast information from the ECMWF EPS. *Meteorol. Appl.*, **9,** 255-271.

Mesoscale & Microscale Meteorology (MMM) Division, National Center for Atmospheric Research, 2011: Weather Research & Forecasting ARW version 3 modeling system user's guide, 362 pp. [Available online at

http://www.mmm.ucar.edu/wrf/users/docs/user_guide_V3/contents.html]

Raftery, A.E., F. Balabdaoui, T. Gneiting, and M. Polakowski, 2003: Using Bayesian model averaging to calibrate forecast ensembles. *Technical Report no. 40, Dept. of Statistics, University of Washington; 15 December 2003*. [Available online at http://www.stat.washington.edu/research/reports/2003/tr440.pdf]

Raftery, A.E., F. Balabdaoui, T. Gneiting, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133,** 1155-1174.

Roulston, M.S., and L.A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A,** 16-30.

Skamarock, W.C., J.B. Klemp, J. Dudhia, D.O. Gill, D.M. Barker, M.G. Duda, X.-Y. Huang, W. Wang, and J.G. Powers, 2008: A description of the Advanced Research WRF Version 3. *NCAR Technical Note NCAR/TN-475+STR*. 113 pp.

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.

Wilks, D.S., 2006: Statistical methods in the atmospheric sciences, 2nd ed., Academic Press, 626 pp.

Witten, I.H., and E. Frank, 2005: Data mining: Practical machine learning tools and techniques, 2nd ed., Morgan Kaufmann, San Francisco, 525 pp.

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

TABLE 1. Physics schemes for the 42-member WRF multiphysics ensemble. Descriptions and references for schemes are contained in Skamarock et al. (2008).

Exp. #	Microphysics	Longwave Radiation	Shortwave Radiation	Land Surface	Surface Layer	Boundary Layer	Cumulus
CTL-01	WSM-5	RRTM	Dudhia	Noah	MM5 Sim.	YSU	Kain-Fritsch
CTL-02	Thompson	RRTM	Dudhia	RUC	Eta Sim.	MYJ mod.	Grell-Devenyi
10	Thompson	RRTM	Dudhia	Therm. Diff.	MM5 Sim.	YSU	Kain-Fritsch
11	Morrison	New Goddard	New Goddard	Therm. Diff.	MM5 Sim.	YSU	Grell-Devenyi
12	WSM-6	RRTMG	RRTMG	Therm. Diff.	MM5 Sim.	YSU	NSAS
13	Eta (Ferrier)	New Goddard	New Goddard	Noah	MM5 Sim.	YSU	Kain-Fritsch
14	Thompson	RRTMG	RRMTG	Noah	MM5 Sim.	YSU	Grell-Devenyi
15	Morrison	RRTM	Dudhia	Noah	MM5 Sim.	YSU	NSAS
16	WSM-6	New Goddard	New Goddard	Noah	MM5 Sim.	YSU	Kain-Fritsch
17	Eta (Ferrier)	RRTM	Dudhia	RUC	MM5 Sim.	YSU	Grell-Devenyi
18	Thompson	New Goddard	New Goddard	RUC	MM5 Sim.	YSU	NSAS
19	Morrison	RRTMG	RRTMG	RUC	MM5 Sim.	YSU	Kain-Fritsch
20	Thompson	RRTM	Dudhia	Therm. Diff.	Eta Sim.	MYJ mod.	Kain-Fritsch
21	Morrison	New Goddard	New Goddard	Therm. Diff.	Eta Sim.	MYJ mod.	Grell-Devenyi
22	WSM-6	RRTMG	RRTMG	Therm. Diff.	Eta Sim.	MYJ mod.	NSAS
23	Eta (Ferrier)	New Goddard	New Goddard	Noah	Eta Sim.	MYJ mod.	Kain-Fritsch
24	Thompson	RRTMG	RRMTG	Noah	Eta Sim.	MYJ mod.	Grell-Devenyi
25	Morrison	RRTM	Dudhia	Noah	Eta Sim.	MYJ mod.	NSAS
26	WSM-6	New Goddard	New Goddard	Noah	Eta Sim.	MYJ mod.	Kain-Fritsch
27	Eta (Ferrier)	RRTM	Dudhia	RUC	Eta Sim.	MYJ mod.	Grell-Devenyi
28	Thompson	New Goddard	New Goddard	RUC	Eta Sim.	MYJ mod.	NSAS
29	Morrison	RRTMG	RRTMG	RUC	Eta Sim.	MYJ mod.	Kain-Fritsch
30	Thompson	RRTM	Dudhia	Therm. Diff.	MYNN	MYNN-2.5	Kain-Fritsch
31	Morrison	New Goddard	New Goddard	Therm. Diff.	MYNN	MYNN-2.5	Grell-Devenyi
32	WSM-6	RRTMG	RRTMG	Therm. Diff.	MYNN	MYNN-2.5	NSAS
33	Eta (Ferrier)	New Goddard	New Goddard	Noah	MYNN	MYNN-2.5	Kain-Fritsch
34	Thompson	RRTMG	RRMTG	Noah	MYNN	MYNN-2.5	Grell-Devenyi
35	Morrison	RRTM	Dudhia	Noah	MYNN	MYNN-2.5	NSAS
36	WSM-6	New Goddard	New Goddard	Noah	MYNN	MYNN-2.5	Kain-Fritsch
37	Eta (Ferrier)	RRTM	Dudhia	RUC	MYNN	MYNN-2.5	Grell-Devenyi
38	Thompson	New Goddard	New Goddard	RUC	MYNN	MYNN-2.5	NSAS
39	Morrison	RRTMG	RRTMG	RUC	MYNN	MYNN-2.5	Kain-Fritsch
40	Thompson	RRTM	Dudhia	Therm. Diff.	Pleim-Xu	ACM2	Kain-Fritsch
41	Morrison	New Goddard	New Goddard	Therm. Diff.	Pleim-Xu	ACM2	Grell-Devenyi
42	WSM-6	RRTMG	RRTMG	Therm. Diff.	Pleim-Xu	ACM2	NSAS
43	Eta (Ferrier)	New Goddard	New Goddard	Noah	Pleim-Xu	ACM2	Kain-Fritsch
44	Thompson	RRTMG	RRMTG	Noah	Pleim-Xu	ACM2	Grell-Devenyi
45	Morrison	RRTM	Dudhia	Noah	Pleim-Xu	ACM2	NSAS
46	WSM-6	New Goddard	New Goddard	Noah	Pleim-Xu	ACM2	Kain-Fritsch
47	Eta (Ferrier)	RRTM	Dudhia	RUC	Pleim-Xu	ACM2	Grell-Devenyi
48	Thompson	New Goddard	New Goddard	RUC	Pleim-Xu	ACM2	NSAS
49	Morrison	RRTMG	RRTMG	RUC	Pleim-Xu	ACM2	Kain-Fritsch

