12.3 EVALUATING NWP ENSEMBLE CONFIGURATIONS FOR AT&D APPLICATIONS

Jared A. Lee*, Walter C. Kolczynski, Tyler C. McCandless, Kerrie J. Long,
Sue Ellen Haupt, David R. Stauffer, and Aijun Deng
The Pennsylvania State University, University Park, PA

1. INTRODUCTION

Due to the chaotic nature of the atmosphere, there are inherent limitations to forecasting a single realization of the future atmospheric state. While numerical weather prediction (NWP) models have become more advanced in recent years in representing and predicting 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. Grimit and Mass (2002) state that there is a strong positive correlation between the ensemble spread and forecast error for short-range mesoscale NWP. For example, if the spread across the ensemble is low, the forecast error (and hence, the forecast uncertainty) is generally low, and vice-versa.

There are a few different approaches to ensemble initialization that are documented in the literature. Each approach attempts to account for uncertainty in the forecast. Several approaches are designed specifically to account for uncertainty in the initial conditions. Some of these include using bred vectors, singular vectors, an ensemble Kalman filter, ensemble transform Kalman filter, or the ensemble transform approach. There are several studies in the literature that compare the relative performance of ensembles that are initialized with these different methods, but the ensemble Kalman filter has generally performed best (Wang and Bishop 2003; Buizza et al. 2005; Bowler 2006; Descamps and Talagrand 2006; Wei et al. 2006).

Another source of uncertainty for regional NWP (or limited-area models) arises from the specification of the lateral boundary conditions. Warner et al. (1997) summarize many of the issues that modelers must consider with regard to LBCs. Notably, errors that arise due to the LBCs "sweep" across the model domain through the integration period, which can degrade model results and constrain ensemble dispersion (Nutter et al. 2004). Therefore, if there is a particular region that is of interest to the modeler, that region should be placed far enough away from the lateral boundaries so that errors from the boundary will not have advected into that region at the time of interest.

* *Corresponding author address:* Jared A. Lee, The Pennsylvania State University, Department of Meteorology, 503 Walker Building, University Park, PA 16802; e-mail: jal488@meteo.psu.edu.

In addition to ICs and LBCs, another source of uncertainty in NWP results from the model physics parameterizations. Because NWP models are unable to resolve many physical processes that occur, such as convection, cloud and ice microphysics, atmospheric radiation, as well as processes in the atmospheric boundary layer (ABL), parameterization schemes for these processes are necessary. The land surface must be represented by discrete categories of typical soil and vegetation types, as well as soil moisture profiles. All of these processes must be approximated. These approximations introduce some amount of error to NWP solutions that is unavoidable. Judging by the focus of the major national centers in the 1990s on creating IC ensembles (*e.g.*, Toth and Kalnay 1993; Molteni *et al.* 1996), it was once thought that IC uncertainty dwarfed physics uncertainty in importance, at least for global forecast models. Many studies in recent years, however, have shown the importance of physics uncertainty in NWP, particularly in limited-area models. As an example, studies investigating the impacts of physics parameterization schemes on NWP forecasts have been conducted for a variety of situations, including for predictions of the southwest monsoon (Bright and Mullen 2002), mesoscale convective systems (Jankov *et al.* 2005), and the passage of a mid-latitude cyclone (Deng and Stauffer 2006), just to name a few.

Several studies have also investigated the performance of ensembles that incorporate multiple sources of uncertainty. Warner et al. (2000) and Jones et al. (2007) showed that physics variability is important in cases with weak synoptic forcing, and IC variability is important in cases with strong synoptic forcing. Physics variability can also be more important than IC variability in increasing ensemble spread in the very short term (6-12 h) (Stensrud et al. 2000), and can have a greater impact on near-surface meteorological parameters than IC variability (Eckel and Mass 2005). Fujita et al. (2007) found that spread was greater for dynamic variables in an IC ensemble, and greater for thermodynamic variables in a physics ensemble. They concluded that because the distribution of spread in the IC and physics ensembles they created was so different, the ensembles likely covered different portions of the probability density function (PDF) of the atmospheric state. They recommended using a combined IC/physics ensemble as the best choice for severe weather and boundary layer forecasting, since it incorporates variability from multiple sources of uncertainty.

One of the main messages that can be gleaned from these studies is that it is critically important to include both IC/LBC and physics uncertainty in short-range (0-48 h) NWP ensembles, in order to sample the forecast PDF of the atmospheric state more accurately.

Current operational short-range ensemble forecast (SREF) systems, such as NCEP SREF (Du et al. 2004) and the UK Met Office Global and Regional Ensemble Prediction System (Bowler et al. 2008), incorporate these multiple sources of uncertainty by varying or perturbing the ICs, LBCs and physics schemes. Another common thread among nearly all the ensemble studies discussed above is that they deal primarily with obtaining spread in precipitation forecasts. It is not clear, however, whether the strategies employed in these studies would be the best strategies for obtaining spread in atmospheric transport and dispersion (AT&D) applications.

To obtain appropriate spread in concentration predictions from AT&D models, there should be good spread in low-level wind direction and ABL depth, as these are two of the most important parameters affecting uncertainty in AT&D predictions (Lewellen and Sykes 1989). A previous study by Lee et al. (2009) modeled concentration predictions from an actual tracer release with an NWP "ensemble of opportunity" that varied only certain physics parameterizations and data assimilation schemes. They found that variability in the wind angle in this physics "ensemble" was insufficient to yield enough spread in the concentration predictions to encompass the concentration observations. It is expected that incorporating IC uncertainty in an NWP ensemble would yield increased spread in the wind direction (Fujita et al. 2007), and hence in the concentration predictions as well (Peltier et al. 2009). Therefore, we hypothesize that for AT&D forecasting purposes, the best NWP ensemble configuration would be one that varies ICs/LBCs in addition to physics parameterizations, with a focus on obtaining appropriate spread in the ABL.

