Journal of Climatology & Weather Forecasting

Lee et al., J Climatol Weather Forecasting 2016, 4:1
http://dx.doi.org/10.4172/2332-2594.1000156

Research Article Open access

Down-Selecting Numerical Weather Prediction Multi-Physics Ensembles with Hierarchical Cluster Analysis

Jared A Lee[1,2*], **Haupt SE**[1,2] **and Young GS**[2]

[1]*Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO, USA*

[2]*Department of Meteorology, The Pennsylvania State University, University Park, PA, USA*

[*]**Corresponding author:** Jared A Lee, Research Applications Laboratory, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO, USA, Tel: 303-497-1740; E-mail: jaredlee@ucar.edu

Received date: Feb 09, 2016; **Accepted date:** Feb 19, 2016; **Published date:** Feb 29, 2016

Abstract

The goal of ensemble down selection is to retain the subset of ensemble members that span the uncertainty space of the forecast while eliminating those that are most redundant. There are hundreds of combinations of physics schemes that can be used in typical numerical weather prediction (NWP) models. Limited computational resources, however, force us to constrain the size of NWP ensembles, and to choose what combinations of physics schemes to use. Ensemble down selection can help guide those choices, and also yield information about how many ensemble members are necessary. In this study we examine the use of hierarchical cluster analysis (HCA) as an objective down selection technique.

To test the performance of HCA across multiple seasons, a 42 member multi physics ensemble is configured and run, with 48 h forecasts initialized every fifth day for twelve months. HCA is performed on forecast errors of low level temperature and wind components over training periods of one, two, and three months. How the ensemble members cluster is found to change by season. The full and subset ensembles are then calibrated using Bayesian model averaging (BMA). The uncalibrated and calibrated ensembles are verified over one month periods. Statistical tests indicate a likelihood that the subset ensemble comes from same distribution as the full ensemble, and have verification scores nearly the same as the full ensemble. Furthermore, intelligently down selecting a subset ensemble with HCA outperforms random down selection.

Keywords: Numerical weather prediction; Radiation; Land surface; Surface layer; Boundary layer; Microphysics

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 models is an area of active research in the community [1-7]. Limited computing resources force sacrifices to be made in balancing several considerations including ensemble size, model resolution, and geographic coverage.

Model error is a key contributor to forecast error in NWP ensembles, particularly for short range forecasts in the atmospheric boundary layer [2,3,8] types of model error include uncertainty arising from the way physical processes are being represented in any given parameterization scheme, and scale truncation (a low pass filter) associated with discretization and numerical scheme. Common approaches for representing model uncertainty include multi-model, multi-physics, and stochastic perturbation ensembles, or combinations thereof [1,5,6].

When constructing a multi-physics ensemble, it is usually unclear what sets of physics schemes are best, or how many members to include. The large number of available physics options exacerbates this problem. For instance, in the Advanced Research Weather Research and Forecasting (WRF-ARW) NWP model [9], for each class of physics scheme (shortwave radiation, long wave radiation, land surface, surface layer, boundary layer, microphysics, and cumulus / convection) there are several options, resulting in thousands of possible combinations of physics schemes from which to choose (over 21,600 possible physics configurations in WRF-ARW v3.3). It is entirely impractical to run a several hundred or several thousand member multi-physics ensemble. Therefore, some guidance in how to configure a multi-physics ensemble would be helpful to many in the community.

In an ideal ensemble, each ensemble member would be an equally likely future state of the atmosphere. Because parameterizations are by nature imperfect, however, there is no assurance in a multi-physics ensemble that each ensemble member will be equally likely. Each individual physics parameterization scheme has its own biases and errors that in turn can depend on region, season, or weather situation (i.e., flow dependent errors). Likewise, each combination of physics schemes has its own typical errors as a result of compensating biases and interactions, and some combinations may be generally more accurate than others.

One approach to choosing a smaller set of ensemble members for computational efficiency could be to select the physics configurations that produce the lowest mean error over some study period. Unless a particular physics configuration can be shown to be consistently an outlier, however, excluding the ensemble members with higher mean errors may not result in a better ensemble forecast distribution. Such an approach may be helpful for predicting the mean, but may be detrimental for quantifying forecast uncertainty.

Lee et al. [7] proposed an objective method using principal component analysis (PCA) 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. That study examined down selection from a 24 member ensemble over a single winter season. Lee [10] compared PCA with two other down selection techniques, K means cluster analysis (KCA) and hierarchical cluster analysis (HCA), and found HCA to be the preferred technique of the three. In this study we extend HCA testing to a 42 member multi-physics ensemble and to a 12 month period to examine the robustness of intelligently down selecting a multi-physics ensemble over different seasons, and with different training periods for the down selection technique. In this study we also demonstrate that down selecting with HCA yields improved verification scores over simply doing a random down selection. We apply Bayesian model averaging (BMA) [11,12] to calibrate the forecasts for both the full ensemble and the HCA subset ensembles (section 2c). The actionable result from this study is the minimum training period required for successful multi-physics ensemble down selection via hierarchical clustering and Bayesian model averaging.

