129

# DOWN-SELECTION OF NWP ENSEMBLE CONFIGURATIONS

Jared A. Lee[1,2,3,*], Walter C. Kolczynski[1], Tyler C. McCandless[1,2], Sue Ellen Haupt[1,2,3], David R. Stauffer[1], Aijun Deng[1], and Kerrie J. Schmehl[2]
[1]Department of Meteorology, The Pennsylvania State University, University Park, PA
[2]Applied Research Laboratory, The Pennsylvania State University, State College, PA
[3]Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO

## 1. INTRODUCTION

There are inherent limitations to forecasting a single realization of the future state of the atmosphere due to its chaotic nature. While numerical weather prediction (NWP) models have become more sophisticated in recent years and better represent and predict the atmospheric state, they are still limited because of imperfect model numerics, imperfect parameterizations of unresolved physical processes, and interpolations of input data that is sparsely located compared to current model grid resolutions. In recognition of these difficulties, contemporary NWP uses ensembles of simulations. Members in these ensembles often differ by imposed initial conditions (ICs), lateral and lower boundary conditions (LBCs), model physics parameterization schemes, and even the choice of NWP modeling system. While the relationship between ensemble spread and forecast error is not linear, Grimit and Mass (2007) state that with larger ensemble spread, there is a larger probability of the forecast errors being larger, and vice-versa.

There are both inherent and practical limitations to ensemble forecasting as well. The first limitation is that NWP ensembles are computationally expensive to run. When faced with limited computing resources, trade-offs must be made when configuring the ensemble. How fine can the horizontal and vertical resolution be? How many members can there be in the ensemble? How big can the forecast domain be, and should it be nested in a larger and coarser outer domain? How long can the forecast duration be? All of these considerations are important, but they all compete for the same limited resources.

* *Corresponding author address:* Jared A. Lee, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80507-3000;
e-mail: jal488@meteo.psu.edu.

A second limitation is that the forecast variances in ensemble forecast systems tend to be uncalibrated (Raftery et al. 2005). If an ensemble is perfectly calibrated, then the forecast error variance and ensemble-mean variance will have a 1:1 ratio (*e.g.*, Grimit and Mass 2007; Kolczynski et al. 2009). In other words, if a meteorological ensemble is uncalibrated, then the forecast uncertainty cannot be properly diagnosed from the ensemble spread. Kolczynski et al. (2011) show that, using a stochastic ensemble, even perfectly constructed ensembles with fewer than hundreds of members will be uncalibrated. Unfortunately, given current computing resources at most institutions, including operational centers, it is not practical to run an NWP ensemble with hundreds of members, so steps must be taken to calibrate ensembles before forecast uncertainty can be properly assessed. This is beyond the need for calibration due to deficiencies in ensemble construction, which could necessitate calibration no matter how many members are used.

There are many applications for which NWP ensemble forecasting is useful. One of these applications is atmospheric transport & dispersion (AT&D) forecasting, as AT&D models such as the Second-order Closure Integrated Puff (SCIPUFF) model (Sykes et al. 2004) are often driven by NWP ensemble model output (e.g., Warner et al. 2002; Lee et al. 2009). Another application is wind power forecasting where uncertainty metrics are needed (Liu et al. 2007). To obtain appropriate spread in concentration predictions from AT&D models, there should be ”good spread” in low-level wind direction and atmospheric boundary layer (ABL) depth, as these are two of the most important parameters affecting uncertainty in AT&D predictions (Lewellen and Sykes 1989). By “good spread” we mean increased ensemble variance that improves the calibration, reliability, and resolution of the ensemble (Eckel and Mass 2005). Therefore, when evaluating the performance of ensemble configurations for this study, emphasis is placed on NWP ensemble verification against observations of parameters in

the ABL, such as 10-m winds and 2-m temperature. For other applications there may need to be good spread in other parameters in order to provide useful ensemble forecasts, and other variables may be more important for verification purposes.

It is not clear *a priori* how best to configure a useful ensemble for AT&D applications, or if that particular configuration would be the most useful for other applications, such as quantitative precipitation forecasting. Therefore, some testing is necessary. There are a large number of possible choices of IC, LBC, and physics perturbations that can be included in an NWP ensemble for any given application. The primary aim of this study is to propose an objective methodology to "down-select," or determine which subset of members should be included in an ensemble, assuming that it is impractical to include all the members. In this study we configure an ensemble to predict low-level winds, although our methodology should be useful for other applications as well. Details about our ensemble and the verification data that we used are discussed in section 2.

Naturally, down-selection may alter the ensemble mean often used as the "best guess" forecast, as well as exacerbate the calibration problem. These downsides result both because we will have fewer members and because, by preferentially choosing specific members, we further alter the natural probability distribution function (PDF) of the ensemble. Therefore, for down-selection to be successful, we need to employ post-processing methods that will help maintain an accurate best guess, and ideally calibrate the ensemble variance as well.

The first post-processing method used is principal component analysis (PCA). In this study PCA is our primary method for defining candidate down-selected ensemble subsets. PCA will be discussed further in section 3. The ensemble calibration method used in this study is Bayesian model averaging (BMA). BMA is a statistical post-processing method introduced to the atmospheric sciences by Raftery et al. (2003, 2005), and is used to correct for the underdispersive, and thus uncalibrated, nature of forecast ensembles. Our BMA application will be discussed in section 4 of this paper. Finally, section 5 will include an overall summary of the study and list avenues of future research.

## 2. DATA

### 2.1. *Ensemble configuration*

While we recognize the likely importance of IC/LBC perturbations even in short-range mesoscale NWP, we choose to test our methodology on a physics ensemble for simplicity. Additionally, if we assume that we could use an IC/LBC perturbation method that results in equally likely perturbations, then ignoring IC/LBC perturbation does not affect the performance of our physics members. Indeed, removing the random signal of IC/LBC perturbation may provide clearer results.

Our 24-member physics ensemble for this study is created with version 3.2 of the Weather Research and Forecasting (WRF) Advanced Research WRF (ARW) NWP model (Skamarock et al. 2008). The microphysics and atmospheric radiation schemes (both longwave and shortwave) are the same for each ensemble member, but the land surface, surface layer, boundary layer and cumulus scheme configuration varied for each member, as detailed in Table 1. There are 45 full vertical levels in each simulation, with the lowest full level at 24 m AGL, 9 full levels below 500 m AGL, 16 full levels below 1 km AGL and 24 full levels below 2 km AGL. The model top is at 50 hPa. Such high vertical resolution in the lowest portions of the troposphere is chosen because this study focuses on processes occurring in the ABL. Two model domains are used. The coarse domain uses a horizontal grid spacing of 36 km and a time step of 180 s, while the nested domain uses 12-km grid spacing and a 60-s time step. The coarse domain encompasses the continental United States (CONUS), and the nested domain covers the Great Lakes, Ohio Valley, Mid-Atlantic and Northeast, as shown in Figure 1.

For each month of June-July-August 2009, six forecast periods are randomly chosen, with three being initialized at 0000 UTC and three being initialized at 1200 UTC, to avoid biasing our results with the diurnal cycle. In total there are 18 forecast times chosen for this summer evaluation period, with a forecast period of 48 h in all cases. No data assimilation is used during the model integration for this study, because we desire to simulate a forecasting system, rather than a hindcasting system. The LBCs for all 24 members in this study come from the 0.5°×0.5°-resolution Global Forecast System (GFS) forecast cycles initialized at each of the randomly chosen forecast times. For the ICs the 0-h GFS analysis is

blended with standard WMO observations to provide a more accurate initial state.

### 2.2. *Verification and quality control*

To verify our WRF ensemble forecasts, we use standard WMO observations. These observations are quality-controlled against the WRF Pre-processing System (WPS)-interpolated GFS forecast fields, using the Obsgrid software developed and released by the National Center for Atmospheric Research (NCAR). We implement some additional quality control checks, including rejecting any surface observations with a reported elevation higher than 600 hPa, and rejecting surface observations where the reported elevation differed from the model terrain by more than 200 m. We find that blending observations into the initial conditions with Obsgrid improved our verification scores, and that implementing these additional quality control checks improve our verification scores even further.

Of the 18 forecasts created for the summer evaluation period, six are randomly chosen to be set aside for verification purposes. The remaining 12 forecasts are used as training data for both the down-selection process and for calibration purposes. Table 2 lists which forecast periods are used for the training period and which were used for the verification period. Three different diagnosed quantities are used for verification: 10-m above ground level (AGL) zonal wind (u), 10-m AGL meridional (v) wind, and 2-m AGL temperature.

