3.4 INVESTIGATING SEASONAL IMPACTS ON CLUSTERING AND ENSEMBLE DOWN-SELECTION

Jared A. Lee[1,2,3,*], Sue Ellen Haupt[1,2], and George S. Young[2]

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

1. INTRODUCTION

A common approach to quantifying uncertainty in a forecast is to use ensembles of numerical weather prediction (NWP) models. How best to configure NWP ensembles is an area of active research in the community (e.g., Eckel and Mass 2005; Fujita et al. 2007; Clark et al. 2008; Berner et al. 2011; Hacker et al. 2011; Lee et al. 2012a). Limited computing resources also force sacrifices to be made in balancing several considerations, including ensemble size, model resolution, and domain footprint.

Model error is a key portion of the error in NWP ensembles, particularly for short-range forecasts in the atmospheric boundary layer (ABL) (Stensrud et al. 2000; Fujita et al. 2007; Clark et al. 2008). Types of model error include uncertainty stemming from lack of knowledge about the processes that are being modeled, uncertainty in the values of model parameters, and scale truncation (a low-pass filter) associated with discretization and numerical scheme. Approaches for representing model error include multi-model, multi-physics, and stochastic perturbation ensembles, or combinations thereof (Eckel and Mass 2005; Hacker et al. 2011).

*Corresponding author address: Jared A. Lee, Naval Postgraduate School, Department of Meteorology, 589 Dyer Rd, Monterey, CA 93943, USA. Email: jaredlee@ucar.edu.

When constructing a multi-physics ensemble, it is not clear *a priori* what sets of physics schemes are best to choose, or how many members to include. The large number of available physics options exacerbates this problem. For instance, for the Weather Research and Forecasting (WRF) Advanced Research WRF (ARW) NWP model (Skamarock et al. 2008), there are hundreds of possible combinations of physics schemes from which to choose.

Previous research proposes objective methods using principal component analysis to choose, or "down-select," a smaller subset of ensemble members that represent the forecast probability density function (PDF) nearly as well as the full ensemble (Lee et al. 2012a). A second down-selection method is *K*-means cluster analysis, explored in Lee et al. (2012b). In this study we employ a third down-selection method, hierarchical cluster analysis (HCA), which performs comparably to principal component analysis and *K*-means cluster analysis (not shown). The goal of our ensemble down-selection technique is to retain the subset of ensemble members that spans the uncertainty space of the forecast, while eliminating those members that are most redundant. We do this because ensembles are most useful when members each sample a different portion of the atmospheric PDF.

In this study we construct a large WRF multi-physics ensemble, and compare the performance of the HCA subset ensemble to

that of the full, large ensemble using several metrics. We also use Bayesian model averaging (BMA; Raftery et al. 2003, 2005) to calibrate the forecasts for the full ensemble and to dress the down-selected subset ensembles.

We choose to configure a multi-physics ensemble that uses the same ICs/LBCs to isolate the effects of model uncertainty for three reasons. First, as mentioned above, physics variability is a crucial source of uncertainty for low-level and short-range forecasts. Second, it is not clear how any down-selection approach would be physically meaningful if it were applied to an ensemble with only equally likely IC/LBC perturbations, because members would then be exchangeable and statistically indistinguishable (Fraley et al. 2010). Third, one of the goals of this study is to define a small set of physics members for potential use in a later ensemble that would also include IC/LBC variability. Thus, we take the approach that because we are dressing the ensemble PDF using BMA, it is sufficient to consider only physics uncertainty here.

We discuss our ensemble configuration and verification procedures in section 2. In section 3 we describe the down-selection procedure with HCA. We present verification results in section 4, and section 5 summarizes the study.

2. DATA

2.1 Ensemble configuration

Our 42-member physics ensemble is created with version 3.3 of the WRF-ARW model. The two control members we use (CTL-01 and CTL-02) are Developmental Testbed Center (DTC) Reference Configurations for WRF-ARW v3 (Wolff et al. 2009; http://www.dtcenter.org/config/). For the remaining forty members we use at least three different options in the ensemble for each class of physics scheme (i.e., microphysics, radiation, land surface, surface layer, boundary layer, and cumulus schemes), as detailed in Table 1. Skamarock et al. (2008) contains details and references for all the parameterization schemes that we employ. We use a slightly modified version of the Mellor-Yamada-Janjic (MYJ) ABL scheme, as in Lee et al. (2012a).

We initialize the 48-h forecasts every fifth day at 0000 UTC starting on 1 Dec 2009, and continuing through Nov 2010, for a total of 18 forecast periods during each season. Table 2 lists all the initialization dates for these forecasts. We choose to space the forecasts evenly through time every five days instead of a more frequent spacing in order to reduce temporal correlations in consecutive forecast periods. Even forecast spacing also allows us to sample synoptic regimes fairly throughout the year.

The coarse domain uses a horizontal grid spacing of 36 km, while the one-way nested fine domain uses 12-km grid spacing. The geographic area spanned by the domains can be seen in Fig. 1. While all of our analysis focuses on the inner 12-km domain, we nest that domain inside the outer 36-km domain to reduce interpolation errors from the IC and LBC data, which we describe further below. There are 45 full vertical levels in each simulation, with high vertical resolution in the lowest 2 km (24 full levels) so that we can resolve processes in the ABL well. In this study we use time steps of 90 s and 30 s for the coarse and fine domains, respectively. We find such small time steps to be necessary to preserve model stability on simulation day 1 Dec 2009 because of a small, powerful vorticity maximum near the Texas Gulf Coast

(not shown) and retain that temporal resolution for consistency.

The LBCs for all 42 members in this study come from the 0.5˚x0.5˚-resolution Global Forecast System (GFS; Environmental Modeling Center 2003) forecast cycles initialized at each of the simulation times. We use sea surface temperature (SST) analyses from the National Centers for Environmental Prediction (NCEP) real-time global 0.083˚ dataset. We use daily snow analyses from the National Environmental Satellite, Data, and Information Service (NESDIS).

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

2.2 Verification and metrics

We perform our down-selection, verification, and analysis on the inner 12-km domain. This approach excludes the detrimental impact of boundary artifacts near the edge of the outer 36-km domain.

We use standard WMO surface and upper-air observations to verify our WRF ensemble forecasts. We examine the forecasts for down-selection and verification at four lead times: 12 h, 24 h, 36 h, and 48 h, those times for which standard radiosonde observations are available (0000 UTC and 1200 UTC). We quality control these observations against the GFS analysis fields that are interpolated by the WRF Pre-processing System (WPS), using Obsgrid as described above.

We divide our year-long forecast dataset into roughly month-long groups of six forecast periods each. For each experiment listed in Table 3 we use one month for verification, while using the previous one, two, or three months for training data, so that we can explore what impact training period length has on results.

The observations used in this study are temperature and wind at four levels: the surface and the mandatory levels of 925 hPa, 850 hPa, and 700 hPa. We choose these levels because in this study we are primarily concerned with factors relevant to forecasting in the lower troposphere, and in particular the ABL. Additionally, choosing a consistent set of mandatory levels serves the purpose of maximizing the number of usable sounding observations, while also not introducing interpolation error into the observations on which we train the down-selection techniques and against which we verify the forecasts. Model predictions are horizontally and vertically interpolated to the observation locations. In the horizontal we use bilinear interpolation from the four surrounding model grid points, and in the vertical we use linear interpolation from the grid points immediately above and below the verification pressure level, with the natural log of pressure as our vertical coordinate for interpolation. We perform verification on wind direction, wind speed, vector wind difference, and the zonal (u) and meridional (v) components of the wind.

The primary verification metric we use is the continuous ranked probability score (CRPS). The CRPS is a probabilistic metric,

and assesses both accuracy and sharpness and is defined as (Wilks 2006):

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

where $p_i^f(x)$ is the forecast cumulative probability of the forecast variable being $\leq x$ at the space-time location of observation i, $p_i^o(x)$ is the CDF of the observation o, f is the forecast value at the time and location of observation i, and N is the total number of observations. CRPS is negatively oriented, with zero representing a perfect score.

To compare the relative performance of the CRPS between a subset and full ensemble, we take the ratio CRPSR of the CRPS for the subset ensemble to that of the full ensemble:

$$\mathrm{CRPSR} = \frac{\mathrm{CRPS}_{subset}}{\mathrm{CRPS}_{full}} \qquad (2)$$

CRPSR is similar to a skill score, except that a score higher than 1 represents a worse CRPS for the subset ensemble compared to the full ensemble, while a score lower than 1 represents a better CRPS for the subset ensemble.