We discuss our ensemble configuration, calibration, and verification procedures in section 2. In section 3 we describe the general down-selection procedure with HCA. In section 4 we examine the seasonal effects on clustering and down-selection using HCA. Section 5 provides a summary and conclusions.

Methods

Ensemble configuration

Using WRF-ARW v3.3 we create a 42-member physics ensemble. The two control members (CTL 01 and CTL 02) are Developmental Testbed Center (DTC) Reference Configurations for WRF-ARW v3 [13]. For the remaining forty members we choose combinations of physics schemes by selecting one option from each class of physics scheme (i.e., microphysics, radiation, land surface, surface layer / boundary layer, and cumulus schemes), as detailed in (Table 1). In WRF, each boundary layer scheme generally only works with a particular surface layer scheme, so those schemes are used as matched pairs. Additionally, particular pairings of long wave and shortwave radiation schemes are generally recommended. Each set of ten members has a different pair of surface layer / boundary layer schemes, but otherwise identical combinations of microphysics, radiation, land surface, and cumulus schemes. These repeating sets of combinations of physics schemes were chosen both to include a systematic variety of combinations, and to make patterns in the cluster analysis easier to discern at a glance.

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 1: Physics schemes for the 42 member WRF multiphysics ensemble. Descriptions and references for schemes are contained in [9]. We use the same slightly modified version of the Mellor Yamada Janjic (MYJ) ABL scheme as in [7].

Because our aim is to isolate the effects of model uncertainty, our multi-physics ensemble uses the same initial conditions (ICs) and lateral boundary conditions (LBCs) for all members. We made this decision for two reasons. First, as mentioned above, physics variability is a crucial source of uncertainty for low level, short range forecasts. Second, no down selection approach would be physically meaningful if applied to an ensemble with only equally likely IC / LBC perturbations, because members would then be statistically indistinguishable and exchangeable [14].

We initialize the 48 h forecasts every fifth day at 0000 UTC from 1 Dec 2009 through 26 Nov 2010, for a total of eighteen 48 h forecasts during each season (Table 2). We choose this frequency in order to reduce temporal correlations between consecutive forecasts and to reduce the computational burden of the experiments.

The model's coarse domain (Figure 1) uses a horizontal grid spacing of 36 km, while the one-way nested inner domain uses 12 km grid spacing. There are 45 full vertical levels, with high vertical resolution in the lowest 2 km (24 full levels) to resolve ABL processes. We use time steps of 90 s and 30 s for the coarse and fine domains, respectively. Such small time steps were 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 were retained for consistency for all the other model runs.

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 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.

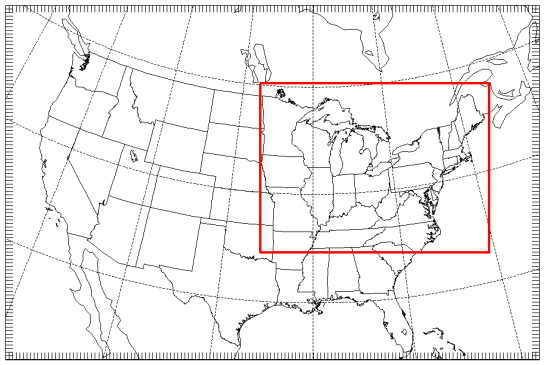

Figure 1: WRF domains used in this study. The outer domain has a 36 km horizontal resolution, and the inner domain (outlined in red) has a 12 km horizontal resolution.

The LBCs for all members come from the 0.5° × 0.5° resolution Global Forecast System (GFS) [15] 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 (RTG) 0.083° dataset, and daily snow cover analyses from the National Ice Center.

The ICs use the 0h GFS forecast and are blended with standard World Meteorological Organization (WMO) observations via the Obsgrid objective analysis software. This blending yields an improved initial state. Obsgrid is part of the WRF modeling system, and uses multiple passes of the objective analysis scheme to modify the first-guess field [16]. In Obsgrid we use the Cressman objective analysis scheme, assigning each observation a distance-weighted flow-dependent radius of influence [17]. We note that Obsgrid does not operate on the GFS fields directly, but on the GFS fields interpolated to our WRF grids, which allows for a better fit to the observations. This process has been shown to improve the initial conditions in other mesoscale modeling studies, including [18,19].

Ensemble calibration

To correct for biases in the first and second moments of the raw ensemble distribution (i.e., to calibrate), we use Bayesian model averaging [12]. BMA estimates the optimal weights and standard deviations for each member of the ensemble by training these parameters to best match the observations during a training period (6, 12 or 18 consecutive forecasts in this study, corresponding to one, two, and three-month training periods, respectively). The BMA weights and standard deviations are then applied to forecasts in a verification period (six consecutive forecasts) to create a better ensemble PDF.