It is not clear *a priori* what NWP ensemble configuration would be best for AT&D applications, especially with the number of options for physics parameterizations that are available in NWP models. Therefore, some testing is necessary. This study is a preliminary examination of the first of a series of historical evaluation periods that will be conducted with many ensemble members. The aim of this study is to explore methods that will help determine a useful ensemble configuration for AT&D applications.

2. ENSEMBLE DESIGN

Our 18-member physics ensemble for this study was created using version 3.1.1 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) were 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 were 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 was at 50 hPa. Such high vertical resolution in the lowest portions of the troposphere was chosen because this study focuses on processes occurring in the ABL. A single model domain was used with a horizontal grid spacing of 36 km and a time step of 180 s. The domain encompassed the continental United States (CONUS), as shown in Figure 1. No data assimilation was used during the model integration for this study. The ICs/LBCs for all 18 members in this study come from the 0.5°x0.5°-resolution Global Forecast System (GFS) forecast cycles initialized at 0000 UTC daily for the two-week period of 4-17 January 2009. That two-week evaluation period was chosen because the synoptic regime between the two weeks. During the first week (4-10 Jan) there was a deep, digging trough moving across the U.S., and during the second week (11-17 Jan) a persistent ridge in the western U.S. and trough in the eastern U.S. set up. By evaluating ensemble performance in different synoptic regimes, our results will be more robust than if the ensemble forecasts were from a single synoptic regime. In the future we would like to do similar two-week evaluation periods in the other three seasons as well, to investigate if our results are seasonally dependent.

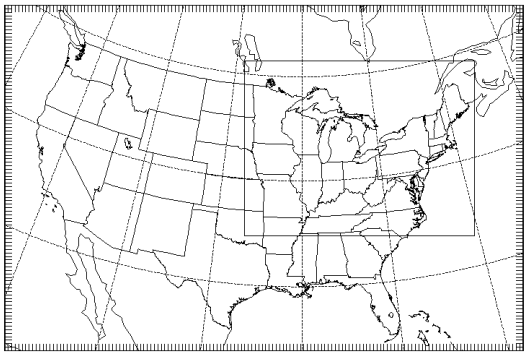

FIG. 1. Geographical domain used by the WRF-ARW ensemble.

3. ENSEMBLE DOWN-SELECTION

Our 18-member ensemble in this study varied only physics parameterizations. By first examining the performance of many combinations of physics options, a more intelligent choice can be made to down-select which combinations of physics would be good candidates for a small number of "control" members in a future ensemble that incorporated physics and IC/LBC uncertainty. One potential way to incorporate IC/LBC uncertainty in that future ensemble would be to use an ensemble Kalman filter to generate IC/LBC perturbations around each of the control physics members. If, for example, four control physics members were used, each with five IC/LBC perturbations around them, this would result in an ensemble of 20 members, which would be comparable in size to NCEP SREF (21 members). Down-selecting to a smaller number of ensemble members is also important due to computational restraints.

TABLE 1. Physics parameterizations for the ensemble members used in this project.

Member	Microphysics	Longwave Radiation	Shortwave Radiation	Land Surface	Surface Layer	Boundary Layer	Cumulus
1	WSM 5-class	RRTM	Dudhia	Thermal Diff.	MM5 Similarity	YSU	Kain-Fritsch
2	WSM 5-class	RRTM	Dudhia	Thermal Diff.	MM5 Similarity	YSU	Grell-Devenyi
3	WSM 5-class	RRTM	Dudhia	Noah	MM5 Similarity	YSU	Kain-Fritsch
4	WSM 5-class	RRTM	Dudhia	Noah	MM5 Similarity	YSU	Grell-Devenyi
5	WSM 5-class	RRTM	Dudhia	RUC	MM5 Similarity	YSU	Kain-Fritsch
6	WSM 5-class	RRTM	Dudhia	RUC	MM5 Similarity	YSU	Grell-Devenyi
7	WSM 5-class	RRTM	Dudhia	Thermal Diff.	Eta Similarity	MYJ	Kain-Fritsch
8	WSM 5-class	RRTM	Dudhia	Thermal Diff.	Eta Similarity	MYJ	Grell-Devenyi
9	WSM 5-class	RRTM	Dudhia	Noah	Eta Similarity	MYJ	Kain-Fritsch
10	WSM 5-class	RRTM	Dudhia	Noah	Eta Similarity	MYJ	Grell-Devenyi
11	WSM 5-class	RRTM	Dudhia	RUC	Eta Similarity	MYJ	Kain-Fritsch
12	WSM 5-class	RRTM	Dudhia	RUC	Eta Similarity	MYJ	Grell-Devenyi
13	WSM 5-class	RRTM	Dudhia	Thermal Diff.	Pleim-Xu	ACM2	Kain-Fritsch
14	WSM 5-class	RRTM	Dudhia	Thermal Diff.	Pleim-Xu	ACM2	Grell-Devenyi
15	WSM 5-class	RRTM	Dudhia	Noah	Pleim-Xu	ACM2	Kain-Fritsch
16	WSM 5-class	RRTM	Dudhia	Noah	Pleim-Xu	ACM2	Grell-Devenyi
17	WSM 5-class	RRTM	Dudhia	RUC	Pleim-Xu	ACM2	Kain-Fritsch
18	WSM 5-class	RRTM	Dudhia	RUC	Pleim-Xu	ACM2	Grell-Devenyi

In order to down-select to a smaller number of ensemble members, the first step was to build specific datasets for a feature selection process. A feature selection process evaluates the predictive ability of each of the features, which in this case are ensemble members. Therefore, the datasets must consist of an observation and the corresponding ensemble member.forecasts. However, the ensemble member forecasts are valid at computational grid points, so a distance-weighted interpolation from the four surrounding grid points to the observation location was used.