Comparing the CRPS of two ensembles does not directly indicate how similar the distributions of the two ensembles are, however. Thus, we use the two-sample Kolmogorov-Smirnov (K-S) test to assess the similarity of the empirical cumulative distribution functions (CDFs) of the full and down-selected ensembles. The null hypothesis for the K-S test is that the two populations of data being compared come from the same distribution. The two-sample K-S test statistic finds the greatest absolute difference between the empirical CDFs of two populations, n_1 observations of x_1 and n_2 observations of x_2 (Wilks 2006):

$$\mathrm{D_S} = \max_x \left|\mathrm{F_n}(x_1) - \mathrm{F_m}(x_2)\right| \qquad (3)$$

The null hypothesis for the two-sample K-S test is rejected at the 95% confidence level is (Wilks 2006):

$$\mathrm{D_S} = \sqrt{-\frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\ln\left(\frac{0.95}{2}\right)} \qquad (4)$$

The two-sample K-S test is computed for every observation in the verification period for the various experiments. The percentage of observations for which the null hypothesis is not rejected in the verification period indicates how well the full ensemble and subset ensemble forecast distributions match.

2.3 Ensemble calibration

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

We use BMA to dress the full ensemble and all down-selected ensembles. BMA estimates the weights and parameters for each ensemble member, and then during a training period (6, 12, or 18 forecast periods in this study), these weights and parameters are trained to best match the observations. The BMA weights and standard deviations are then applied to forecasts in a verification period to create a better ensemble PDF.

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

3. ENSEMBLE DOWN-SELECTION WITH HCA

The down-selection technique we use is hierarchical cluster analysis. HCA is used in several studies to group together similar members in an NWP ensemble forecast (e.g., Legg et al. 2002; Alhamed et al. 2002; Yussouf et al. 2004; Johnson et al. 2011).

Here we perform down-selection via HCA on bias-corrected, normalized temperature errors and normalized vector wind differences (VWD) over the forecast training period, combining data from all four forecast lead times (12, 24, 36, and 48 h) at multiple levels (surface, 925 hPa, 850 hPa, and 700 hPa). We combine the normalized temperature errors and VWDs to perform multivariate down-selection, so that for each experiment there is a single subset ensemble chosen for all lead times, levels, and variables.

In HCA each data vector starts out as a singleton cluster, and at each step of the algorithm, the two clusters that are closest to each other according to some distance metric are combined. This process continues until all the data vectors are combined into a single cluster.

The version of HCA we use is Ward's minimum variance method, known more simply as Ward's method (Wilks 2006). Ward's method combines the two clusters that have the smallest sum of squares – that is, the sum of squares of distances between each point in the cluster and the cluster centroid. The distance metric $d(r,s)$ that Ward's method uses in the MATLAB® Statistics Toolbox is:

$$d(r,s) = \sqrt{\frac{2n_r n_s}{(n_r + n_s)}} \left\| \overline{x_r} - \overline{x_s} \right\|_2 \qquad (5)$$

where $\| \|_2$ is the Euclidean distance, $\overline{x_r}$ and $\overline{x_s}$ are the centroids of clusters r and s, and n_r and n_s are the numbers of elements in clusters r and s. Ward's method is used frequently in other studies that employ HCA (e.g., Alhamed et al. 2002; Yussouf et al. 2004; Johnson et al. 2011). Additionally, we find that alternate versions of HCA yield results that are no better than Ward's method (not shown).

To determine the number of clusters present in the data for each case, we find the maximum number of clusters for which each

cluster has at least three members. We choose this clustering criterion to exclude singleton clusters as not meaningful, and also so that each cluster will have a single member that is closest to the centroid. The HCA subset ensembles are all comprised of the members that are the nearest to their respective cluster centroids. A dendrogram from the DJF experiment is shown in Fig. 2 as an example visualization of how the ensemble members cluster. The higher up the vertical axis (the distance metric from Eq. 5) that two clusters join, the more dissimilar they are. The colored branches of the dendrogram represent the ten clusters determined by HCA for the DJF experiment.

To demonstrate that down-selection using HCA has value, we also compare HCA to a random down-selection method. For the DJF (winter) and JJA (summer) experiments, we randomly choose ten sets of subset ensembles for each ensemble size ranging from 2-15 members (i.e., ten random subset ensembles of 2 members each, ten random subset ensembles of 3 members each, etc.). By examining a range of ensemble sizes we can also assess whether there is an ensemble size above which additional members no longer add forecast skill. We use HCA to determine single subset ensembles for each of those ensemble sizes (2-15 members), but allowed for singleton or two-member clusters in order to make this comparison. For two-member clusters both members are equidistant from the cluster centroid, so in those cases we randomly choose which member becomes part of the HCA subset.

4. RESULTS

The HCA clusters that result from the one-month training experiments are shown in Table 4, the clusters from the two-month training experiments are listed in Table 5, and the clusters from the three-month training experiments can be seen in Table 6. There are several insights that can be drawn from those clustering experiments.

First, members cluster differently in different seasons. Tables 4-6 show this clearly, with several identical clusters in the experiments that share a common month(s) in the training period. This high degree of "overlapping" of clusters (i.e., shared clusters) is seen within each season, though somewhat more strongly in winter and summer than in the transition seasons.

In each experiment, every cluster has at least one physics scheme that is common among all members of that cluster. The right-most column of Tables 4-6 indicates whether the cluster shares the same land surface scheme (L), boundary layer scheme (B), cumulus scheme (C), longwave and shortwave radiation schemes (R), microphysics scheme (M), or some combination thereof. For the vast majority of the clusters in all the experiments, the cluster shares a common land surface scheme. This result is unsurprising because there are an order of magnitude more surface than upper-air observations in the verification dataset and roughly 20% more temperature than wind observations, and also because of the large effect that the land surface scheme has on near-surface parameters (Wyngaard 2010; Warner 2011).

Boundary layer and cumulus parameterizations generally appear to be of secondary importance to the clustering. As can be seen in Tables 4-6, clusters that share the same microphysics and/or radiation schemes also all share the same cumulus scheme in this ensemble, but the converse is often not true; thus it appears that the cumulus scheme has greater importance with regard to determining clusters than do either

the microphysics or radiation schemes, at least for the region studied here. In this region, cumulus parameterization schemes often have a more direct impact on model temperatures and winds than do microphysics and radiation schemes. Therefore it makes physical sense that cumulus schemes would be more relevant for clustering than microphysics or radiation. It should be noted, however, that in other regions, such as the U.S. west coast, for example, microphysics and radiation schemes are likely to have a larger impact on surface variables than cumulus schemes due to the modeling of marine stratus.

In the summer the clusters tend to share a cumulus and/or land surface scheme, but typically not a boundary layer scheme. In the transition seasons the clusters frequently share a boundary layer and/or land surface scheme, but not a cumulus scheme. A plausible meteorological explanation for this behavior is that there is more convection across the 12-km domain (see Fig. 1) in summer, and in the transition seasons of spring and autumn the effects of surface heating are increasing and decreasing, respectively. In winter there are many synoptic systems moving across the domain with forcing strong enough to trigger convection despite the weak land surface forcing, and because boundary layer schemes in WRF-ARW have variable performance in cold and stable regimes in different regions (Gilliam and Pleim 2010).

Second, the length of the training period for clustering and calibration has little impact on the verification scores, whether the training period is one, two, or three months long (Figs. 3, 4, and 5, respectively). In Figs. 3-5, the CRPSR (Eq. 2) of the HCA subset to the full ensemble for each variable changes very little with training period length (e.g., compare the data points for FM from Fig. 3, JFM from Fig. 4, and DJFM from Fig. 5).

Furthermore, for experiments in which the training period overlapped, there tends to be considerable overlap in the members that comprise the HCA down-selected subset ensembles (Table 7). The subset membership overlap occurs because similar or identical clusters tend to form in experiments with overlapping training periods (see Tables 4-6). In other words, when the same cluster is found in multiple experiments, there is one ensemble member that is frequently closest to the centroid of that cluster. Therefore, at least for this ensemble, a longer training period, which requires several hours more computation time for calibration with BMA in order to dress the ensemble, appears to confer little if any tangible benefit. As a result, a one-month training period is considered to be sufficient and practical. A training period of about a month for BMA is similar to the findings of Raftery et al. (2005), though they use a daily NWP ensemble, instead of an every-fifth-day ensemble.

Third, across all experiments in all months and seasons, the HCA subsets have CRPS values that are within about 4% of the CRPS values of the full ensemble. This result can be seen in Figs. 3-5, for the one, two, and three-month training experiments. Thus, it can be said that down-selection via HCA is effective year-round, not just in one particular season. It is also worth noting that the CRPSR ratios are marginally closer to 1 for the one-month training experiments than for the two and three-month training experiments. That the one-month training experiments have a CRPSR that changes least from 1 is likely caused by training on data that is most similar to the verification data, compared to the other experiments with longer training periods.

Fourth, down-selection using HCA usually results in better verification scores than if down-selection is done randomly. We calculate the CRPSR for both HCA-determined subsets and the average of ten randomly-determined subsets (comparing both to the full ensemble), for ensemble sizes ranging from two to fifteen members, for both a winter experiment (DJF) and a summer experiment (JJA). The advantage of HCA down-selection over random down-selection is more pronounced in the winter case, as seen in Figs. 6 and 7 for 2-m T and 10-m u-wind, respectively, than it is for the summer case (Figs. 8 and 9). Yet even in the summer case the HCA subsets still tend to perform slightly better than the random subsets.

