Improving week two forecasts with multi-model re-forecast ensembles

Jeffrey S. Whitaker* **and Xue Wei**

NOAA-CIRES Climate Diagnostics Center, Boulder, CO

Frédéric Vitart

Seasonal Forecasting Group, ECMWF, Reading, United Kingdom

submitted to *Mon. Wea. Rev.*

October 21, 2005

*NOAA-CIRES Climate Diagnostics Center, 325 Broadway R/CDC1, Boulder, CO 80305-3328, Jeffrey.S.Whitaker@noaa.gov

Abstract

It has been recently demonstrated that model output statistics (MOS) computed from a long retrospective dataset of ensemble 're-forecasts' from a single model can significantly improve the skill of probabilistic week two forecasts (with the same model). In this study we extend the technique to a multi-model re-forecast dataset consisting of forecasts from ECMWF and NCEP global models. Even though the ECMWF model is more advanced than the version of the NCEP model used (it has more the double the horizontal resolution and is about five years newer), the forecasts produced by the multi-model MOS technique are more skillful than those produced by the MOS technique applied to either the NCEP or ECMWF forecasts alone.

These results demonstrate that the MOS re-forecast approach yields benefits for week two forecasts which are just as large for high-resolution state-of-the-art models as they are for relatively low-resolution, out-of-date models. Furthermore, operational forecast centers can benefit by sharing both retrospective 're-forecast' datasets and real-time forecasts.

1. Introduction

Operational weather prediction centers strive to improve forecasts by continually improving weather forecast models, and the analyses used to initialize those models. In a recent study, Hamill et al. (2004) investigated an alternative approach to improving forecasts, namely using the statistics of past forecast errors to correct errors in independent forecasts. They used a dataset of retrospective ensemble forecasts (or 're-forecasts') with a 1998 version of the operational global forecast model from the National Centers for Environmental Prediction (NCEP, part of the U.S. National Weather Service). Probabilistic forecasts of surface temperature and precipitation in week two (days 8-14) were not skillful relative to climatology when computed directly from the model output, but were more skillful than the operationally produced forecasts from NCEP's Climate Prediction Center when computed using forecast-error statistics from the re-forecast dataset.

Ensembles of forecasts with a single model suffer from a deficiency in spread partly because they do not represent the error in the forecast model itself. This causes probabilistic forecasts derived from such ensembles to be overconfident. Statistical post-processing using re-forecast datasets can ameliorate this problem to some extent, yielding reliable probabilities (e.g. Hamill et al. 2004). However, the use of multi-model ensembles (ensembles consisting of forecasts from different models, perhaps initialized from different analyses) have also been shown to significantly increase the skill of probabilistic forecasts (Krishnamurti and coauthors 1999; Palmer and coauthors 2004; Rajagopalan et al. 2002; Mylne et al. 2002), even without statistical post-processing. In this note, we examine whether the results of Hamill

et al. (2004) can be improved upon if the re-forecast dataset used to statistically correct the forecasts consists of more than one model. Here we combine the re-forecast dataset described in Hamill et al. (2005) using a 1998 version of the NCEP Global Forecast System (GFS) model, with a re-forecast dataset generated at the European Center for Medium Range Forecasts (ECMWF) to support monthly forecasting Vitart (2004). The ECMWF forecasts are run at roughly 1°grid spacing, as compared to 2.5°grid spacing used for the NCEP 1998 model forecasts.

In this note we address two questions. The first concerns whether he large gains in skill reported by Hamill et al. (2004) were a consequence of the fact that the forecast model was run at relatively low-resolution and is now nearly 6 years behind the state-of-the art. In other words, do the benefits of statistically correcting forecasts using re-forecasts datasets found in that study also apply to newer, higher-resolution models? Secondly, we investigate whether the methods described in Hamill et al. (2004) can be applied to multi-model ensembles. Specifically, we seek to address the question of whether statistically corrected multi-model forecasts are more skillful than the forecasts generated from the component models (using the statistics from the component re-forecast datasets).

