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

Assessing Performance of Calibrated Multi-model Ensembles in the 3–4 Week Forecast Period

Kyle MacRitchie[1,2], Dan Collins[2] and Jon Gottschalck[2]
[1]*Innovim LLC., Greenbelt, Maryland*
[2]*Climate Prediction Center, NOAA/NWS/NCEP, College Park, Maryland*

1. Introduction

The NOAA Climate Prediction Center (CPC) began experimental week 3-4 probabilistic forecasts of below and above normal temperature and precipitation in September 2015. These forecasts are issued weekly on Friday afternoon. The CPC uses a number of tools to aid in its forecast creation including three dynamical models: the NCEP Climate Forecast System (CFS), the Japan Meteorological Agency (JMA) model, and the European Centre for Medium-Range Weather Forecasts (ECMWF) model.

Ensemble calibration (Unger et al., 2009), trained on the model reforecast data, is helpful to produce reliable real-time forecasts and to generate a full probability distribution from which forecast probabilities can be calculated. We conducted this study to examine the improvements that ensemble calibration yields over raw model forecasts.

2. Data and methodology

The CFS reforecast dataset includes hindcasts initiated daily from 1999–2010. Each run of the CFS includes four ensemble members which we increased to eight by including the previous forecast each day. The ECMWF reforecast dataset spanned 1996–2014 and included five ensemble members run once per week. Our JMA reforecast data included runs of five ensemble members on the 10th, 20th, and final day of each month from 1991–2010. All reforecast data was on a 1° x 1° horizontal grid.

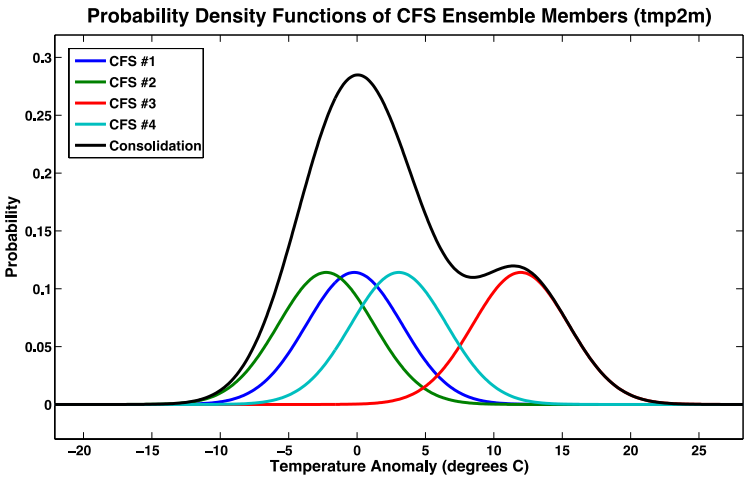

Fig. 1 Schematic showing Ensemble Regression based on Unger *et al*. (2009).

We evaluated the hindcasts using Brier Skill Scores, and reliability diagrams, which are commonly used at CPC.

Brier Skill Scores (BSS) measure the accuracy of probabilistic forecasts. The squared term in the BSS ensures that large errors are penalized more than small errors.

$$BS = \frac{1}{N}\sum_{i=1}^{N} (F_i - O_i)^2 \quad , \qquad BSS = 1 - BS_f / BS_r$$

where N is the total number of forecasts, F is the forecast probability for above normal temperature or precipitation, and O is the observed probability. BS_f is the forecast's Brier score and BS_r is the reference score for predicting climatology which is 0.5 in this case. Values range from -∞ to 1. 1 indicates perfect forecast and 0 no skill when compared to the reference forecast.