Fifth, there appears to be little additional forecast skill gained by increasing ensemble size beyond roughly 7-10 members. The CRPSR in Figs. 6-9 all decrease with increasing ensemble size until about 7-10 members, at which point the CRPSR remains roughly flat and approximately 1.0 for larger ensemble sizes. Because only a few ensemble members can deliver nearly equivalent forecast performance as a much larger ensemble, it is likely that such a large multi-physics ensemble contains much redundancy in representing model error.

Sixth, down-selection is more effective for calibrated ensembles. In Figs. 6-9 the CRPSR for calibrated ensembles (solid lines) is smaller (i.e., better) than the CRPSR for uncalibrated ensembles (dashed lines). Thus, fewer ensemble members are required to achieve forecast skill equivalent to that of the full ensemble when the full and subset ensembles are both calibrated. Furthermore, as in Lee et al. (2012a,b), calibration improves the CRPS by 10-15% at all levels for all variables and lead times (not shown).

Finally, the two-sample K-S test indicates that the full and HCA subset ensembles are generally similar. Table 8 details, for each experiment and for each lead time-variable combination, the percentage of observation locations for which the two-sample K-S test determined the full and subset ensemble CDFs to be similar at the 95% confidence level. For most experiments and lead time-variable combinations, the full and subset ensemble CDFs are statistically similar to each other for 95% or more of the forecast locations in the verification period. Table 8 only shows K-S test results for surface variables, but results are similar for above-surface variables as well.

5. SUMMARY

This study demonstrates the performance of an ensemble down-selection methodology using hierarchical cluster analysis on a year-long, 42-member NWP multi-physics ensemble dataset. Training for both the calibration and down-selection is done over one, two, and three-month periods, with one month of forecast data used for verification. The length of training period is shown to have little impact on verification results, however, so it is practical to use a shorter training period for computational efficiency.

The ensemble members cluster differently in different seasons, but always share at least one common physics parameterization scheme. To account for model uncertainty in a multi-physics framework, the classes of physics schemes in which diversity is most important change with season.

Down-selection with HCA, particularly after calibration with BMA, is effective year-round at representing the forecast distribution of a 42-member multi-physics ensemble with just 7-10 members. In all seasons this study demonstrates that increasing ensemble size beyond about 10

members would simply be gratuitous computing, and that resources would be more wisely spent on increasing model resolution or the size of the model domain.

ACKNOWLEDGMENTS

We gratefully acknowledge computing resource grants provided by both the Extreme Science and Engineering Discovery Environment (XSEDE) and the Computational Information Systems Laboratory (CISL) at the National Center for Atmospheric Research (NCAR) for allowing the computation and storage of the WRF ensemble data. XSEDE and NCAR are both supported by the National Science Foundation (NSF). We also gratefully thank Greg Thompson and Laurie Carson of NCAR for providing additional computing resources on NCAR's Bluefire supercomputer for this work. We thank Aijun Deng of Penn State for providing the NESDIS snow analyses for the WRF initial conditions, and we also thank David Stauffer of Penn State, Walter Kolczynski and Joshua Hacker of Naval Postgraduate School, and Tressa Fowler, Tom Hopson, and Luca Delle Monache of NCAR for helpful discussions during the course of this project. Author Lee is also grateful to Xcel Energy, NASA, Telvent Corporation, and the National Research Council for funding that partially supported this project.

REFERENCES

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

Berner, J., S.-Y. Ha, J.P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system. *Mon. Wea. Rev.*, **139,** 1972-1995.

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

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

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

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

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

Gilliam, R.C. and J.E. Pleim, 2010: Performance assessment of new land surface and planetary boundary layer physics in the WRF-ARW. *J. Appl. Meteor. Climatol.*, **49,** 760-774.

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

Hacker, J.P., S.-Y. Ha, C. Snyder, J. Berner, F.A. Eckel, E. Kuchera, M. Pocernich, S. Rugg, J. Schramm, and X. Wang, 2011: The U.S. Air Force Weather Agency's mesoscale

ensemble: Scientific description and performance results. *Tellus*, **63A,** 625-641.

Johnson, A., X. Wang, M. Xue, and F. Kong, 2011: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part II: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139,** 3694-3710.

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

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

Lee, J.A., W.C. Kolczynski, T.C. McCandless, and S.E. Haupt, 2012a: An objective methodology for configuring and down-selecting an NWP ensemble for low-level wind prediction. *Mon. Wea. Rev.*, **140,** 2270-2286.

Lee, J.A., S.E. Haupt, G.S. Young, W.C. Kolczynski, and T.C. McCandless, 2012b: Evaluating methods for down-selecting NWP multiphysics ensembles for wind prediction. 10th Conf. on Artificial Intelligence at 92nd AMS Annual Meeting, New Orleans, LA, 24 Jan 2012.

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

Mesoscale & Microscale Meteorology (MMM) Division, National Center for Atmospheric Research, 2011: Weather Research & Forecasting ARW version 3 modeling system user's guide, 362 pp. [Available online at http://www.mmm.ucar.edu/wrf/users/docs/user_guide_V3/contents.html]

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

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

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

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

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

Warner, T.T., 2011: Numerical weather and climate prediction, Cambridge University Press, 526 pp.

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

Wolff, J.K., L. Nance, L. Bernardet, and B. Brown, 2009: WRF reference configurations. 23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction at 89th AMS Annual Meeting, Phoenix, AZ. [Available online at http://www.dtcenter.org/config/WRF_RCs.pdf]

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

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

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

TABLE 2. Initialization dates for the WRF ensemble from Dec 2009-Nov 2010, in YYYY-MM-DD format. All forecasts are initialized at 0000 UTC on these dates. Also shown are the "month"-long blocks of six forecast periods each into which the ensemble dataset is divided.

Winter	**Spring**	**Summer**	**Autumn**
D	M	J	S
2009-12-01	2010-03-01	2010-06-04	2010-09-02
2009-12-06	2010-03-06	2010-06-09	2010-09-07
2009-12-11	2010-03-11	2010-06-14	2010-09-12
2009-12-16	2010-03-16	2010-06-19	2010-09-17
2009-12-21	2010-03-21	2010-06-24	2010-09-22
2009-12-26	2010-03-26	2010-06-29	2010-09-27
J	A	J	O
2009-12-31	2010-03-31	2010-07-04	2010-10-02
2010-01-05	2010-04-05	2010-07-09	2010-10-07
2010-01-10	2010-04-10	2010-07-14	2010-10-12
2010-01-15	2010-04-15	2010-07-19	2010-10-17
2010-01-20	2010-04-20	2010-07-24	2010-10-22
2010-01-25	2010-04-25	2010-07-29	2010-10-25
F	M	A	N
2010-01-30	2010-04-30	2010-08-03	2010-11-01
2010-02-04	2010-05-05	2010-08-08	2010-11-06
2010-02-09	2010-05-10	2010-08-13	2010-11-11
2010-02-14	2010-05-15	2010-08-18	2010-11-16
2010-02-19	2010-05-20	2010-08-23	2010-11-21
2010-02-24	2010-05-25	2010-08-28	2010-11-26

TABLE 3. Abbreviations for each experiment conducted in this study, with the corresponding "month(s)" used for training and verification (see Table 2).

Experiment Name	Training "month(s)"	Verification "month"
DJ	Dec	Jan
JF	Jan	Feb
FM	Feb	Mar
MA	Mar	Apr
AM	Apr	May
MJ	May	Jun
JJ	Jun	Jul
JA	Jul	Aug
AS	Aug	Sep
SO	Sep	Oct
ON	Oct	Nov
DJF	Dec-Jan	Feb
JFM	Jan-Feb	Mar
FMA	Feb-Mar	Apr
MAM	Mar-Apr	May
AMJ	Apr-May	Jun
MJJ	May-Jun	Jul
JJA	Jun-Jul	Aug
JAS	Jul-Aug	Sep
ASO	Aug-Sep	Oct
SON	Sep-Oct	Nov
DJFM	Dec-Jan-Feb	Mar
JFMA	Jan-Feb-Mar	Apr
FMAM	Feb-Mar-Apr	May
MAMJ	Mar-Apr-May	Jun
AMJJ	Apr-May-Jun	Jul
MJJA	May-Jun-Jul	Aug
JJAS	Jun-Jul-Aug	Sep
JASO	Jul-Aug-Sep	Oct
ASON	Aug-Sep-Oct	Nov

TABLE 4. Listing of all the clusters formed throughout the one-month training experiments from Table 3, the experiments in which those clusters are found, and also what classes of physics scheme are shared throughout the cluster (L = land surface scheme, B = boundary layer scheme, C = cumulus scheme, R = radiation schemes, M = microphysics scheme).