2. Datasets and methodology.

a. The Analyses

Forecasts for 850 hPa temperature are presented in this note. These forecasts have been verified using both the NCEP/NCAR Reanalysis (Kistler and Coauthors 2001) and the ERA-40 reanalysis (Uppala and Coauthors 2005). Both of these analyses are on a 2.5° grid, and only those points poleward of $20°N$ are included. Since the verification statistics computed with the NCEP/NCAR and ERA-40 analyses are so similar, all of the results described here have been computed using the ERA-40 reanalyses unless otherwise noted.

b. The Forecasts

A re-forecast dataset was created at the NOAA Climate Diagnostics Center (CDC), using a T62 resolution (roughly 2.5°grid spacing) version of NCEP's Global Forecasting System (GFS) model, which was operational until January 1998. This model was run with 28 vertical sigma levels. A 15-member ensemble run out to 15 days lead time is available for every day from 1979 to the present, starting from 0000 UTC initial conditions. The ensemble initial conditions consisted of a control initialized with the NCEP-NCAR reanalysis (Kistler and Coauthors 2001) and a set of 7 bred pairs of initial conditions (Toth and Kalnay 1997) , centered each day on the reanalysis initial condition. The breeding method and the forecast model are the same as that used operationally at NCEP in January, 1998. Sea surface conditions were specified from the NCEP-NCAR reanalysis, and were held fixed to their

initial values throughout the forecast. Further details describing this dataset are available in Hamill et al. (2004) and Hamill et al. (2005).

An independent set of re-forecasts has been generated at ECMWF to support operational monthly forecasting (Vitart 2004). The atmospheric component of the forecast model is a T159 (roughly 1°grid) version of the ECMWF Integrated Forecast System (IFS) with 40 levels in the vertical. The oceanic component is the Hamburg Ocean Primitive Equation Model (HOPE). Atmospheric initial conditions were taken from the ERA-40 reanalysis, and ocean conditions were taken from the ocean data assimilation system used to produce seasonal forecasts at ECMWF. A five-member ensemble was run out to 32 days once every two weeks for a 12 year period from March 27, 1990 to June 18, 2002. Initial atmospheric perturbations were generated using the same singular vector technique used in ECMWF operations (Buizza and Palmer 1995). Further details describing the ECMWF monthly forecasting system can be found in Vitart (2004).

A combined re-forecast dataset was created by sub-sampling five members from the NCEP/CDC re-forecasts for just those dates in December, January and February for which the ECMWF forecasts were available (a total of 84). Only week two (8-14 day average) forecast means for 850 hPa temperature in the Northern Hemisphere poleward of $20°N$ were included.

c. Methodology

Three category probability forecasts for 850hPa temperature in week two were produced for the Northern Hemisphere poleward of $20°N$. The three categories are the lower, middle and upper tercile of the climatological distribution of analyzed anomalies, so that the climatological forecast for each category is always 33 percent. The climatological distribution is defined for all winter (December - February) seasons from 1971-2000. The climatological mean for each day is computed by smoothing the 30-year mean analyses for each calendar day with a 31 day running mean. The terciles of the climatological distribution are calculated using the analyzed anomalies over a 31 day window centered on each calendar day. Further details may be found in Hamill et al. (2004).

Tercile probability forecasts are generated from the raw 5-member ensemble output for each model by simply counting the number of members that fall into each category at each grid point, and then dividing that number by 5. The NCEP 15-member ensemble was sub-sampled by taking the control and the first two (out of seven) positive and negative perturbations generated by the breeding method (Toth and Kalnay 1997). These uncorrected, or 'raw' ensemble probabilities are compared to probabilities computed using a technique called logistic regression (Wilks 1995), which is described in detail in Hamill et al. (2004). In short, as implemented here, the logistic regression predicts the tercile probabilities given the ensemble mean forecast anomaly (relative to the observed climatology). As discussed in Hamill et al. (2004), we have found adding other predictors, such as ensemble spread, into the logistic regression does not improve forecast skill. Software from the Numerical Algorithms