Correspondence to: Kyle MacRitchie, Innovim LLC., Greenbelt, MD and Climate Prediction Center, NOAA/NWS/NCEP, 5830 University Research Court, College Park, MD; E-mail: kyle.macritchie@noaa.gov

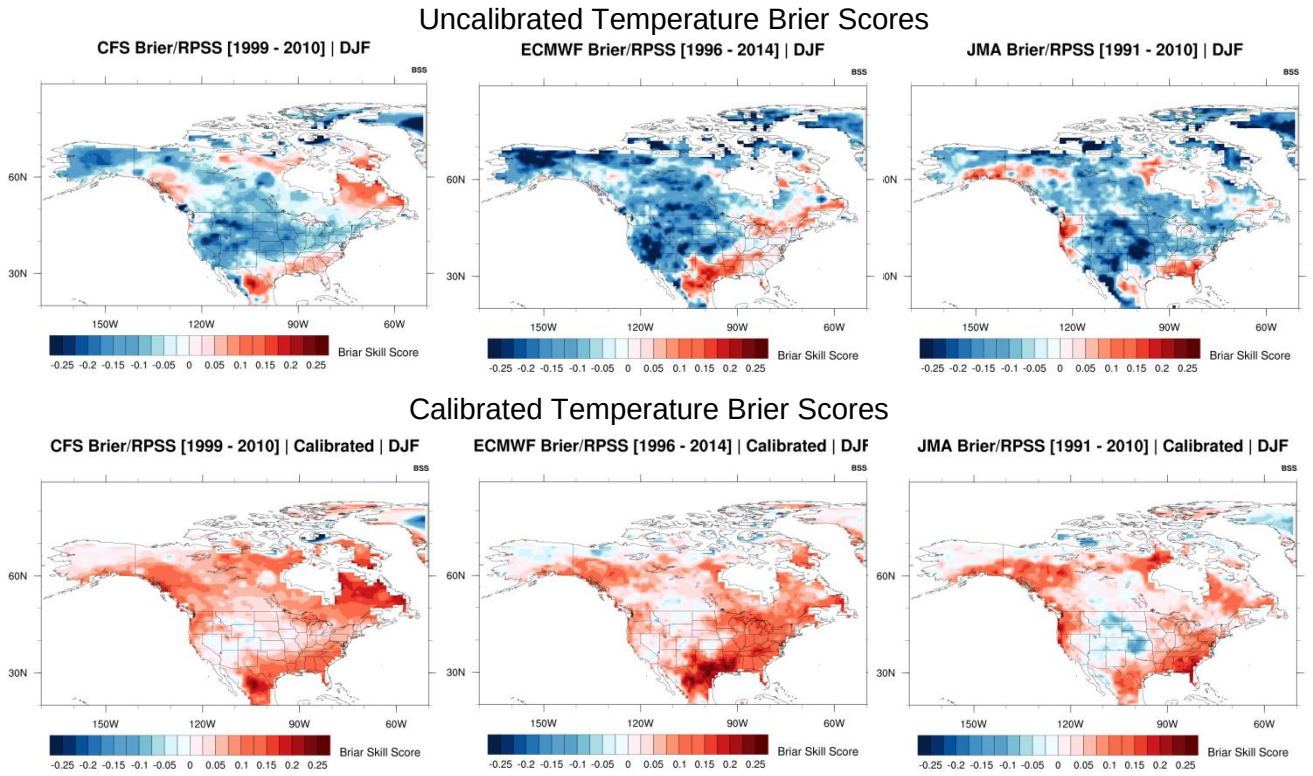

Fig. 2 Comparison between uncalibrated and calibrated Brier Skill scores for temperature.