TABLE 2. Initialization dates in YYYY-MM-DD format for the forecasts used in both the training and verification periods in this study. All forecasts are initialized at 0000 UTC.

Training Dataset		***Verification Dataset***
2009-12-01	2009-12-31	*2010-01-30*
2009-12-06	2010-01-05	*2010-02-04*
2009-12-11	2010-01-10	*2010-02-09*
2009-12-16	2010-01-15	*2010-02-14*
2009-12-21	2010-01-20	*2010-02-19*
2009-12-26	2010-01-25	*2010-02-24*

TABLE 3. Members that are included in each candidate down-selection subset. Naming convention is as follows: P or K refers to a subset determined by PCA or k-means clustering; the number is the number of members included in that subset; and W or T refers to whether the subset was determined from VWD or temperature errors. Refer to Table 1 for the physics schemes for each member.

Subset	**Ensemble Members**
P07W	CTL01, 11, 16, 23, 24, 25, 44
P07T	10, 13, 16, 23, 26, 33, 46
K10W	10, 15, 18, 21, 25, 29, 30, 36, 39, 43
K08T	CTL02, 13, 33, 37, 42, 44, 46, 48

TABLE 4. Clusters of ensemble members from the VWD data. One member from each of these clusters is randomly chosen for subset K10W (see Table 3).

Cluster	**Ensemble Members**
Cluster 1	CTL01, 13, 14, 15, 16
Cluster 2	CTL02, 27, 28, 29
Cluster 3	10, 11, 12
Cluster 4	17, 18, 19
Cluster 5	20, 21, 22
Cluster 6	23, 24, 25, 26
Cluster 7	30, 31, 32, 40, 41, 42
Cluster 8	33, 34, 35, 36
Cluster 9	37, 38, 39, 47, 48, 49
Cluster 10	43, 44, 45, 46

TABLE 5. Clusters of ensemble members from the temperature error data. One member from each of these clusters is randomly chosen for subset K08T (see Table 3).

Cluster	**Ensemble Members**
Cluster 1	01, 13, 14, 15, 16
Cluster 2	02, 27, 28, 29
Cluster 3	10, 11, 12, 20, 21, 22, 30, 31, 32, 40, 41, 42
Cluster 4	17, 37, 47
Cluster 5	18, 19, 38, 39, 48, 49
Cluster 6	23, 33, 43
Cluster 7	24, 25, 34, 35, 44, 45
Cluster 8	26, 36, 46