For the ensemble member down-selection we used a data mining and analysis program called *RapidMiner* (Mierswa et al. 2006). The datasets examined by this program were the interpolated ensemble member forecasts and the corresponding observations for each of three prognostic meteorological variables: 2-m temperature, 10-m u-wind, and 10-m v-wind. The first ten days of the 00 UTC ensemble forecasts (4-13 January 2009) were used to build these datasets, at forecast lead times of 24 h, 36 h, and 48 h.

Principal Component Analysis (PCA) weighting was then applied to each dataset individually. PCA is a mathematical procedure that reduces a large dataset of correlated variables to a smaller dataset of uncorrelated variables, or principal components. This operator uses the factors from the first principal component, which describes the most variability in the dataset, to calculate ensemble member weights. The ensemble members that contribute most to the first principal component are thus given a higher weight. The ensemble member weights are a quantitative analysis of the predictive ability of each of the ensemble members.

In order to evaluate the deterministic prediction of the ensemble, a linear regression with ten-fold cross-validation was used. Ten-fold cross-validation splits the dataset into ten different subsets. The linear regression model is trained on nine of the subsets and then predicts the tenth set. This is repeated on each of the ten subsets to assess how the results of the linear regression technique will generalize to an independent data set (Wilks 2005). Linear regression is used to predict the weather variables. This method was selected because it is a simple and effective way of relating predictors (ensemble member forecasts) to the predictand (observations), and because it is similar to Model Output Statistics (MOS) (Glahn and Lowry 1972). A RapidMiner operator called Weight Guided Feature Selection then determines the optimal selection of ensemble members to maximize the predictive ability of the linear regression model based on the ensemble member weights. The goal is to produce a subset of the ensemble members that are judged to be the most important for capturing the variance and predictive ability of the ensemble.

We down-selected the ensemble from eighteen members to twelve members, although the process can be tuned to yield a different number of members. The results for all three variables and all lead times (24, 36, and 48 h) are shown in Table 2**.** The final column shows the root mean square error (RMSE) of the linear regression model on the down-selected ensembles. These results indicate similar forecast errors for all lead times for each forecast parameter.

For the 2-m temperature forecast, the selected subset of ensemble members were consistent for all lead times, and none of the ensemble members that were selected used the Thermal Diffusion land surface model (LSM) (see Table 1). The Thermal Diffusion LSM is a simpler scheme than either the Rapid Update Cycle (RUC) or Noah LSMs because of a simpler treatment of soil moisture and no modeling of explicit vegetation effects (Skamarock et al. 2008). As evidenced by the

TABLE 2. Summary of ensemble members selected for all lead times and all forecast parameters. The final column is the root mean square error (RMSE) of the linear regression model that computes a deterministic forecast.

Parameter	Ensemble Members Selected												RMSE
2-m T, 24 h	3	4	5	6	9	10	11	12	15	16	17	18	3.010 K
2-m T, 36 h	3	4	5	6	9	10	11	12	15	16	17	18	3.351 K
2-m T, 48 h	3	4	5	6	9	10	11	12	15	16	17	18	3.257 K
10-m u, 24 h	1	2	3	4	5	6	7	8	9	10	11	12	3.397 m s^{-1}
10-m u, 36 h	1	2	3	4	5	6	7	8	9	10	11	12	3.042 m s^{-1}
10-m u, 48 h	1	2	3	4	5	6	7	8	9	10	11	12	3.328 m s^{-1}
10-m v, 24 h	1	2	3	4	5	6	7	8	9	10	11	12	3.423 m s^{-1}
10-m v, 36 h	1	2	3	4	7	8	9	10	11	12	13	14	3.177 m s^{-1}
10-m v, 48 h	1	2	3	4	7	8	9	10	11	12	13	14	3.334 m s^{-1}

ensemble member selection, a better representation of the land surface is important for better predictions of 2-m temperature.

For the 10-m u-wind and v-wind forecasts, different members were selected and the results were not uniform across all lead times. Ensemble members 1-12 were selected for most cases. The exceptions were for the v-wind at 36- and 48-h lead times that selected ensemble members 13 and 14 instead of 5 and 6. Ensemble members 13-18 all use the Pleim-Xu surface layer scheme and the Asymmetric Convective Model (ACM2) boundary layer scheme. Ensemble members 13 and 14 use the Thermal Diffusion LSM while ensemble members 5 and 6 have the RUC LSM. These results indicate that different representations of boundary layer processes are important to predictions of 10-m winds.

Scatterplots of ensemble member forecasts vs. observations for each forecast variable at each forecast lead time were created (not shown). The scatterplots for temperature showed that the chosen subset of ensemble members was well correlated with observations. The scatterplots for u- and v-wind, however, showed poor correlations. This result indicates that this down-selected ensemble poorly forecasts u- and v-wind speeds.

The selection of ensemble members appears directly related to the forecast variable as the process selected different ensemble members for temperature and wind speed. Ensemble members 3, 4, 10, 11, and 12 were always selected. If down-selecting to five ensemble members, these five have added value to all of our forecast parameter and lead times and would likely be an accurate down-selection number. Down-selecting even further to members 4, 10, and 12 would eliminate redundant forecasts from being used as "control" physics members in a future IC/LBC/physics ensemble, as the differences in the cumulus schemes appear to have very little effect in this time period. This is possibly due to convection not being as widespread in January as in warmer months, and the results may change as we test the method for other seasons.

As redundant ensemble members were not eliminated in this case, it seems that a calibrated down-selected ensemble is unlikely to result from the PCA weighting ensemble member selection method. However, the PCA weighting method did provide guidance about which ensemble members were the most valuable for a deterministic forecast. The PCA weighting method evaluated the deterministic predictive ability of the ensemble, while other methods, such as Bayesian Model Averaging, examine both the deterministic and probabilistic skill of the ensemble.