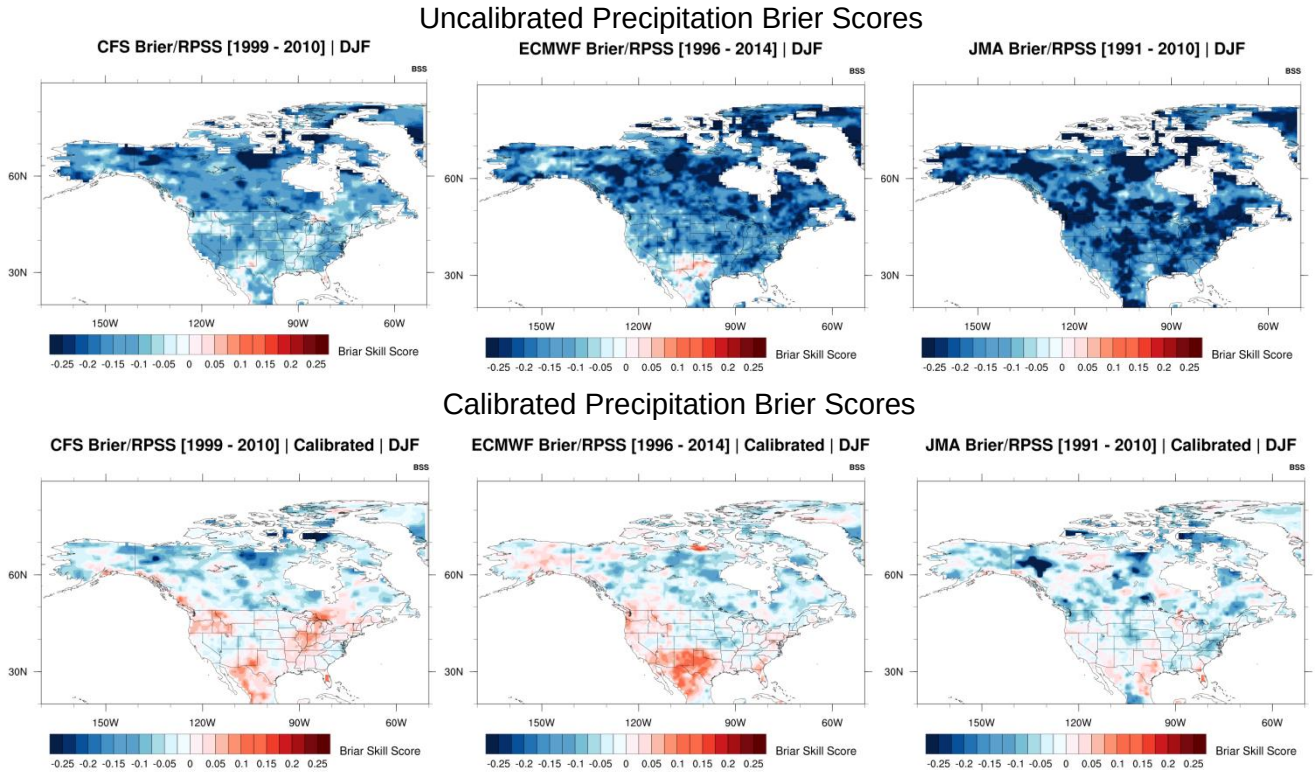

Fig. 3 Comparison between uncalibrated and calibrated Brier Skill scores for precipitation.

