11A.3 THE REGIME DEPENDENCE OF OPTIMALLY WEIGHTED ENSEMBLE MODEL CONSENSUS FORECASTS

Steven J. Greybush
Sue Ellen Haupt*
George S. Young
The Pennsylvania State University

1. INTRODUCTION

Ensemble modeling is proving to be a viable approach to addressing uncertainty in numerical weather prediction. Ensemble modeling creates a distribution of potential forecast solutions by varying the initial conditions, boundary conditions, model physics, parameterization schemes, and the models themselves. Ensemble models are run operationally at several meteorological centers, including the National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-Range Weather Forecasts (ECWMF).

Recent research has elucidated methods for optimally combining the forecasts from the set of ensemble member solutions to create a single deterministic consensus forecast. These methods include using the simple ensemble mean, linear regression / Ensemble Model Output Statistics (EMOS) (Gneiting et al., 2005), performance-based metrics (Woodcock and Engel, 2005), and Bayesian Model Averaging (BMA) (Raftery et al., 2005). In its simplest expression, the consensus forecast can be stated as a linear combination of individual ensemble member forecasts plus a bias correction term. The weight of a member in the combination typically is determined by considering the performance of the member relative to the other ensemble members over some window of time. Most methods employ a sliding window of dates that directly precede the forecast date whose length *w* is between one week and two months; Eckel and Mass (2005) prefer 14 days but Gneiting et al. (2005) prefer 40 days. Thus, the underlying assumption is that if member A performs relatively better than member B on days *n-w* to *n*, then member A can be expected to continue to perform better than member B on day *n+1*.

Forecasters have long understood that model performance is linked to the characteristics of the weather situation being forecast (Fritsch et al., 2000). Forecasters frequently weight sources of information depending upon these weather characteristics (Roebber 1998); for example, model A might be superior to model B when predicting heavy precipitation events. Although model output statistics (MOS) address the correlations between the predicted variable of interest and the other atmospheric fields, it does not relate the fields to the relative predictability of the variable amongst different models or ensemble members. Clustering, the process through which similar elements of a dataset are grouped together, can be used to separate a time series of model forecasts into weather situations or regimes. Here, regimes are defined broadly as any configuration of atmospheric variable fields such that membership in a specific regime can be determined objectively.

The logical amalgamation of the ideas of consensus ensemble forecasting and weather regimes is explored in this paper. The relative performance of ensemble members should be linked to the weather regime present; different relative performances imply that varying the weights of ensemble members depending upon regimes should produce a more accurate consensus forecast. A hypothesis can thus be stated: clustering forecast dates by weather regime prior to assigning weights to ensemble members results in reduced forecast error for a consensus deterministic forecast. The regime-independent methods serve as benchmarks with which to judge the success of the new post-processing technique.

2. ENSEMBLE MODEL AND VERIFICATION DATA

The University of Washington Mesoscale Ensemble (UWME+) (Eckel and Mass, 2005) is used as the source of ensemble forecasts for this project. The system is based upon the 5th-generation MM5 mesoscale model and contains eight members. Each member has different initial conditions as well as different physical / parameterization schemes. This project uses data from a nested 12-km-spaced grid roughly centered over the states of Washington and

* *Corresponding author address:* Sue Ellen Haupt, Applied Research Lab, P.O. Box 30, Penn State University, State College, PA 16804; email: haupts2@asme.org

Oregon. Table 1 summarizes the data downloaded from the UWME archives.