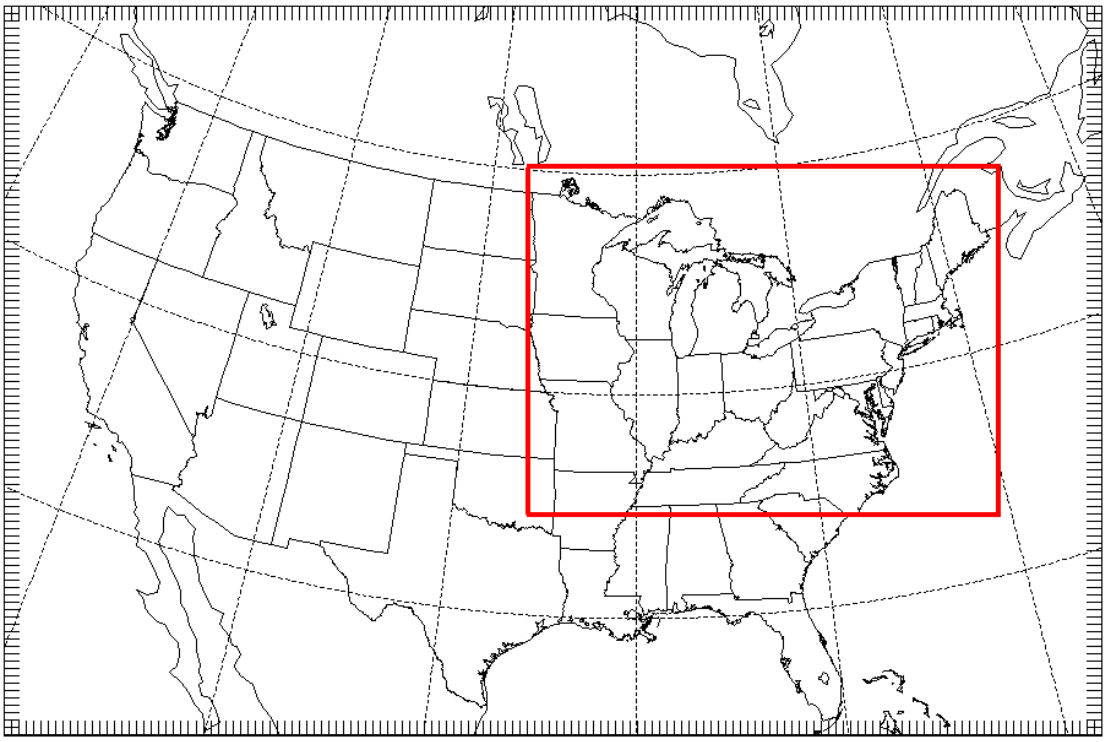

FIG. 1. WRF domains used in this study. The outer domain has a 36-km horizontal resolution, and the inner domain (outlined in red) has a 12-km horizontal resolution.

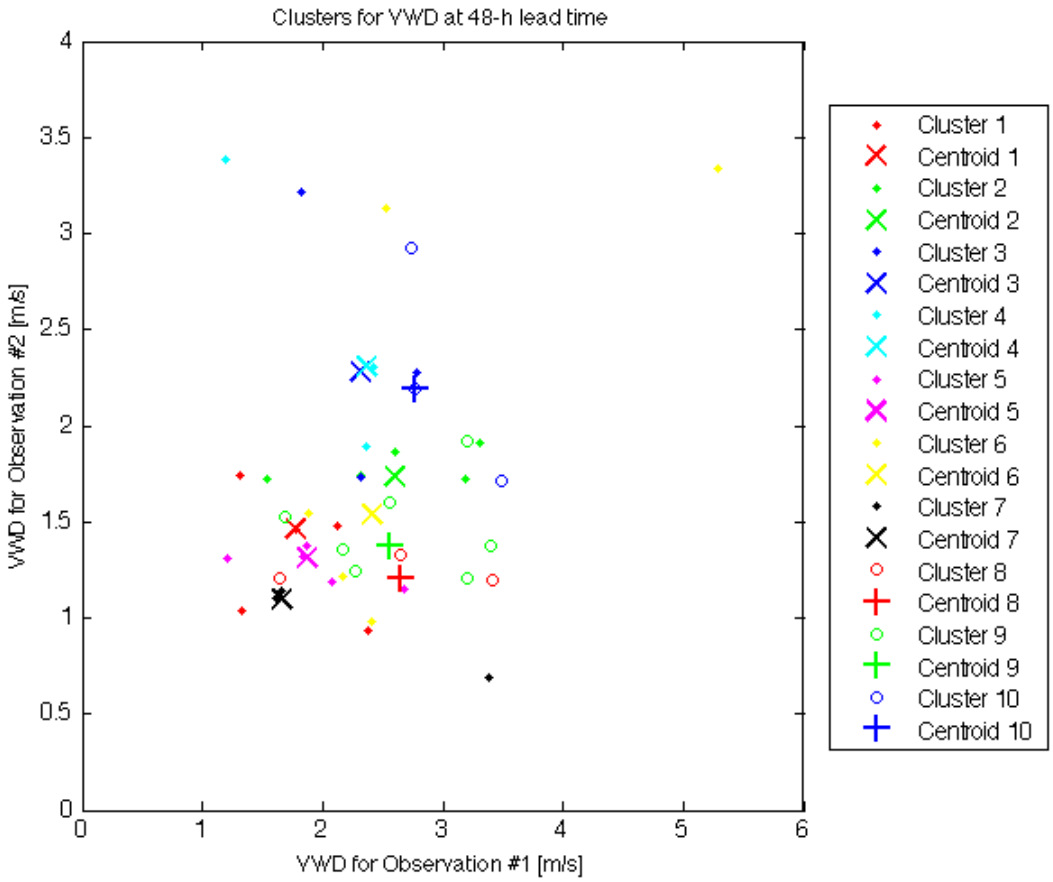

FIG. 2. Clusters and cluster centroids for VWD at a lead time of 48 h for just two of the 3746 total observations. Each closed or open circle represents an ensemble member, and each rotated or upright cross represents the position of the centroid in these two dimensions.

