Science and Technology Infusion Climate Bulletin
NOAA's National Weather Service
41st NOAA Annual Climate Diagnostics and Prediction Workshop
Orono, ME, 3-6 October 2016

Assessment of Ensemble Regression to Combine and Weight Seasonal Forecasts from Multiple Models of the NMME

Dan C. Collins
Climate Prediction Center, NOAA/NWS/NCEP, College Park, Maryland

1. Introduction

Ensemble and multi-model ensemble prediction systems have become state-of-art tools for climate forecasts on subseasonal and seasonal timescales. Ensemble prediction systems are meant to identify the forced climate signal, through generation of multiple realizations of the forecast, the differences of which can be attributable to uncertainty in the outcome due to the chaotic nature of the climate system. From ensemble predictions, it is possible to derive the probability of uncertain future events. The use of several models in a multi-model ensemble (MME) is known to improve on the forecasts of a single model ensemble, through the chance cancellation of individual model errors (Becker *et al.*, 2014; Kirtman *et al.*, 2014). The North American Multi-Model Ensemble (NMME) has been used as guidance for real-time seasonal forecasts by the Climate Prediction Center since August of 2011. Hindcasts over multiple years can be used to identify and correct the systematic biases of each model system. However, in addition to model biases, it is possible to discern the skill of each model as a function of region and season. Identifying models and regions of lower and higher skill, and utilizing this information to combine and weight the forecasts, should improve the skill of probabilistic forecasts. We seek to intelligently combine models to extract forced signals from the NMME and eliminate poor forecasts when possible.

2. Methodology and data

In this study, NMME seasonal forecasts for North America are calibrated and consolidated using a regression methodology to improve the reliability of probabilities and weight individual models according to their skill. Calibration entails relating the probabilities determined from an ensemble of forecasts to the skill of the forecast system, such that probabilities assigned are a reliable representation of the expected frequency of an event's occurrence. Reliable probabilities are considered skillful when they also resolve differences between cases with high and low probabilities. The spread of ensemble members should inform the probabilities of events, though models can be unreliable in representing the actual frequency of events.

Regression is widely used for correction of dynamical model forecasts and has been successfully applied at the Climate Prediction Center to dynamical model forecasts of temperature and precipitation for subseasonal lead times from 2 to 4 weeks. In this study, we apply the ensemble regression (EREG) method to NMME seasonal forecasts (Unger *et al.*, 2009). EREG retains the ensemble spread to represent conditional uncertainty of forecasts, to the extent that spread is found to be a reliable indicator of the average mean square error of a model's forecasts. EREG uses the expected value of the mean square error of hindcasts to adjust the model probability density function (PDF) and improve the reliability of probabilistic forecasts – collapsing the spread when either skill is low, or spread is a poor indicator of skill. EREG maintains the resolution of categorical forecasts, when the model spread is a good indicator of skill, while minimizing the mean square error of the ensemble mean.

Initially, each NMME model is calibrated individually using the EREG methodology. To combine the individual EREG-corrected model forecasts, we test two possible methods: 1. Combinations with no further adjustment of the MME PDF such that regressed model anomalies are weighted by their correlation to observations; and 2. Combinations with additional weighting of each individual model probability by its correlation to observations. The ensemble regression technique is compared to forecasts made by estimating the probability from the count of ensemble members (CE) in each category after bias corrections of model means and variances. Forecasts are verified using the Brier skill score (BSS), as well as assessed for reliability.

Correspondence to: Dan C. Collins, Climate Prediction Center, NOAA/NWS/NCEP, 5830 University Research Court, College Park, MD; E-mail: dan.collins@noaa.gov

For this study, we test calibration and weighting using the November 1st initializations from the model hindcasts for predictions of DJF Temperature, *i.e.*, lead-1 winter forecasts. Hindcasts are available from DJF 1982-83 to DJF 2010-11. GHCN and CAMS 2-m temperature observations are used for verification. We consider the skill and calibration of probability forecasts for terciles (above-normal, below-normal and near-normal, defined as the lower-third, mid-third and upper-third, respectively, of the climatological distribution for the 1982-2011 period). Extreme, above-normal and below-normal, seasonal forecasts at the 10th and 90th percentile respectively, consistent with the tercile forecasts, were also derived but are not shown in this study. The regression-calibration EREG methodology follows Unger *et al.*, 2009, "Ensemble regression". All statistics – the climatologies of model and observational seasonal mean and variances, as well as correlations, reliability and BSS – were cross-validated using a leave-1-year-out methodology, such that dependent data used in training the regression were independent of the verification.

3. Results

The baseline methodology that combines individual model counts of ensemble (CE) members to generate probability forecasts, after bias and variance corrections, indicates that models with regions of poor skill deteriorate the skill of the NMME when models are combined. Figure 1 shows the Brier skill scores for the CFS on the left and the combined NMME on the right, using the CE method to indicate probabilities. Note that in parts of the eastern US, skill in the CFS is higher than the combined NMME.