4. CALIBRATION AND VERIFICATION

Even with our best efforts to make the ensemble represent the possible outcomes of the atmosphere, it is unlikely that the distribution of the ensemble will be the same as the observed distribution of atmospheric states. Therefore, it is necessary to post-process ensemble data in order to more accurately represent the probability density function of potential atmospheric states. The primary technique used in this study is Bayesian Model Averaging (BMA), as detailed in Raftery et al. (2003) and discussed below.

The goal of BMA is to provide a calibrated model that predicts the PDF of a forecast variable given an ensemble of deterministic forecasts from a number of dynamical numerical weather prediction models. This is done by comparing previous ensemble forecasts against a set of verification data for some length of time, a "training period". When applied to forecasts, BMA can yield probabilistic predictions such as those shown in Fig. 2.

If the relationship between an observed value y and a bias-corrected ensemble of predictions of y, $\tilde{f}_k$, is the same as during the training period, the conditional PDF of y conditional on $\tilde{f}_k$ is:

$$p\left(y \mid f_1, \cdots, f_k\right) = \sum_{k=1}^{K} w_k g_k\left(y \mid \tilde{f}_k\right) \quad (1)$$

where $g_k\left(y \mid \tilde{f}_k\right)$ is the posterior probability of y given that k is the best ensemble member and w_k is the probability of k being the best ensemble member, termed the *weight* of ensemble member k. The log-likelihood of this model is given by:

$$\ell\ \vec{\theta}\ = \sum_{s,t} \log\left(\sum_{k=1}^{K} w_k g_k\ \ y_{st} \mid \tilde{f}_{kst}\ \right) \quad (2)$$

where $\vec{\theta}$ represents all of the parameters to be estimated and *s* and *t* represent the spatial and temporal dimensions, respectively. This log-likelihood cannot be solved analytically and would be complex for direct methods, but can be recast in terms of "missing data" which, if known, would make the problem trivial. In this case the "missing data" is z_{kst}, an indicator function that is 1 if ensemble member *k* is the closest to the verification at space and time *st* and 0 otherwise. Using this missing data, we can maximize the likelihood using expectation-maximization (EM).

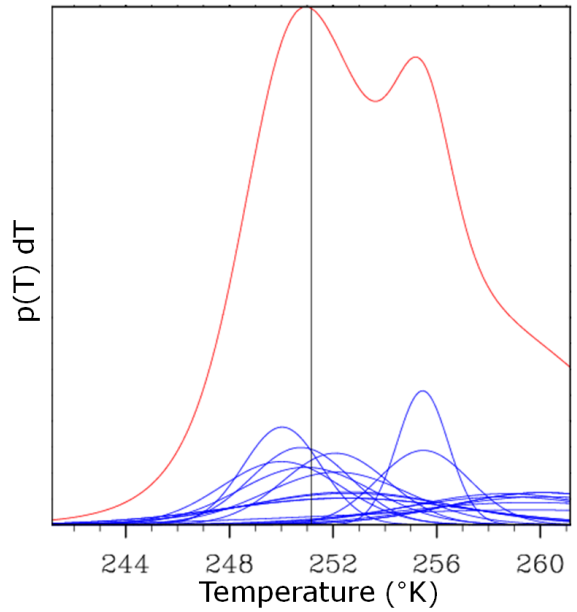

FIG. 2. Example probabilistic forecast that uses BMA ensemble member weights and standard deviations. Each blue curve represents the weighted conditional probability for an ensemble member. The red curve indicates the total prediction probability density function. The black vertical line represents the observation.

In order to proceed, we also need to assume a functional form of the posterior distribution g_k. Here we assume that this distribution is a normal, centered on $\tilde{f}_k$ (since $\tilde{f}_k$ is bias corrected) with a standard deviation σ_k. Thus, for each ensemble member *k*, we have two parameters, w_k and σ_k, which describe the contribution of that ensemble member to the total posterior PDF of *y*. Using this normal assumption together with the BMA model, the expectation (E) step which estimates z_{kst} is:

$$\hat{z}_{kst}^{\ i} = \frac{g\ \ y_{st} \mid \tilde{f}_{kst}, \hat{\sigma}_k^{\ i-1}}{\sum_{k=1}^{K} g\ \ y_{st} \mid \tilde{f}_{kst}, \hat{\sigma}_k^{\ i-1}} \quad (3)$$

The estimates of w_k and σ_k are then updated based on the new $\hat{z}_{kst}$ in the maximization (M) step:

$$\hat{w}_k^{(i)} = \overline{\hat{z}_{k\cdot\cdot}^{\ i}}$$

$$\sigma_k^{2\ i} = \frac{\sum_{s,t}\left[\hat{z}_{kst}^{\ i}\ \ y_{st} - \tilde{f}_{kst}\ ^2 \right]}{\sum_{s,t}\left[\hat{z}_{kst}^{\ i} \right]} \quad (4)$$

EM iteration is stopped when any of the following criteria are achieved:

1. The change in all z are less than the specified tolerance
2. The change in all parameters (w_k and σ_k) are less than their respective specified tolerances
3. The change in the log-likelihood is less than the specified tolerance
4. The maximum number of iterations is reached

As with the member down-selection, we use 00 UTC forecasts during 4-13 Jan 2009 as our training period. BMA weights and standard deviations are determined for forecasts every 12 h from 12 h to 48 h comparing model diagnosed 2-m temperature, 10-m U and 10-m V to WMO surface observations. Figure 3 shows the BMA ensemble weights for 2-m temperature. The six ensemble members that use the thermal diffusion land-surface model (1,2,7,8,13,14) have the lowest weights, indicating these members contribute the least to the optimized ensemble prediction. This result agrees with the results from section 3 where the thermal diffusion members were selected for removal from the ensemble based on temperature. The highest weights were assigned to the members using the RUC land-surface model (5,6,11,12,17,18). As forecast lead time increases, the members with thermal diffusion are weighted even less while the members using RUC LSM have increasing weights.