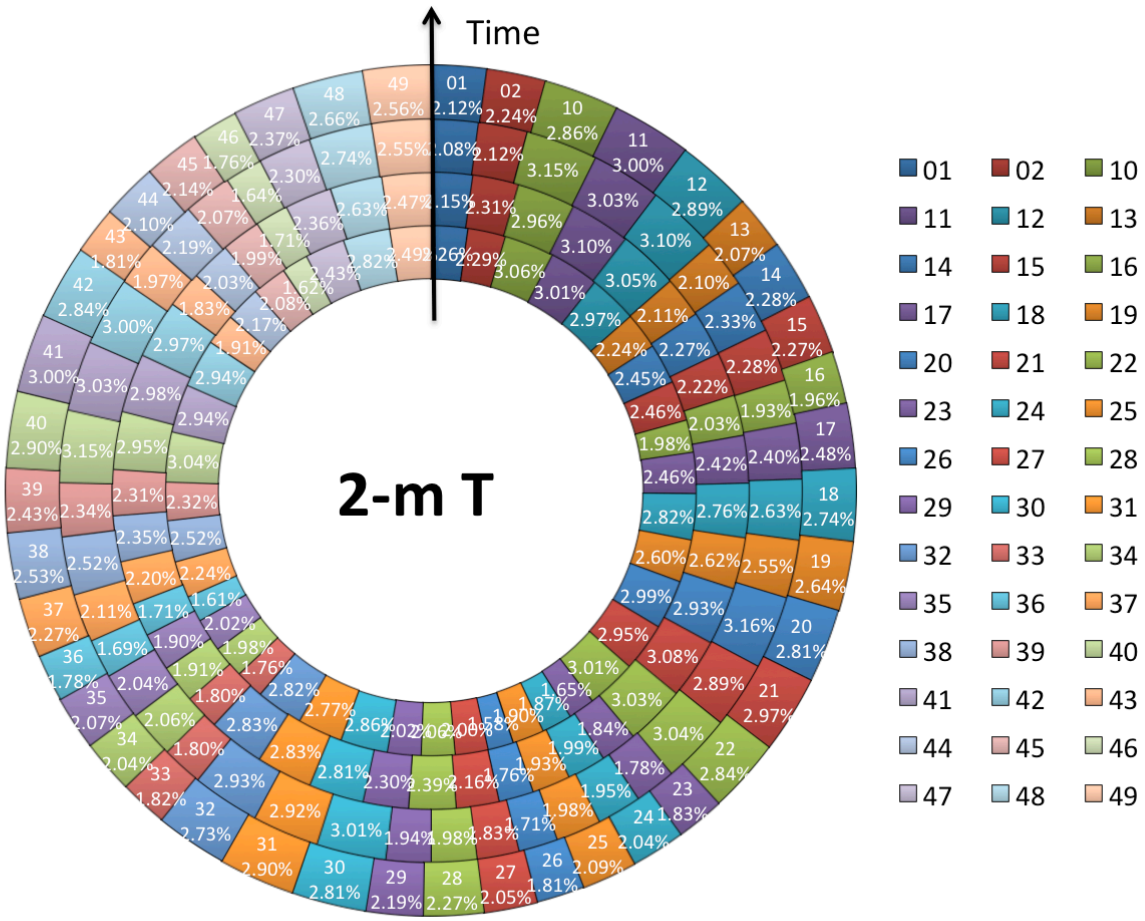

FIG. 3. BMA weights in percent for 2-m temperature forecasts during the training period. The 42 ensemble members are color-coded. The inner ring for 12-h forecast weights, the second ring is for 24-h forecast weights, the third ring for 36-h forecast weights, and the outer ring for 48-h forecast weights.

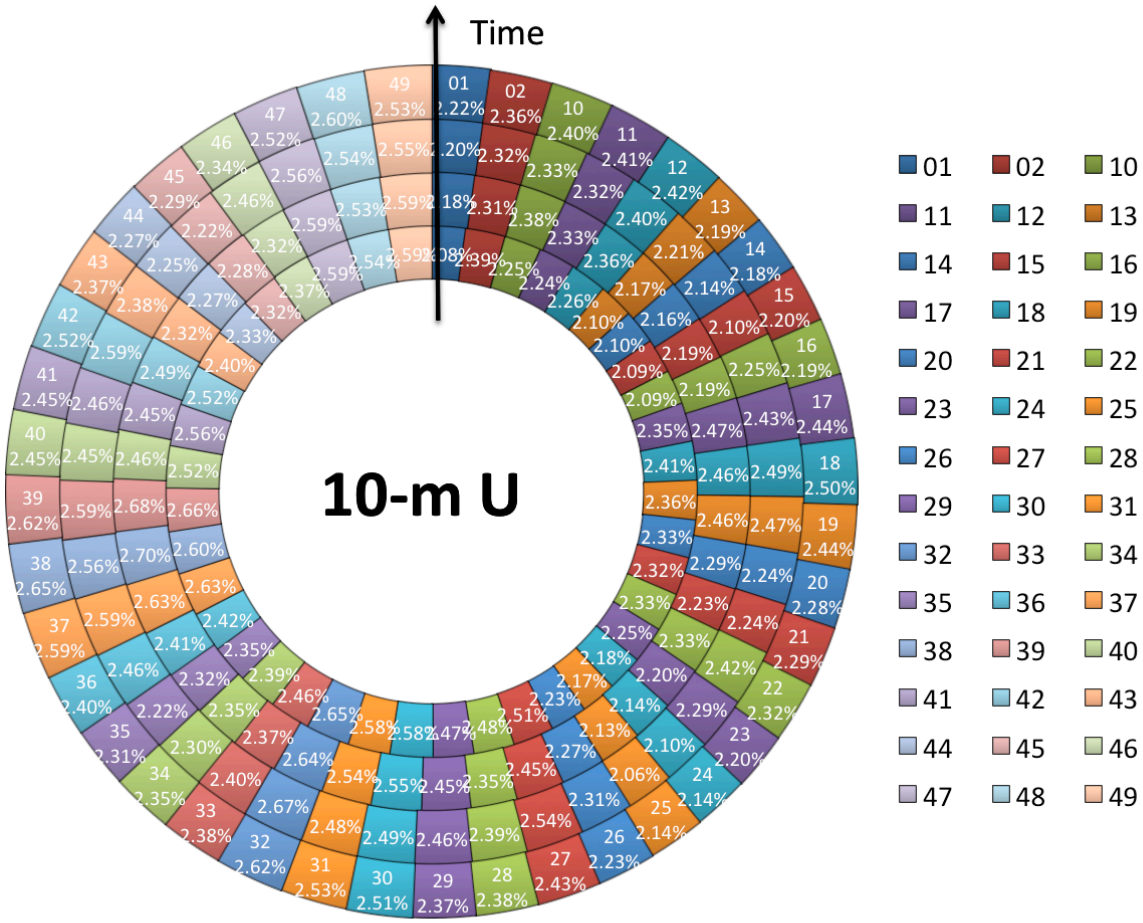

FIG. 4. Same as Fig. 3, but for 10-m u-wind component.