We measure the performance of the calibrated ensemble of both a deterministic forecast and a probabilistic forecast. Here we consider root-mean squared error (RMSE) as our primary deterministic measure and cumulative rank probability score (CRPS) as our primary probabilistic measure. The root-mean squared error is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} f_i - o_i^{\,2}} \qquad (1)$$

where $o$ is the value of observation $i$, $f$ is the forecast value at the time and location of observation $i$, and $N$ is the total number of observations. The cumulative rank probability score is defined as (Wilks 2006):

$$CRPS = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\left[p_i^f(x) - p_i^o(x)\right]^2 dx \qquad (2)$$

$$p_i^o(x) = \begin{cases} 0 & x < o_i \\ 1 & x \geq o_i \end{cases}$$

where $p_i^f(x)$ is the forecast cumulative probability of the variable being $\leq x$ at the time and location of observation $i$ and all other variables are as before.

Rank histograms are another verification tool we use to assess the reliability of the ensemble forecasts. For an ensemble containing $n_{ens}$ members, verification rank histograms are created by binning each verifying observation within the $n_{ens}+1$ -member distribution. These histograms are frequently used to diagnose the bias and dispersion of ensembles, and ensembles that exhibit no biases and are neither underdispersive nor overdispersive have rank histograms that are approximately flat (Wilks 2006). We also use the continuous analog of the verification rank histogram, which is the probability integral transform (PIT) histogram. PIT histograms are defined by evenly spacing the bins throughout the forecast distribution (Raftery et al. 2005).

## 3. ENSEMBLE MEMBER DOWN-SELECTION

### 3.1. *PCA*

In order to down-select to a smaller number of ensemble members, the first step is to calculate the errors of each of the ensemble member forecasts. The goal of our ensemble down-selection technique is to remove redundant members and retain the members that contribute to the forecast accuracy and spread. This goal arises both because computational resources are generally too limited to allow for very large ensembles, and because ensembles are most useful when each member contribute to the forecast variability. To accomplish this, we utilize principal component analysis (PCA). PCA is a useful mathematical technique to apply to large datasets with correlated variables, which reduces the dataset by identifying a smaller number of uncorrelated variables, or principal components (PCs) (Jolliffe 2002; Wilks 2006; Witten and Frank 2005). Principal components are and the eigenstructures ordered so that the lowest PCs contain the maximum amount of variance. We then truncate the set so that the first several PCs

that represent 95% of the variability of the data are used to represent the dataset (Jolliffe 2002). Figure 2 shows a plot of the cumulative variance for our PCs and where we truncate to maintain 95% of the variance, shown as a solid red line for 2-m temperature at a forecast lead time of 12 h. We use PCA here to capture the maximum variability in the forecast errors and then determine which ensemble members contributed most to those PCs.

The goal of the PCA is to determine which ensemble members contribute most to the variability in the forecast errors; thus, the first step was to perform PCA on the dataset of forecast errors to identify the number of ensemble members that contributed to the 95% cumulative variance. In addition to testing the 95% cumulative variance threshold for the number of principal components to use, we also experimented with the 90% threshold as a cut-off. The freely available off-the-shelf software program RapidMiner is used to perform the PCA here (Mierswa et al. 2006).

For the three different weather variables we verify against in this study, 10-m zonal (u) wind, 10-m meridional (v) wind, and 2-m temperature, each at four different forecast lead times, 12-h, 24-h, 36-h, and 48-h, PCA identifies between four and seven PCs that account for 95% of the variability during the training period. After determining the appropriate number of PCs for each variable at each lead time, the next step is to determine which ensemble member contributed most to each of those PCs.

The number of times each ensemble member contributes to each PC is tallied. Four different subsets of ensemble members are defined: subset A, containing the top-contributing members to PCs that as a group account for 95% of cumulative forecast variability; subset B, containing the top-contributing members to PCs that each account for at least 2.0% of forecast variability to PCs; subset C, containing the top-contributing members to PCs that as a group account for 90% of cumulative forecast variability; and subset D, containing the most frequent top-contributing members to PCs that account for 95% of cumulative forecast variability. Details of which ensemble members are included in each candidate subset for the nested 12-km domain are listed in Table 3. The size and membership of each subset is unique to this particular domain configuration; the PCA method selects different a different set and number of members for the 36-km domain, both over the full domain and the 12-km sub-domain. Thus ensemble performance depends upon both the specific region covered by and the resolution of the simulations.

### 3.2. *Correlation Analysis*

When down-selecting to a subset of ensemble members, it is desirable to try to exclude members that provide redundant information. One method that can shed light on whether certain members are providing redundant information is correlation analysis. We choose to correlate the forecast errors for each ensemble member.

We compute correlations of the errors between all ensemble members for each forecast parameter at each forecast lead time. Error correlations for 24-h forecasts of 2-m temperature and 10-m zonal wind are shown in Tables 4 and 5, respectively. The tables are symmetric, so half of each table is color-coded as a visual aid, with warmer colors highlighting stronger correlations, and cooler colors highlighting weaker correlations between the forecast errors in the ensemble members. The results discussed below are for the 12-km domain, but results are similar for both the same area with 36-km resolution and for the full 36-km domain. Thus we have greater confidence when attributing physical explanations to our results.

Some interesting patterns appear in the error correlations. For 2-m temperature forecasts, error correlations tend to be grouped according to what land surface model was used in each member (Table 4). For all members, the highest correlations are with all other members that share the same land surface scheme (see Table 1 for the physics configuration used for each ensemble member). The next-highest correlations are between pairs of land surface schemes. Members that used the Thermal Diffusion land surface scheme and the Pleim-Xu land surface model had highly correlated errors. The same is true for members that used the Noah and Rapid Update Cycle (RUC) land surface models, although those correlations are slightly weaker. The weakest correlations are between members that used either the Noah or RUC land surface models and those that use either the Thermal Diffusion or Pleim-Xu land surface schemes. These results indicate that the choice of land surface scheme has a substantial impact on 2-m temperature forecasts. This makes sense because different land surface models represent surface energy and moisture fluxes somewhat differently. This impacts surface temperature because processes in the surface layer are dominated by interactions with the land surface (Wyngaard 2010). Another

pattern worth noting from Table 4 is that correlations are also quite high between members that are identical except for the cumulus scheme. This is not an entirely surprising result, as the cumulus scheme only has an indirect effect on the model surface temperature by producing precipitation and downdrafts. Thus, to achieve greater variability in 2-m temperature forecasts, ensembles should contain diversity in land surface schemes.

For 10-m wind forecasts, error correlations tend to be grouped according to the choice of boundary layer scheme (Table 5). Correlations are generally strongest between members that share the same boundary layer scheme, and especially between members that share both the same boundary layer and cumulus schemes. This result indicates that the choice of ABL scheme, and the cumulus scheme to a lesser extent, has a substantial impact on 10-m wind forecasts. This is due to the fact that the ABL schemes differ in how they model the dynamics and structure of the ABL. That the cumulus scheme appears to have a secondary effect on 10-m wind forecast error correlations because low-level model winds are generally only affected by the cumulus scheme in regions where the model produces precipitation. Thus, to achieve greater variability in 10-m wind forecasts, ensembles should contain diversity in boundary layer schemes and, somewhat less importantly, cumulus schemes.

## 4. ENSEMBLE CALIBRATION

Our calibration technique is Bayesian Model Averaging (BMA) (Raftery et al. (2003). BMA improves ensemble forecasts by estimating the best weights and parameters for each ensemble member to make a smooth PDF. Weights and parameters are trained to best match the observations during some training period, then applied to future forecasts to create a full PDF of the ensemble forecast.

A first step in BMA determines the functional form of the posterior distributions of the ensemble. For temperature, we use a normal distribution, as in Raftery et al. (2005). For vector quantities, however, selecting the distribution is more difficult because the vectors are described by two (or more) different scalar quantities that are related. Here we define the horizontal wind in terms of zonal (u) and meridional (v) components. Sloughter et al. (2010) use an alternate approach by decomposing wind into speed and direction, then using a gamma distribution for the wind speed, which allows us to use a normal posterior distribution for each wind component.

While earlier studies using BMA have focused on calibrating for specific forecast locations independently, we have chosen to calibrate on all locations simultaneously. This is important for our future applications, as we intend to apply BMA over an entire forecast region for insertion into an AT&D model, rather than forecasting at specific point locations.

Each of the three variables (2-m AGL temperature, 10-m AGL u, 10-m AGL v) is calibrated independently. Each forecast lead time (12, 24, 36 and 48 h) is also calibrated separately.