By calibrating probabilities from the individual models of the NMME using ensemble regression (EREG), models with areas of negative or zero skill are effectively removed from the final NMME forecast, improving

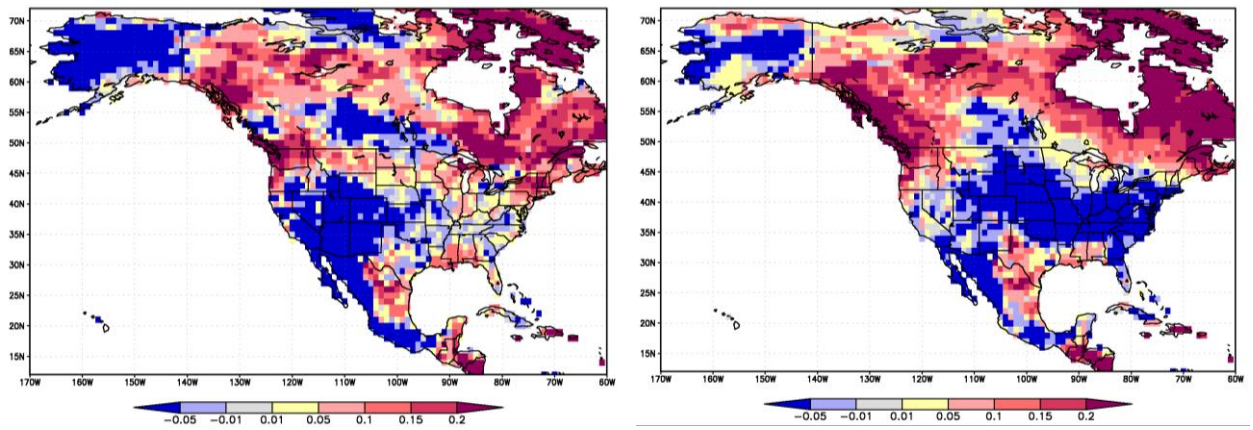

Fig. 1 Cross-validated Brier skill scores at each grid point for the uncalibrated count of ensemble members from the CFS (left) and for the combined NMME (right). Negative skill is depicted in blue.

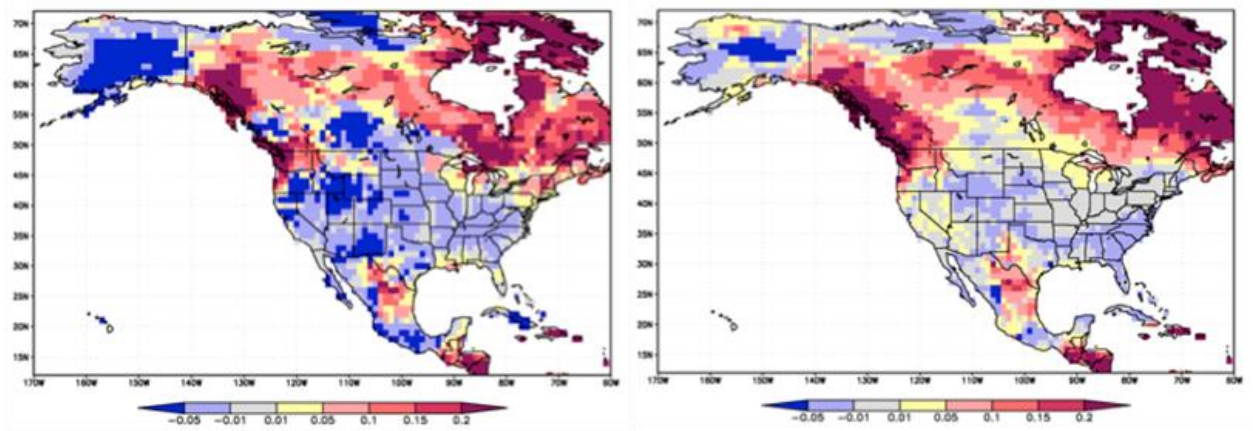

Fig. 2 Cross-validated Brier skill scores at each grid point for the EREG calibrated CFS (left) and for the combined NMME (right). Negative skill is depicted in blue.

the combined skill. Figure 2 shows the Brier skill scores for the CFS and NMME as in Figure 1, but for the EREG-calibrated probabilities. The consolidated NMME forecast improves upon the skill of the CFS alone in nearly all regions.

Using counts of ensemble members, individual models of the NMME are found to be under-dispersive or over-confident, while calibrating probability forecasts using ensemble regression produces more reliable forecast probabilities (Figure 3). It is shown that the calibrated probabilities using EREG are an improvement over CE for almost all individual models (Figure 4). However, the Brier skill score of the combined EREG-calibrated NMME is only a slight improvement over the Brier of the combined CE model probabilities. The combined calibrated forecasts are found to be slightly under-confident (Figure 3). Further work is needed to account for the additional skill obtained from combining multiple models to obtain a calibrated NMME forecast.