Group (NAG - `http://www.nag.co.uk`) was used to compute the regression coefficients (NAG library routine G02GBF). These coefficients are computed at each grid point using cross-validation, that is for each of the 84 forecast times, the other 83 forecast/verification pairs are used to compute the regression coefficients. Since the forecasts are separated by two weeks, the correlation between adjacent forecasts is quite small and each forecast is essentially independent of the others. Probabilities are computed using logistic regression using each model's ensemble mean anomaly separately, and then a multi-model forecast probability is produced by using a combined ECMWF/NCEP ensemble mean anomaly as the predictor in the logistic regression. The combined ensemble mean anomaly is computed using

$$F_{combined} = b_0 + b_{NCEP}F_{NCEP} + b_{ECMWF}F_{ECMWF}, \tag{1}$$

where the F_{NCEP} is the NCEP ensemble mean, F_{ECMWF} is the ECMWF ensemble mean, $F_{combined}$ is the multi-model ensemble mean, b_{NCEP} is the weight given to the NCEP ensemble mean, b_{ECMWF} is the weight given to the ECMWF ensemble mean and b_0 is a constant term. The coefficients $(b_0, b_{NCEP}, b_{ECMWF})$ are estimated at each Northern Hemisphere grid point using an iterative least-squares regression algorithm (NAG library routine G02GAF).

Instead of combining the ensemble means and using the result as a predictor in the logistic regression, a multi-model forecast can be computed by simply using each ensemble mean as a separate predictor in the logistic regression. We have found that forecasts obtained with a two-predictor logistic regression are slightly less skillful, so only results using the two-step procedure (using the combined ensemble mean as a single predictor in the logistic regression)

are presented here.

3. Results

Figure 1 is a reliability diagram summarizing the skill of the raw, uncorrected week two tercile probability forecasts for each model in the Northern Hemisphere poleward of $20°N$. For both the NCEP and ECMWF models, the reliability of the forecasts is poor (the forecasts are overconfident) and the Ranked Probability Skill Score (RPSS, see Wilks (1995)) is near zero, indicating no skill relative to the climatological forecast of 33 percent in each category. Richardson (2001) has show that the Brier Skill Score (BSS, which is the same as the RPSS for forecasts of a single category) depends on ensemble size according to

$$B_\infty = \frac{B_M + M^{-1}}{1 + M^{-1}}, \tag{2}$$

where B_M is the BSS for an ensemble of size M and $B_\infty$is the BSS for an infinite ensemble. This equation assumes the ensemble prediction system is perfect, i.e. the verifying analysis is statistically indistinguishable from a randomly chosen ensemble member. According to this formula, if a five-member ensemble with a perfect model has no skill ($B_M = 0$) then an infinite ensemble should have a BSS of 0.166. However, we have found that the RPSS of the NCEP forecast is still negative for an ensemble size of 15 (the total number of ensemble members available). This suggests that model error is primarily responsible for the low skill of the NCEP model forecasts, not the small ensemble size used to generate the probabilities.

Although we cannot test the sensitivity of the RPSS to ensemble size with the ECMWF model (since the re-forecasts were only run with five members), the fact that the reliability diagrams shown in Figure 1 are so similar suggests that the same holds true for the ECMWF forecasts. The lack of reliability shown in Figure 1 is a consequence of the spread deficiency in both models - on average the ensemble spread is nearly a factor of two smaller than the ensemble mean error (not shown).

Applying a logistic regression to the ensemble means produces forecasts which are more reliable and have higher skill for both ensembles (Figure 2). The corrected ECMWF model forecasts are significantly more skillful than the corrected NCEP model forecasts (RPSS of 0.155 for ECMWF versus 0.113 for NCEP) . The relative improvement in the ECMWF forecasts is nearly the same as the NCEP forecasts, demonstrating that the value of having a re-forecast dataset is just as large for the ECMWF model as it was for the NCEP model. This is despite the fact that the ECMWF model has the benefit of twice the resolution and five extra years of model development. The improvement in skill is limited by the small sample size - if all 25 years of forecasts available for the NCEP model are used to compute the logistic regression, the RPSS increases to 0.15. The multi-model forecast, using the combined ensemble mean in the logistic regression, produces a forecasts with an average RPSS of 0.17 (Figure 3), about a 10 percent improvement over the corrected ECMWF forecast alone. The fact that the multi-model ensemble forecast is more skillful than the ECMWF forecast, even though the ECMWF forecast is, on average, superior to the NCEP forecast, is a consequence of the fact that there are places where the NCEP model is consistently more skillful than the

ECMWF model (Krishnamurty and coauthors 2000). This is reflected in the weights used to combine the two ensemble means (Figure 4). Although the ECMWF model receives more weight on average, there are regions where the NCEP model receives a weight of greater than 0.5 (the red regions in the left panel of Figure 4). In other words, the NCEP model, though inferior on average to the ECMWF model, supplies independent information that can be used to improve upon the ECMWF forecast. This is obviously only possible if a re-forecast dataset is available for both models.