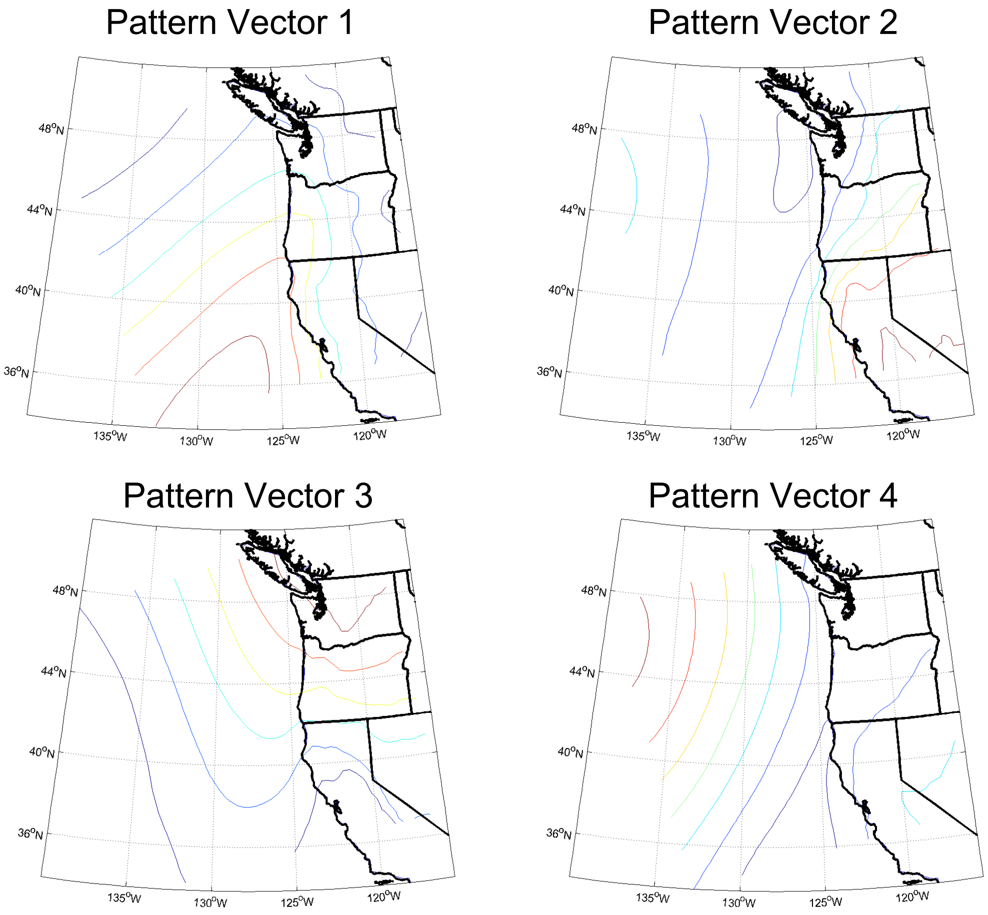
Table 1: UWME+ dataset summary.

Number of Members	8 members
Temporal Domain	12 months (2005-09-01 – 2006-08-30)
Initialization Time(s)	00 UTC (4 pm PST)
Forecast Time(s)	+48 hours
Domain Size	12-km-spaced nested grid (Pacific Northwest)
Forecast Variable(s)	2 m temperature (T_{sfc})

To provide a relatively simple methodological test bed, point data from Portland, OR (latitude 45°35' N, longitude 122°36' W) are extracted from the UWME+ ensemble forecast dataset. Specifically, 2-m temperature data are linearly interpolated between the two closest spatial grid points to Portland, OR to create two time series representing model forecasts from each ensemble member extending for the period from September 1, 2005 to August 31, 2006.

Next, ensemble model forecasts are compared with the actual state of the atmosphere – a verification dataset. Surface temperature verifications were thus taken from the Portland airport Automated Surface Observing System (ASOS) reports at 00Z each day, and collected into a time series extending from September 1, 2005 to August 31, 2006.

Upon comparing the ensemble forecasts with observations, the mean absolute error (MAE) was 2.5 deg C, with a bias of -0.50 deg C. Applying bias correction to each ensemble member resulted in a new MAE of 2.2 deg C, an improvement of more than 10%.