We perform calibration with BMA on the forecasts directly so that we modify the forecast PDF. 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, 850 and 700 hPa). As in [7,12] we assume a normal distribution for the temperature. Whereas [7] assumed a separate normal distribution for each wind component, here we assume a bivariate normal distribution for the wind components, similar to the approach of [20], and perform BMA on the u-wind and v-wind together at each level and lead time. A single domain wide calibration is performed for each variable at each lead time at each level, using observations as ground truth. We also calibrate both the full and subset ensembles, and we calculate verification statistics on both the calibrated and uncalibrated ensembles in order to compare how well the down selection procedure works both with and without calibration.

The ensemble member weights generated by BMA for each variable, vertical level, and lead time generally are similar; there is not a small subset of members that have substantially larger weights than the rest of the ensemble members. The relatively even BMA weights indicate that all 42 members of our physics ensemble are of roughly comparable quality [10].

Verification and metrics

We perform down selection, calibration, 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. Before any down selection, calibration, or verification, for each ensemble member's forecasts we apply a single domain wide average bias correction for each forecast variable at each vertical level and forecast lead time, as in [7]. When we perform the down selection we also normalize the errors by subtracting the mean and then dividing by the standard deviation for each variable, lead time and vertical level during the training period so that errors of variables with different units can be put on the same magnitude scale, for fairer treatment.

Standard WMO surface and upper air observations are used to verify our ensemble forecasts. We perform down selection and verification at four lead times: 12, 24, 36, and 48 h, i.e., those times for which standard radiosonde observations are available (0000 and 1200 UTC). To quality control these observations against the GFS analysis

fields that are interpolated by the WRF Pre-processing System (WPS), Obsgrid is used as described above.

We divide our yearlong forecast dataset into roughly month long groups of six forecasts 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 the impact of training period length.

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

The observations used in this study are temperature and horizontal wind components at four levels: the surface and the mandatory upper-air levels of 925 hPa, 850 hPa, and 700 hPa. We choose these levels because we are primarily concerned with factors relevant to forecasting in the lower troposphere, and in particular the ABL. Additionally, by choosing a consistent set of mandatory levels we maximize the number of usable sounding observations, and avoid introducing interpolation error into the observations. Model predictions are horizontally and vertically interpolated to the observation locations. In the horizontal we use bilinear interpolation, and in the vertical we use linear interpolation between 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 temperature, wind direction, wind speed, vector wind difference, and the zonal (u) and meridional (v) components of the wind.

We use both the standard root mean squared error (RMSE) and continuous ranked probability score (CRPS) as verification metrics. The CRPS assesses both the accuracy and sharpness of a probabilistic forecast distribution and is defined as [21,22].

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

where N is the total number of observations, $p_i^f(x)$ is the cumulative distribution function (CDF) 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 (a Heaviside function). Both RMSE and CRPS are negatively oriented metrics (i.e., lower scores are better) with a perfect score of 0. RMSE and CRPS are also both suitable verification metrics for continuous predictands like temperature and wind. CRPS is also a strictly proper scoring rule [23].

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

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:

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

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, and would imply that the subset ensemble is sharper and / or more accurate than the full ensemble. Our values of RMSE, CRPS, and their associated subset to full ratio scores, are calculated using N = 1000 bootstrap resamples with replacement, so that we can also compute sample standard deviations with our sample mean values of these statistics [24].

Comparing the RMSE or CRPS of two ensembles does not directly indicate the similarity of the distributions of the two ensembles, however. Thus, we use the two-sample Komolgorov-Smirnov (K-S) test

to assess the similarity of the empirical CDFs of the full and down-selected subset ensembles. The null hypothesis for the K-S test is that the two samples 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 samples, n_1 observations of x_1 and n_2 observations of x_2 [22]:

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

The null hypothesis for the two-sample K - S test is rejected at the 95% confidence level if [22]

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

The two-sample K-S test is computed for every observation location in the verification period for the various experiments. We choose not to merge all locations for the K-S test because we do not expect the null hypothesis to be equally true everywhere. If the null hypothesis is supported by the vast majority of observation locations, then the likelihood is quite small that the full ensemble and subset ensemble forecast distributions differ.

Ensemble Down-Selection with HCA

We use HCA as our ensemble down selection technique. HCA has been used in several studies to group together similar members in an NWP ensemble forecast [25-27] and in an air quality ensemble [28]. Each of those studies applied HCA to multi-model ensembles, with the general finding that ensemble members clustered together by model. In contrast, in this study we apply HCA to a single model, multi physics ensemble to focus on the question of how many ensemble members are needed to represent model error from a single modeling system (i.e., WRF). An HCA data vector is defined initially as the normalized forecast errors of an ensemble member that corresponds to a singleton cluster. At each step of the algorithm, the two clusters that are closest to each other according to some distance metric are combined. If uninterrupted, this process continues until eventually all of the data vectors are combined into a single cluster. In this study we stop the clustering procedure while there are still several clusters, and the criterion we use to do that is described below.

The version of HCA we use is Ward's minimum variance method, known more simply as Ward's method [22]. 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 to compare clusters r and s.