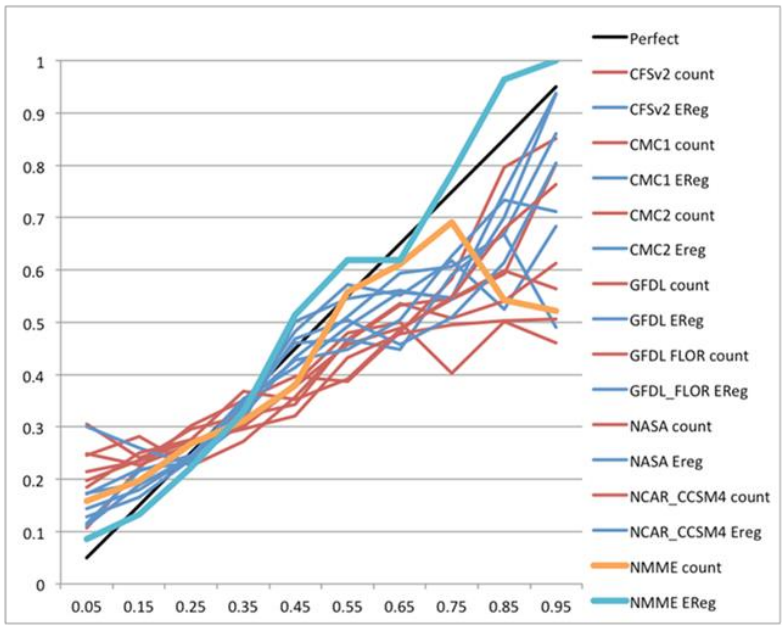

Fig. 3 Reliability diagram showing all models for the count of ensemble members (red) and EREG-calibrated probabilities (blue), reliability of the combined NMME count of ensembles (orange) and NMME EREG probabilities (cyan).

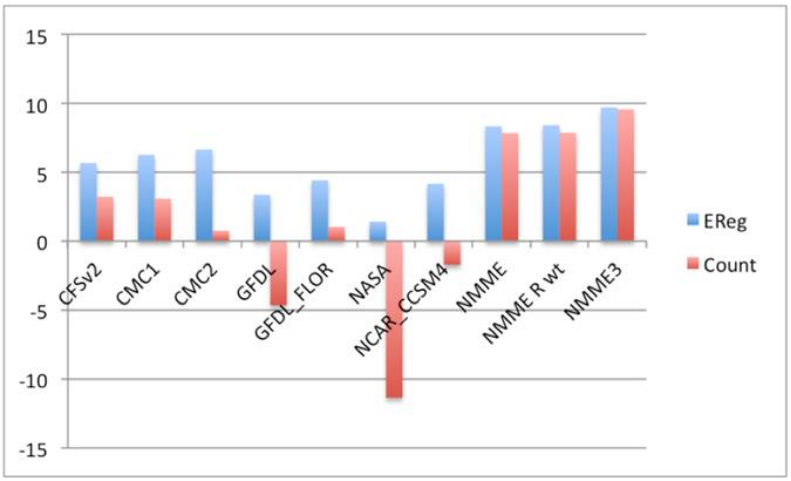

Fig. 4 Average Brier skill scores (in percent) over North America for individual models, the combined NMME, correlation-weighted NMME (NMME R wt), and a 3-model NMME (NMME3), for the count of ensemble members (red) and EREG calibrated probabilities (blue).

4. Summary and conclusions

It is found that ensemble regression (EREG) for individual model forecasts is often an improvement on counts of ensembles (CE) and climatology forecasts. Also, use of EREG to combine and weight models of the NMME virtually removes individual models in areas and seasons with no skill, improving the NMME forecast skill.

In the winter season (DJF) lead-1 temperature forecasts, skill is not significantly changed by weighting models using correlation in the seasonal NMME system beyond initial gains from regression. Skill for correlation-weighted model probabilities in Figure 4 ("NMME R wt") is nearly identical to the combined calibrated model forecasts without additional weighting of probabilities. It is noteworthy that regression effectively weights the anomalies of each model by its correlation to observations, prior to calculation of probabilities. This produces improvements in both the individual model Brier skill scores and the combined NMME skill over counts of ensemble members. Further analyses showed that the results are generally true for other seasons. Selecting and combining the three models with the highest average cross-validated skill (see far right bars of Figure 4), skill of a 3-model MME

is greater than the full NMME. This suggests that evidence-based selection of models might be used to optimize skill.

References

Becker, E., H. van den Dool, and Q. Zhang, 2014: Predictability and forecast skill in NMME. *J. Climate*, **27**, 5891-5906.

Kirtman, B. P., and Coauthors, 2014: The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585-601.

Unger, D. A., and Coauthors, 2009: Ensemble regression. *Mon. Wea. Rev*. **137**, 2365-2379.