Cluster members	Experiments	Shared
01, 13, 14, 15, 16	DJ JF FM MA AM MJ SO ON	L B
02, 27, 28, 29	DJ JF FM MA AM ON	L B
10, 11, 12	DJ JF FM MA AM MJ	L B
17, 18, 19	DJ AM MJ	L
20, 21, 22, 30, 31, 32, 40, 41, 42	DJ ON	L
23, 24, 25, 33, 34, 35, 43, 44, 45	DJ	L C R M
26, 36, 46	DJ JF FM MA	L
37, 38, 39, 47, 48, 49	DJ AM ON	L
17, 19, 37, 39, 47, 49	JF MA	L C R M
18, 38, 48	JF MA	L
20, 21, 30, 31, 40, 41	JF FM MA	L
22, 32, 42	JF FM MA	L C R M
23, 33, 43	JF FM	L C R M
24, 25, 34, 35, 44, 45	JF FM	L
17, 37, 47	FM JJ	L C R M
18, 19, 38, 39, 48, 49	FM	L
23, 24, 25	MA	L B
33, 34, 35, 43, 44, 45	MA	L
20, 21, 22	AM MJ	L B
23, 24, 25, 26	AM ON	L B
30, 31, 32, 40, 41, 42	AM MJ ON	L
33, 34, 35, 36, 43, 44, 45, 46	AM	L B
02, 27, 28	MJ	L B
23, 24, 26	MJ	L B
25, 35, 45	MJ SO	L C R M
29, 37, 38, 39, 47, 48, 49	MJ SO	L
33, 34, 36, 43, 44, 46	MJ	L
01, 13, 16	JJ AS	L B C
02, 21, 27	JJ	B C
10, 19, 20, 29, 30, 39, 40, 49	JJ	C R
11, 31, 41	JJ	L C R M
12, 18, 22, 28	JJ	C
14, 24, 34, 44	JJ JA AS	L C R M
15, 25, 35, 45	JJ JA AS	L C R M
23, 26, 33, 36, 43, 46	JJ AS	L C R
32, 38, 42, 48	JJ	C
01, 13, 16, 23, 26, 33, 36, 43, 46	JA	L C
02, 17, 27, 37, 47	JA	L C R M
10, 20, 30, 40	JA AS	L C R M
11, 21, 31, 41	JA AS	L C R M
12, 22, 32, 42	JA AS	L C R M
18, 28, 38, 48	JA AS	L C R M
19, 29, 39, 49	JA	L C R M
02, 17, 19, 27, 29, 37, 39, 47, 49	AS	L
02, 22, 27, 28	SO	B
10, 11, 12, 17, 18, 19	SO ON	B
20, 21, 30, 31, 32, 40, 41, 42	SO	L
23, 24, 26, 33, 34, 36, 43, 44, 46	SO	L

TABLE 5. Listing of all the clusters formed throughout the two-month training experiments from Table 3, the experiments in which those clusters are found, and also what classes of physics scheme are shared throughout the cluster (L = land surface scheme, B = boundary layer scheme, C = cumulus scheme, R = radiation schemes, M = microphysics scheme).

Cluster Members	Experiments	Shared
01, 13, 14, 15, 16	DJF JFM FMA MAM AMJ SON	L B
02, 27, 28, 29	DJF JFM FMA MAM AMJ SON	L B
10, 11, 12	DJF JFM FMA MAM AMJ	L B
17, 19, 37, 39, 47, 49	DJF	L
18, 38, 48	DJF	L C R M
20, 21, 30, 31, 40, 41	DJF JFM FMA	L
22, 32, 42	DJF JFM FMA	L C R M
23, 33, 43	DJF JFM	L C R M
24, 25, 34, 35, 44, 45	DJF JFM	L B
26, 36, 46	DJF JFM FMA	L C R M
17, 37, 47	JFM FMA	L C R M
18, 19, 38, 39, 48, 49	JFM FMA	L
23, 24, 25	FMA	L B
33, 34, 35, 43, 44, 45	FMA	L
17, 18, 19	MAM AMJ ASO	L B
20, 21, 22	MAM AMJ	L B
23, 24, 25, 26	MAM AMJ	L B
30, 31, 32, 40, 41, 42	MAM AMJ	L
33, 34, 35, 36, 43, 44, 45, 46	MAM AMJ SON	L
37, 38, 39, 47, 48, 49	MAM AMJ SON	L
01, 13, 14, 16	MJJ ASO	L B
02, 27, 29, 37, 39, 47, 49	MJJ	L
10, 11, 17, 19	MJJ	B
12, 18, 22, 28, 38, 48	MJJ	C
15, 25, 35, 45	MJJ JJA JAS ASO	L C R M
20, 21, 30, 31, 32, 40, 41, 42	MJJ	L
23, 24, 26, 33, 34, 36, 43, 44, 46	MJJ ASO	L
01, 13, 16, 23, 26, 33, 36, 43, 46	JJA JAS	L C
02, 17, 27, 37, 47	JJA	L C R
10, 19, 20, 29, 30, 39, 40, 49	JJA	C R
11, 21, 31, 41	JJA JAS ASO	L C R M
12, 22, 32, 42	JJA JAS ASO	L C R M
14, 24, 34, 44	JJA JAS	L C R M
18, 28, 38, 48	JJA JAS	L C R M
02, 17, 19, 27, 29, 37, 39, 47, 49	JAS	L
10, 20, 30, 40	JAS ASO	L C R M
02, 27, 28	ASO	L
29, 37, 38, 39, 47, 48, 49	ASO	L
10, 11, 12, 17, 18, 19	SON	B
20, 21, 22, 30, 31, 32, 40, 41, 42	SON	L
23, 24, 25, 26	SON	L B

TABLE 6. Listing of all the clusters formed throughout the three-month training experiments from Table 3, the experiments in which those clusters are found, and also what classes of physics scheme are shared throughout the cluster (L = land surface scheme, B = boundary layer scheme, C = cumulus scheme, R = radiation schemes, M = microphysics scheme).

Cluster Members	Experiments	Shared
01, 13, 14, 15, 16	DJFM JFMA FMAM MAMJ AMJJ ASON	L B
02, 27, 28, 29	DJFM JFMA FMAM MAMJ ASON	L B
10, 11, 12	DJFM JFMA FMAM MAMJ AMJJ	L B
17, 19, 37, 39, 47, 49	DJFM FMAM	L
18, 38, 48	DJFM FMAM	L C R M
20, 21, 30, 31, 40, 41	DJFM JFMA	L
22, 32, 42	DJFM JFMA	L C R M
23, 33, 43	DJFM JFMA	L C R M
24, 25, 34, 35, 44, 45	DJFM JFMA	L B
26, 36, 46	DJFM JFMA	L C R M
20, 21, 22, 30, 31, 32, 40, 41, 42	FMAM	L
23, 24, 25, 26, 36, 46	FMAM	L
33, 34, 35, 43, 44, 45	FMAM	L
17, 18, 19	MAMJ AMJJ ASON	L B
20, 21, 22	MAMJ	L B
23, 24, 25, 26	MAMJ	L B
30, 31, 32, 40, 41, 42	MAMJ	L
33, 34, 35, 36, 43, 44, 45, 46	MAMJ	L
37, 38, 39, 47, 48, 49	MAMJ ASON	L
02, 27, 28	AMJJ	L B
20, 30, 40	AMJJ	L C R M
21, 22, 31, 32, 41, 42	AMJJ	L
23, 24, 26, 33, 34, 36, 43, 44, 46	AMJJ JASO ASON	L
25, 35, 45	AMJJ ASON	L C R M
29, 37, 38, 39, 47, 48, 49	AMJJ	L
01, 13, 16, 23, 26, 36, 36, 43, 46	MJJA JJAS	L C
02, 17, 19, 27, 29, 37, 39, 47, 49	MJJA JJAS JASO	L
10, 11, 20, 21, 30, 31, 40, 41	MJJA ASON	L
12, 22, 32, 42	MJJA JJAS JASO ASON	L C R M
14, 24, 34, 44	MJJA JJAS	L C R M
15, 25, 35, 45	MJJA JJAS JASO	L C R M
18, 28, 38, 48	MJJA JJAS JASO	L C R M
10, 20, 30, 40	JJAS JASO	L C R M
11, 21, 31, 41	JJAS JASO	L C R M
01, 13, 14, 16	JASO	L B

TABLE 7. A list of the ensemble members chosen by the HCA method in each experiment (grouped by verification month).