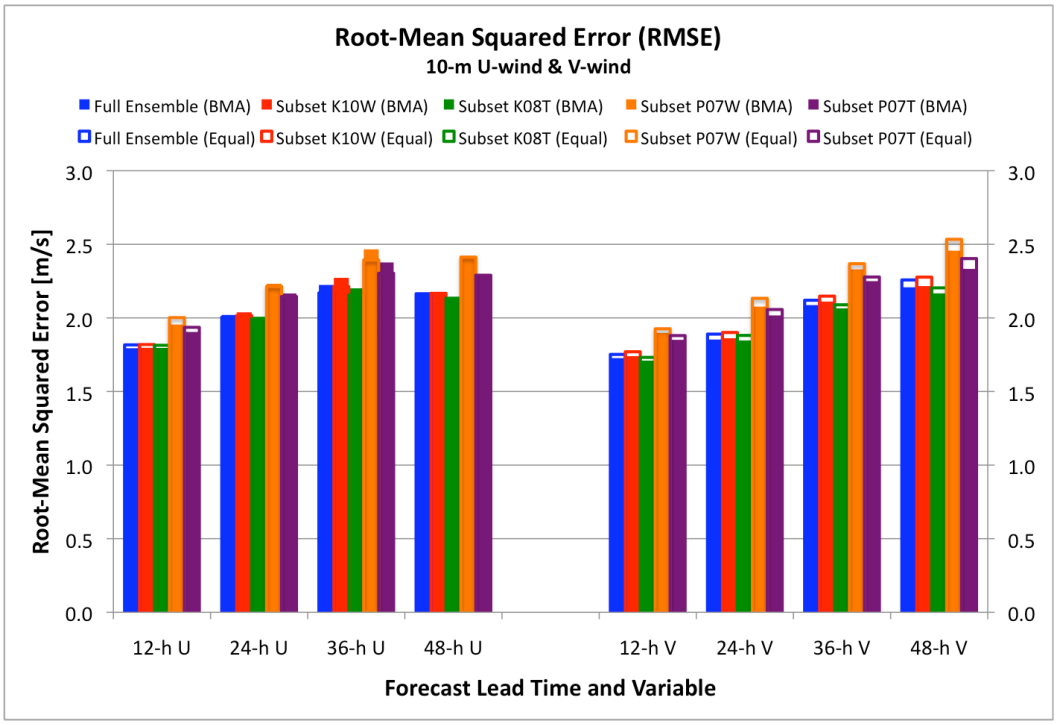

FIG. 5. RMSE for 10-m u-wind and v-wind components at four forecast lead times over the verification period, for the full ensemble and all four subsets. Filled bars are for the BMA-weighted ensembles, and unfilled bars are for the equal-weighted ensembles.

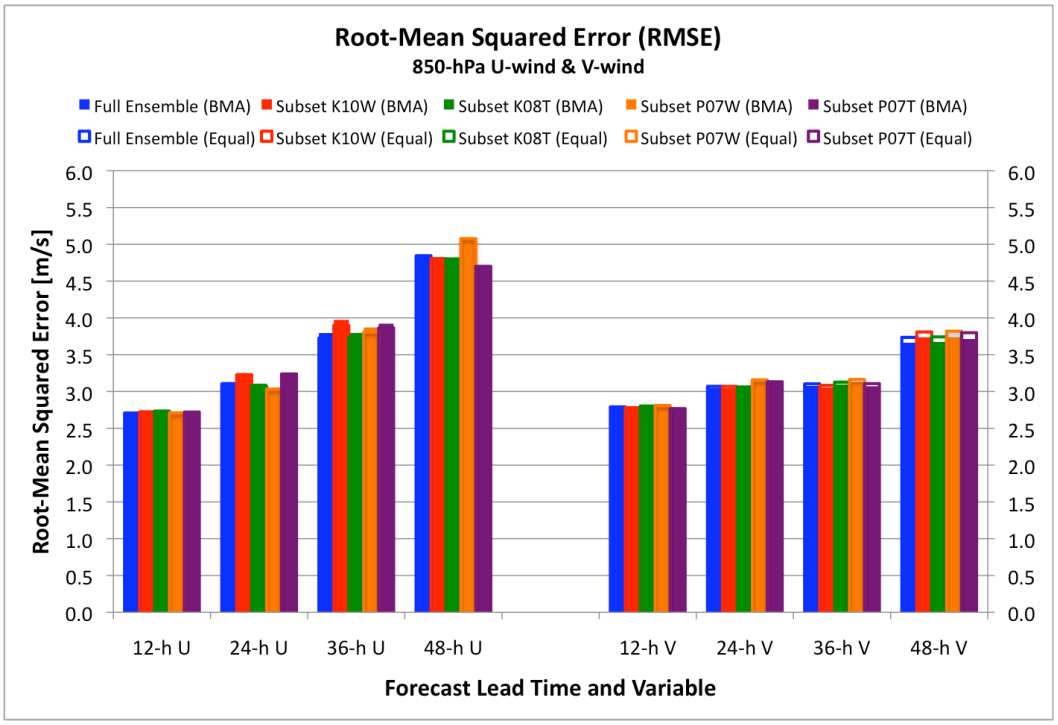

FIG. 6. Same as Fig. 5, but for 850-hPa u-wind and v-wind components.