Pattern Vector 1

Pattern Vector 2

Pattern Vector 3

Pattern Vector 4

Figure 1: First Four Obliquely Rotated Eigenvectors from PCA of MSLP, 2005-09-01 – 2006-08-31

3. ATMOSPHERIC WEATHER REGIMES

The North American Regional Reanalysis (NARR) (Messinger et al., 2006) provides a rich dataset from which to characterize atmospheric regimes. Data for 00 UTC each day are obtained for the same 12-month time period (2005-09-01 – 2006-08-30) as the UWME+ ensembles. For a synoptic domain on the order of 1000 x 1000 km consisting of the Pacific states and the adjacent ocean, the fields include 500-mb heights (h500) to represent mid-level atmospheric flow, mean sea level pressure (MSLP), and specific humidity (q700) as a proxy for mid-level cloud cover. For a mesoscale domain on the order of 300 x 300 km, the fields include the u-component and v-component of the wind at 850 mb and 925 mb (uv850 and uv925).

Principal Component Analysis (PCA) is a technique for reducing the dimensionality of a dataset while capturing the most important modes of variability (Wilks 2006). PCA is applied to each of the five NARR data fields individually. This analysis produces pattern vectors (eigenvectors) and a time series of amplitudes called principal components (PCs) for each data field. In order to interpret the pattern vectors physically and to avoid obtaining Buell patterns, the eigenvectors need to be rotated (Richman 1986). Buell patterns are characteristic of the domain shape, and can appear regardless of the field under consideration – even PCs of random data can contain such patterns. Here, rotation is accomplished using the promax (oblique) technique, retaining only the first four eigenvectors. On average, the first four pattern vectors capture 88 % of the variance of the data. These pattern vectors for MSLP are plotted in Figure 1.

The first pattern vector can be interpreted as the Aleutian Low with a high pressure area west of California; the second can be interpreted as a trough of low pressure offshore of and roughly parallel to the Pacific coast. A linear combination of these four patterns describes the actual atmospheric state on a given day. These PCs together can characterize atmospheric regimes, and provide inputs to a regime-dependent consensus forecast.

4. ENSEMBLE MODEL CONSENSUS FORECASTS

A consensus forecast is some combination of individual member forecasts, with the goal that over the long term, the consensus forecasts are more accurate