Experiment (subset)	**Subset members**
DJ (subset H08)	11, 14, 19, 29, 39, 40, 44, 46
JF (subset H10)	10, 14, 29, 40, 42, 43, 45, 46, 48, 49
DJF (subset H10)	11, 14, 29, 40, 42, 43, 45, 46, 48, 49
FM (subset H10)	10, 14, 29, 33, 35, 36, 38, 40, 42, 47
JFM (subset H10)	10, 14, 29, 35, 40, 42, 43, 46, 47, 49
DJFM (subset H10)	11, 14, 29, 40, 42, 43, 45, 46, 48, 49
MA (subset H10)	02, 11, 14, 24, 34, 38, 39, 40, 42, 46
FMA (subset H10)	01, 11, 24, 29, 34, 36, 40, 42, 47, 49
JFMA (subset H10)	11, 14, 29, 35, 40, 42, 43, 46, 47, 49
AM (subset H09)	11, 16, 17, 22, 26, 27, 39, 41, 44
MAM (subset H09)	01, 02, 11, 19, 21, 24, 34, 39, 41
FMAM (subset H08)	01, 02, 11, 26, 41, 44, 48, 49
MJ (subset H10)	02, 10, 16, 19, 21, 23, 30, 33, 45, 49
AMJ (subset H09)	11, 16, 17, 22, 26, 27, 30, 36, 39
MAMJ (subset H09)	01, 02, 11, 19, 21, 24, 34, 39, 41
JJ (subset H10)	02, 12, 16, 31, 32, 34, 35, 36, 37, 39
MJJ (subset H07)	16, 18, 19, 35, 40, 46, 49
AMJJ (subset H09)	02, 11, 16, 19, 40, 41, 45, 46, 49
JA (subset H09)	16, 17, 31, 32, 34, 35, 38, 39, 40
JJA (subset H08)	17, 31, 32, 34, 35, 36, 38, 39
MJJA (subset H07)	32, 34, 35, 36, 38, 40, 49
AS (subset H09)	16, 17, 31, 32, 34, 35, 36, 38, 40
JAS (subset H08)	17, 31, 32, 34, 35, 36, 38, 40
JJAS (subset H08)	19, 31, 32, 34, 35, 36, 38, 40
SO (subset H07)	16, 17, 27, 35, 40, 46, 49
ASO (subset H09)	02, 16, 19, 31, 32, 35, 39, 40, 46
JASO (subset H08)	16, 31, 32, 35, 38, 40, 46, 49
ON (subset H07)	01, 02, 12, 24, 34, 39, 42
SON (subset H07)	12, 16, 23, 27, 38, 42, 46
ASON (subset H08)	16, 19, 27, 32, 35, 39, 40, 46

TABLE 8. For each experiment listed, the percentage of observations of each lead time–surface variable combination for which the two-sided Kolmogorov-Smirnov test indicates that the full and subset ensemble distributions are statistically similar.

Experiment	12-h T	24-h T	36-h T	48-h T	12-h U	24-h U	36-h U	48-h U	12-h V	24-h V	36-h V	48-h V
DJ	100.0	100.0	100.0	99.90	99.86	99.62	99.93	99.65	99.93	99.85	99.93	99.93
JF	100.0	99.95	100.0	100.0	99.58	99.85	99.78	99.06	99.37	99.71	99.78	99.39
FM	100.0	100.0	100.0	100.0	100.0	99.93	99.93	99.93	100.0	100.0	100.0	99.93
MA	100.0	100.0	100.0	99.95	100.0	99.93	99.92	100.0	100.0	99.93	99.69	100.0
AM	100.0	99.90	99.95	99.85	100.0	99.62	99.85	99.92	99.80	99.77	99.93	99.85
MJ	99.95	99.81	99.95	99.94	99.80	100.0	99.78	99.74	100.0	99.93	99.35	99.91
JJ	99.71	99.90	99.95	99.75	98.26	98.77	99.08	99.42	97.70	97.55	99.54	98.54
JA	99.81	99.90	100.0	99.90	97.71	98.12	98.79	99.54	96.60	98.74	99.27	99.27
AS	99.47	99.22	99.65	100.0	94.15	98.53	98.23	98.99	95.35	98.05	98.36	98.99
SO	99.86	99.86	99.37	99.66	99.93	99.67	99.36	100.0	99.86	99.93	99.29	99.47
ON	100.0	99.95	99.95	99.85	99.74	99.87	99.87	99.74	99.80	99.74	99.81	99.80
DJF	100.0	100.0	100.0	100.0	99.86	99.71	99.34	98.99	99.23	99.78	99.56	99.60
JFM	100.0	100.0	100.0	100.0	99.58	99.87	99.67	99.87	99.93	99.67	99.87	99.80
FMA	100.0	100.0	100.0	100.0	100.0	100.0	99.92	100.0	99.91	99.93	99.77	100.0
MAM	99.95	98.67	99.66	99.03	99.33	99.62	99.13	99.69	99.46	99.77	99.42	99.23
AMJ	99.95	99.71	99.71	99.63	99.73	99.60	99.56	99.57	99.67	99.80	99.93	99.74
MJJ	99.56	98.97	99.95	99.26	98.75	99.11	99.08	98.54	97.77	98.77	99.08	98.47
JJA	97.49	99.56	98.03	99.85	85.08	89.96	91.56	97.62	82.72	90.31	93.70	92.65
JAS	99.19	98.79	98.41	99.37	86.24	93.72	93.91	95.84	88.16	94.56	95.55	96.38
ASO	99.95	100.0	100.0	100.0	99.50	99.93	99.79	100.0	99.79	100.0	99.43	99.93
SON	99.18	99.75	99.37	98.97	99.22	99.22	99.03	99.34	99.28	98.89	99.16	98.88
DJFM	100.0	100.0	100.0	100.0	97.97	99.54	99.40	99.67	99.09	99.41	99.67	99.60
JFMA	99.94	100.0	100.0	100.0	99.73	99.86	99.85	99.79	99.91	99.93	99.54	99.71
FMAM	100.0	100.0	100.0	99.90	99.80	99.92	99.93	99.85	99.87	99.92	99.56	99.69
MAMJ	98.94	96.99	98.27	98.41	98.73	98.14	99.20	99.32	98.80	99.20	98.77	99.32
AMJJ	99.95	99.61	99.90	99.85	99.44	99.73	99.54	99.56	99.58	99.86	99.68	99.49
MJJA	98.12	98.54	98.03	99.17	89.52	91.70	93.50	95.10	89.87	92.12	94.84	94.51
JJAS	99.09	99.18	98.31	99.13	87.03	94.21	99.39	96.45	87.03	94.21	99.39	96.45
JASO	99.95	99.04	99.57	99.22	95.99	97.41	97.87	98.54	97.14	97.61	99.01	97.16
ASON	100.0	99.90	99.81	99.90	99.67	99.74	99.68	99.80	99.74	99.67	99.87	99.47

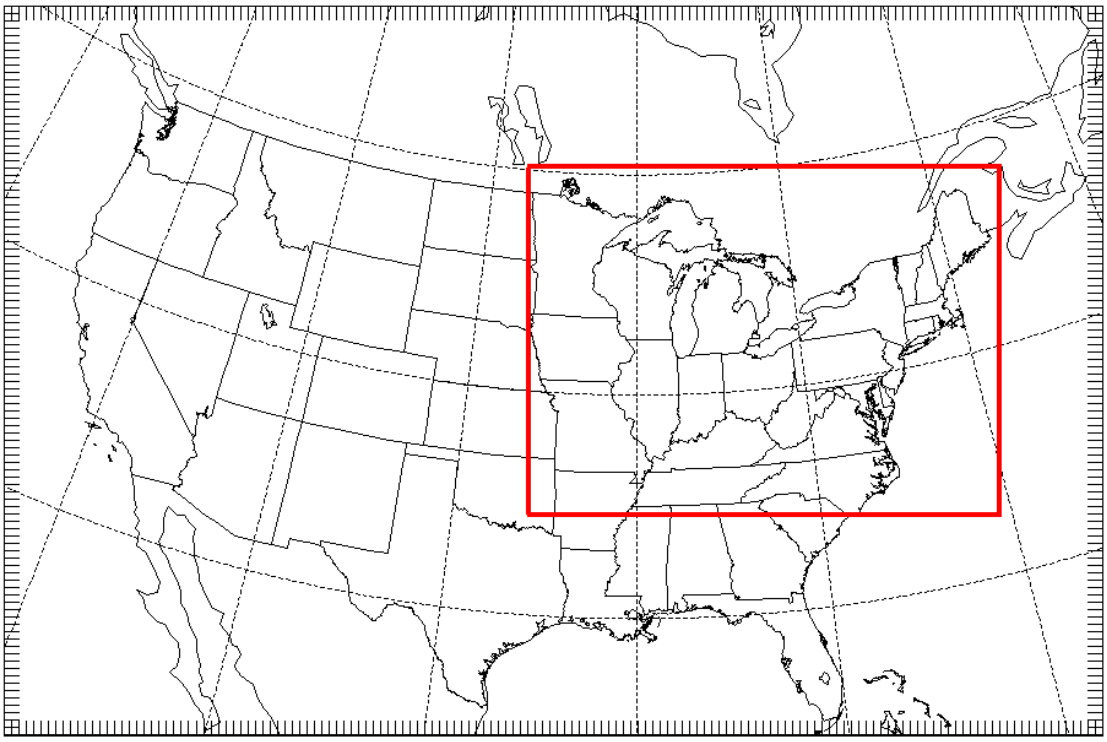

FIG. 1. WRF domains used in this study. The outer domain has a 36-km horizontal resolution, and the inner domain (outlined in red) has a 12-km horizontal resolution.