The members with the strongest weights for 10-m u-wind are those that use the Pleim-Xu land surface and ACM2 PBL scheme (Fig. 4, members 13-18). This is in contrast to section 3, where members 13-18 were nominated for removal based on the 10-m u-wind. The cause of this discrepancy is not immediately obvious, and is a subject that requires further study, though it may be related to the bias removal performed prior to the use of BMA. BMA standard deviations for each ensemble member were also computed, but are not shown here.

In addition to BMA, we also examined the correlations between ensemble members in order to identify possible redundant members. Correlations in the diagnosed 2-m temperature (Table 3) show that

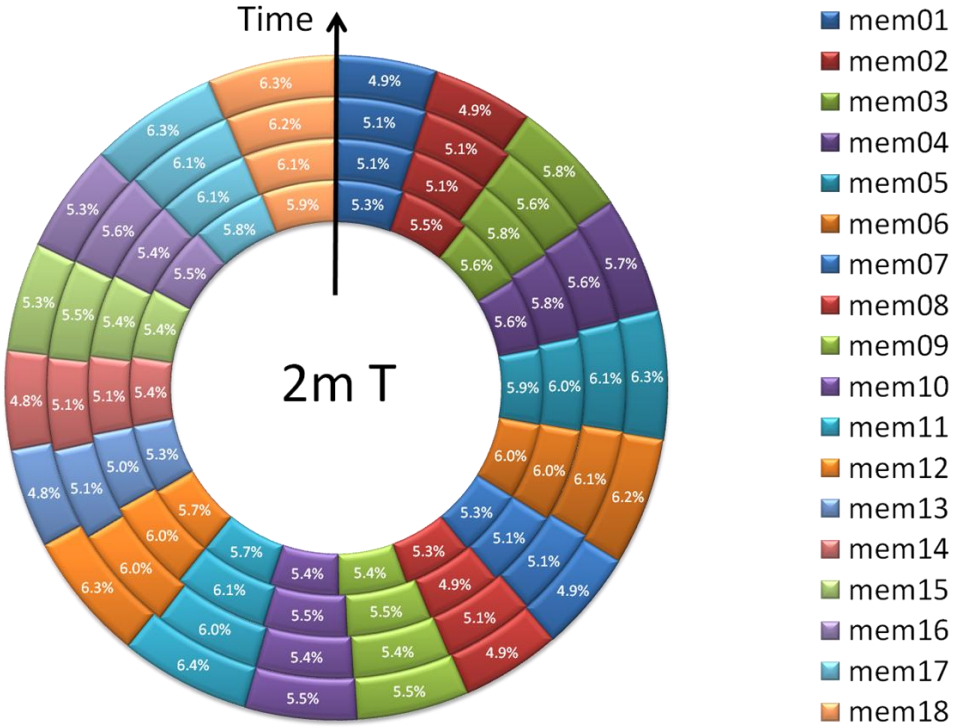

FIG. 3. BMA determined ensemble member weights for 2-m temperature. Radial distance indicates forecast lead time with the innermost ring representing BMA weights for 12-h forecasts, increasing 12-h going outward so that the outermost ring is for 48-h forecasts.

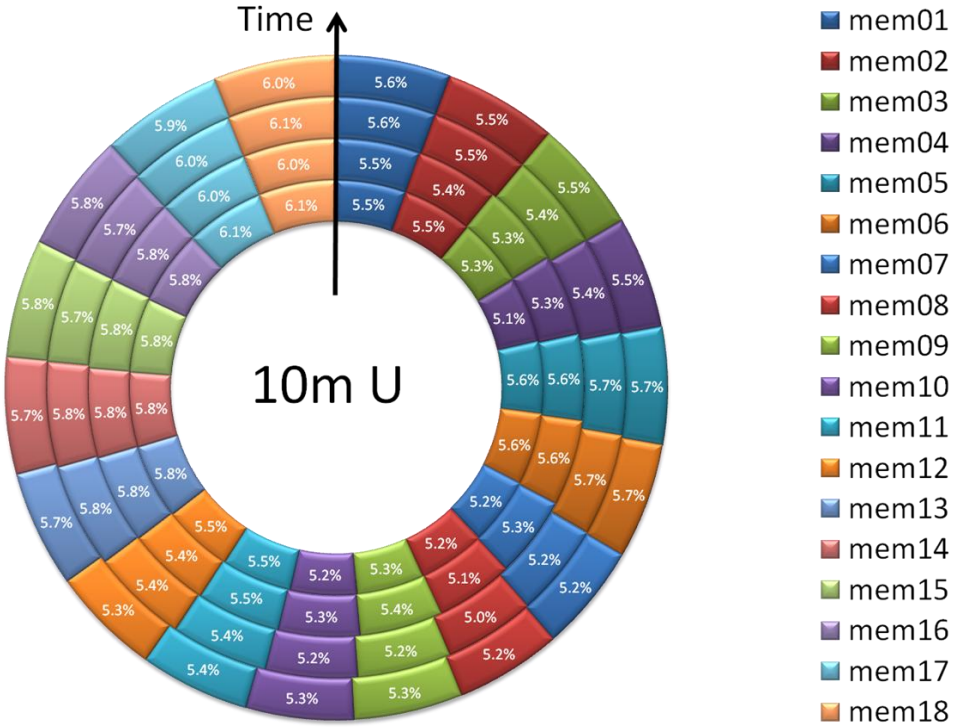

FIG. 4. As in Fig. 3, but for 10-m u-wind.