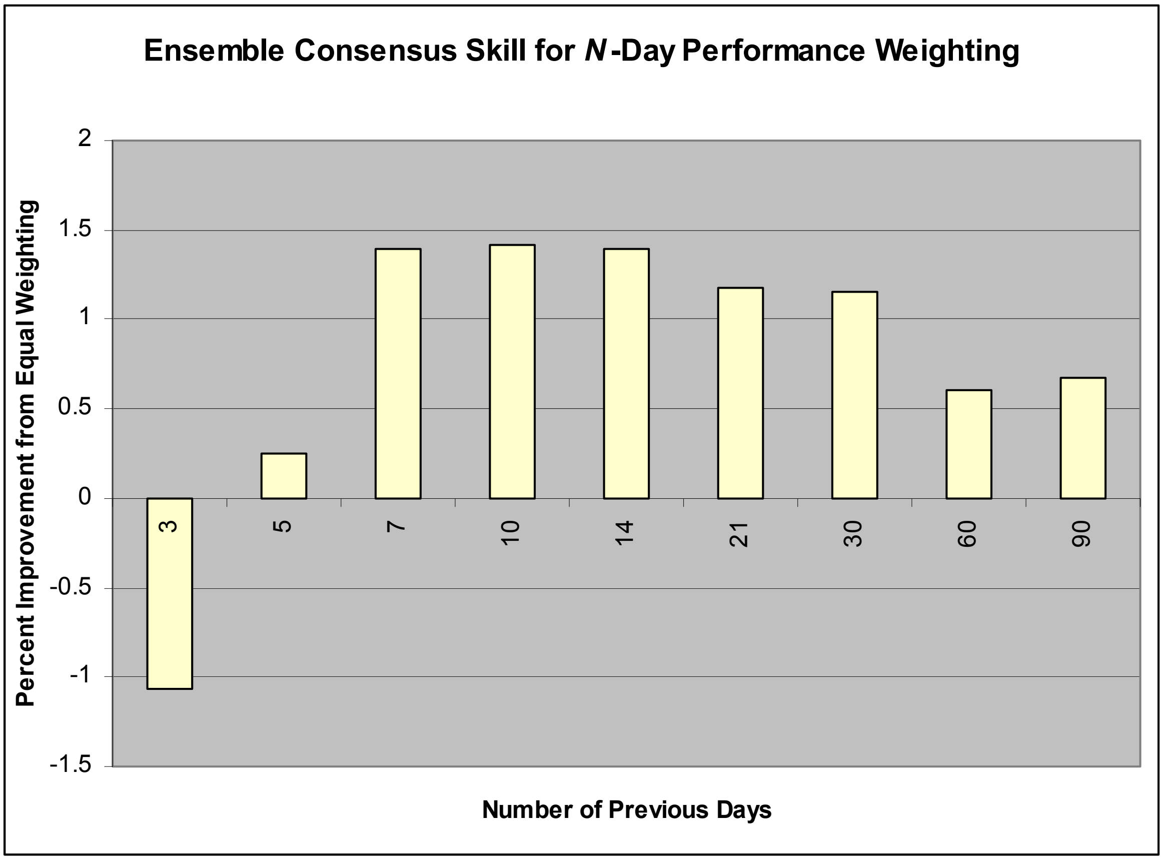


Figure 2: Ensemble Consensus Skill for *N*-Day Performance Weighting using various values for *w,* the performance window size (window is composed of previous days).

than any of the individual member forecasts. If the forecast combination is linear, it can be expressed in the form of:

$$X^* = \sum_{i=1}^{n} a_i X_i \qquad (1)$$

where X^* is the consensus forecast, n is the number of ensemble members, X_i are the individual bias-corrected ensemble member forecasts, and a_i are ensemble member weights. The goal of this work is to select ensemble member weights a_i so that X^* is the optimal consensus forecast. Here, the optimal consensus forecast has the least MAE – the absolute value of a forecast minus an observation.

The simplest consensus forecast is the ensemble mean, where a_i are equal for all i such that their sum is one ($a_i = 1 / n$, where n is the number of ensemble members); this technique is also referred to as the equal-weighted average. An improved choice of the ensemble member weights, a_i, can be obtained by considering the relative performance of ensemble members. Performance-weighted averages (PWA) use the inverse of the MAE of that ensemble member over a window of forecasts W to derive raw weights (Woodcock and Engel, 2005). Thus ensemble members that are performing the best (have the smallest MAE) are given the largest weighting in the linear combination. The weights are then normalized so that their sum is equal to one. The formula for calculating an ensemble member weight is given as follows:

$$a_i = p_i^{-1} \Big/ \sum_{j=1}^{n} p_j^{-1} \qquad (2)$$

Where p_i is a performance measure (here, MAE) for ensemble member i, and n is the number of ensemble members.

Consensus forecast error can be reduced based upon a clever selection of the performance window, W. Here, W refers to a set of dates under which performance is evaluated; w refers to the number of forecast dates in W, typically in the context of the previous w days. The simplest choice of W is the entire set of forecasts from the dependent dataset, henceforth referred to as W_{dep}. Operationally, other studies employ a sliding window of dates that directly precede the date that the forecast is issued. Eckel and Mass (2005) prefer 14 days (W_{14}) but Gneiting et al. (2005) prefer 40 days for the duration, w, of the window. Figure 2 reveals the performance of N-day performance weighting for various performance window sizes (w). Here, the optimum window size was 10 days, with 14 days having slightly less than minimal MAE.

Consensus forecasting using performance-weighted averaging with both W_{dep} and W_{14} offers only a slight improvement (1.4%) over equal-weighted forecast MAE. This 1.4% improvement serves as a benchmark for the performance of the regime-dependent methods shown below; they should be considered successful only if they can produce a greater percent improvement. The goal of the clustering methods discussed in the next section is to identify patterns in the distribution of optimal weights so that a performance-weighted average applied to independent data will reveal a more substantial improvement. The purpose of the clustering is thus an intelligent selection of W in order to minimize MAE.