BMA weights are calculated for each forecast parameter and lead time during the training period on the 12-km domain. The optimal BMA weights for 2-m temperature forecasts are displayed in a donut chart in Figure 3. From that figure we see that the ensemble members that use the Noah land surface model generally have the highest weights early in the forecast period, followed by members that use the RUC land surface model. The members with the smallest BMA weights use the thermal diffusion and Pleim-Xu land surface schemes. As with the forecast error correlation results, this again indicates that the choice of land surface scheme has a large impact on 2-m temperature model predictions. Additionally, these results imply that the Noah and RUC land surface models yield better 2-m temperature forecasts over the training period. It should also be noted that these patterns are most prevalent for 12-h and 24-h forecasts, but the BMA weights tend to become more even at 48-h lead time.

The optimal BMA weights for the 10-m zonal wind forecasts are displayed in Figure 4. The ensemble members that use the ACM2 boundary layer scheme generally have higher weights than those with the MYJ boundary layer scheme. The members that use the YSU boundary layer scheme generally have the lowest weights. Again, as with the forecast error correlation results, this indicates that the choice of boundary layer scheme has a substantial impact on model predictions of 10-m winds.

To assess the value of the calibration provided by BMA, the equal-weighted and BMA-weighted ensembles on the 12-km domain are compared with verification metrics over the verification period. To evaluate the performance of the ensemble as a deterministic forecast, we use RMSE to compare a weighted ensemble mean (using the BMA weights) to the standard (equally weighted) ensemble mean. For each forecast lead time and variable, the RMSE for the BMA-

weighted ensemble is slightly lower than or the same as the RMSE for the equal-weighted ensemble (see Figure 5).

Rank histograms are used to evaluate the distribution of the ensemble. Figure 6a shows a verification rank histogram for the 24-h forecasts of 10-m zonal wind for the equal-weighted ensemble, and the ensemble is clearly quite underdispersive. Figure 6b shows a PIT histogram for the 24-h forecasts of 10-m zonal wind for the BMA-weighted ensemble. By comparing Fig. 6b with Fig. 6a, it is clear that the BMA-weighted ensemble is less underdispersive. In other words, the calibration performed by the BMA made the ensemble more reliable than it was before.

We use CRPS to evaluate the overall probabilistic predictions of the ensemble. CRPS is calculated over the 12-km domain for each forecast lead time and variable. The CRPS values for the equal-weighted and BMA-weighted ensembles are plotted in Figure 7, for the full ensemble and each of the four candidate subsets defined by PCA (see Table 3). For all variables, lead times and ensemble sizes, the BMA-weighted ensembles have CRPS values that are approximately 10-15% lower (better) than for the corresponding equal-weighted ensembles. This result illustrates that the BMA-weighted ensemble forecasts were substantially improved when compared to the equal-weighted ensemble forecasts. For the BMA-weighted ensembles, the CRPS slightly increases (worsens) as ensemble size decreases from 24 to 20 to 14 to 12 members, and is approximately 5-10% worse yet when ensemble size becomes very small (5 members). It is not surprising that the CRPS is noticeably poorer for mini-ensembles, because there are only a few forecast-specific data contributing to the forecast ensemble PDF.

## 5. DISCUSSION AND CONCLUSIONS

The main goal of this study is to propose an objective methodology to "down-select," or determine which subset of members should be included in an ensemble used for forecasting applications for which short-range low-level wind prediction is of primary importance. Then we use Bayesian model averaging to fill in the details of the PDF. The NWP dataset on which we demonstrate this proposed methodology is a 24-member WRF-ARW physics ensemble over 18 forecast periods during summer 2009. Our down-selection methodology centers on using principal component analysis to determine the ensemble members whose forecast errors were contributing most to variability in the forecast. Using PCA, we define four candidate ensemble subsets of various sizes, ranging from 5 to 20 members, for the nested 12-km domain. Somewhat different candidate subsets are defined using the 36-km domain.

To improve forecast reliability by statistically filling in the PDF, we then calibrate the full ensemble and each subset ensemble using Bayesian model averaging. The BMA-calibrated ensembles perform as well as or better than the equal-weighted, uncalibrated ensemble by several metrics that are computed over each forecast lead time (12, 24, 36, and 48 h) and forecast variable (2-m AGL temperature, 10-m AGL zonal wind component, and 10-m meridional wind component). First, the RMSE for the BMA-calibrated ensemble shows no change or a slight improvement for the equal-weighted ensemble. Second, rank histograms indicate that the BMA-weighted ensemble is substantially less underdispersive than the equal-weighted ensemble. Third, the CRPS for the BMA-weighted ensemble show a 10-15% improvement compared to the equal-weighted ensemble. Therefore, we conclude that the BMA calibration improves the quality of the ensemble forecasts substantially.

Correlations of both forecast errors across the ensemble and BMA weights for all the ensemble members reveal that model predictions of 2-m temperature are greatly influenced by the choice of land surface scheme, and that model predictions of 10-m winds are greatly influenced by the choice of boundary layer scheme, and to a lesser extent by the choice of cumulus parameterization scheme. Therefore, we conclude that for any subset to have sufficient variability in the structure and dynamics of the ABL, it must include diversity in land surface and boundary layer schemes, and should also have diversity in cumulus schemes.

Additionally, for the BMA-calibrated ensembles, the CRPS is roughly the same for the full ensemble (24 members), subset A (20 members), subset B (14 members), and subset C (12 members), but 5-10% worse for subset D (5 members). Thus, BMA appears to be successful at reconstructing the PDF of the down-selected ensemble, as long as a sufficient number of members is included to capture the primary sources of variability. Based on this result and the importance of including a diversity of land surface, boundary layer, and cumulus parameterization schemes, we recommend subset C as an ensemble for applications concerned with low-

level wind forecasting. Subset C does contain a diversity of physics options, and gave forecasts of similar quality as the full ensemble for half the computational cost (12 vs. 24 members).

This study is preliminary and we must acknowledge several caveats. First, it is unknown how these results depend on seasonality. The down-selection results are somewhat different for the different domains and resolutions; thus the results may differ somewhat by season as well. Second, when creating this ensemble we assumed that diversity in microphysics or radiation schemes would not add substantial variability to the ensemble. This assumption should be tested. Therefore, we plan to explore the seasonal dependence of this down-selection method by creating a larger WRF-ARW ensemble that varies additional physics schemes over a four-season period. Third, while we calibrate each of the three forecast variables independently in this study, we recognize this may not be an ideal approach, especially with the related zonal and meridional wind fields. This is a logical initial approach that we hope to refine in future research. We also plan to explore other post-processing techniques, such as k-means clustering and self-organizing maps. We also plan to investigate additional calibration techniques. Additionally, we plan to verify ensemble predictions against other observations in the ABL, rather than just surface temperature and wind.

This study describes a way to use statistical post-processing to down-select the members of an ensemble that contribute the most variability, then to "dress" the PDF of the ensemble using BMA. The combination of post-processing ensemble forecasts with PCA and BMA over a short evaluation period provides an objective method for determining which subset ensemble members should be included in a longer-term forecast ensemble. This is particularly relevant given the nearly ubiquitous constraint on computational resources that most companies, universities and research centers face. By filling in the ensemble PDF with BMA, we regain the appropriate reliability as demonstrated in the PIT histograms. While in this study we focused on down-selecting an ensemble for low-level wind forecasting applications, the same principles outlined in this study could be applied to other forecasting applications, by verifying against the relevant variables.

**ACKNOWLEDGEMENTS**

The authors would like to thank Chuck Ritter of the Penn State University Applied Research Lab for invaluable computational support, and Christian Pagé of MeteoCentre for providing the meteorological observations we used for verification. This study was sponsored by the Defense Threat Reduction Agency, contract DTRA01-03-D-0010, John Hannan, CIV, Contract Monitor. Authors Jared Lee and Tyler McCandless are also grateful for funding from the Penn State University Applied Research Lab Exploratory & Foundational Program to support their graduate studies.

**References**

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

Grimit, E.P., and C.F. Mass, 2007: Measuring the ensemble spread-error relationship with a probabilistic approach: Stochastic ensemble results. *Mon. Wea. Rev.*, **135**, 203-221.

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

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. Submitted to *Mon. Wea. Rev.*, December 2010.

Lee, J.A., L.J Peltier, S.E. Haupt, J.C. Wyngaard, D.R. Stauffer, and A. Deng, 2009: Improving SCIPUFF dispersion forecasts with NWP ensembles. *J. Appl. Meteor. Climat.*, **48**, 2305-2319.