TABLE 3. Correlation of 2-m temperature between each pair of ensemble members. The lower triangle is color-coded for ease of interpretation, with warmer colors corresponding to larger correlations.

	mem01	mem02	mem03	mem04	mem05	mem06	mem07	mem08	mem09	mem10	mem11	mem12	mem13	mem14	mem15	mem16	mem17	mem18
mem01	1.000	1.000	0.989	0.989	0.984	0.984	0.997	0.997	0.987	0.987	0.983	0.983	0.997	0.997	0.986	0.986	0.981	0.981
mem02	1.000	1.000	0.989	0.989	0.984	0.984	0.997	0.997	0.987	0.987	0.983	0.983	0.997	0.997	0.986	0.986	0.981	0.981
mem03	0.989	0.989	1.000	1.000	0.996	0.995	0.984	0.984	0.994	0.994	0.991	0.991	0.984	0.984	0.994	0.994	0.991	0.991
mem04	0.989	0.989	1.000	1.000	0.995	0.996	0.984	0.984	0.994	0.994	0.991	0.991	0.984	0.984	0.994	0.994	0.991	0.991
mem05	0.984	0.984	0.996	0.995	1.000	1.000	0.980	0.981	0.992	0.992	0.995	0.995	0.980	0.980	0.992	0.992	0.996	0.996
mem06	0.984	0.984	0.995	0.996	1.000	1.000	0.980	0.981	0.991	0.992	0.995	0.995	0.979	0.980	0.992	0.992	0.996	0.996
mem07	0.997	0.997	0.984	0.984	0.980	0.980	1.000	1.000	0.989	0.989	0.985	0.985	0.998	0.998	0.986	0.986	0.981	0.981
mem08	0.997	0.997	0.984	0.984	0.981	0.981	1.000	1.000	0.989	0.989	0.985	0.985	0.998	0.998	0.986	0.986	0.981	0.981
mem09	0.987	0.987	0.994	0.994	0.992	0.991	0.989	0.989	1.000	1.000	0.996	0.996	0.988	0.988	0.998	0.998	0.994	0.994
mem10	0.987	0.987	0.994	0.994	0.992	0.992	0.989	0.989	1.000	1.000	0.996	0.996	0.987	0.988	0.998	0.998	0.994	0.994
mem11	0.983	0.983	0.991	0.991	0.995	0.995	0.985	0.985	0.996	0.996	1.000	1.000	0.982	0.983	0.993	0.993	0.998	0.998
mem12	0.983	0.983	0.991	0.991	0.995	0.995	0.985	0.985	0.996	0.996	1.000	1.000	0.982	0.983	0.993	0.993	0.998	0.998
mem13	0.997	0.997	0.984	0.984	0.980	0.979	0.998	0.998	0.988	0.987	0.982	0.982	1.000	1.000	0.987	0.987	0.981	0.981
mem14	0.997	0.997	0.984	0.984	0.980	0.980	0.998	0.998	0.988	0.988	0.983	0.983	1.000	1.000	0.987	0.987	0.981	0.981
mem15	0.986	0.986	0.994	0.994	0.992	0.992	0.986	0.986	0.998	0.998	0.993	0.993	0.987	0.987	1.000	1.000	0.994	0.994
mem16	0.986	0.986	0.994	0.994	0.992	0.992	0.986	0.986	0.998	0.998	0.993	0.993	0.987	0.987	1.000	1.000	0.994	0.994
mem17	0.981	0.981	0.991	0.991	0.996	0.996	0.981	0.981	0.994	0.994	0.998	0.998	0.981	0.981	0.994	0.994	1.000	1.000
mem18	0.981	0.981	0.991	0.991	0.996	0.996	0.981	0.981	0.994	0.994	0.998	0.998	0.981	0.981	0.994	0.994	1.000	1.000

TABLE 4. As in Table 3, but for 10-m u-wind.

	mem01	mem02	mem03	mem04	mem05	mem06	mem07	mem08	mem09	mem10	mem11	mem12	mem13	mem14	mem15	mem16	mem17	mem18
mem01	1.000	0.995	0.993	0.989	0.986	0.981	0.976	0.975	0.969	0.969	0.965	0.965	0.976	0.974	0.972	0.970	0.966	0.963
mem02	0.995	1.000	0.989	0.993	0.982	0.986	0.974	0.976	0.966	0.970	0.963	0.966	0.975	0.976	0.971	0.972	0.964	0.964
mem03	0.993	0.989	1.000	0.995	0.989	0.985	0.972	0.971	0.974	0.973	0.969	0.968	0.969	0.967	0.974	0.972	0.967	0.964
mem04	0.989	0.993	0.995	1.000	0.986	0.989	0.970	0.972	0.971	0.974	0.967	0.969	0.968	0.969	0.973	0.974	0.966	0.966
mem05	0.986	0.982	0.989	0.986	1.000	0.995	0.967	0.966	0.969	0.969	0.974	0.974	0.967	0.965	0.973	0.971	0.977	0.974
mem06	0.981	0.986	0.985	0.989	0.995	1.000	0.964	0.967	0.966	0.970	0.971	0.975	0.965	0.966	0.970	0.971	0.974	0.974
mem07	0.976	0.974	0.972	0.970	0.967	0.964	1.000	0.995	0.992	0.989	0.989	0.985	0.976	0.973	0.973	0.969	0.965	0.962
mem08	0.975	0.976	0.971	0.972	0.966	0.967	0.995	1.000	0.987	0.992	0.984	0.989	0.976	0.976	0.972	0.971	0.965	0.964
mem09	0.969	0.966	0.974	0.971	0.969	0.966	0.992	0.987	1.000	0.995	0.995	0.990	0.968	0.965	0.978	0.975	0.970	0.966
mem10	0.969	0.970	0.973	0.974	0.969	0.970	0.989	0.992	0.995	1.000	0.990	0.994	0.968	0.968	0.979	0.977	0.970	0.969
mem11	0.965	0.963	0.969	0.967	0.974	0.971	0.989	0.984	0.995	0.990	1.000	0.995	0.965	0.962	0.974	0.970	0.975	0.972
mem12	0.965	0.966	0.968	0.969	0.974	0.975	0.985	0.989	0.990	0.994	0.995	1.000	0.966	0.965	0.974	0.973	0.976	0.974
mem13	0.976	0.975	0.969	0.968	0.967	0.965	0.976	0.976	0.968	0.968	0.965	0.966	1.000	0.994	0.985	0.983	0.981	0.979
mem14	0.974	0.976	0.967	0.969	0.965	0.966	0.973	0.976	0.965	0.968	0.962	0.965	0.994	1.000	0.983	0.984	0.978	0.981
mem15	0.972	0.971	0.974	0.973	0.973	0.970	0.973	0.972	0.978	0.979	0.974	0.974	0.985	0.983	1.000	0.993	0.989	0.986
mem16	0.970	0.972	0.972	0.974	0.971	0.971	0.969	0.971	0.975	0.977	0.970	0.973	0.983	0.984	0.993	1.000	0.986	0.988
mem17	0.966	0.964	0.967	0.966	0.977	0.974	0.965	0.965	0.970	0.970	0.975	0.976	0.981	0.978	0.989	0.986	1.000	0.992
mem18	0.963	0.964	0.964	0.966	0.974	0.974	0.962	0.964	0.966	0.969	0.972	0.974	0.979	0.981	0.986	0.988	0.992	1.000