5. REGIME-BASED CONSENSUS FORECASTS

The process known as clustering divides a dataset into a number of subsets or groups called clusters. The goal of traditional clustering methods, such as K-means, is to group similar elements together while separating dissimilar ones (Wilks 2006). Similarity is determined based upon some distance formula between data points. Here, each cluster is a group of forecast dates with its own W_{clstr}, the window of forecast dates used for performance-weighted averaging. The goal is to find clusters such that the sum of the MAEs of the clusters (weighted by the number of members in the cluster) is minimized. Examination of the scatter plots of the principal components did not reveal any readily discernable patterns; simple structure (Richman 1986) was not attained during the PC rotations. Therefore it is likely that traditional clustering methods would not be appropriate for characterizing patterns in this dataset. Here the goal is not merely to find patterns in the data, but to minimize MAE. Thus a new method is presented that uses a genetic algorithm to attain that goal.

Consider a simple cluster rule *R1*: given a single PC (here, a time series of data points x_1), place all $x_1 > r$ into the first cluster, all $x_1 \leq r$ into a second. Thus r is a boundary that subdivides the phase space of the PC. A second cluster rule *R2* can be applied to another PC. Together, these rules can divide the set of forecast dates into up to four clusters. If both rules *R1* and *R2* apply to the same PC x_1, only three clusters are produced. In general, b unique boundaries produce up to 2^b clusters.

The boundaries (cluster rules) that minimize MAE can be determined using a genetic algorithm (GA). A genetic algorithm is a technique for optimization that intelligently samples a large portion of the solution space before eventually converging upon the optimum location (Haupt and Haupt, 2004). A GA begins with a random pool of possible solutions called chromosomes. As in biological evolution, the solution pool evolves. Less fit members are removed, and offspring of the more fit members take their place. Here, the fitness of a chromosome (each chromosome specifies a cluster rule) is determined by the total MAE of all the clusters using PWA, with each individual cluster MAE rescaled by the size of the cluster. Crossover blends information from different solutions, while mutations of the solutions prevent the gene pool from being trapped in local minima. This GA simultaneously optimizes several parameters, including the selection of which PC to apply a cluster boundary (a discrete variable) and the exact value *r* of the boundary (a continuous variable). The GA can select from 20 predictors (5 atmospheric fields times 4 PCs each). The number of boundaries is determined before the GA is run, and is repeated for a varying number of boundaries (1-5) which correspond to between 2 and 32 clusters. The 100 benchmark runs used 500 GA iterations with 20 chromosomes, a mutation rate of 0.25, and a survival rate of 0.5. Once optimal regimes are defined by the GA, performance-weighted averaging using optimal weights for each regime (cluster) is applied to generate consensus forecasts.

An alternative technique for regime-based consensus forecasts uses multivariate linear regression for considering model performance under various weather regimes. Whereas clustering divides a dataset into discrete groups, regression maintains the continuous nature of the principal components. With the 20 atmospheric PCs as inputs, forward screening multivariate linear regression predicts the errors (MAEs) of each ensemble member. These predicted errors are used to calculate the performance weights using the PWA equation. Here, the performance window consists of the set of predicted errors according to the regression equation. Although this regression approach uses atmospheric regime information (the PCs), it does not explicitly classify a given forecast date as belonging to a specific regime as in the clustering approach.

6. DISCUSSION

Figure 3 compares the performance of several methods for generating consensus forecasts. Using the principles of cross-validation, the methods were trained on the dependent dataset (243 days), and applied to an independent dataset (122 days). To mitigate the random fluctuations of arbitrarily dividing the data into independent and dependent sets, these methods were simulated 50 times with different dataset divisions.