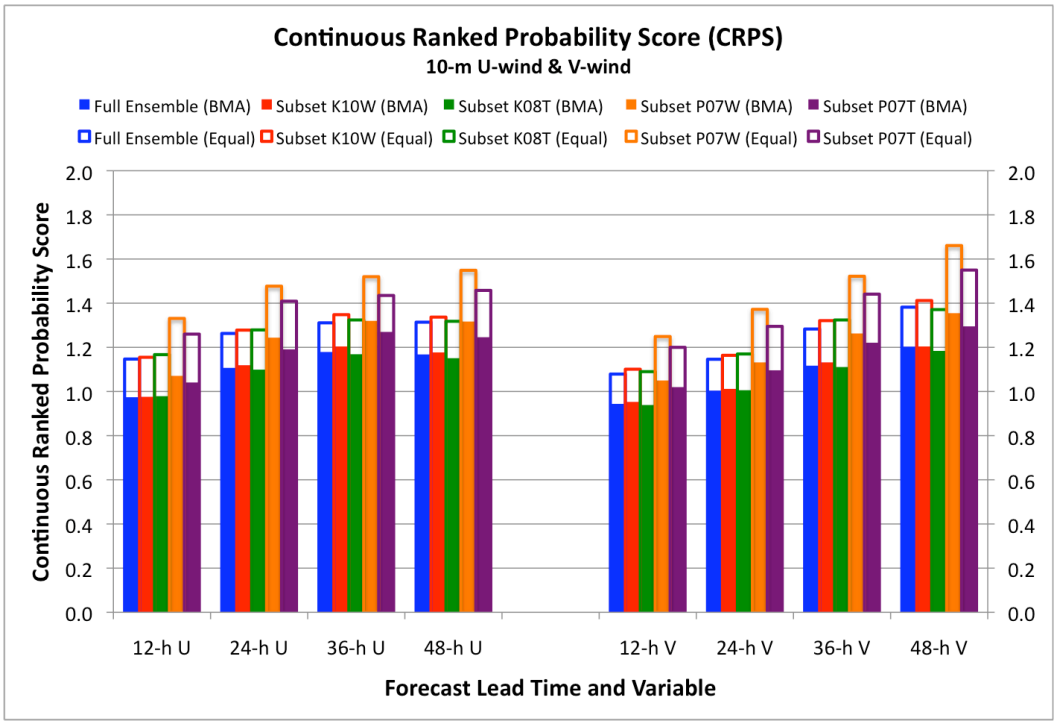

FIG. 7. CRPS for 10-m u-wind and v-wind components at four forecast lead times over the verification period, for the full ensemble and all four subsets. Filled bars are for the BMA-weighted ensembles, and unfilled bars are for the equal-weighted ensembles.

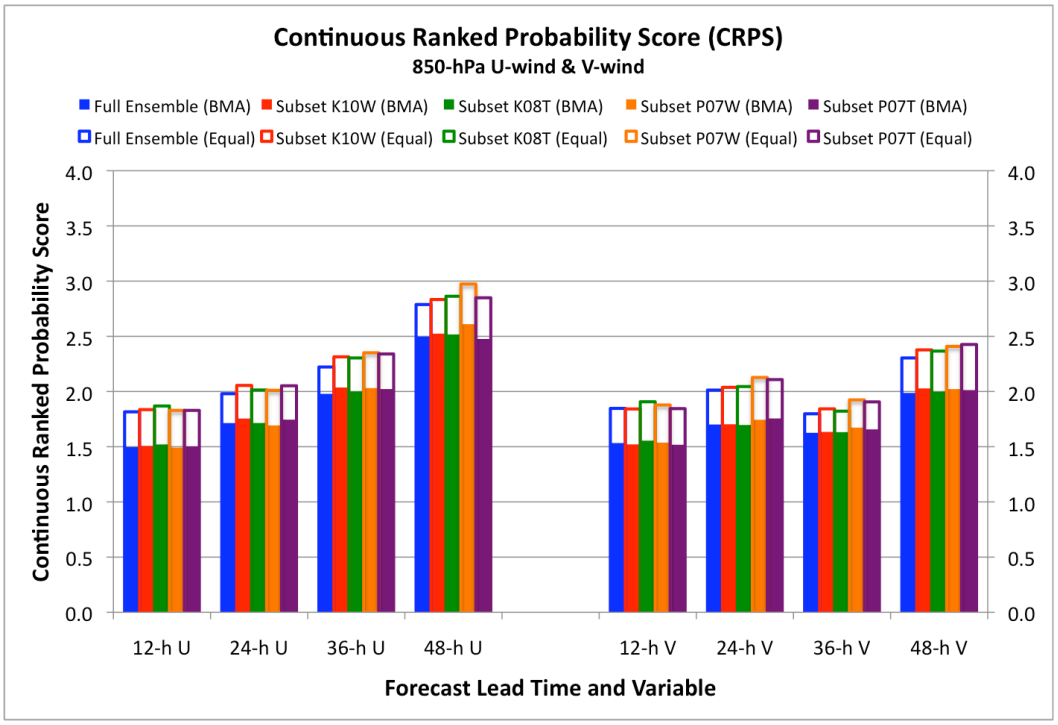

FIG. 8. Same as Fig. 7, but for 850-hPa u-wind and v-wind components.