All of the results discussed in this section have been computed using the ERA-40 reanalysis as the verifying analysis. The results are not substantively changed if the NCEP-NCAR reanalysis is used instead (not shown).

4. Discussion

The study of Hamill et al. (2004) has been extended to include more than one forecast model. The results show that the benefits of re-forecasts apply equally to older, lower resolution forecasts systems as they do to higher resolution state-of-the-art ones. In this case, logistic regression was used to combine the NCEP and ECMWF week two forecasts. The resulting multi-model forecast was more skillful than the statistically corrected ECMWF forecast, which was run at more than twice the horizontal resolution as the other model, a 1998 version of the NCEP global forecast model. These results clearly demonstrate that all operational centers can benefit by sharing both re-forecast datasets and real-time forecasts. This is true even if the models run by the various centers differ substantially in resolution, as

long as they some skill and provide independent information to the statistical scheme used to combine the forecasts. For week two tercile probability forecasts of 850 hPa temperature, the benefits of multi-model ensembles cannot be realized unless all the models have corresponding re-forecast datasets with which to estimate regression equations to combine (and calibrate) the forecasts.

Acknowledgments

Fruitful discussions with Tom Hamill are gratefully acknowledged. The NOAA "Weather-Climate Connection" program funded this project.

References

Buizza, R. and T. N. Palmer, 1995: The singular vector structure of the atmospheric global circulation. *J. Atmos. Sci.*, **52**, 1434–1456.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2005: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, to appear.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447.

Kistler, R. and Coauthors, 2001: The NCEP-NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation. *Bull. Amer. Meteor. Soc.*, **82**, 247–268.

Krishnamurti, T. N. and coauthors, 1999: Improved weather and climate forecasts from multimodel superensembles. *Science*, **285**.

Krishnamurty, T. N. and coauthors, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196–4216.

Mylne, K. R., R. E. Evans, and R. T. Clark, 2002: Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Quart. J. Roy. Meteor. Soc.*, **128**, 361–384.

Palmer, T. N. and coauthors, 2004: Development of a European multimodel ensemble system for seasonal to interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.*, **130**.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489.

Toth, Z. and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Bull. Amer. Meteor. Soc.*, **125**, 3297–3319.

Uppala, S. M. and Coauthors, 2005: The ERA-40 reanalysis. *Quart. J. Roy. Meteor. Soc.*, accepted.

Vitart, F., 2004: Monthly forecasting at ECMWF. *Mon. Wea. Rev.*, **132**, 2761–2779.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

Figure Captions

Figure 1: Reliability diagrams for week two tercile probability forecasts of 850 hPa temperature. 84 forecasts during December-February 1990 to 2002 are used, and forecasts are evaluated of the Northern Hemisphere poleward of $20°N$ using the ERA-40 reanalysis for verification. The reliability curve shown is for both upper and lower tercile probability forecasts. Inset histograms indicate frequency with which extreme tercile probabilities were issued. (A) Forecasts from the five member ECMWF model ensemble, (B) forecasts from the five member NCEP model ensemble.

Figure 2: As in Figure 1, but for forecasts corrected with a logistic regression.

Figure 3: As in Figure 1, but for multi-model ensemble. The NCEP and ECMWF forecast ensemble means are combined using linear regression, and then the combined ensemble mean is used to predict tercile probabilities using logistic regression. The combined ensemble is more skillful than the logistic regression correction applied to either the ECMWF or NCEP models alone (Figure 2).

Figure 4: A map of the weights used to combine the ECMWF and NCEP ensemble mean forecasts. These maps correspond to b_{NCEP} and b_{ECMWF} in equation 1. The mean term (b_0) is not shown.