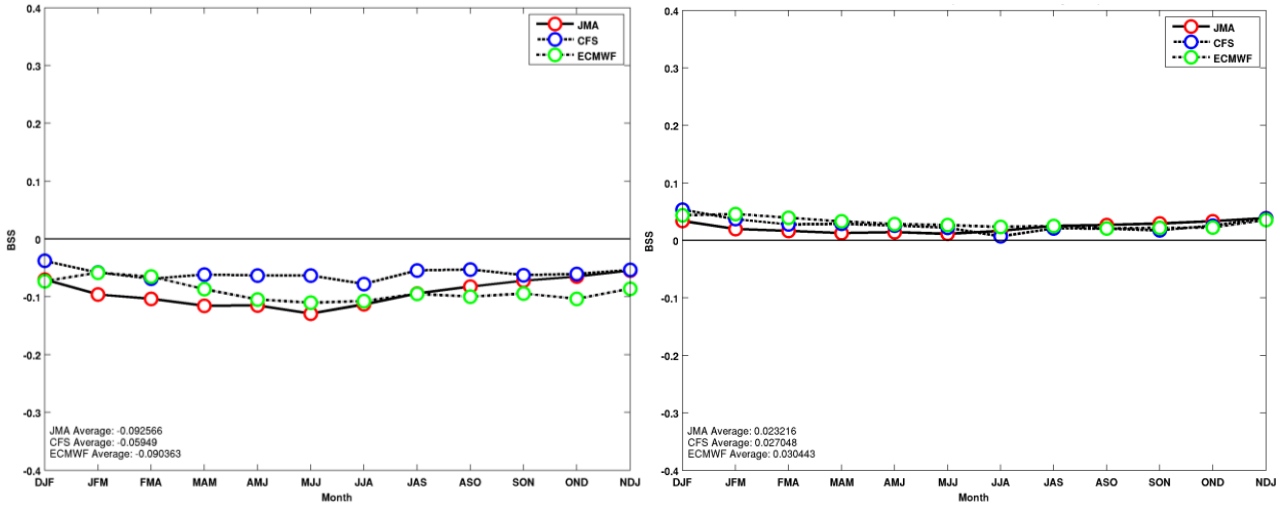

Fig. 4 Brier skill scores averaged over North America for uncalibrated (left) and calibrated (right) week 3+4 2m temperature forecast.

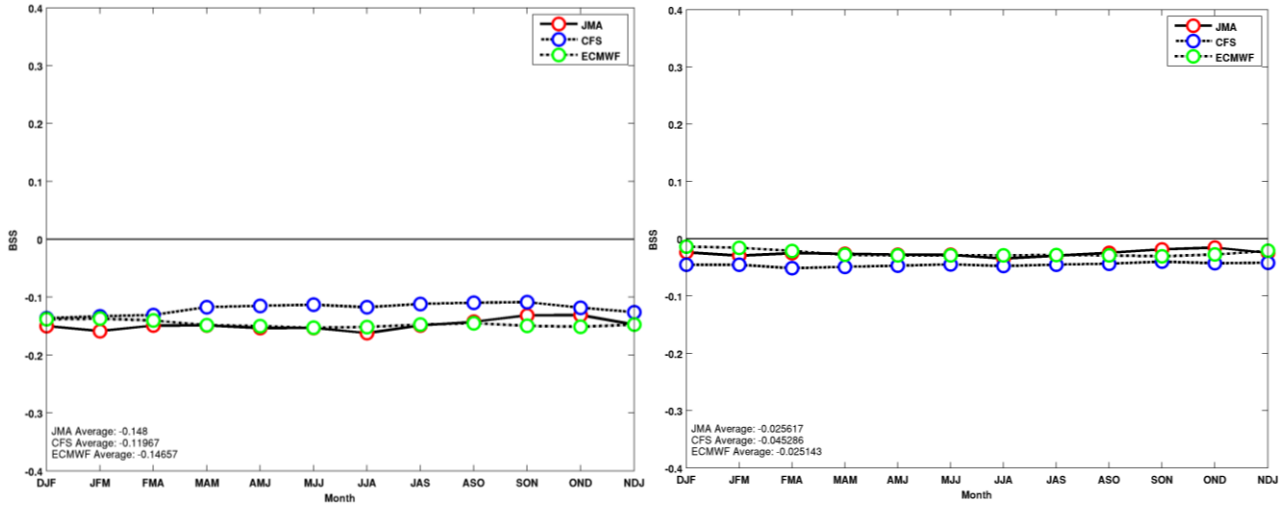

Fig. 5 Same as Fig. 4 but for precipitation.

Ensemble regression (Unger et al. 2009) was developed at CPC as a method to fit a calibrated probability density function to each ensemble member based on the model's reforecast performance. In a simplistic sense, ensemble regression weights forecasts based on the correlations between the model's reforecasts and verifications. The calibrated PDFs are assigned to each ensemble member and can then be combined for use in forecasting, as shown in Fig. 1.

3. Results

Our results show that Ensemble Regression improves model temperature and precipitation forecasts throughout the Continental United States (CONUS). Figures 2 and 3 show Brier skill scores over the CONUS for DJF temperature and precipitation forecast, respectively. Each column represents output from a specific model, the top rows show uncalibrated forecasts, and the bottom rows show forecasts after calibration with Ensemble Regression.

Skill increases from calibration are evident nearly everywhere on the map for both temperature and precipitation forecasts, although temperature forecast improvements are uniformly better than precipitation forecast improvements. This is also evident in Figs 4 and 5 which show the Brier skill scores averaged over the CONUS (weighted by the cosine of latitude).