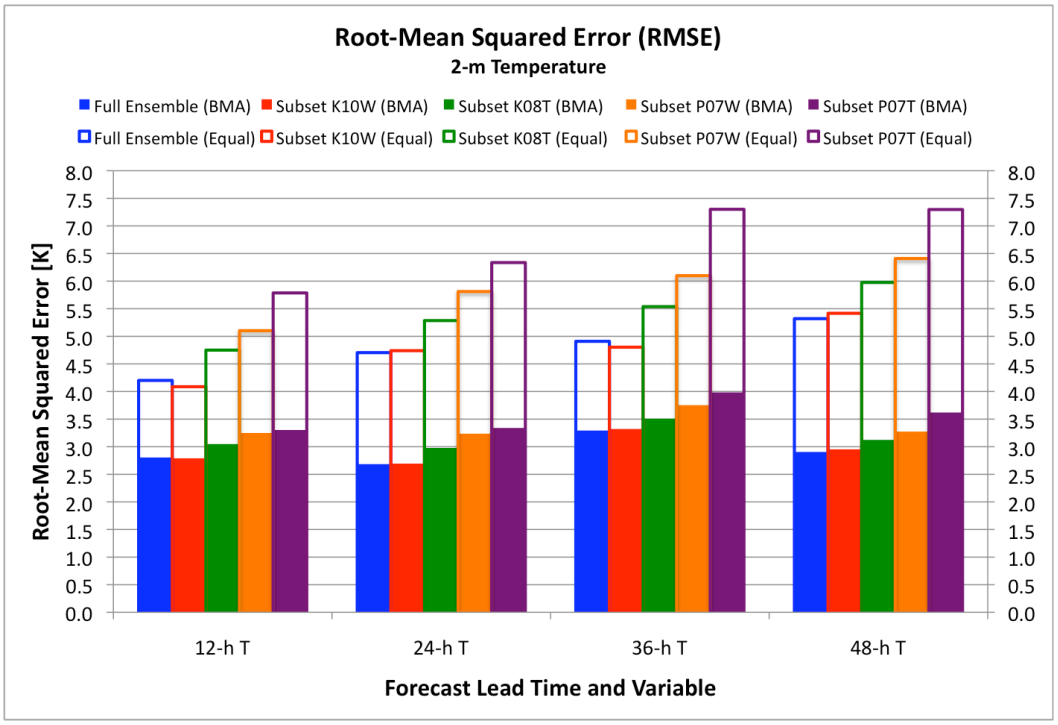

FIG. 9. RMSE for 2-m temperature at four forecast lead times over the verification period, for the full ensemble and all four subsets. Filled bars are for the BMA-weighted ensembles, and unfilled bars are for the equal-weighted ensembles.

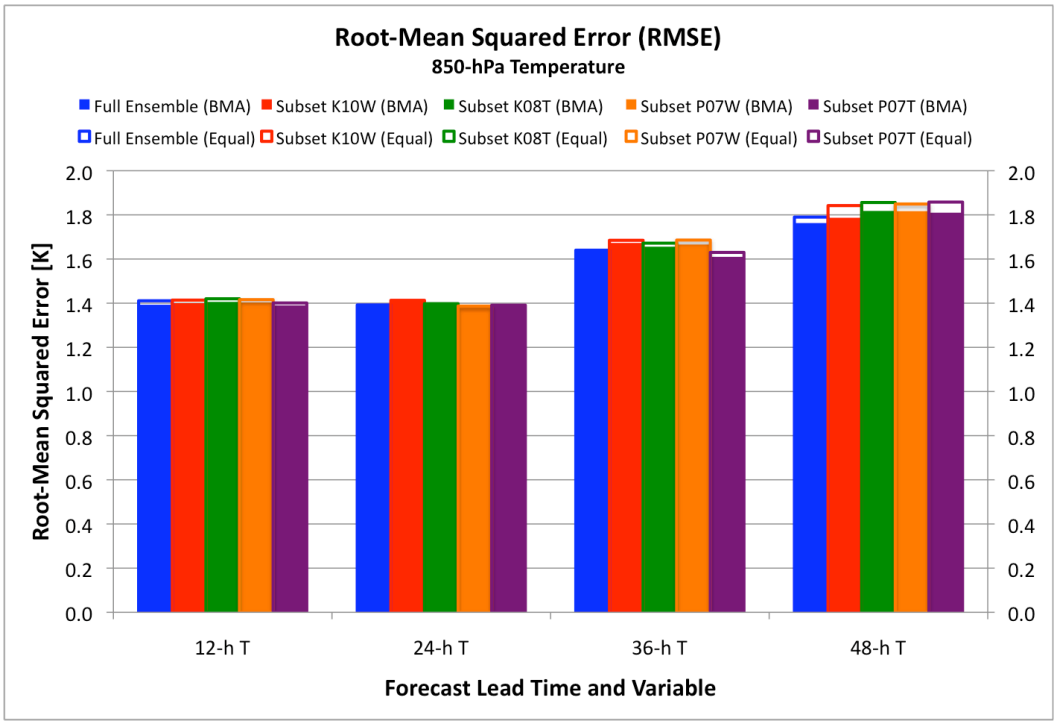

FIG. 10. Same as Fig. 9, but for 850-hPa temperature.