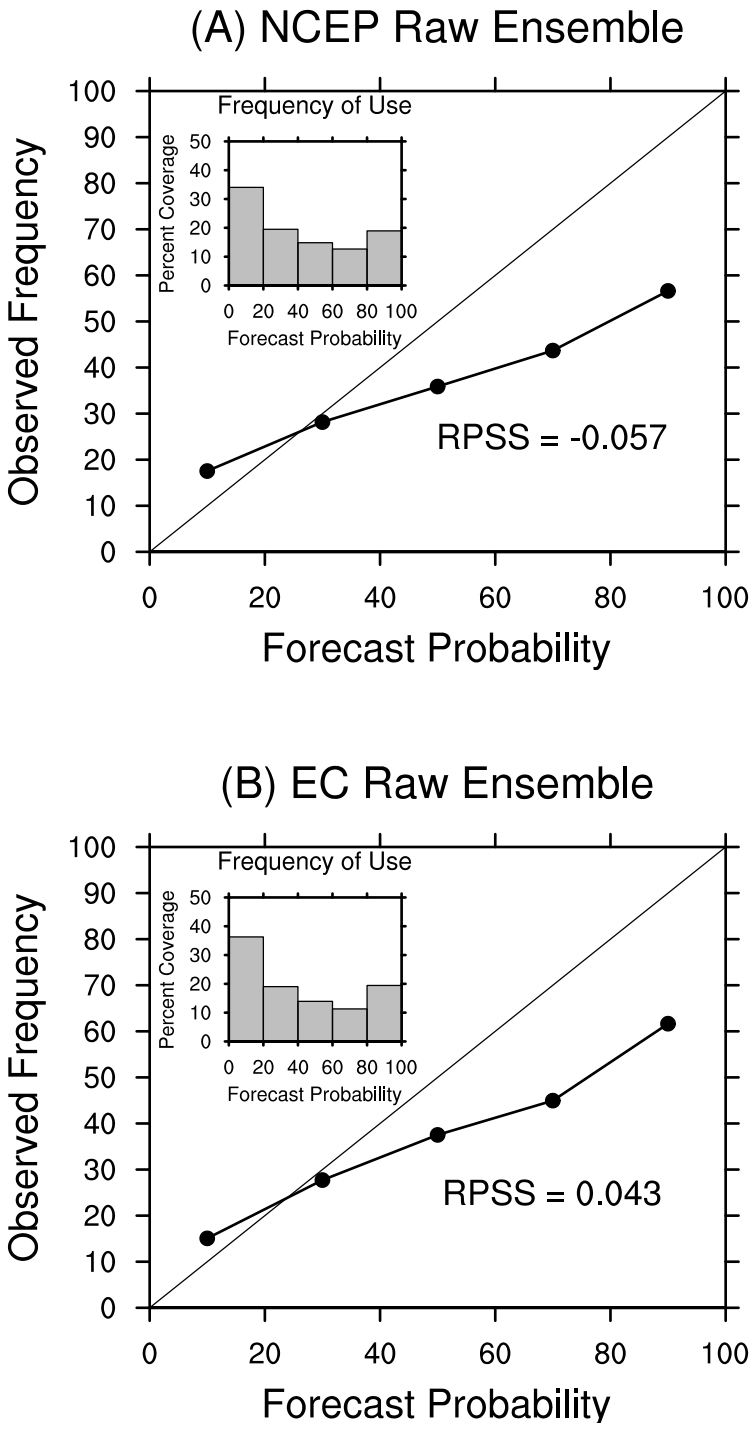

Figure 1: Reliability diagrams for week two tercile probability forecasts of 850 hPa temperature. 84 forecasts during December-February 1990 to 2002 are used, and forecasts are evaluated of the Northern Hemisphere poleward of $20°N$ using the ERA-40 reanalysis for verification. The reliability curve shown is for both upper and lower tercile probability forecasts. Inset histograms indicate frequency with which extreme tercile probabilities were issued. (A) Forecasts from the five member ECMWF model ensemble, (B) forecasts from the five member NCEP model ensemble.