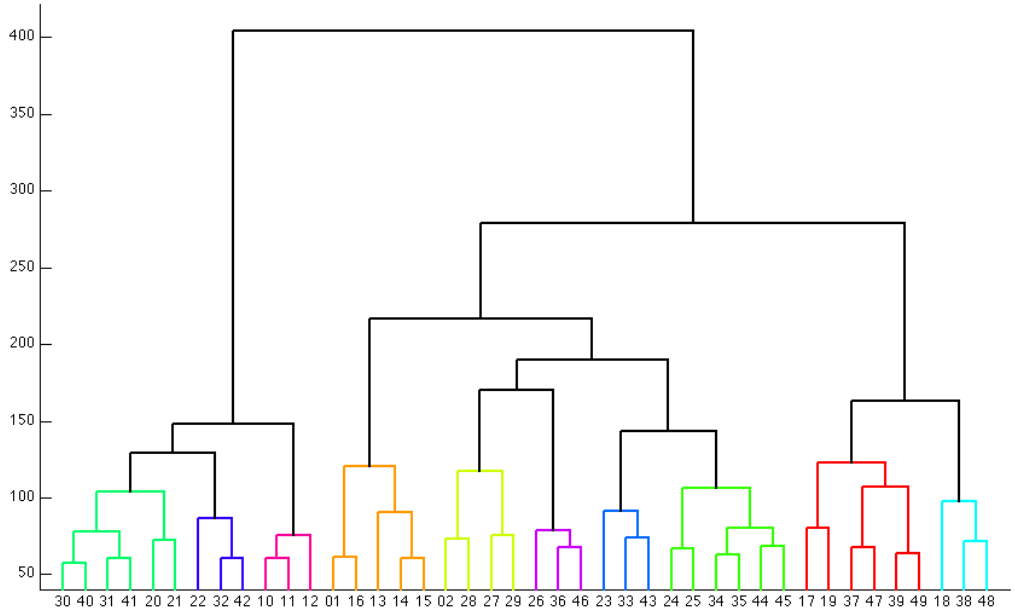

FIG. 2. HCA dendrogram illustrating how the ensemble groups into ten clusters for the DJF experiment. The colored branches of the dendrogram are the ten clusters determined by HCA for this experiment. The vertical axis is the distance metric from Eq. 5.

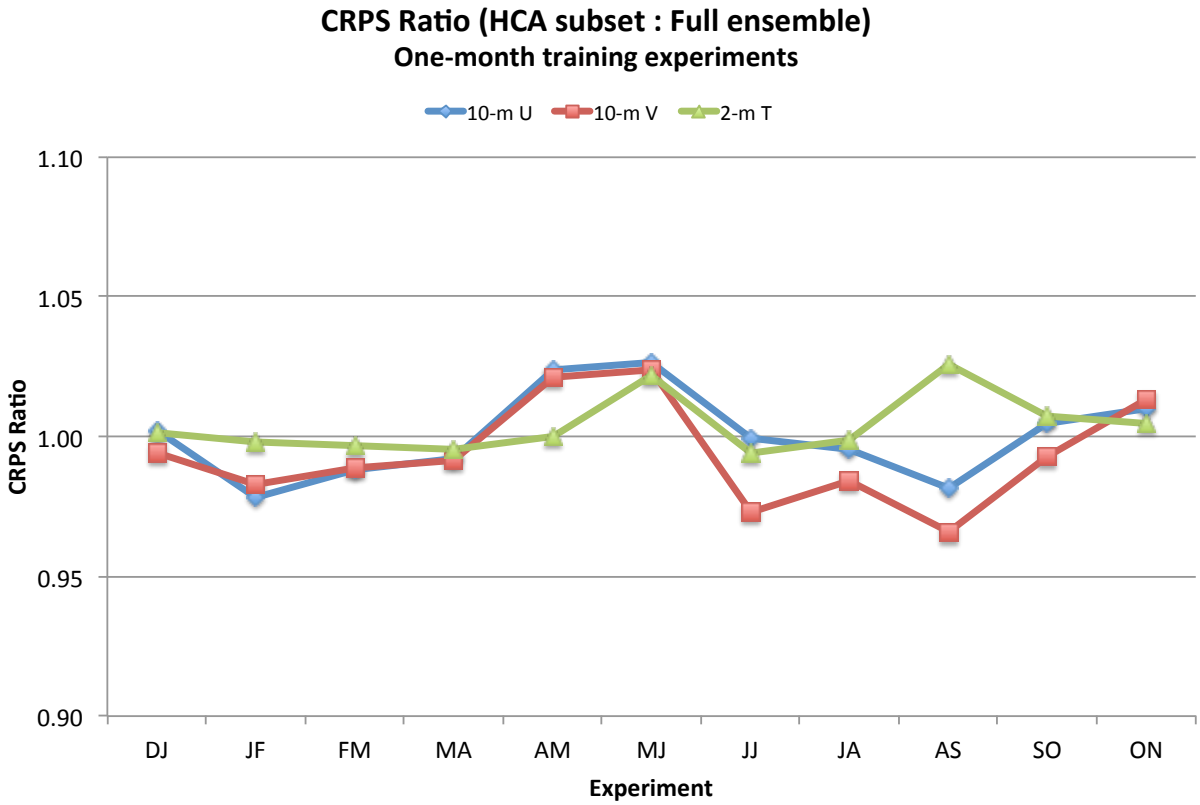

FIG. 3. Ratios of CRPS for HCA subset ensembles to the CRPS of the full 42-member ensemble, averaged over all forecast lead times for 10-m U (blue line), 10-m V (red line), and 2-m T (green line) over all one-month training experiments. These ratios are calculated for BMA-calibrated ensembles.

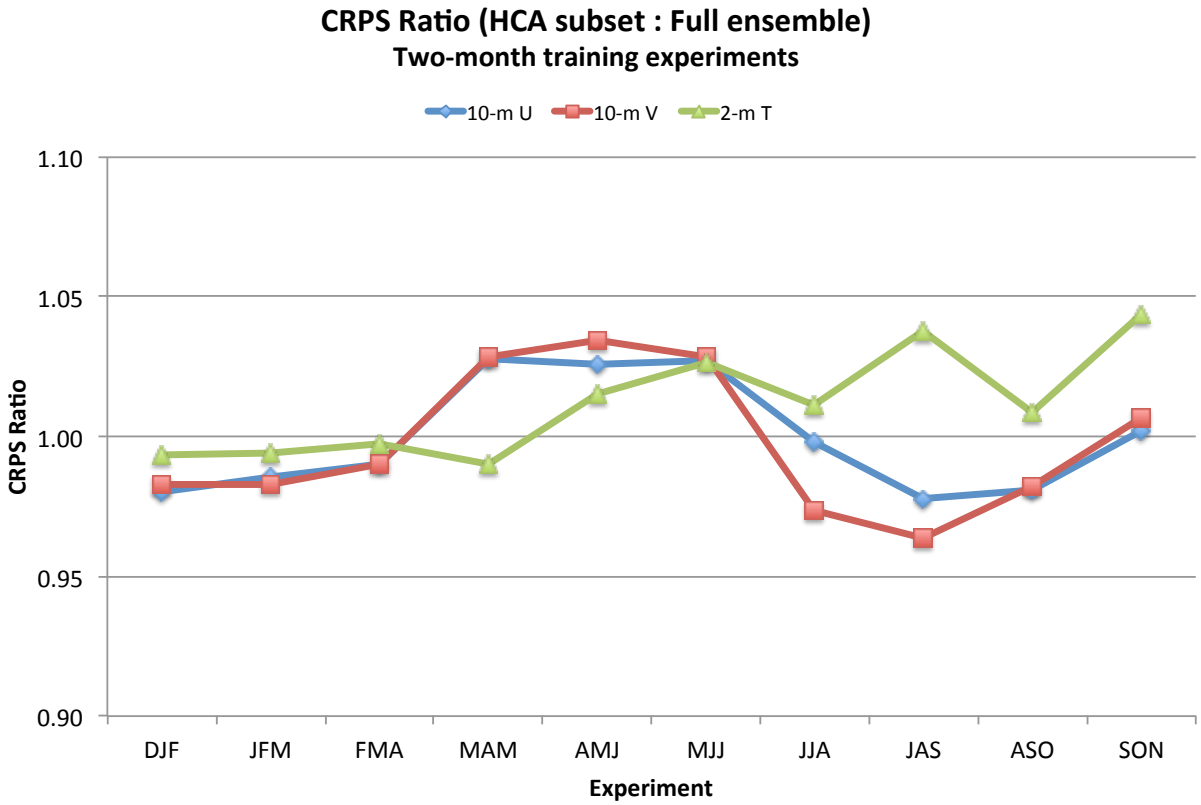

FIG. 4. Ratios of CRPS for HCA subset ensembles to the CRPS of the full 42-member ensemble, averaged over all forecast lead times for 10-m U (blue line), 10-m V (red line), and 2-m T (green line) over all two-month training experiments. These ratios are calculated for BMA-calibrated ensembles.