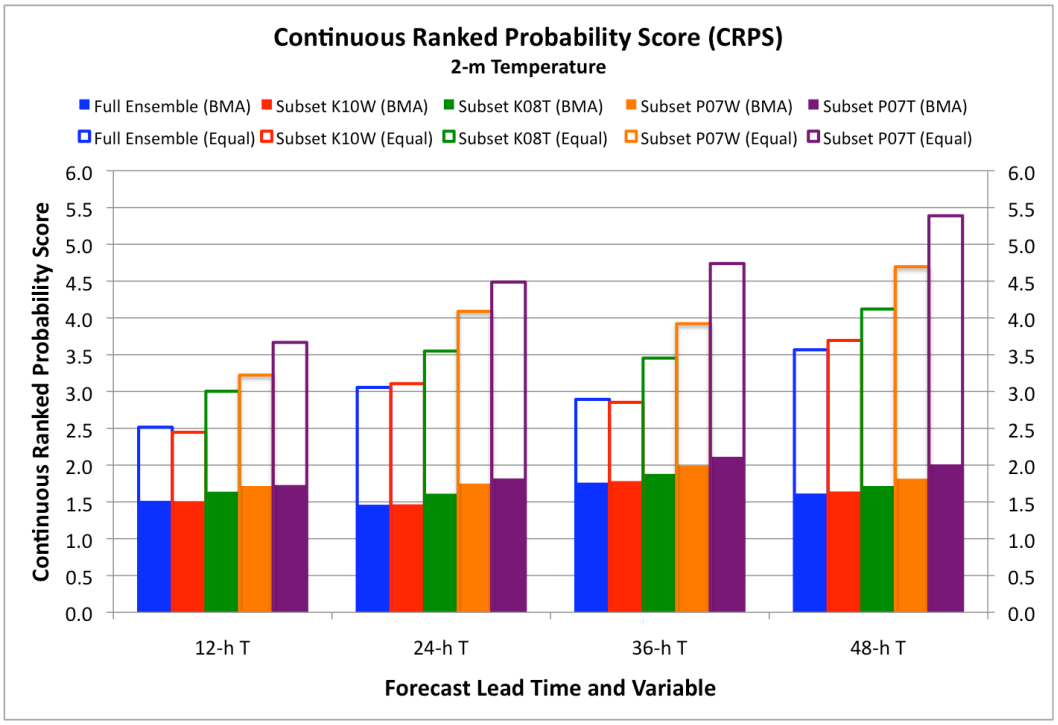

FIG. 11. CRPS for 2-m temperature at four forecast lead times over the verification period, for the full ensemble and all four subsets. Filled bars are for the BMA-weighted ensembles, and unfilled bars are for the equal-weighted ensembles.

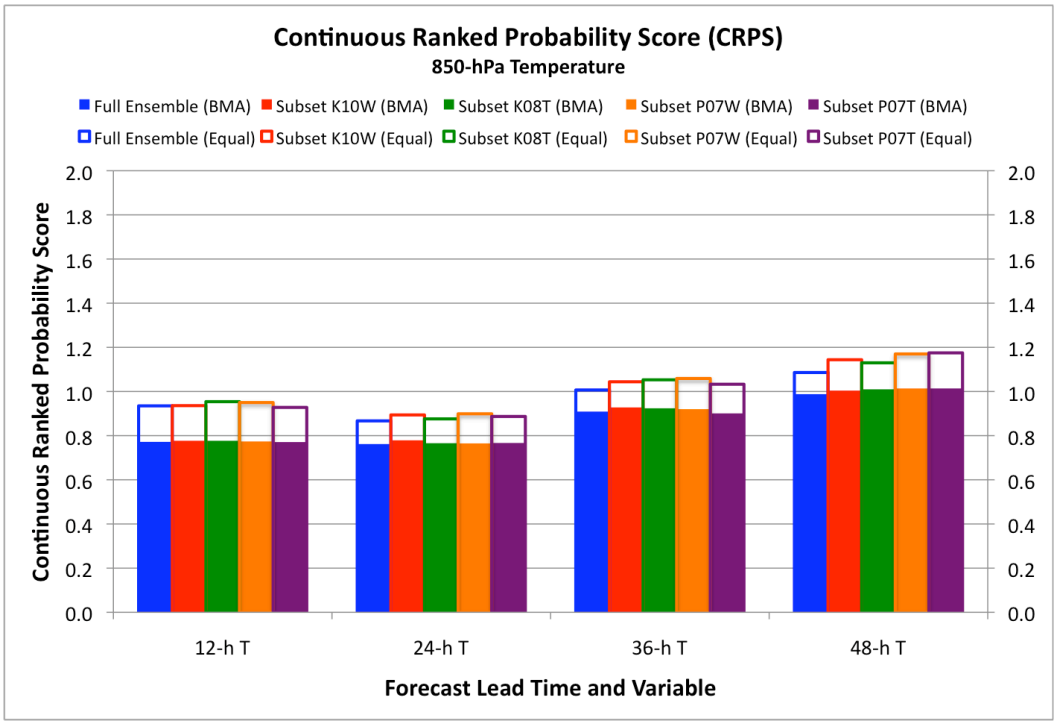

FIG. 12. Same as Fig. 11, but for 850-hPa temperature.