correlations between members with the same land-surface model are highest. This result is consistent with the BMA weights for temperature, where weights for ensemble members with the same land-surface model are similar. Also note that correlations between each ensemble pair where only the cumulus scheme varies are one within the precision shown in the table, meaning the cumulus scheme has little impact on the 2-m temperature. This is not surprising given that the training period is in January, when there is little convective activity.

The 10-m u-wind correlations (Table 4) also show high correlations between members that vary only in cumulus scheme. However, for 10-m u-wind the surface layer/PBL parameterizations seem to be the dominant influence, with the highest correlations between members that use the same surface layer/PBL pairing. These correlations reinforce the BMA results above, which estimated similar weights for members with the same surface layer and PBL.

In addition to these insights into the dominant model process in determining the surface conditions, we would like a measure of the performance of the calibrated ensemble, both for the deterministic forecast and the probabilistic forecast. Here we consider root-mean squared error (RMSE) as our deterministic measure. Verifications are conducted over the period 13-17 Jan, 2009, which keeps our verification period independent of our training period. However, because the first week of the total forecast period, which encompasses much of the training period, has a very different weather regime than the verification period, calibration methods (including BMA) may perform worse than would otherwise be expected.

The deterministic ensemble forecast is determined by weighting the forecast of each ensemble by its BMA-estimated weight. This is compared to a simple ensemble mean where all of the weights are identical. The probabilistic ensemble forecast is determined by calculating the conditional PDF of each ensemble member using a normal distribution with a mean of the member's forecast value and a standard deviation as determined by BMA, then multiplying each by the ensemble weight and adding. Neither ensemble prediction removes the member biases, which may adversely affect the performance of the BMA forecast.

The root-mean squared error is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} f_i - v_i^{\,2}} \qquad (5)$$

where v 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 RMSE of the BMA deterministic forecast is lower than the RMSE of the simple ensemble mean for every variable and forecast lead time combination studied, with the largest improvement coming in predicted 2-m temperature forecasts (Fig. 5). This indicates that the BMA yields improved ensemble forecasts, at least for this case. In the future we also plan to compute the continuous ranked probability score (CRPS) (Wilks 2005) as a probabilistic measure to evaluate the BMA-weighted ensemble, and compare it against the discrete CRPS for equal-weighted ensemble members. This will provide another measure of ensemble performance.

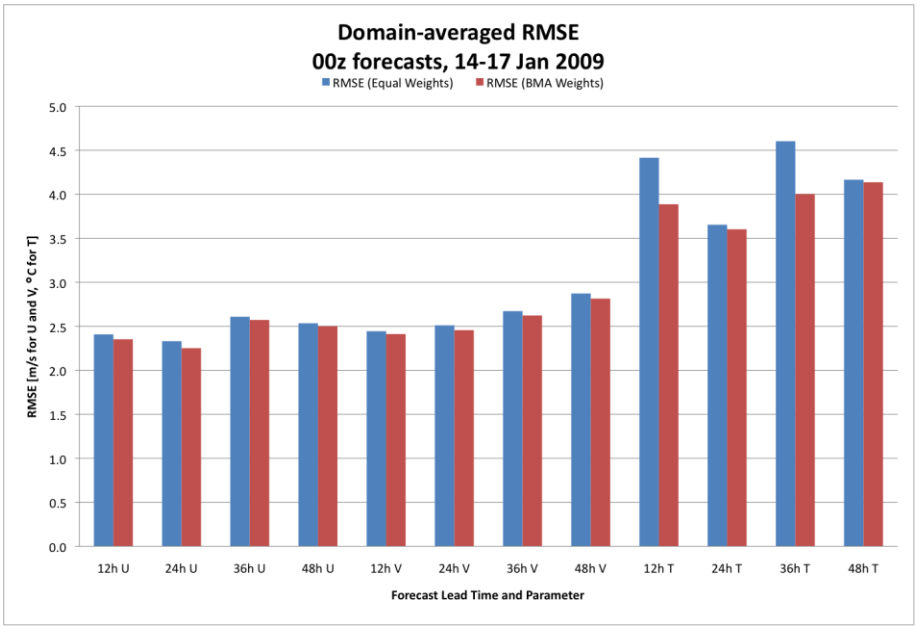

FIG. 5. Root-mean squared error of the deterministic ensemble forecast using equal ensemble weights (blue) and BMA-determined weights (red) in the verification period of 14-17 Jan 2009. Units are m s^{-1} for wind speed RMSE and K for temperature RMSE.