Lewellen, W.S., and R.I. Sykes, 1989: Meteorological data needs for modeling air quality uncertainties. *J. Atmos. Ocean Tech.*, **6**, 759-768.

Liu, Y., M. Xu, J. Hacker, T. Warner, and S. Swerdlin, 2007: A WRF and MM5-based 4-D mesoscale ensemble data analysis and prediction

system (E-RTFDDA) developed for ATEC operational applications. *18th Conf. on Numerical Weather Prediction*, AMS, 25-29 June 2007. Park City, UT. 8pp.

Mierswa, I., M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, 2006: YALE: Rapid Prototyping for Complex Data Mining Tasks. In Tina Eliassi-Rad and Lyle H. Ungar and Mark Craven and Dimitrios Gunopulos (eds.), *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, 935-940, New York, USA, ACM Press.

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.

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.

Sloughter, J.M, T. Gneiting and A.E. Raftery, 2010. Probabilistic Wind Speed Forecasting Using Ensembles and Bayesian Model Averaging. *J. Amer. Stat. Assoc.*, **105**, 25-35.

Sykes, R. I., S. F. Parker, and D. S. Henn, 2004: SCIPUFF version 2.0, technical documentation. A.R.A.P. Tech. Rep. 727, Titan Corporation, Princeton, NJ, 284 pp.

Warner, T. T., R.-S. Sheu, J. F. Bowers, R. I. Sykes, G.C. Dodd, and D. S. Henn, 2002: Ensemble simulations with coupled atmospheric dynamic and dispersion models: Illustrating uncertainties in dosage simulations. *J. Appl. Meteor.*, **41**, 488–504.

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.

Wyngaard, J.C., 2010: Turbulence in the atmosphere, Cambridge University Press, Cambridge, UK, 393 pp.

**TABLES AND FIGURES**

TABLE 1. Physics parameterizations for the WRF-ARW ensemble members used in this study.

| Member | Microphysics | Longwave Radiation | Shortwave Radiation | Land Surface | Surface Layer | Boundary Layer | Cumulus |
|---|---|---|---|---|---|---|---|
| 01 | WSM 5-class | RRTMG | RRTMG | Thermal Diff. | MM5 Similarity | YSU | Kain-Fritsch |
| 02 | WSM 5-class | RRTMG | RRTMG | Thermal Diff. | MM5 Similarity | YSU | Grell-Devenyi |
| 03 | WSM 5-class | RRTMG | RRTMG | Noah | MM5 Similarity | YSU | Kain-Fritsch |
| 04 | WSM 5-class | RRTMG | RRTMG | Noah | MM5 Similarity | YSU | Grell-Devenyi |
| 05 | WSM 5-class | RRTMG | RRTMG | RUC | MM5 Similarity | YSU | Kain-Fritsch |
| 06 | WSM 5-class | RRTMG | RRTMG | RUC | MM5 Similarity | YSU | Grell-Devenyi |
| 07 | WSM 5-class | RRTMG | RRTMG | Pleim-Xu | MM5 Similarity | YSU | Kain-Fritsch |
| 08 | WSM 5-class | RRTMG | RRTMG | Pleim-Xu | MM5 Similarity | YSU | Grell-Devenyi |
| 09 | WSM 5-class | RRTMG | RRTMG | Thermal Diff. | Eta Similarity | MYJ | Kain-Fritsch |
| 10 | WSM 5-class | RRTMG | RRTMG | Thermal Diff. | Eta Similarity | MYJ | Grell-Devenyi |
| 11 | WSM 5-class | RRTMG | RRTMG | Noah | Eta Similarity | MYJ | Kain-Fritsch |
| 12 | WSM 5-class | RRTMG | RRTMG | Noah | Eta Similarity | MYJ | Grell-Devenyi |
| 13 | WSM 5-class | RRTMG | RRTMG | RUC | Eta Similarity | MYJ | Kain-Fritsch |
| 14 | WSM 5-class | RRTMG | RRTMG | RUC | Eta Similarity | MYJ | Grell-Devenyi |
| 15 | WSM 5-class | RRTMG | RRTMG | Pleim-Xu | Eta Similarity | MYJ | Kain-Fritsch |
| 16 | WSM 5-class | RRTMG | RRTMG | Pleim-Xu | Eta Similarity | MYJ | Grell-Devenyi |
| 17 | WSM 5-class | RRTMG | RRTMG | Thermal Diff. | Pleim-Xu | ACM2 | Kain-Fritsch |
| 18 | WSM 5-class | RRTMG | RRTMG | Thermal Diff. | Pleim-Xu | ACM2 | Grell-Devenyi |
| 19 | WSM 5-class | RRTMG | RRTMG | Noah | Pleim-Xu | ACM2 | Kain-Fritsch |
| 20 | WSM 5-class | RRTMG | RRTMG | Noah | Pleim-Xu | ACM2 | Grell-Devenyi |
| 21 | WSM 5-class | RRTMG | RRTMG | RUC | Pleim-Xu | ACM2 | Kain-Fritsch |
| 22 | WSM 5-class | RRTMG | RRTMG | RUC | Pleim-Xu | ACM2 | Grell-Devenyi |
| 23 | WSM 5-class | RRTMG | RRTMG | Pleim-Xu | Pleim-Xu | ACM2 | Kain-Fritsch |
| 24 | WSM 5-class | RRTMG | RRTMG | Pleim-Xu | Pleim-Xu | ACM2 | Grell-Devenyi |

TABLE 2. Randomly chosen forecast initialization times for the training set (italics) and verification set (bold).

| | | |
|---|---|---|
| **2009-06-06_12** | **2009-07-11_00** | **2009-08-01_12** |
| *2009-06-07_00* | *2009-07-19_12* | *2009-08-06_12* |
| *2009-06-16_00* | *2009-07-22_12* | *2009-08-15_00* |
| *2009-06-17_00* | *2009-07-23_00* | *2009-08-17_00* |
| *2009-06-23_12* | **2009-07-27_12** | **2009-08-18_00** |
| *2009-06-29_12* | *2009-07-31_00* | **2009-08-28_12** |

TABLE 3. Summary of which members were included in each candidate ensemble subset for the 12-km domain. See text for description of how each subset was defined.

| Subset | Ensemble Size | Ensemble Members |
|---|---|---|
| **A** | 20 members | 01, 02, 03, 04, 05, 06, 08, 09, 10, 12, 13, 14, 16, 17, 18, 19, 20, 22, 23, 24 |
| **B** | 14 members | 01, 03, 05, 06, 08, 13, 14, 16, 18, 19, 20, 22, 23, 24 |
| **C** | 12 members | 01, 03, 05, 06, 13, 14, 16, 19, 20, 22, 23, 24 |
| **D** | 5 members | 05, 06, 13, 14, 22 |

TABLE 4. Error correlations between every ensemble member for 24-h forecasts of 2-m temperature on the 12-km domain. Warm colors (red and orange) denote stronger correlations, and cool colors (yellow and green) denote weaker correlations.

| | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **01** | 1.000 | 0.944 | 0.833 | 0.779 | 0.780 | 0.736 | 0.873 | 0.833 | 0.905 | 0.861 | 0.781 | 0.754 | 0.846 | 0.798 | 0.862 | 0.821 | 0.899 | 0.881 | 0.748 | 0.734 | 0.823 | 0.798 | 0.854 | 0.826 |
| **02** | 0.944 | 1.000 | 0.820 | 0.836 | 0.771 | 0.790 | 0.856 | 0.874 | 0.874 | 0.884 | 0.768 | 0.782 | 0.840 | 0.828 | 0.857 | 0.856 | 0.879 | 0.896 | 0.748 | 0.778 | 0.813 | 0.826 | 0.836 | 0.846 |
| **03** | 0.833 | 0.820 | 1.000 | 0.933 | 0.787 | 0.751 | 0.744 | 0.722 | 0.776 | 0.771 | 0.861 | 0.841 | 0.794 | 0.755 | 0.727 | 0.704 | 0.775 | 0.756 | 0.849 | 0.818 | 0.800 | 0.783 | 0.662 | 0.634 |
| **04** | 0.779 | 0.836 | 0.933 | 1.000 | 0.763 | 0.799 | 0.717 | 0.754 | 0.733 | 0.780 | 0.819 | 0.861 | 0.766 | 0.772 | 0.706 | 0.730 | 0.740 | 0.757 | 0.813 | 0.846 | 0.767 | 0.796 | 0.634 | 0.644 |
| **05** | 0.780 | 0.771 | 0.787 | 0.763 | 1.000 | 0.944 | 0.684 | 0.674 | 0.695 | 0.699 | 0.742 | 0.741 | 0.850 | 0.821 | 0.690 | 0.679 | 0.711 | 0.714 | 0.761 | 0.755 | 0.861 | 0.855 | 0.651 | 0.649 |
| **06** | 0.736 | 0.790 | 0.751 | 0.799 | 0.944 | 1.000 | 0.661 | 0.704 | 0.664 | 0.717 | 0.722 | 0.765 | 0.829 | 0.837 | 0.678 | 0.707 | 0.689 | 0.723 | 0.742 | 0.781 | 0.825 | 0.868 | 0.629 | 0.661 |
| **07** | 0.873 | 0.856 | 0.744 | 0.717 | 0.684 | 0.661 | 1.000 | 0.952 | 0.870 | 0.850 | 0.700 | 0.684 | 0.797 | 0.758 | 0.931 | 0.898 | 0.841 | 0.836 | 0.654 | 0.650 | 0.740 | 0.719 | 0.925 | 0.901 |
| **08** | 0.833 | 0.874 | 0.722 | 0.754 | 0.674 | 0.704 | 0.952 | 1.000 | 0.836 | 0.869 | 0.673 | 0.706 | 0.777 | 0.781 | 0.906 | 0.927 | 0.823 | 0.848 | 0.642 | 0.677 | 0.724 | 0.742 | 0.891 | 0.916 |
| **09** | 0.905 | 0.874 | 0.776 | 0.733 | 0.695 | 0.664 | 0.870 | 0.836 | 1.000 | 0.929 | 0.807 | 0.759 | 0.865 | 0.807 | 0.928 | 0.876 | 0.932 | 0.903 | 0.732 | 0.703 | 0.811 | 0.778 | 0.886 | 0.852 |
| **10** | 0.861 | 0.884 | 0.771 | 0.780 | 0.699 | 0.717 | 0.850 | 0.869 | 0.929 | 1.000 | 0.764 | 0.812 | 0.845 | 0.874 | 0.896 | 0.927 | 0.903 | 0.912 | 0.719 | 0.728 | 0.803 | 0.805 | 0.846 | 0.854 |
| **11** | 0.781 | 0.768 | 0.861 | 0.819 | 0.742 | 0.722 | 0.700 | 0.673 | 0.807 | 0.764 | 1.000 | 0.912 | 0.814 | 0.724 | 0.766 | 0.707 | 0.760 | 0.738 | 0.920 | 0.878 | 0.751 | 0.735 | 0.703 | 0.668 |
| **12** | 0.754 | 0.782 | 0.841 | 0.861 | 0.741 | 0.765 | 0.684 | 0.706 | 0.759 | 0.812 | 0.912 | 1.000 | 0.774 | 0.787 | 0.736 | 0.757 | 0.738 | 0.752 | 0.881 | 0.897 | 0.738 | 0.762 | 0.671 | 0.675 |
| **13** | 0.846 | 0.840 | 0.794 | 0.766 | 0.850 | 0.829 | 0.797 | 0.777 | 0.865 | 0.845 | 0.814 | 0.774 | 1.000 | 0.918 | 0.842 | 0.806 | 0.843 | 0.829 | 0.763 | 0.735 | 0.877 | 0.850 | 0.791 | 0.771 |
| **14** | 0.798 | 0.828 | 0.755 | 0.772 | 0.821 | 0.837 | 0.758 | 0.781 | 0.807 | 0.874 | 0.724 | 0.787 | 0.918 | 1.000 | 0.795 | 0.837 | 0.820 | 0.837 | 0.714 | 0.728 | 0.853 | 0.860 | 0.746 | 0.759 |
| **15** | 0.862 | 0.857 | 0.727 | 0.706 | 0.690 | 0.678 | 0.931 | 0.906 | 0.928 | 0.896 | 0.766 | 0.736 | 0.842 | 0.795 | 1.000 | 0.937 | 0.883 | 0.872 | 0.703 | 0.690 | 0.766 | 0.747 | 0.941 | 0.913 |
| **16** | 0.821 | 0.856 | 0.704 | 0.730 | 0.679 | 0.707 | 0.898 | 0.927 | 0.876 | 0.927 | 0.707 | 0.757 | 0.806 | 0.837 | 0.937 | 1.000 | 0.855 | 0.874 | 0.666 | 0.693 | 0.745 | 0.763 | 0.894 | 0.915 |
| **17** | 0.899 | 0.879 | 0.775 | 0.740 | 0.711 | 0.689 | 0.841 | 0.823 | 0.932 | 0.903 | 0.760 | 0.738 | 0.843 | 0.820 | 0.883 | 0.855 | 1.000 | 0.962 | 0.760 | 0.735 | 0.859 | 0.824 | 0.864 | 0.837 |
| **18** | 0.881 | 0.896 | 0.756 | 0.757 | 0.714 | 0.723 | 0.836 | 0.848 | 0.903 | 0.912 | 0.738 | 0.752 | 0.829 | 0.837 | 0.872 | 0.874 | 0.962 | 1.000 | 0.740 | 0.769 | 0.834 | 0.850 | 0.848 | 0.866 |
| **19** | 0.748 | 0.748 | 0.849 | 0.813 | 0.761 | 0.742 | 0.654 | 0.642 | 0.732 | 0.719 | 0.920 | 0.881 | 0.763 | 0.714 | 0.703 | 0.666 | 0.760 | 0.740 | 1.000 | 0.937 | 0.775 | 0.762 | 0.677 | 0.648 |
| **20** | 0.734 | 0.778 | 0.818 | 0.846 | 0.755 | 0.781 | 0.650 | 0.677 | 0.703 | 0.728 | 0.878 | 0.897 | 0.735 | 0.728 | 0.690 | 0.693 | 0.735 | 0.769 | 0.937 | 1.000 | 0.742 | 0.788 | 0.664 | 0.686 |
| **21** | 0.823 | 0.813 | 0.800 | 0.767 | 0.861 | 0.825 | 0.740 | 0.724 | 0.811 | 0.803 | 0.751 | 0.738 | 0.877 | 0.853 | 0.766 | 0.745 | 0.859 | 0.834 | 0.775 | 0.742 | 1.000 | 0.951 | 0.719 | 0.700 |
| **22** | 0.798 | 0.826 | 0.783 | 0.796 | 0.855 | 0.868 | 0.719 | 0.742 | 0.778 | 0.805 | 0.735 | 0.762 | 0.850 | 0.860 | 0.747 | 0.763 | 0.824 | 0.850 | 0.762 | 0.788 | 0.951 | 1.000 | 0.694 | 0.723 |
| **23** | 0.854 | 0.836 | 0.662 | 0.634 | 0.651 | 0.629 | 0.925 | 0.891 | 0.886 | 0.846 | 0.703 | 0.671 | 0.791 | 0.746 | 0.941 | 0.894 | 0.864 | 0.848 | 0.677 | 0.664 | 0.719 | 0.694 | 1.000 | 0.954 |
| **24** | 0.826 | 0.846 | 0.634 | 0.644 | 0.649 | 0.661 | 0.901 | 0.916 | 0.852 | 0.854 | 0.668 | 0.675 | 0.771 | 0.759 | 0.913 | 0.915 | 0.837 | 0.866 | 0.648 | 0.686 | 0.700 | 0.723 | 0.954 | 1.000 |

TABLE 5. Same as Table 4, but for 10-m zonal wind.

| | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **01** | 1.000 | 0.920 | 0.895 | 0.838 | 0.875 | 0.827 | 0.910 | 0.855 | 0.845 | 0.799 | 0.815 | 0.784 | 0.784 | 0.745 | 0.835 | 0.795 | 0.881 | 0.842 | 0.860 | 0.800 | 0.840 | 0.797 | 0.854 | 0.799 |
| **02** | 0.920 | 1.000 | 0.860 | 0.884 | 0.835 | 0.876 | 0.868 | 0.889 | 0.821 | 0.825 | 0.799 | 0.813 | 0.775 | 0.775 | 0.810 | 0.819 | 0.851 | 0.859 | 0.838 | 0.835 | 0.816 | 0.829 | 0.824 | 0.824 |
| **03** | 0.895 | 0.860 | 1.000 | 0.900 | 0.946 | 0.879 | 0.904 | 0.848 | 0.871 | 0.810 | 0.868 | 0.828 | 0.829 | 0.792 | 0.856 | 0.804 | 0.877 | 0.825 | 0.891 | 0.803 | 0.870 | 0.806 | 0.836 | 0.756 |
| **04** | 0.838 | 0.884 | 0.900 | 1.000 | 0.868 | 0.916 | 0.847 | 0.864 | 0.832 | 0.835 | 0.836 | 0.849 | 0.804 | 0.801 | 0.814 | 0.839 | 0.827 | 0.838 | 0.839 | 0.861 | 0.821 | 0.842 | 0.795 | 0.803 |
| **05** | 0.875 | 0.835 | 0.946 | 0.868 | 1.000 | 0.906 | 0.888 | 0.825 | 0.850 | 0.797 | 0.828 | 0.799 | 0.828 | 0.792 | 0.832 | 0.787 | 0.857 | 0.811 | 0.855 | 0.791 | 0.890 | 0.827 | 0.808 | 0.745 |
| **06** | 0.827 | 0.876 | 0.879 | 0.916 | 0.906 | 1.000 | 0.835 | 0.864 | 0.817 | 0.843 | 0.813 | 0.835 | 0.821 | 0.832 | 0.806 | 0.821 | 0.821 | 0.831 | 0.822 | 0.825 | 0.840 | 0.871 | 0.781 | 0.785 |
| **07** | 0.910 | 0.868 | 0.904 | 0.847 | 0.888 | 0.835 | 1.000 | 0.911 | 0.863 | 0.806 | 0.830 | 0.784 | 0.795 | 0.751 | 0.876 | 0.822 | 0.866 | 0.825 | 0.858 | 0.788 | 0.827 | 0.776 | 0.889 | 0.828 |
| **08** | 0.855 | 0.889 | 0.848 | 0.864 | 0.825 | 0.864 | 0.911 | 1.000 | 0.827 | 0.841 | 0.810 | 0.815 | 0.778 | 0.783 | 0.844 | 0.861 | 0.829 | 0.831 | 0.823 | 0.806 | 0.789 | 0.806 | 0.850 | 0.856 |
| **09** | 0.845 | 0.821 | 0.871 | 0.832 | 0.850 | 0.817 | 0.863 | 0.827 | 1.000 | 0.875 | 0.940 | 0.861 | 0.911 | 0.824 | 0.959 | 0.871 | 0.885 | 0.839 | 0.854 | 0.799 | 0.848 | 0.805 | 0.858 | 0.792 |
| **10** | 0.799 | 0.825 | 0.810 | 0.835 | 0.797 | 0.843 | 0.806 | 0.841 | 0.875 | 1.000 | 0.855 | 0.913 | 0.841 | 0.896 | 0.857 | 0.929 | 0.833 | 0.853 | 0.811 | 0.814 | 0.797 | 0.820 | 0.801 | 0.811 |
| **11** | 0.815 | 0.799 | 0.868 | 0.836 | 0.828 | 0.813 | 0.830 | 0.810 | 0.940 | 0.855 | 1.000 | 0.889 | 0.916 | 0.827 | 0.919 | 0.850 | 0.858 | 0.824 | 0.875 | 0.819 | 0.844 | 0.809 | 0.841 | 0.781 |
| **12** | 0.784 | 0.813 | 0.828 | 0.849 | 0.799 | 0.835 | 0.784 | 0.815 | 0.861 | 0.913 | 0.889 | 1.000 | 0.846 | 0.879 | 0.842 | 0.879 | 0.821 | 0.837 | 0.836 | 0.846 | 0.815 | 0.832 | 0.792 | 0.798 |
| **13** | 0.784 | 0.775 | 0.829 | 0.804 | 0.828 | 0.821 | 0.795 | 0.778 | 0.911 | 0.841 | 0.916 | 0.846 | 1.000 | 0.868 | 0.899 | 0.834 | 0.830 | 0.802 | 0.814 | 0.769 | 0.841 | 0.810 | 0.785 | 0.742 |
| **14** | 0.745 | 0.775 | 0.792 | 0.801 | 0.792 | 0.832 | 0.751 | 0.783 | 0.824 | 0.896 | 0.827 | 0.879 | 0.868 | 1.000 | 0.812 | 0.866 | 0.797 | 0.801 | 0.777 | 0.777 | 0.804 | 0.817 | 0.739 | 0.745 |
| **15** | 0.835 | 0.810 | 0.856 | 0.814 | 0.832 | 0.806 | 0.876 | 0.844 | 0.959 | 0.857 | 0.919 | 0.842 | 0.899 | 0.812 | 1.000 | 0.876 | 0.872 | 0.822 | 0.848 | 0.783 | 0.842 | 0.790 | 0.877 | 0.807 |
| **16** | 0.795 | 0.819 | 0.804 | 0.839 | 0.787 | 0.821 | 0.822 | 0.861 | 0.871 | 0.929 | 0.850 | 0.879 | 0.834 | 0.866 | 0.876 | 1.000 | 0.825 | 0.834 | 0.801 | 0.807 | 0.792 | 0.803 | 0.815 | 0.838 |
| **17** | 0.881 | 0.851 | 0.877 | 0.827 | 0.857 | 0.821 | 0.866 | 0.829 | 0.885 | 0.833 | 0.858 | 0.821 | 0.830 | 0.797 | 0.872 | 0.825 | 1.000 | 0.926 | 0.942 | 0.867 | 0.926 | 0.862 | 0.911 | 0.839 |
| **18** | 0.842 | 0.859 | 0.825 | 0.838 | 0.811 | 0.831 | 0.825 | 0.831 | 0.839 | 0.853 | 0.824 | 0.837 | 0.802 | 0.801 | 0.822 | 0.834 | 0.926 | 1.000 | 0.883 | 0.909 | 0.858 | 0.909 | 0.857 | 0.882 |
| **19** | 0.860 | 0.838 | 0.891 | 0.839 | 0.855 | 0.822 | 0.858 | 0.823 | 0.854 | 0.811 | 0.875 | 0.836 | 0.814 | 0.777 | 0.848 | 0.801 | 0.942 | 0.883 | 1.000 | 0.887 | 0.923 | 0.852 | 0.910 | 0.834 |
| **20** | 0.800 | 0.835 | 0.803 | 0.861 | 0.791 | 0.825 | 0.788 | 0.806 | 0.799 | 0.814 | 0.819 | 0.846 | 0.769 | 0.777 | 0.783 | 0.807 | 0.867 | 0.909 | 0.887 | 1.000 | 0.834 | 0.899 | 0.843 | 0.879 |
| **21** | 0.840 | 0.816 | 0.870 | 0.821 | 0.890 | 0.840 | 0.827 | 0.789 | 0.848 | 0.797 | 0.844 | 0.815 | 0.841 | 0.804 | 0.842 | 0.792 | 0.926 | 0.858 | 0.923 | 0.834 | 1.000 | 0.891 | 0.861 | 0.787 |
| **22** | 0.797 | 0.829 | 0.806 | 0.842 | 0.827 | 0.871 | 0.776 | 0.806 | 0.805 | 0.820 | 0.809 | 0.832 | 0.810 | 0.817 | 0.790 | 0.803 | 0.862 | 0.909 | 0.852 | 0.899 | 0.891 | 1.000 | 0.808 | 0.848 |
| **23** | 0.854 | 0.824 | 0.836 | 0.795 | 0.808 | 0.781 | 0.889 | 0.850 | 0.858 | 0.801 | 0.841 | 0.792 | 0.785 | 0.739 | 0.877 | 0.815 | 0.911 | 0.857 | 0.910 | 0.843 | 0.861 | 0.808 | 1.000 | 0.892 |
| **24** | 0.799 | 0.824 | 0.756 | 0.803 | 0.745 | 0.785 | 0.828 | 0.856 | 0.792 | 0.811 | 0.781 | 0.798 | 0.742 | 0.745 | 0.807 | 0.838 | 0.839 | 0.882 | 0.834 | 0.879 | 0.787 | 0.848 | 0.892 | 1.000 |

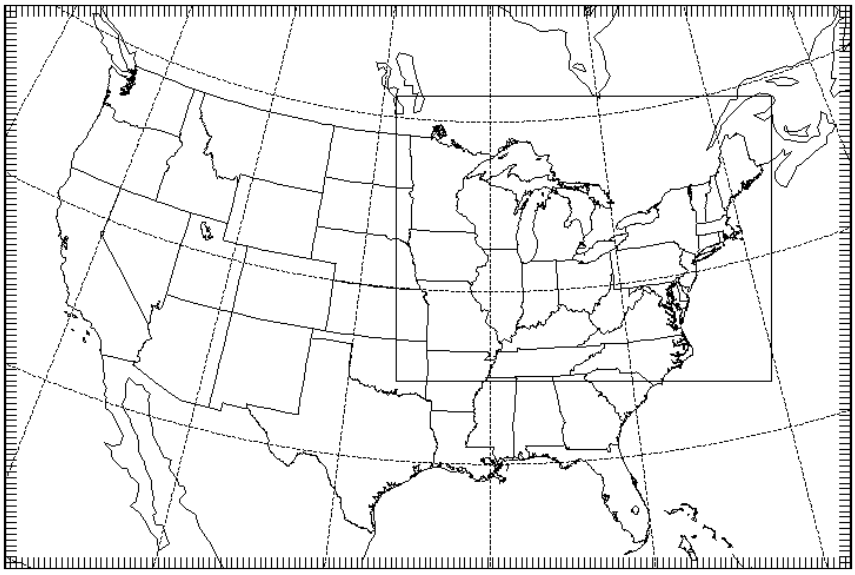

FIG. 1. Geographical domains used by the WRF-ARW ensemble.