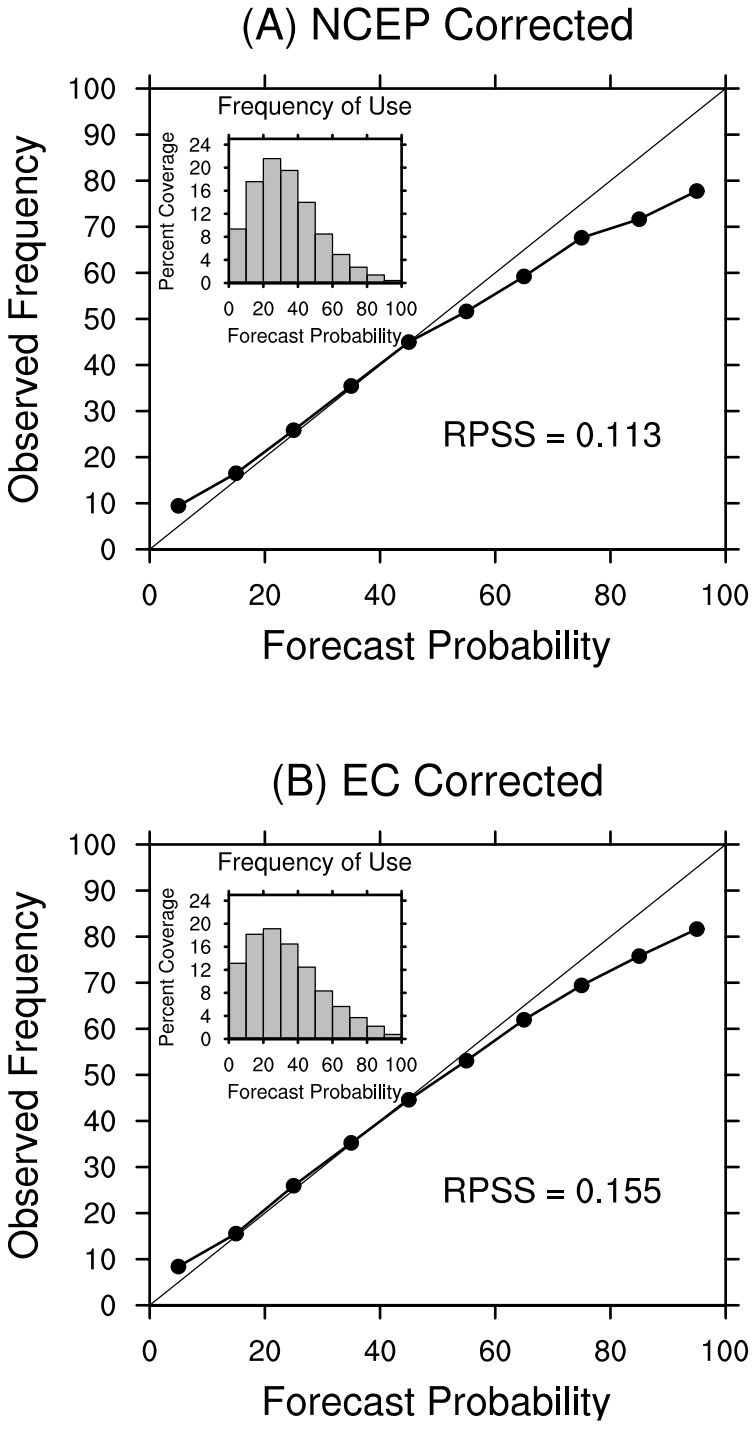

Figure 2: As in Figure 1, but for forecasts corrected with a logistic regression.

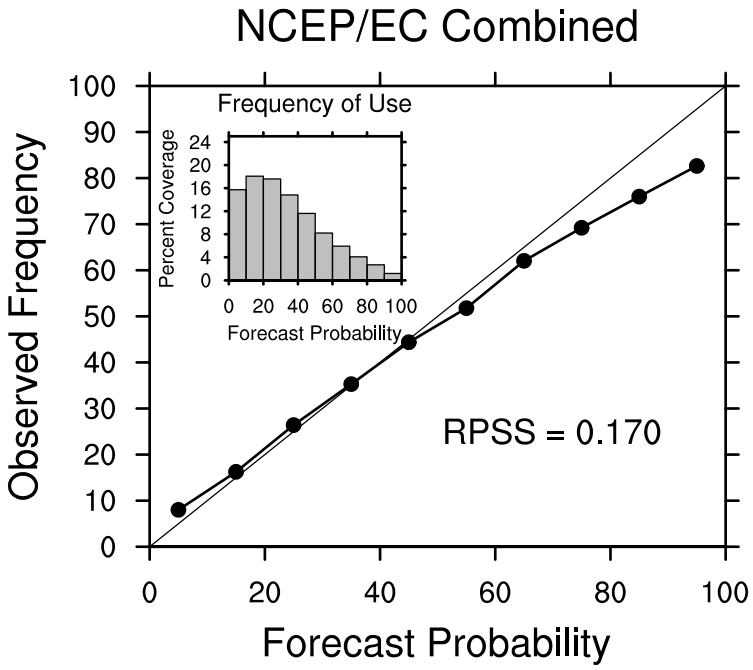

Figure 3: As in Figure 1, but for multi-model ensemble. The NCEP and ECMWF forecast ensemble means are combined using linear regression, and then the combined ensemble mean is used to predict tercile probabilities using logistic regression. The combined ensemble is more skillful than the logistic regression correction applied to either the ECMWF or NCEP models alone (Figure 2).