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

where $\|\|_2$ is the Euclidean distance, $\overline{x_r}$ and $\overline{x_s}$ are the cluster centroids, and n_r and n_s are the numbers of elements in the clusters. Ward's method is used frequently in studies that employ HCA [25-27]. Additionally, we find that alternate versions of HCA yield results that are no better than Ward's method (not shown).

Dendrograms display the order in which sub-clusters merge together in HCA; two sub-clusters that merge at a relatively low height on the dendrogram are considered to be similar to one another. An example dendrogram for the DJF experiment is shown in (Figure 2). To determine the number of clusters present in the data for each experiment, we use the height on the dendrogram at which each cluster has at least three members. The HCA down-selected subset ensembles are comprised of the ensemble members that are closest to each cluster centroid.

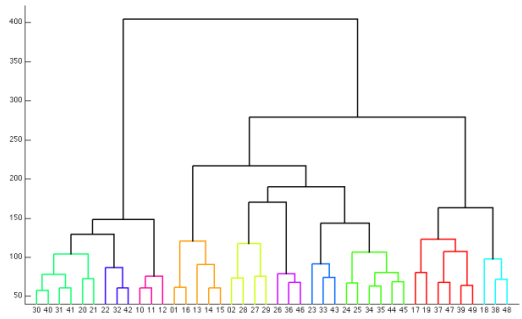

Figure 2: HCA dendrogram illustrating how the ensemble groups into ten clusters for the DJF experiment. The colored branches of the dendrogram are the ten clusters determined by HCA for this experiment. Within each of the ten clusters, the member closest to the cluster centroid is chosen to form subset H10. The vertical axis is the distance metric from Eqn 5.

The HCA is performed on bias corrected, normalized temperature errors and normalized vector wind differences (VWD) over the training period for each experiment, combining data from all four forecast lead times at all four levels. The down-selection is performed prior to calibrating the ensemble with BMA to avoid adding an unnecessary layer of complexity to assessing the impact of down-selection. Down-selection can be performed in either a univariate framework (on temperature errors and VWD separately) or in a multivariate framework (normalizing and then combining temperature errors and VWD into a single data vector). Lee [10] demonstrated that there is little change in verification results between subset ensembles from univariate down-selection and multivariate down-selection. Therefore, we perform a single, multivariate down-selection for each experiment here, as it is more straightforward to analyze. For additional simplification, we combine the model data from all four lead times for down-selection, instead of performing a separate down-selection on each lead time, because there is little difference between the two approaches [10].

For additional demonstration that down-selection using HCA has value, we also compare HCA to a random down-selection method. By relaxing our requirement that each cluster have at least three members, we can examine a range of ensemble sizes and also assess whether there is an ensemble size above which additional members no longer add forecast skill.

Results

Sensitivity to training window and season

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.

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 4: Listing of all the clusters (see Table 1 for member descriptions) formed throughout the one-month training experiments from Table 3, the experiments in which those clusters are found, and also what class(es) of physics scheme(s) is in common throughout the members of the cluster (L = land surface scheme, B = boundary layer / surface 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 5: Listing of all the clusters formed throughout the two months training experiments from Table 3, the experiments in which those clusters are found, and also what class(es) of physics scheme(s) is in common throughout the members of the cluster.

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 6: Listing of all the clusters formed throughout the three-months training experiments from Table 3, the experiments in which those clusters are found, and also what class(es) of physics scheme(s) is in common throughout the members of the cluster.

First, members cluster differently in different seasons. Tables 4-6 show this clearly, with several identical clusters in the experiments that have overlapping training periods. This behavior is seen within each season, though somewhat more strongly in winter and summer than in 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 members share 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 members share 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, thus making the clustering sensitive to the large effect that the land surface scheme has on near-surface parameters [29,30].

Boundary layer and cumulus parameterizations generally appear to be of secondary importance to the clustering. As can be seen in Tables 4-6, cluster members 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 geographic 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, especially for low-level temperature and wind. It should be noted, however, that in other regions, such as the U.S. west coast, microphysics and radiation schemes are likely to have a larger impact on surface variables than cumulus schemes because of the modeling of marine stratus.

In the summer the cluster members tend to have a common cumulus and / or land surface scheme, but typically not a boundary layer scheme. In the transition seasons the cluster members frequently share a common 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 Figure 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. Boundary layer schemes in WRF-ARW also have variable performance in cold and stable regimes in different regions [18].

A second major finding is that the length of the training period of clustering and calibration generally has little impact on the verification scores, whether the training period is one, two, or three months long. This can be seen from examining the CRPS for 2 m temperature in Figure 3 and 10 mu-wind in Figure 4, which show the CRPS for both the BMA-calibrated (solid lines) and uncalibrated (dashed lines) full ensembles (the 10 mv-wind plot is quite similar to the 10 mu-wind plot, and so is not shown), for the one-month (blue), two-month (red), and three-month (green) training experiments. Occasionally there are some significant differences between the different training periods for 2 m temperature (with the one-month training being the best of the experiments in the majority of those instances), but for the 10 m wind components there is no statistically significant difference between the training periods. The same general observations are true when comparing the CRPS for the HCA subset ensembles (not shown). It can also be noted from Figures 3 and 4 that the calibrated ensembles have statistically significantly lower (better) CRPS values than do the uncalibrated ensembles, which is a finding consistent with dozens of other ensemble modeling studies [12,20,31].