For regime-dependent methods, these simulations used PCs of NARR atmospheric fields for the same time and date that the forecasts are valid. Operationally, this would assume that the centroid of

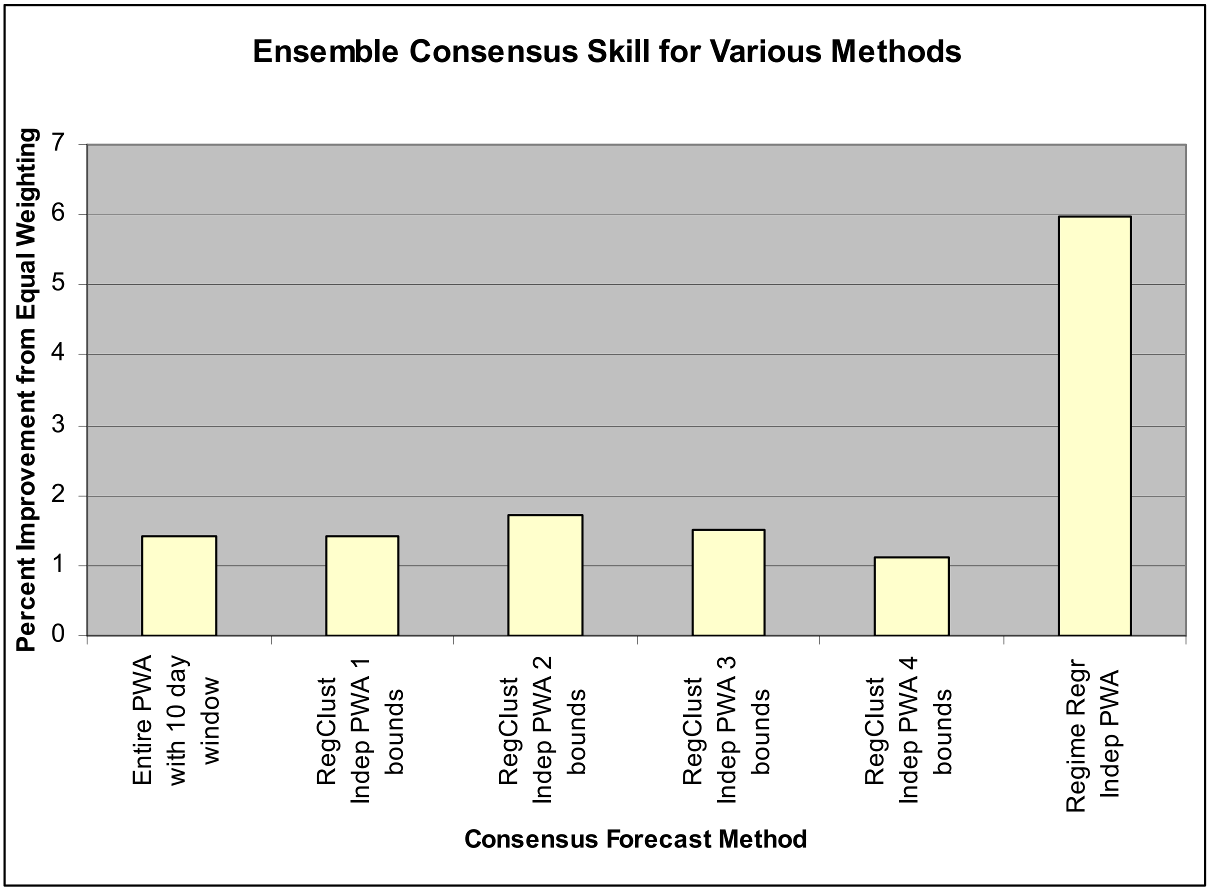


Figure 3: Ensemble Consensus Skill for various performance-weighting schemes on an independent dataset. They include a 10-day sliding window (the optimal sliding window size); regime-based clustering using a genetic algorithm with 1,2,3, and 4 boundaries (2,4,8,16 clusters); and regime regression.