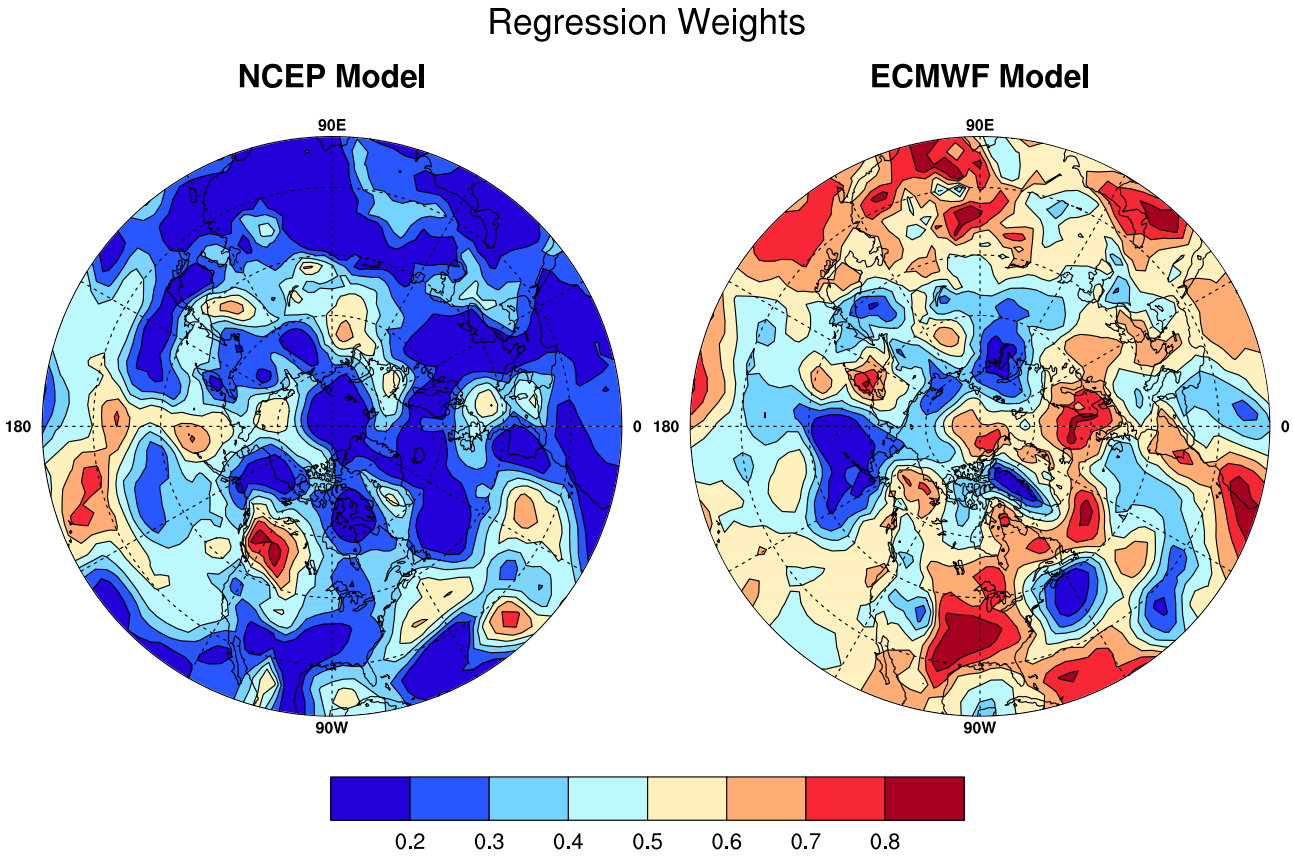

Figure 4: A map of the weights used to combine the ECMWF and NCEP ensemble mean forecasts. These maps correspond to b_{NCEP} and b_{ECMWF} in equation 1. The mean term (b_0) is not shown.