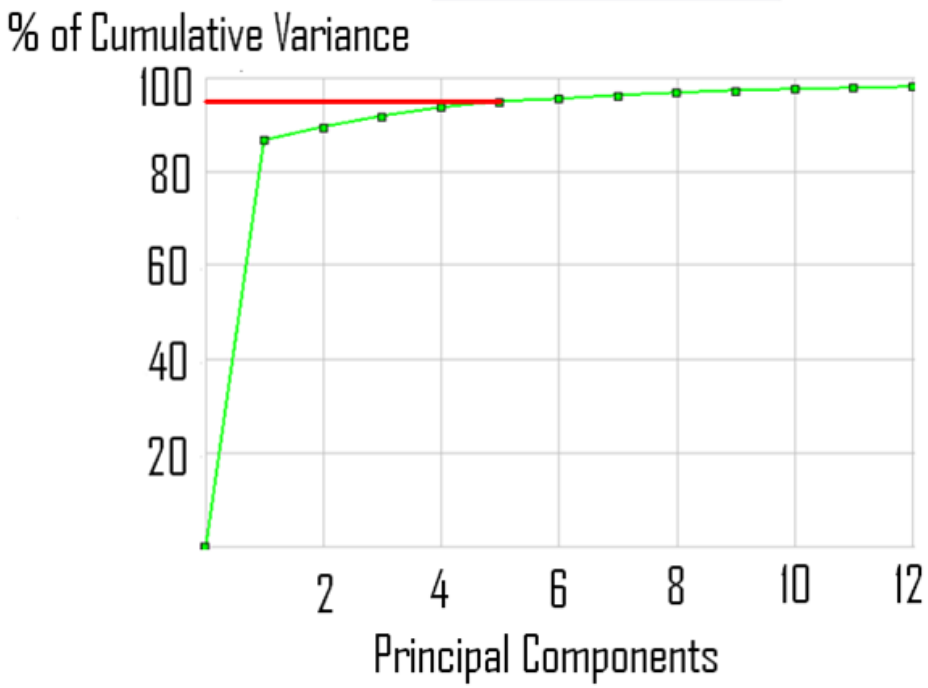


FIG. 2. Plot of the cumulative variance accounted for by the principal components for the 12-h 2-m AGL temperature forecasts. The solid red line indicates the 95% level used by subsets A and D.

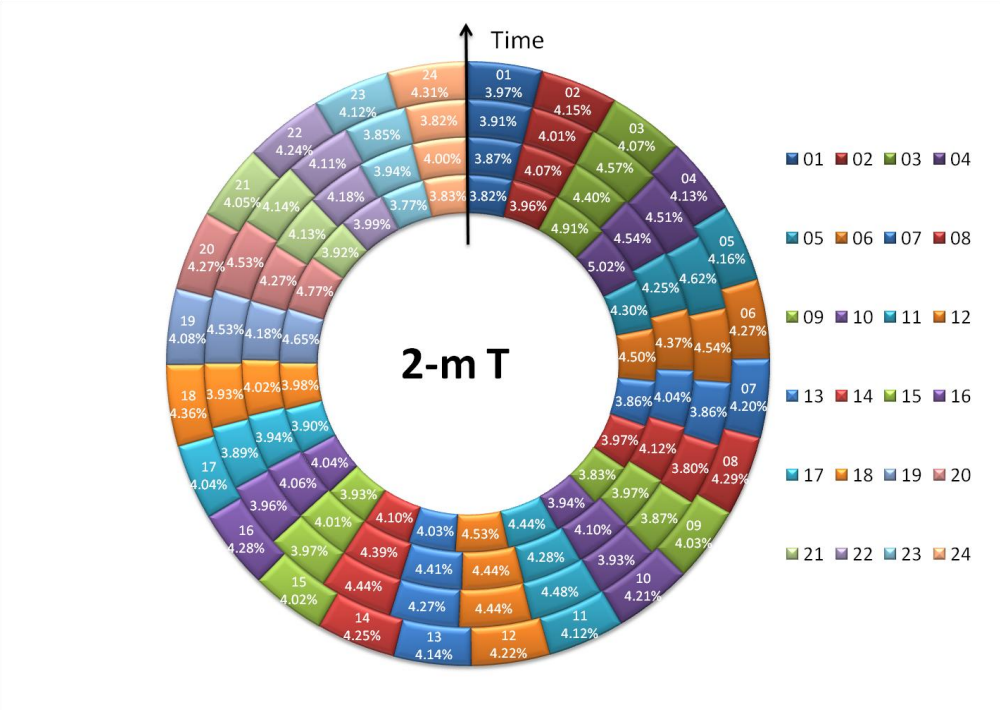


FIG. 3. Donut chart showing the BMA weights for each ensemble member for 2-m temperature forecasts on the 12-km domain. The innermost ring is for 12-h forecasts, the second ring for 24-h forecasts, the third ring for 36-h forecasts, and the outermost ring is for 48-h forecasts.

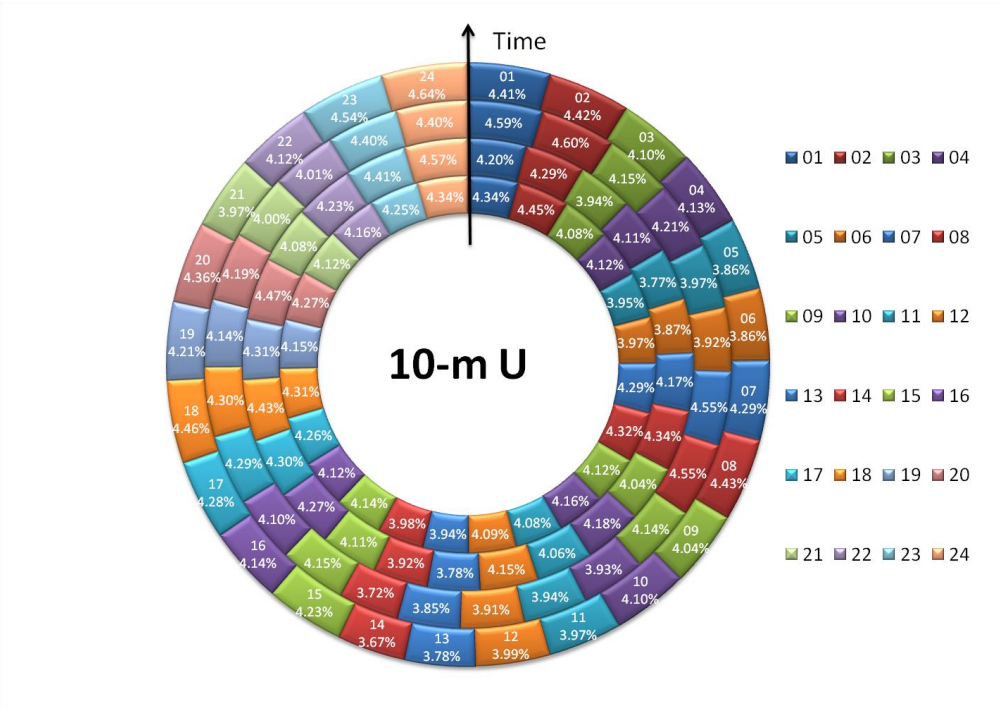


FIG. 4. Same as Fig. 3, but for 10-m zonal wind forecasts.