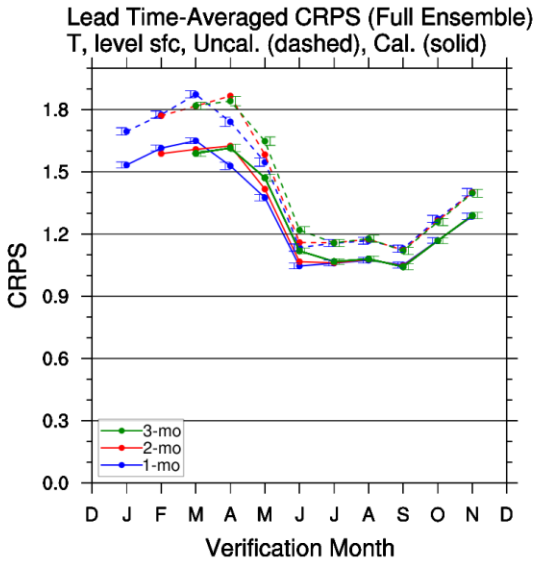

Figure 3: CRPS of the full 42 member ensemble for each of the experiments with one-month (blue), two-month (red), and three-month (green) training periods. The solid lines denote the CRPS for the BMA-calibrated ensembles, and the dashed lines for the uncalibrated ensembles. The confidence intervals that are shown denote 1 standard deviation in the bootstrapped estimated value of CRPS. Confidence intervals for the two-month data points are hidden for clarity, but are of similar magnitudes. CRPS in this figure is in units of K.

Likewise, the CRPSR, the ratio between the HCA subset ensemble and full ensemble CRPS values, is generally similar regardless of training period length. The CRPSR for 2 m temperature and 10 mu-wind are shown in Figures 5 and 6, respectively, for both the calibrated (solid line) and uncalibrated (dashed line) ensembles, for the one-month (blue), two-month (red), and three-month (green) training experiments. When the confidence intervals include 1.0, then the null hypothesis that the CRPS of the subset and full ensembles are statistically equivalent cannot be rejected. This condition is met for most of the experiments when the full and subset ensembles are both calibrated. When examining the uncalibrated ensembles for all experiments, however, the CRPS for the subset ensemble is more frequently significantly degraded from the full ensemble CRPS (i.e., the CRPSR is significantly larger than 1.0). This finding indicates that calibration should also be used, in addition to an intelligent down-selection method, to achieve minimal degradation in average forecast performance.

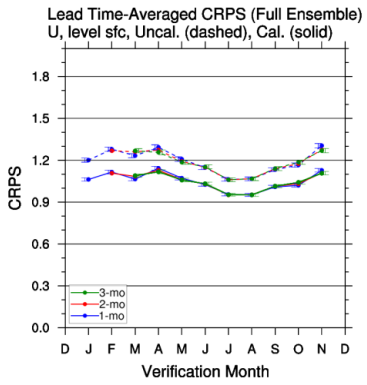

Figure 4: CRPS of the full 42 member ensembles for each of the experiments with one month (blue), two-month (red), and three-month (green) training periods but for the 10 mu-wind component. The solid lines denote the CRPS for the BMA-calibrated ensembles, and the dashed lines for the uncalibrated ensembles. The confidence intervals that are shown denote 1 standard deviation in the bootstrapped estimated value of CRPS. Confidence intervals for the two month data points are hidden for clarity, but are of similar magnitudes. CRPS in this figure is in units of ms-1.

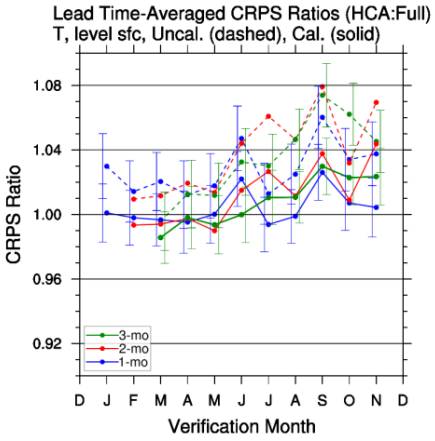

Figure 5: Ratios of CRPS for HCA subset ensembles to the CRPS of the full 42 member ensemble, averaged over all forecast lead times for 2 m temperature over all one-month (blue), two month (red), and three-month (green) training experiments. The solid lines denote the CRPSR for the BMA-calibrated ensembles, and the dashed lines for the uncalibrated ensembles. The confidence intervals that are shown denote 1 standard deviation in the bootstrapped estimated value of CRPSR. Confidence intervals for the two month data points are hidden for clarity, but are of similar magnitudes.

