Science and Technology Infusion Climate Bulletin
NOAA's National Weather Service
36th NOAA Annual Climate Diagnostics and Prediction Workshop
Fort Worth, TX, 3-6 October 2011

Recalibrating and Combining Ensemble Predictions

Michael K. Tippett, Tony Barnston, Lisa Goddard, Simon Mason
International Research Institute for Climate & Society (IRI), Columbia University

Malaquias Pena Mendez
SAIC-Environmental Modeling Center, NOAA/NWS/NCEP

Huug van den Dool
Climate Prediction Center, NOAA/NWS/NCEP

1. Introduction

The "model output statistics" (MOS) approach has long been used in forecasting to correct systematic errors of numerical models and to predict quantities not included in the model (Glahn and Lowry 1972). The MOS procedure is based on capturing the statistical relation between model outputs and observations and, in its simplest form, consists of a linear regression between these quantities. In theory, this procedure optimally calibrates the model forecast and provides reliable forecasts.

In practice, the regression parameters must be estimated from data. In seasonal forecasting, forecast histories are short, and skill is modest. Both factors lead to substantial sampling errors in the estimates. This work examines two problems where sampling error affects the reliability of regression-calibrated forecasts and provides solutions based on two "penalized" methods: ridge regression and lasso regression (Hoerl and Kennard 1988; Tibshirani 1996). The first problem comes from the observation that, even in a bivariate setting, ordinary least squares estimates lead to unreliable forecasts. The second problem arises in the context of multivariate MOS and is that common methods of predictor selection lead to negative skill and unreliable forecasts.

2. Are regression forecasts reliable?

The task of a forecaster is to make the best estimate of a future observation given available model output. In the case of probabilistic forecasts, the uncertainty of the estimate is also needed. Modeling the forecast f and its verifying observation o as random variables, the goal of the forecaster is to obtain the conditional distribution $p(o|f)$ which is defined to be the probability distribution of the verifying observation o given that the forecast f is known to have a particular value (DelSole 2005). The mean of the conditional distribution is the "best" estimate in the sense that it minimizes the expected squared error. Uncertainty information such as the forecast variance can

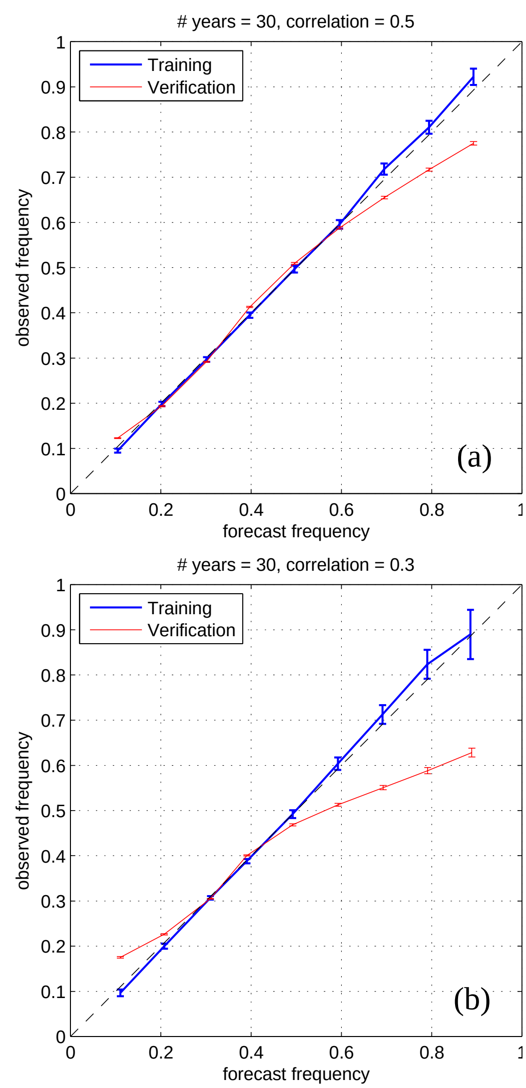

Fig. 1 Reliability diagrams for regression forecasts in (a) moderate skill (r=0.5) and (b) low skill (r=0.3) setting. The blue line is the in-sample reliability and the red line is the out-of-sample reliability.