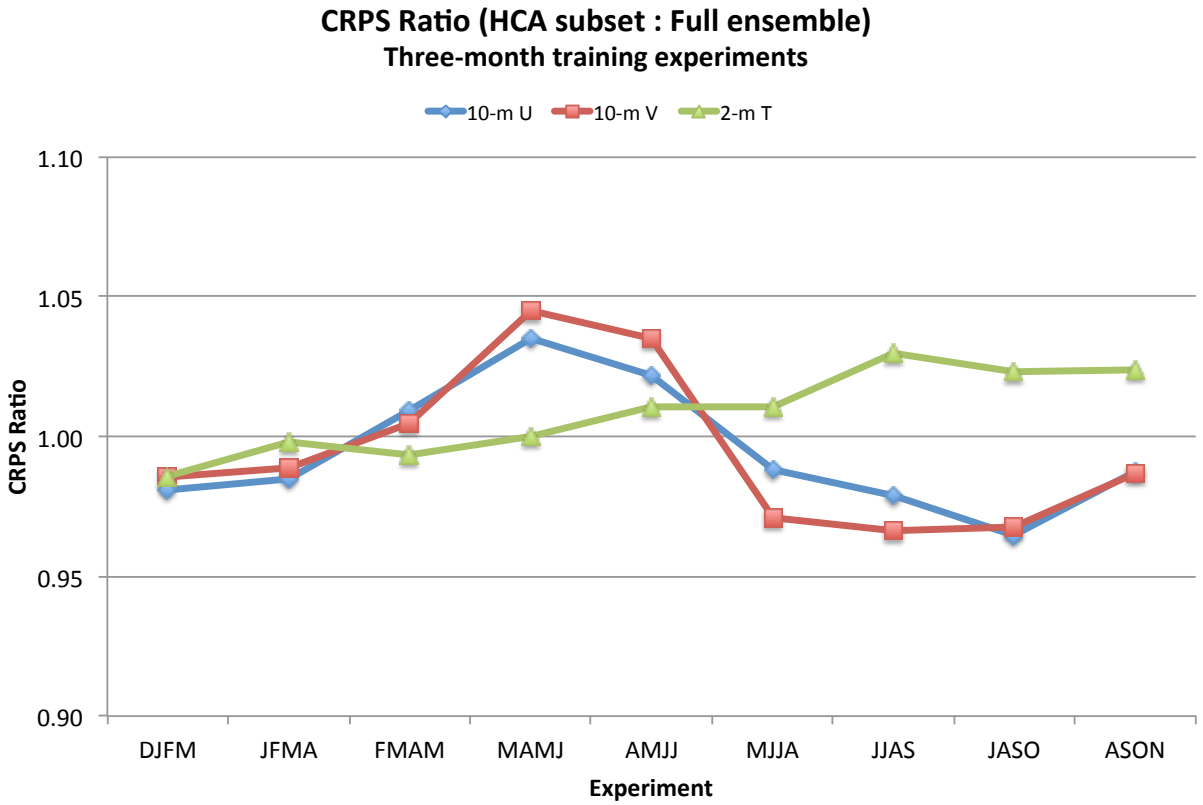

FIG. 5. Ratios of CRPS for HCA subset ensembles to the CRPS of the full 42-member ensemble, averaged over all forecast lead times for 10-m U (blue line), 10-m V (red line), and 2-m T (green line) over all three-month training experiments. These ratios are calculated for BMA-calibrated ensembles.

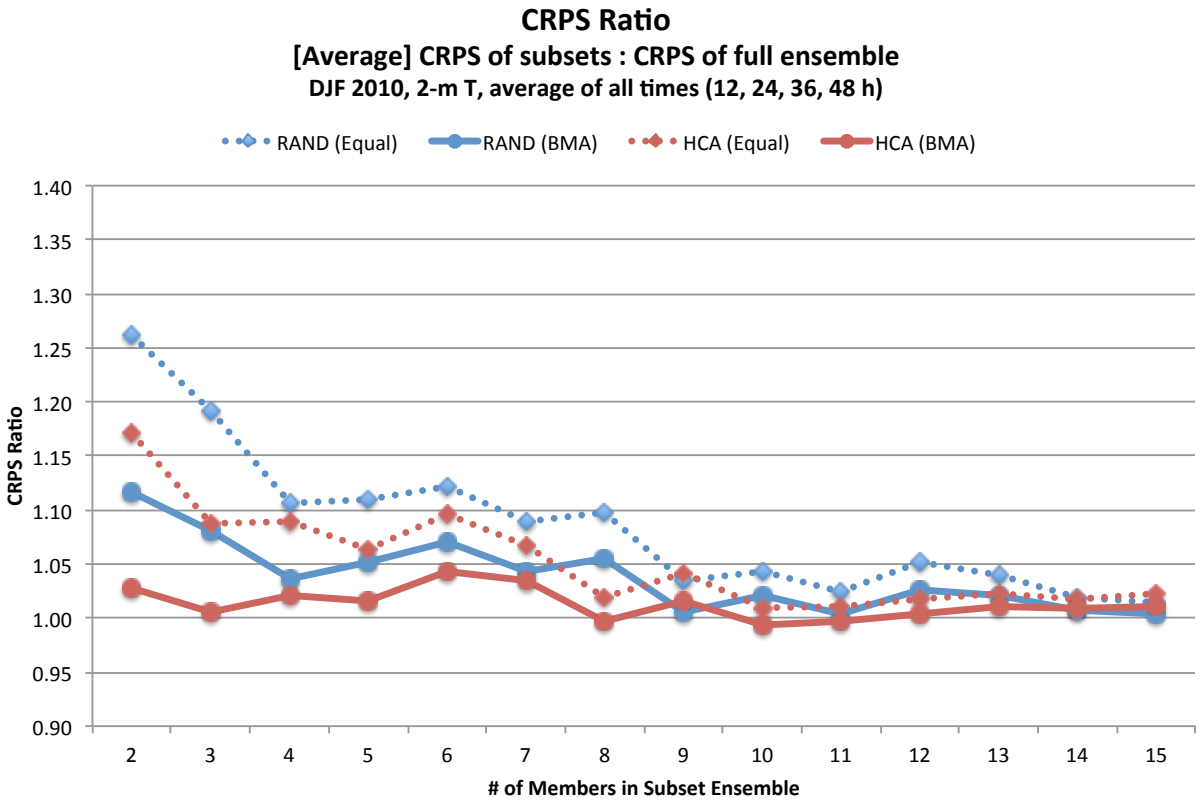

FIG. 6. Ratios of CRPS for subset ensembles of a range of sizes to the CRPS of the full 42-member ensemble, averaged over all forecast lead times for 2-m T in the DJF experiment. The blue line corresponds to the average CRPS ratio for ten random subset ensembles of each ensemble size, while the red line is for the CRPS ratio of HCA-determined subset ensembles. The dashed lines are the ratios for the uncalibrated (equal-weighted) ensembles, while the solid lines are the ratios for the calibrated (BMA-weighted) ensembles.