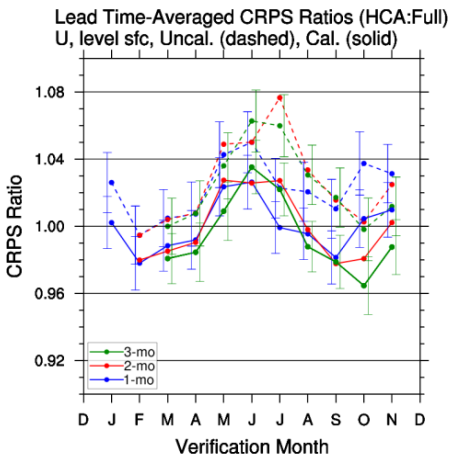

Figure 6: Ratios of CRPS for HCA subset ensembles to the CRPS of the full 42 member ensemble, averaged over all forecast lead times for 10 mu-wind component over all one-month (blue), two-month (red), and three-month (green) training experiments. The solid lines denote the CRPSR for the BMA-calibrated ensembles, and the dashed lines for the uncalibrated ensembles. The confidence intervals that are shown denote 1 standard deviation in the bootstrapped estimated value of CRPSR. Confidence intervals for the two-month data points are hidden for clarity, but are of similar magnitudes.

A likely explanation for why there is little difference in CRPS and CRPSR values across different training period lengths is that there tends to be considerable overlap in the members that comprise the HCA subset ensembles for experiments in which the training period overlapped (Table 7). When the same cluster is found in multiple experiments, there is one ensemble member that is frequently closest to the centroid of that cluster across several experiments. Therefore, at least for the ensemble in this study, a longer training period, which requires several hours more computation time for calibration with BMA in order to calibrate the ensemble, appears to confer little if any tangible benefit. As a result, a one-month training period, encompassing six consecutive forecasts every fifth day, is sufficient and practical. A training period of about a month is similar to the findings of Raftery et al. [12], though they use a daily NWP ensemble instead of an every-fifth-day ensemble like we use here. By examining ensemble forecasts every fifth day, temporal correlations are reduced. This allows calibration to be both possible and effective over a similar time span, while requiring less data.

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

Third, because the CRPS values for the calibrated subset and full ensembles are generally not significantly different from each other for all experiments Figures 5 and 6, it can be said that down-selection via HCA is effective year round, not just in one particular season. This finding increases the potential utility of the HCA down-selection method.

Finally, the two-sample K-S test indicates that the full and HCA subset ensembles are quite unlikely to come from different distributions. 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 null hypothesis that the two distributions are the same could not be rejected at the 95% confidence level. For most experiments and lead time variable combinations, the full and subset ensemble CDFs are likely to be similar to each other for 95% or more of the forecast locations in the respective verification periods. Table 8 only shows K-S test results for surface variables, but results are similar for above surface variables as well. The K-S test results further indicate that the HCA subsets are providing good approximations to the full ensemble.

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

Table 8: For each experiment, the percentage of observation locations of each lead time surface variable combination for which the two-sided Kolmogorov Smirnov test indicates that the null hypothesis of statistical similarity between the full and subset ensemble distributions is supported. High percentages indicate that it is unlikely that the two distributions differ.

HCA vs. random down-selection

To assess the value of down selection using HCA, it is important to demonstrate that HCA adds value over a random down selection process. We find that down-selection using HCA usually does result in better verification scores than if down selection is done randomly. For the DJF experiment, 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.). We then 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 compare with random subsets. 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.

We calculate the CRPSR for both the HCA determined subsets and the mean of the ten randomly determined subsets (comparing both to the full ensemble). For both 2 m temperature (Figure 7) and 10 mu wind (Figure 8) the HCA subsets are generally statistically significantly better than random down-selection for most ensemble sizes. Therefore, down selecting using HCA adds value compared to randomly choosing a subset ensemble.

At least two other main observations can be drawn from the data in Figures 7 and 8. First, there appears to be little additional forecast skill gained by increasing ensemble size beyond roughly 7-10 members. The CRPSR in these figures all generally decrease with increasing ensemble size until about 7-10 members, at which point the CRPSR remains roughly flat for larger ensemble sizes. Because only a few ensemble members can deliver nearly equivalent forecast skill as a much larger ensemble, it is likely that such a large multi-physics ensemble contains much redundancy.

Second, down selection is more effective when both the full and subset ensembles are calibrated than when both are uncalibrated for nearly all subset sizes, not just the objectively chosen one. In Figures 7 and 8, the CRPSR for calibrated ensembles (solid lines) are significantly smaller (i.e., better) than the CRPSR for uncalibrated ensembles (dashed lines), though these improvements tend not to be statistically significant for ensembles larger than about 8 members. 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, the CRPSR for the HCA subset ensembles is sometimes less than 1.0, indicating an even lower (better) CRPS than for the full ensemble.