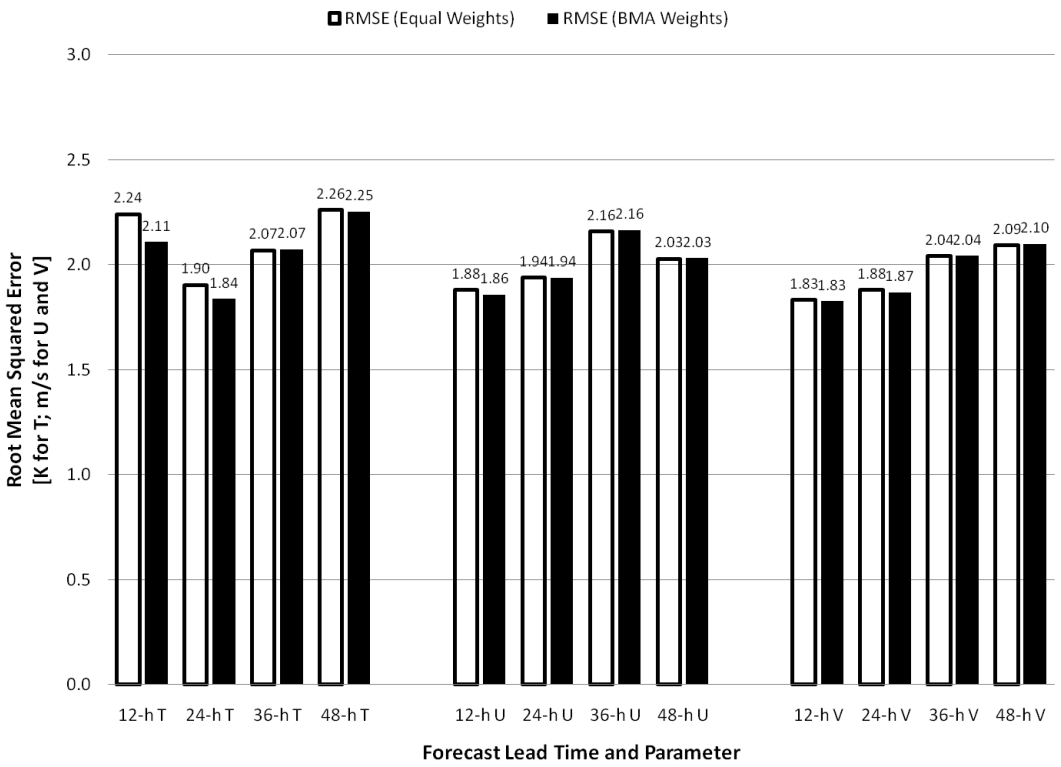


FIG. 5. Root-mean-square error over the 12-km domain during the verification period, for each forecast variable and lead time, for the equal-weighted ensemble (unfilled) and BMA-weighted ensemble (filled).

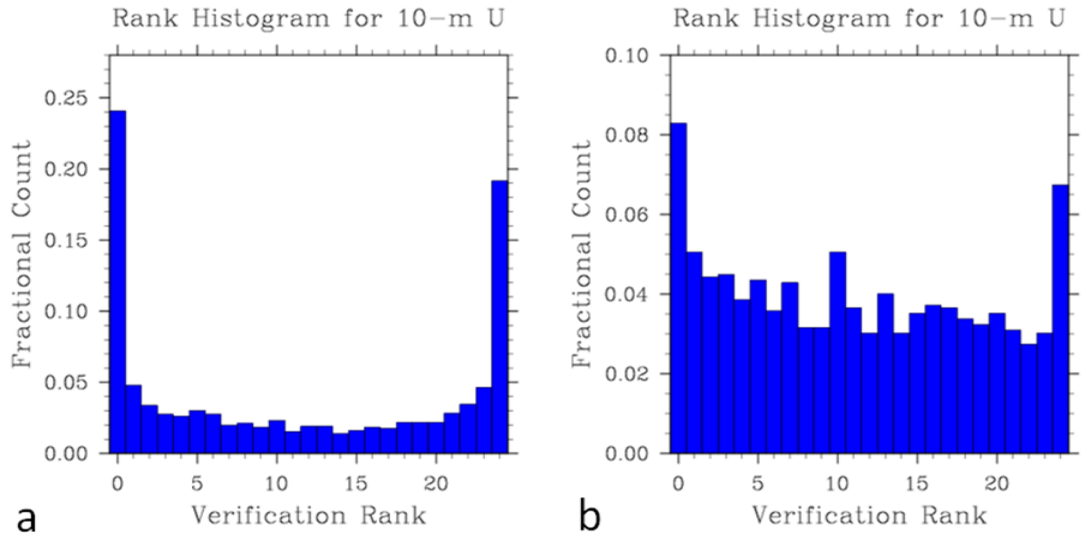


FIG. 6. Histograms for the 10-m zonal wind forecasts on the 12-km domain during the verification period. (a) is a verification rank histogram for the equal-weighted ensemble, and (b) is a probability integral transform (PIT) histogram for the BMA-weighted ensemble. In both (a) and (b), because the ensemble size is 24, each of the 25 bins would be at 0.04 if the histogram were flat.

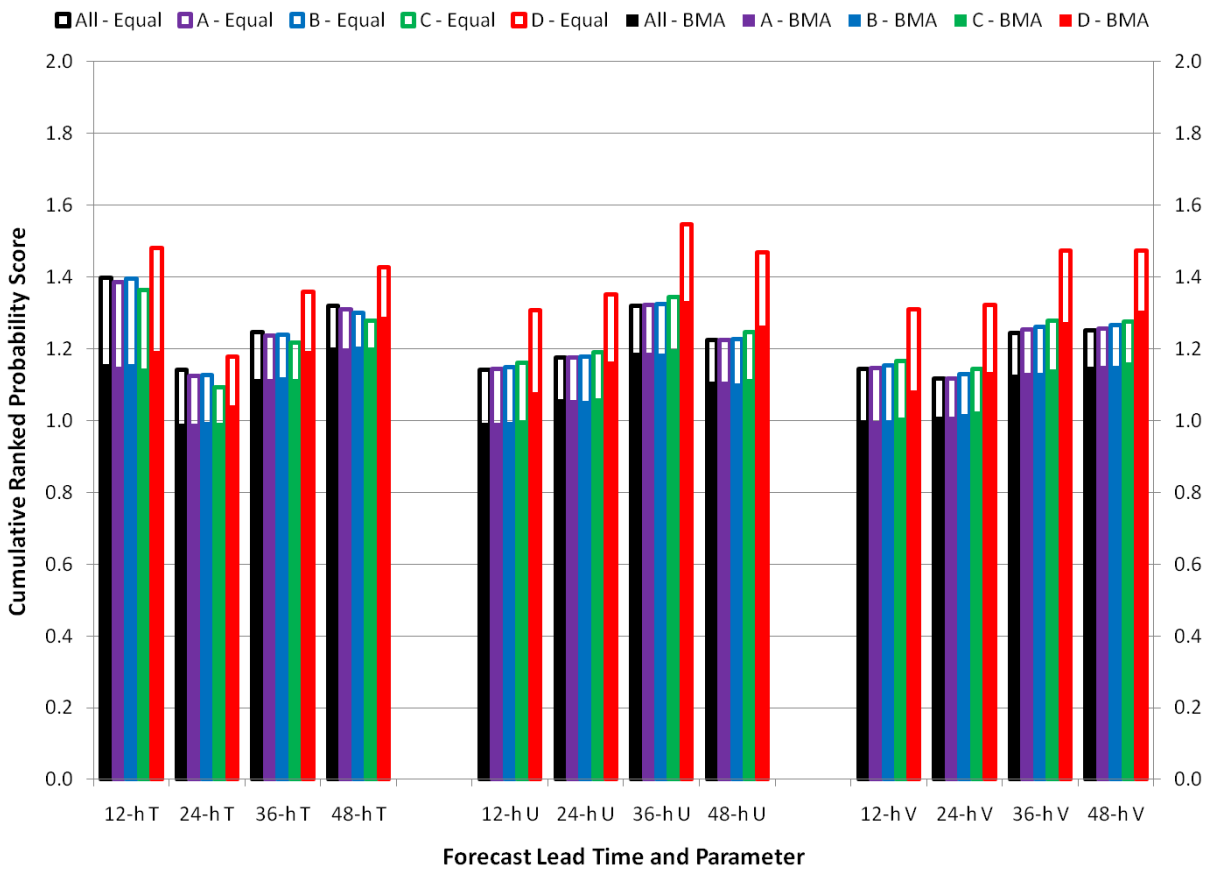


FIG. 7. Cumulative ranked probability scores for each forecast lead time (12 h, 24 h, 36 h, and 48 h) and parameter (2-m temperature, 10-m zonal and 10-m meridional winds). Unfilled bars correspond to equally-weighted ensembles, and filled bars correspond to BMA-weighted ensembles. The black bars are for the full ensemble, purple for subset A, blue for subset B, green for subset C, and red for subset D.