Correspondence to: Michael K. Tippett, International Research Institute for Climate & Society (IRI), Columbia University, Palisades, NY; E-mail: tippett@iri.columbia.edu.

be computed from the conditional distribution and used to make probabilistic forecasts.

The challenge is to obtain the conditional distribution, which, in general, requires a complete description of the statistical relation between the forecast and the verifying observations. However, when forecast and observations have a joint Gaussian distribution, this only requires knowing the means and variances of o and f, and their correlation. In this case, the conditional distribution is itself Gaussian, and moreover, the conditional mean is simply given by linear regression; the conditional variance is the error variance of the regression. Such regression forecasts are known to be reliable when the regression parameters are known (Johnson and Bowler 2009). In practice, regression parameters must be estimated from data, and in the case of seasonal climate forecasts, from fairly short records. Therefore, an important issue is the impact of sampling error on the quality of forecasts, and in particular, on their reliability.

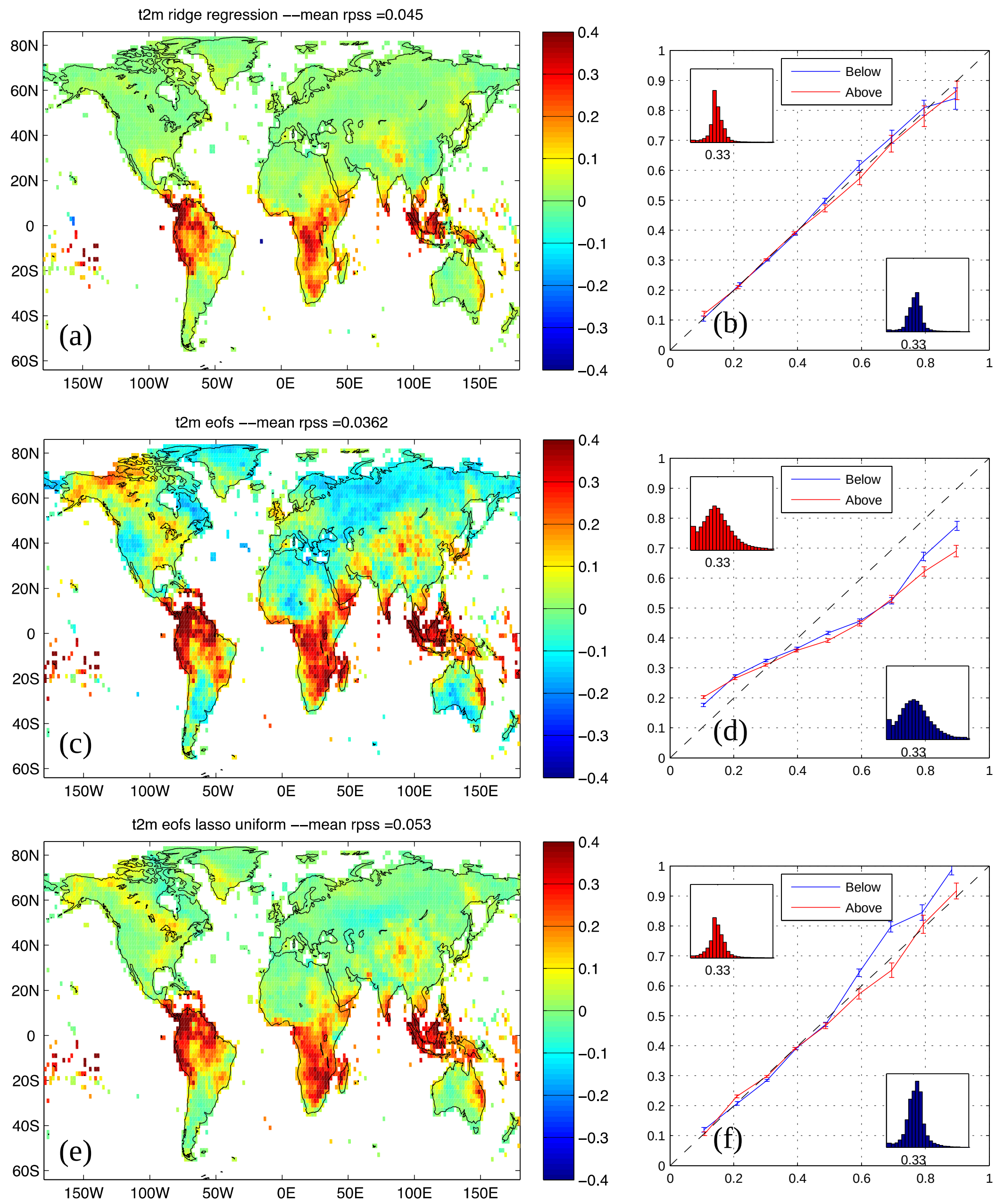

Fig. 2. RPSS and reliability diagrams for DJF t_{2m} forecasts based on gridpoint ridge regression (a,b), EOF predictors (c,d) and lasso applied to EOF predictors.

Monte Carlo experiments with bivariate synthetic Gaussian data show (Fig. 1) that regression forecasts are reliable with respect to the data used to estimate the regression parameters (in-sample). However, when regression forecasts are made for independent data (out-of-sample), they display systematic unreliability in the form of overconfidence. This problem is worse when the training sample is small and the underlying skill is low, both factors contributing to sampling error in the regression coefficient. Additional analysis of the synthetic data, as well as analytical expressions, indicate that the overconfidence is due to too strong signals rather than too small forecast error variance. Shrinkage methods like ridge and lasso reduce signal variance and improve reliability.

3. Selecting predictor patterns

A common multivariate regression forecast approach is to use spatial patterns as predictors. To illustrate our findings we use two-tier hindcasts of Dec-February (DJF) 2-meter temperature using the ECHAM4.5 atmospheric GCM with 24 ensemble members forced by Constructed Analogue (CA) forecast SST. The CA SST forecasts use data through the end of October. We use the period 1958-2001 and University of East Anglia observation data sets. Applying ridge regression on a gridpoint basis give reliable forecasts with skill in some regions as measured by the ranked probability skill score (RPSS; Figs. 2a,b). Using correlation-based EOFs of model output as