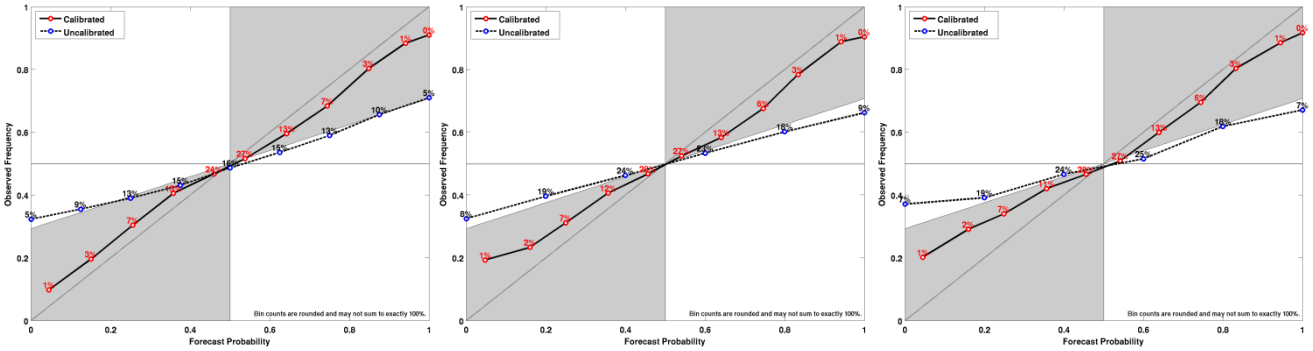

Fig. 6 Reliability diagrams of week 3-4 DJF 2m temperature forecasts over North America for models of CFS (left), ECMWF (middle) and JMA (right). Numbers next to each dot indicate the percentage of events in each bin.

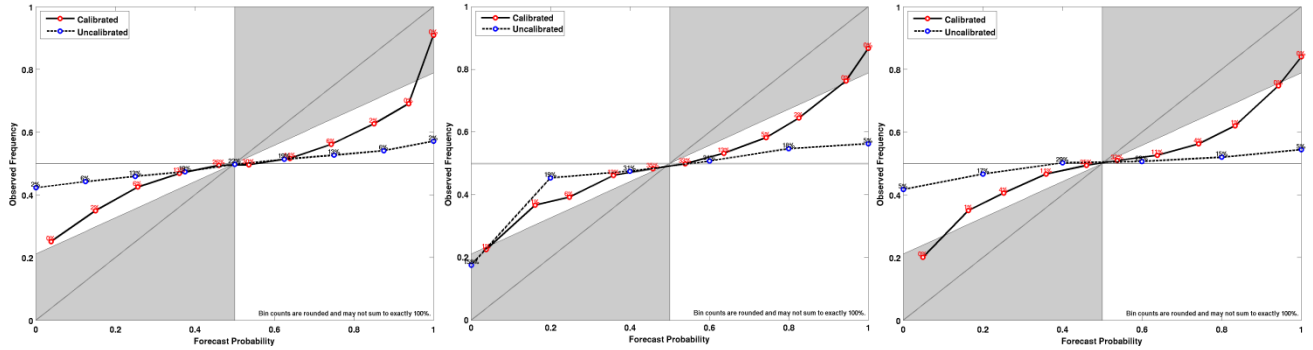

Fig. 7 Same as Fig. 6 but for precipitation.

Each model does noticeably better after calibration, although temperature forecasts are improved more than precipitation forecasts, which are markedly stuck in negative BSS territory.

Figures 6 and 7 show the effects of calibration on model reliability. As expected, calibration increases reliability within each model for both temperature and precipitation forecasts.

4. Summary

Our results show that all three models are more skillful at temperature forecasts than precipitation forecasts with 3-4 week lead times. Calibration using Ensemble Regression yields significant improvements across multiple skill and reliability metrics for both temperature and precipitation. Weeks 3-4 forecasts are inherently difficult and we will continue to improve them with whatever techniques we find.

References

Unger, D. A., H. van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Mon. Wea. Rev.*, **137**, 2365–2379.