Summary and Conclusion

This study demonstrates the performance of hierarchical cluster analysis (HCA) as an ensemble down selection methodology on a 42 member WRF multi-physics ensemble dataset, with forecasts initialized every fifth day for an entire year. The full ensemble and subset ensembles are then calibrated with Bayesian model averaging (BMA) to calibrate the ensemble PDF. The HCA and BMA are trained over varying lengths of consecutive forecasts (six, twelve, and eighteen, covering one, two and three months, respectively). We then verify the wind component and temperature forecasts over six consecutive forecasts, also spaced five days apart, using CRPS as our primary metric, in addition to the CRPS ratio between the subset and full ensembles.

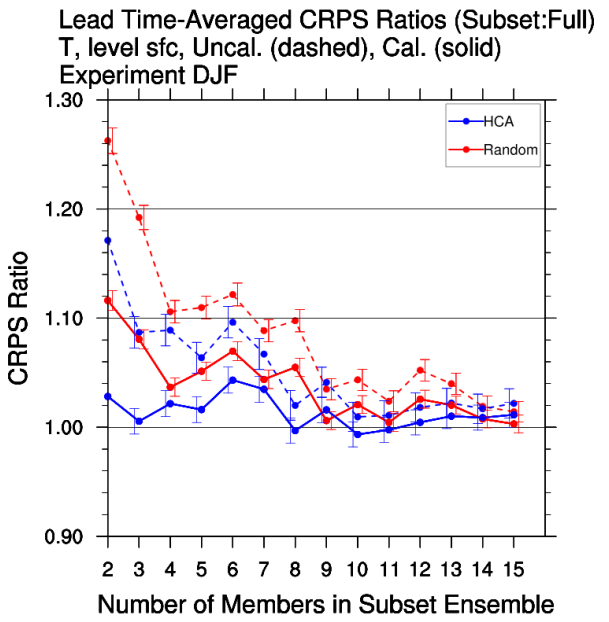

Figure 7: Ratios of CRPS for subset ensembles of a range of sizes to the CRPS of the full 42 member ensemble, averaged over all forecast lead times for 2 m temperature in the DJF experiment. The blue line corresponds to the average CRPS ratio for ten random subset ensembles of each ensemble size, while 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. Confidence intervals denote 1 standard deviation.

The primary conclusions that can be drawn from this study are as follows:

Down selection with HCA, particularly when paired with BMA calibration, is effective year round at representing the forecast distribution of a WRF based 42 member multi physics ensemble with just 7-10 members.

Down-selecting to a subset ensemble using HCA yields statistically significantly better skill than randomly choosing subsets.

The length of training period has little impact on verification results. Therefore, it is practical to use a shorter (one month of forecasts every fifth day) training period for computational efficiency.

The ensemble members cluster differently in different seasons, but the cluster members 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.

While physics uncertainty is the only component of the total forecast uncertainty that we focused on here, we believe that this study can help guide and constrain future efforts in ensemble NWP modeling, particularly with the attention that efforts to better represent physics and model uncertainty is receiving in the community currently. It is clear that for multi-physics ensembles like the one in this study, increasing ensemble size beyond about 10 members would simply be gratuitous computing if we plan to calibrate the ensemble. Therefore, resources would likely be more wisely spent on increasing model resolution, the size of the model domain, or more importantly, representing additional types of forecast error, than by including more than about 10 intelligently chosen physics suites in an ensemble prediction system.

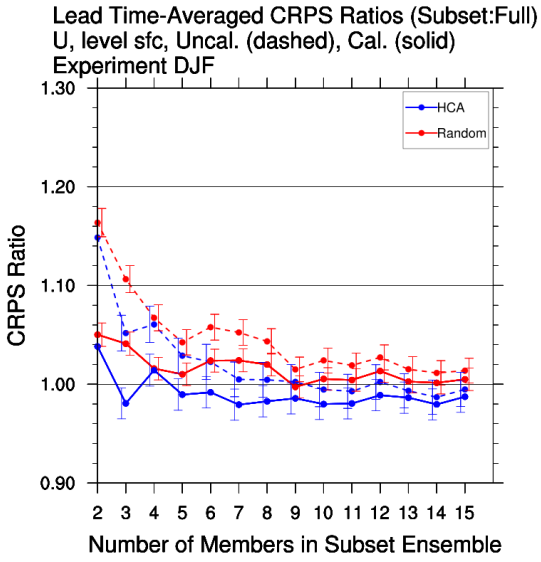

Figure 8: Ratios of CRPS for subset ensembles of a range of sizes to the CRPS of the full 42 member ensemble, averaged over all forecast lead times for 10 mu wind component in the DJF experiment. The blue line corresponds to the average CRPS ratio for ten random subset ensembles of each ensemble size, 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. Confidence intervals denote 1 standard deviation.

Acknowledgement

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 thank Greg Thompson and Laurie Carson of NCAR for providing additional computing resources on NCAR's now-retired Bluefire supercomputer. We thank Walter Kolczynski of NCEP's Environmental Modeling Center for providing the early foundations of the BMA and verification codes for this study, which are mostly, coded using the NCAR Command Language. We thank Aijun Deng (Penn State) for providing the National Ice Center snow analyses for the WRF initial conditions, Michael Richman and Aaron Johnson (Oklahoma) for initially suggesting we investigate HCA as a down-selection technique, and David Stauffer (Penn State), Joshua Hacker, Tom Hopson, and Luca Delle Monache (NCAR) for other helpful discussions during the course of this project. Tressa Fowler and John Halley Gotway (NCAR) provided valuable input on statistical

significance issues, and Luca Delle Monache made helpful comments on the manuscript. JA. Lee is also grateful to NASA (grant NNX11AQ44G), and the National Research Council Research Associateship Program for funding that partially supported this project.