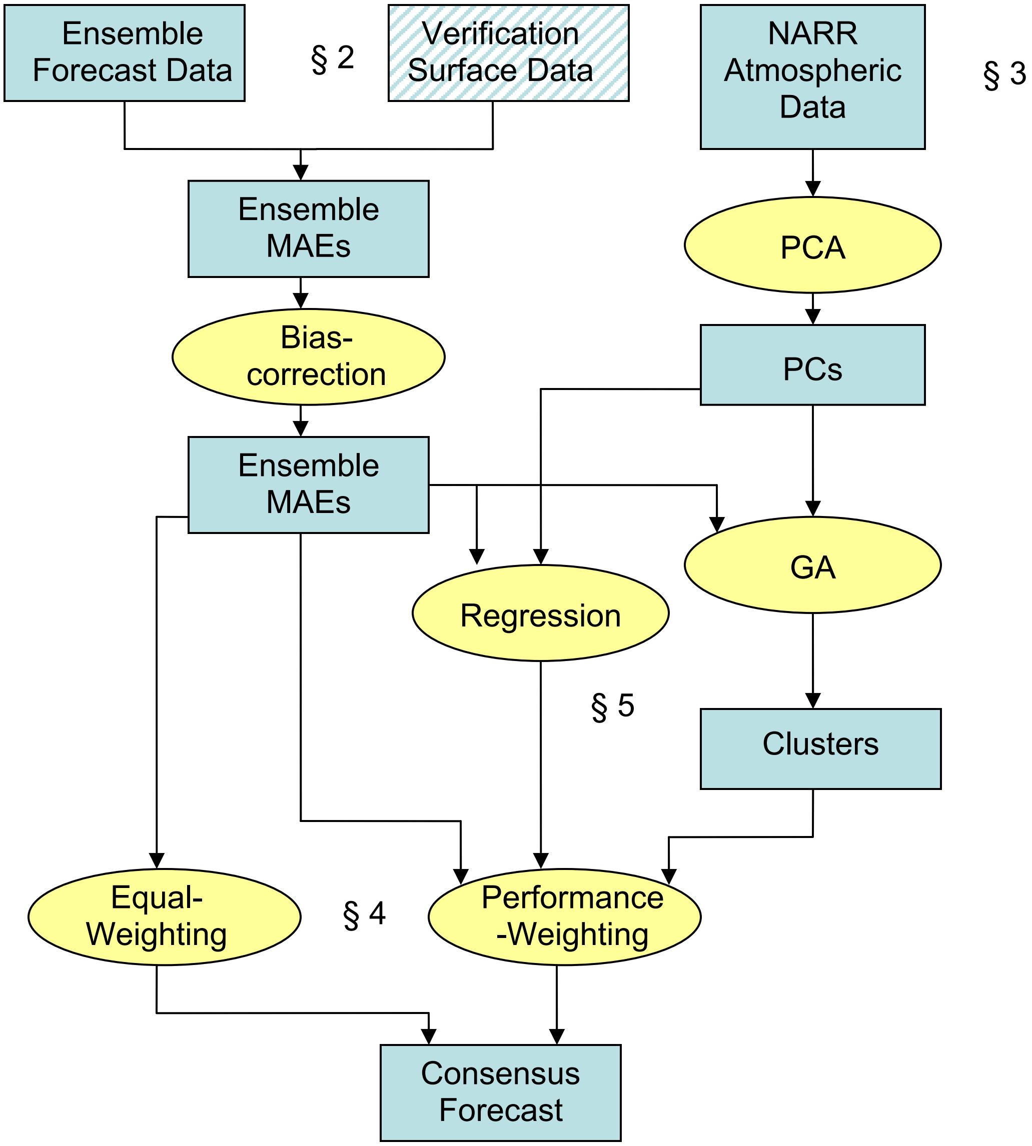


Figure 4: Flow chart of this paper's methodology. Rectangles denote datasets, ovals denote procedures, and arrows denote dependencies. The section numbers for each portion of the technique are noted as well.

the ensemble member 48-hour forecasts agrees perfectly with the actual state of the atmosphere at that time – the "perfect prog" assumption.