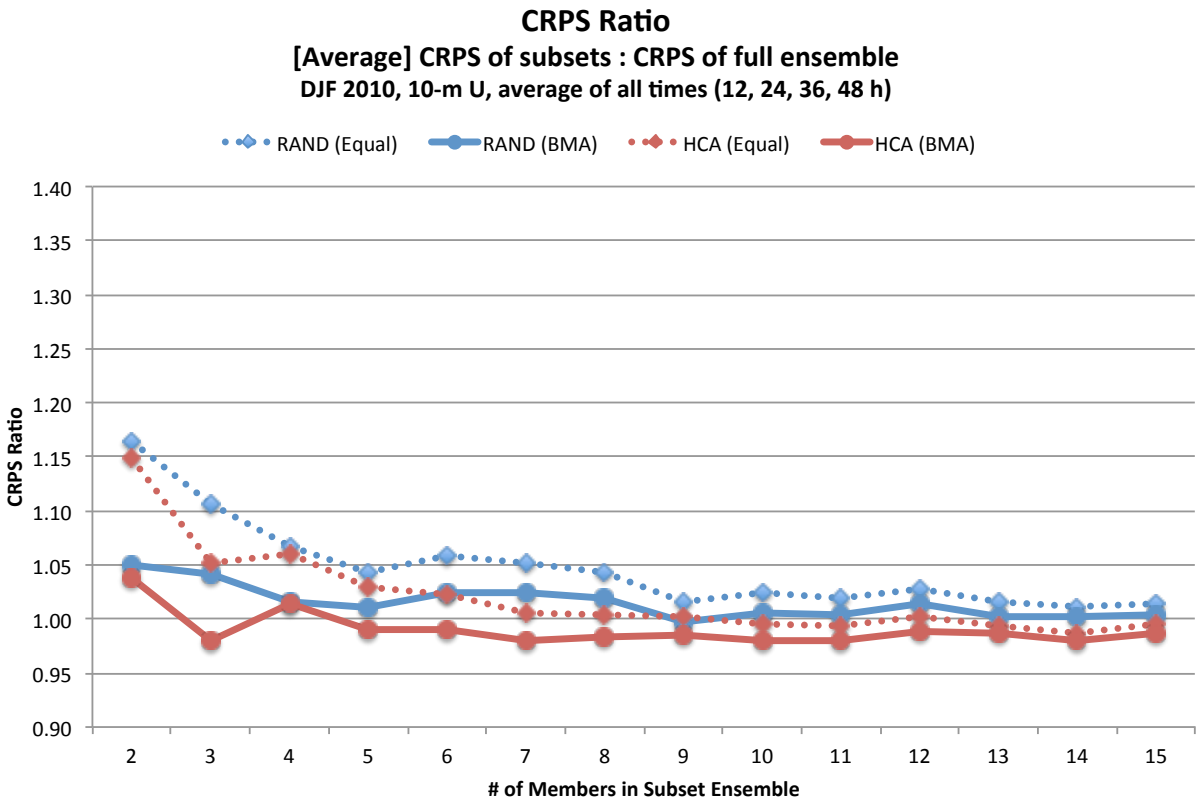

FIG. 7. Ratios of CRPS for subset ensembles of a range of sizes to the CRPS of the full 42-member ensemble, averaged over all forecast lead times for 10-m U in the DJF experiment. The blue line corresponds to the average CRPS ratio for ten random subset ensembles of each ensemble size, while the red line is for the CRPS ratio of HCA-determined subset ensembles. The dashed lines are the ratios for the uncalibrated (equal-weighted) ensembles, while the solid lines are the ratios for the calibrated (BMA-weighted) ensembles.

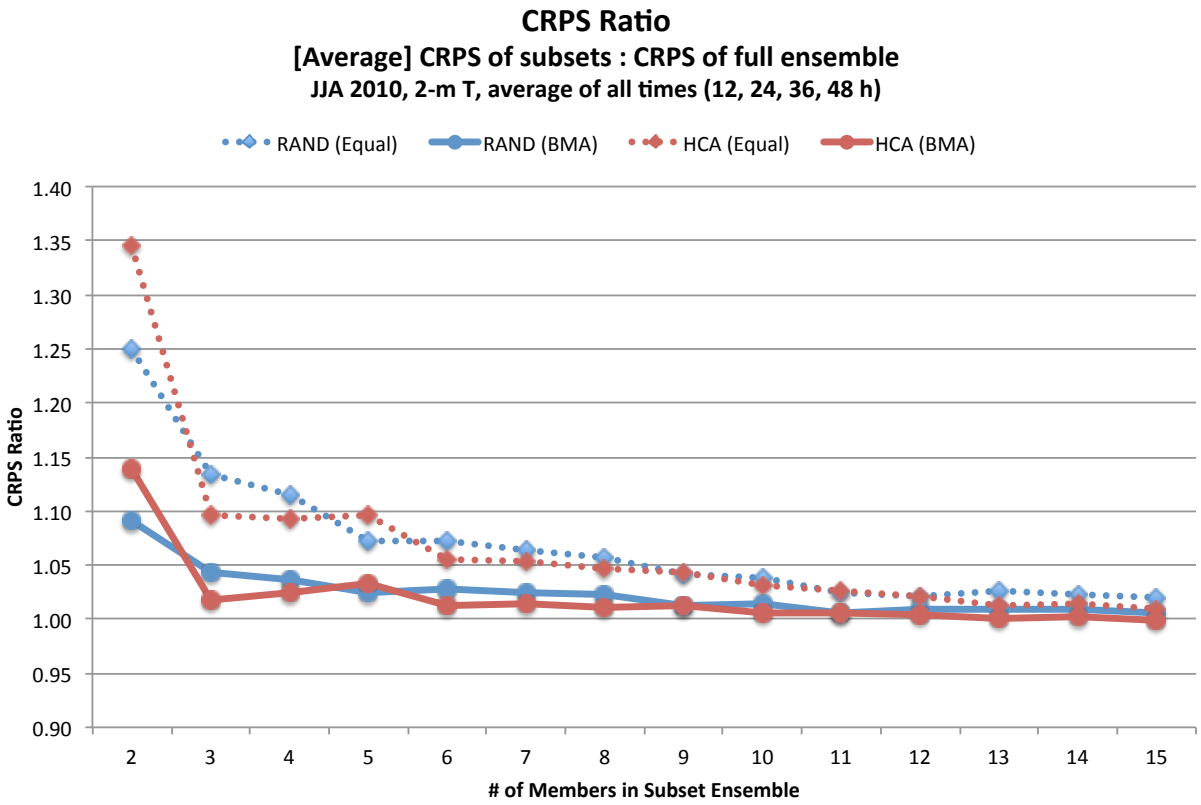

FIG. 8. Ratios of CRPS for subset ensembles of a range of sizes to the CRPS of the full 42-member ensemble, averaged over all forecast lead times for 2-m T in the JJA experiment. The blue line corresponds to the average CRPS ratio for ten random subset ensembles of each ensemble size, while the red line is for the CRPS ratio of HCA-determined subset ensembles. The dashed lines are the ratios for the uncalibrated (equal-weighted) ensembles, while the solid lines are the ratios for the calibrated (BMA-weighted) ensembles.

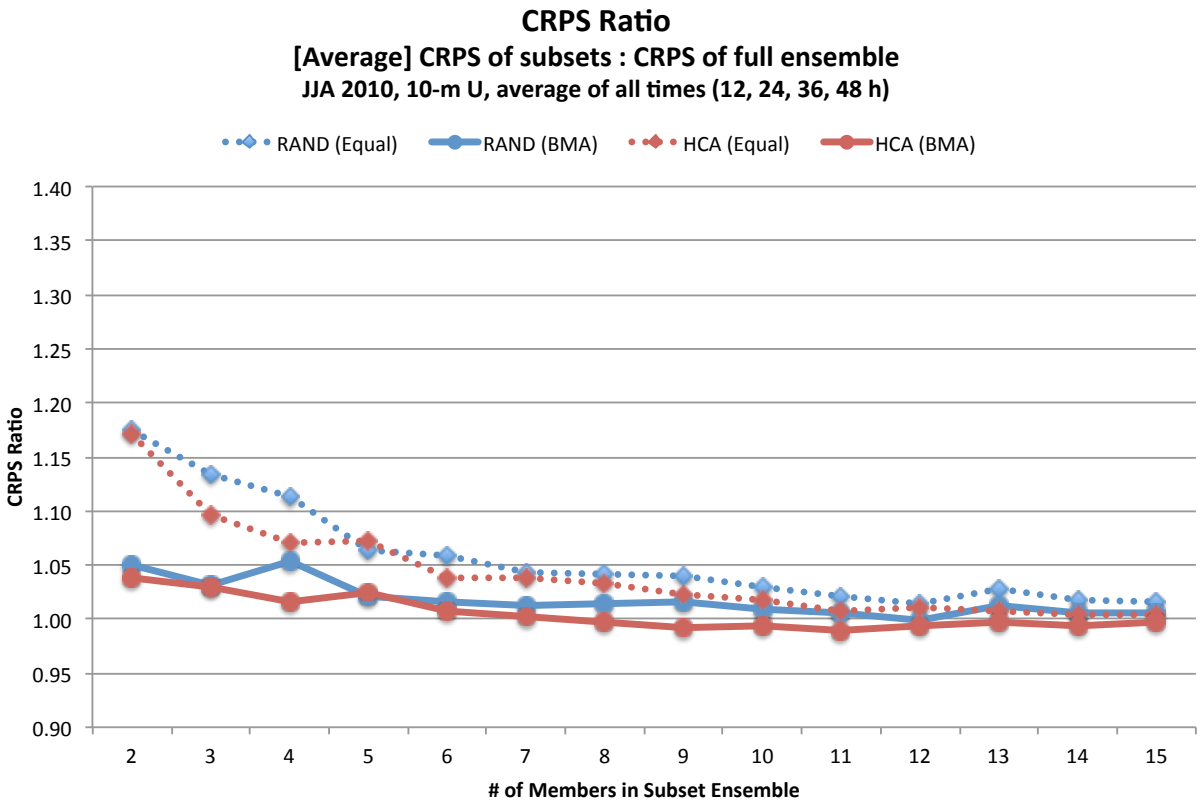

FIG. 9. Ratios of CRPS for subset ensembles of a range of sizes to the CRPS of the full 42-member ensemble, averaged over all forecast lead times for 10-m U in the JJA experiment. The blue line corresponds to the average CRPS ratio for ten random subset ensembles of each ensemble size, while the red line is for the CRPS ratio of HCA-determined subset ensembles. The dashed lines are the ratios for the uncalibrated (equal-weighted) ensembles, while the solid lines are the ratios for the calibrated (BMA-weighted) ensembles.