References

1. Eckel FA, Mass CF (2005) Aspects of effective mesoscale, short-range forecasting. Wea Forecasting 20: 328-350.
2. Fujita T, Stensrud DJ, Dowell DC (2007) Surface data assimilation using an ensemble Kalman filter approach with initial condition and model physics uncertainties. Mon Wea Rev 135: 1846-1868.
3. Clark AJ, Gallus WA, Chen TC (2008) Contributions of mixed physics versus perturbed initial / lateral boundary conditions to ensemble-based precipitation forecast skill. Mon Wea Rev 136: 2140-2156.
4. Berner J, Ha SY, Hacker JP, Fournier A, Snyder C (2011) Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. Mon Wea Rev 139: 1972-1995.
5. Berner J, Fossell KR, Ha SY, Hacker JP, Snyder C (2015) Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. Mon Wea Rev 143: 1295-1320.
6. Hacker JP, Ha SY, Snyder C, Berner J, Eckel FA, et al. (2011) The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. Tellus 63: 625-641.
7. Lee JA, Kolczynski WC, McCandless TC, Haupt SE (2012) An objective methodology for configuring and down-selecting an NWP ensemble for low-level wind prediction. Mon Wea Rev 140: 2270-2286.
8. Stensrud DJ, Bao JW, Warner TT (2000) Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. Mon Wea Rev 128: 2077-2107.
9. Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, et al. (2008) A description of the Advanced Research WRF Version 3. NCAR Technical Note NCAR / TN-475 + STR, pp: 113.
10. Lee JA (2012) Techniques for down-selecting numerical weather prediction ensembles. Ph.D. dissertation, The Pennsylvania State University, pp: 131.
11. Raftery AE, Balabdaoui F, Gneiting T, Polakowski M (2003): Using Bayesian model averaging to calibrate forecast ensembles. Technical Report no. 440, Dept of Statistics, University of Washington 1-28.
12. Raftery AE, Balabdaoui F, Gneiting T, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Mon Wea Rev 133: 1155-1174.
13. Wolff JK, Nance L, Bernadet L, Brown B (2009) WRF Reference Configurations. 23rd Conf on Weather Analysis and Forecasting / 19th Conf on Numerical Weather Prediction at the 89th Amer. Meteor Soc Annual Meeting, Omaha.
14. Fraley C, Raftery AE, Gneiting T (2010) Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. Mon Wea Rev 138: 190-202.
15. Environmental Modeling Center (2003) The GFS atmospheric model. NCEP Office Note 442, pp: 14.
16. NCAR (2011) ARW Version 3 Modeling System User's Guide, pp: 372.
17. Cressman GP (1959) An operational objective analysis system. Mon Wea Rev 87: 367-374.
18. Gilliam RC, Pleim JE (2010) Performance assessment of new land surface and planetary boundary layer physics in the WRF-ARW. J Appl Meteor Climatol 49: 760-774.
19. Rogers RE, Deng A, Stauffer DR, Gaudet BJ, Jia Y, et al. (2013) Application of the Weather Research and Forecasting model for air quality modeling in the San Francisco Bay area. J Appl Meteor Climatol 52: 1953-1973.
20. Sloughter JM, Gneiting T, Raftery AE (2013) Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. Mon Wea Rev 141: 2107-2119.
21. Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Wea Forecasting 15: 559-570.
22. Wilks DS (2006) Statistical methods in the atmospheric sciences (2nd eds.) Academic Press, pp: 626.
23. Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Amer Stat Assoc 102: 359-378.
24. Efron B (1979) Bootstrap methods: Another look at the jackknife. Ann Stat 7: 1-26.
25. Alhamed A, Lakshivarahan S, Stensrud DJ (2002) Cluster analysis of multimodel ensemble data from SAMEX. Mon Wea Rev 130: 226-256.
26. Yussouf N, Stensrud DJ, Lakshmivarahan S (2004) Cluster analysis of multimodel ensemble data over New England. Mon Wea Rev 132: 2452-2462.
27. Johnson A, Wang X, Xue M, Kong F (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.
28. Solazzo E, Riccio A, Kioutsioukis I, Galmarini S (2013) Pauci ex tanto numero: Reducing redundancy in multi-model ensembles. Atmos Chem Phys Discuss 13: 4989-5038.
29. Wyngaard JC (2010) Turbulence in the atmosphere. Cambridge University Press, pp: 393.
30. Warner TT (2011) Numerical weather and climate prediction. Cambridge University Press, pp: 526.
31. Sloughter JM, Gneiting T, Raftery AE (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. J Amer Stat Assoc 105: 25-35.