Predicting ensemble member forecast errors based upon linear regression of atmospheric regimes is the best method for reducing the MAE of ensemble model consensus forecasts, presenting a 6 % improvement on independent data. Both regime clustering with a genetic algorithm and regime regression of principal components have less MAE than the best of the regime-independent methods. This result demonstrates the importance of considering atmospheric regime information when generating optimal ensemble model consensus forecasts.

7. CONCLUSIONS

This paper analyzes the relative merits of several post-processing techniques for the creation of a deterministic consensus forecast from ensemble model output. Figure 4 presents a flow chart of the methods used in this study. The test-bed consists of a 12-month time series of UWME+ ensemble forecasts for surface temperature at Portland, OR. One technique uses PCs of other atmospheric fields to group the forecast dates into clusters; a genetic algorithm determines the optimum cluster rules with the goal of reducing MAE of the resulting consensus forecasts. This technique reduces MAE below that of 14-day moving window performance weighted averaging. Predicting ensemble member forecast errors using multivariate linear regression of atmospheric regime principal components had the greatest success in reducing PWA MAE.

Regime-based consensus forecasts show promise for use in operational forecasting. Additional studies, however, need to confirm their superior performance on longer datasets and with other locations. An independent dataset from a different year would ensure that no autocorrelations in weather patterns are artificially inflating scores. Although currently used for deterministic forecasts, the clustering paradigm can be applied eventually to probabilistic forecasting such as Bayesian Model Averaging (Raftery et al., 2005). Tuning of the genetic algorithm parameters, as well as generalizing the cluster rules using Fisher's Linear Discriminant (Wilks 2006), may result in improved performance of the GA clustering technique. Finally, the technique should be expanded to a grid-based forecasting domain, and include several forecast variables simultaneously.

Acknowledgments. The authors wish to thank Dr. John H. E. Clark and Dr. Hampton N. Shirer for providing invaluable feedback on this paper.

Thanks also are due to Ben Root, Rich Grumm, and Paul Knight for facilitating the acquisition of NARR data, and to the University of Washington for enabling public access to their ensemble data.

This research was funded in part by the PSU Applied Research Lab Honors Program.

REFERENCES

Eckel, F.A. and C. Mass, 2005: Aspects of Effective Mesoscale, Short-Range Ensemble Forecasting, Wea. Forecasting, 20, 328-350.

Fritsch, J.M., J. Hilliker, and J. Ross, 2000: Model Consensus, Wea. Forecasting, 15, 571-582.

Gneiting, T., A.E. Raftery, A. Westveld, and T. Goldman, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Mon Wea. Rev., 133, 1098-1118.

Haupt, R.L. and S. E. Haupt, 2004: Practical Genetic Algorithms, John Wiley and Sons, 255 pp.

Messigner, F., G. DiMego, E. Kalnay, K. Mitchell, P.C. Shafran, W. Ebisuzaki, D. Jovic, J. Woollen, E. Rogers, E.H. Berbery, M.B. Ek, Y. Fan, R. Grumbine, W. Higgins, H. Li, Y. Lin, G. Manikin, D. Parrish, and W. Shi., 2006: North American Regional Reanalysis, Bull. Amer. Meteor. Soc., 87, 343-360.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Mon. Wea. Rev., 133, 1155-1174.

Richman, M., 1986: Rotation of Principal Components, J. Climatology, 6, 293-335.

Roebber, P., 1998: The Regime Dependence of Degree Day Forecast Technique, Skill, and Value, Wea. Forecasting, 13, 783-794.

Wilks, D.S., 2006: Statistical Methods in the Atmospheric Sciences, Academic Press, 626 pp.

Woodcock, F. and C. Engel, 2005: Operational consensus forecasts. Wea. Forecasting, 20, 101-111.