predictors, results in increased RPSS in many regions (Fig. 2c), especially in the tropics, but is accompanied by negative RPSS in some regions, especially ones where the gridpoint MOS showed little or no skill. The average RPSS is 0.036, and the reliability diagram shows overconfidence (Fig. 2d). As in the bivariate case, this is due to excessive forecast signal, especially in areas where a climatological forecast of equal odds would be more appropriate. One way to proceed is to cast the problem as model-selection one, where one must choose between a pattern-based regression forecast or a climatological forecast. Methods like cross-validation and Akaike Information Criteria (AIC) can be used to select the model. This model selection approach offers some improvement, but does not entirely eliminate areas of negative skill and forecast overconfidence (not shown). Lasso regression is similar to ridge regression but more aggressively eliminates poor predictors. Lasso regression retains much of the skill improvements of EOF regression while not introducing negative skill, and improving reliability (Figs. 2e,f); average RPSS is 0.053.

4. Summary

Regression methods are often used to post-process model forecasts. When the regression parameters are known precisely, such regression forecasts lead to reliable probability forecasts. However, even in the idealized situation when the forecast and observation distributional forms are known (for instance, joint Gaussian), sampling error leads to unreliable forecast probabilities, with the lack of reliability being in the form of overconfidence. Sampling error is worse when the sample size is small or the underlying skill level is low. This systematic overconfidence is somewhat surprising since estimated regression coefficients are known to be unbiased estimates of the actual coefficients. The explanation for this behavior is that the signal variance of a regression forecast is the sum of the actual signal variance and a term that depends on the variance of the coefficient estimate. To the extent that the latter term is nonzero, the signal variance of the regression forecast is positively biased. This explanation suggests that shrinkage methods like ridge are useful. Shrinkage methods also are useful in the selection of predictors in multivariate pattern-based MOS approaches.

References

DelSole, T., 2005: Predictability and information theory. Part II: Imperfect forecasts. *J. Atmos. Sci.*, **62**, 3368–3381.

Glahn, H. R. and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

Hoerl, A. and R. Kennard, 1988: *Encyclopedia of Statistical Sciences*, **8**, chap. Ridge Regression, 129–136. Wiley, New York.

Johnson, C. and N. Bowler, 2009: On the reliability and calibration of ensemble forecasts. *Mon. Wea. Rev.*, **137**, 1717–1720.

Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267–288.