5. CONCLUSION

This study demonstrates several aspects of ensemble creation and evaluation. We created an 18-member physics ensemble using the WRF-ARW model, and evaluated its performance over a two-week period in January 2009. In section 3 we demonstrated how principal component analysis could be used in conjunction with regression analysis to nominate the elimination of poorly performing ensemble members in order to down-select a smaller ensemble. In section 4 we demonstrated Bayesian Model Averaging as a post-processing method for ensembles that can produce calibrated probabilistic predictions and also be used to identify dominant processes. Section 4 also introduced two evaluation metrics: the root-mean squared error, which is a deterministic measure of skill, and cumulative rank probability score, which is a probabilistic measure of skill. We also calculated error correlations between all of the ensemble members, which showed that, for this two-week period in winter, changing the cumulus parameterization scheme had almost no effect on the forecasts. These error correlations also indicated that changing the land surface model appeared to have the greatest effect on 2-m temperature predictions, while changing the pairings of the surface layer and planetary boundary layer schemes had the greatest effect on 10-m wind predictions. This illustrates the importance of varying these physics options to increase spread in a physics ensemble. However, the ensemble spread in this study was quite low, indicating the need to incorporate initial condition and lateral boundary condition uncertainty in ensemble forecasting, at least in winter, when synoptic patterns tend to be more active. Also, while these methods did not specifically point out redundant ensemble members, it is apparent from several lines of evidence, most notably the error correlations shown in Tables 2 and 3 and the BMA-determined ensemble member weights in Figs. 3 and 4, that changing the cumulus scheme had little effect on the forecasts in this study.

In the future we plan to increase the size of the ensemble with additional combinations of physics options, and to evaluate the physics ensemble over two-week periods in all four seasons. In this study we used the first ten days of the 14-day evaluation period as the training periods for the PCA and BMA methods; in the future we plan to compare the results using a randomly chosen ten days for a training period. The performance of the ensemble should also be investigated for forecasts initialized at 12 UTC in addition to 00 UTC, to account for possible diurnal effects in the forecasts. Recommendations will then be made of which ensemble members to use in a year-long ensemble for atmospheric transport and dispersion forecasting studies.

ACKNOWLEDGMENTS

We gratefully acknowledge funding provided for this project by the Defense Threat Reduction Agency (DTRA), under Contract DTRA-01-03-D-0010-0012 (John Hannan, Contract Monitor). Thanks are also due to Chuck Ritter of the Penn State Applied Research Lab (ARL) for computational support. The lead author also thanks the Penn State ARL Educational & Foundational Program for funding support.

References

Bowler, N.E., 2006: Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model. *Tellus*, **58A,** 538-548.

Bright, D.R., and S.L. Mullen, 2002: The sensitivity of the numerical simulation of the southwest monsoon boundary layer to the choice of PBL turbulence parameterization in MM5. *Wea. Forecasting*, **17,** 99-114.

Buizza, R., P.L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133,** 1076-1097.

Bowler, N.E., A. Arribas, K.R. Mylne, K.B. Robertson, and S.E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Q. J. Roy. Met. Soc.*, **134,** 703-722.

Deng, A., and D.R. Stauffer, 2006: On improving 4-km mesoscale model simulations. *J. Appl. Meteor. and Climat.*, **45,** 361-381.

Descamps, L. and O. Talagrand, 2006: On some aspects of the definition of initial conditions for ensemble prediction. *Mon. Wea. Rev.*, **135,** 3260-3272.

Du, J., J. McQueen, G. DiMego, T. Black, H. Juang, E. Rogers, B. Ferrier, B. Zhou, and Z. Toth, 2004: The NOAA/NWS/NCEP Short Range Ensemble Forecast (SREF) system: Evaluation of an initial condition versus multiple model physics ensemble approach. Preprints, *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 21.3. [Available online at http://www.emc.ncep.noaa.gov/mmb/SREF/srefupdate_2004.pdf]

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.

Glahn, H.R., and D.A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

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

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

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,** 36-55.

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. and 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.

Mierswa, I., M. Wurst, R. Klinkenberg, M. Scholz and E. Timm 2006: YALE: Rapid Prototyping for Complex Data Mining Tasks**.** Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.

Molteni, F., R. Buizza, T.N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Q. J. Royal Met. Soc.*, **122,** 73-119.

Nutter, P., D. Stensrud, and M. Xue, 2004: Effects of coarsely resolved and temporally interpolated lateral boundary conditions on the dispersion of limited-area ensemble forecasts. *Mon Wea. Rev.*, **132,** 2358-2377.

Peltier, L.J., J.C. Wyngaard, S.E. Haupt, D.R. Stauffer, A. Deng, and J.A. Lee, 2009: Computing meteorological uncertainty for dispersion modeling. Submitted to *J. Appl. Meteor. and Climat.*, 17 Sep 2009.

Raftery, A.E, F. Balabdaoui, T. Gneiting and M. Polakowski 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]

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.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Am. Met. Soc.*, **74,** 2317-2330.

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

Warner, T.T., R.A. Peterson, and R.E. Treadon, 1997: A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Amer. Met. Soc.*, **78,** 2599-2617.

Wei, M., Z. Toth, R. Wobus, Y. Zhu, C.H. Bishop, and X. Wang, 2006: Ensemble transform Kalman filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus*, **58A,** 28-44.

Wilks, D.S., 2005: Statistical Methods in the Atmospheric Sciences, 2nd ed., Academic Press, 626 pp.