# Skill of multi-model ENSO probability forecasts

MICHAEL K. TIPPETT* AND ANTHONY G. BARNSTON

*International Research Institute for Climate and Society, Palisades, NY, USA*

October 19, 2007

*Corresponding author address*: M. K. Tippett, International Research Institute for Climate and Society, The Earth Institute of Columbia University, Lamont Campus / 61 Route 9W, Palisades New York 10964, USA. (tippett@iri.columbia.edu)

ABSTRACT

The cross-validated hindcast skills of various multi-model ensemble combination strategies are compared for probabilistic predictions of monthly SST anomalies in the ENSO-related NINO3.4 region of the tropical Pacific Ocean. Forecast data from seven individual models of the DEMETER project are used, spanning the 22-year period of 1980-2001. Skill of the probabilistic forecasts is measured using the ranked probability skill score and rate of return, the latter being an information theory-based measure. Although skill is generally low during boreal summer compared to other times of the year, the advantage of the model forecasts over historical frequencies is greatest at this time. Multi-model ensemble predictions, even those using simple combination methods, generally have higher skill than single model predictions, and this advantage is greater than that expected due to increase in ensemble size. Overall, slightly better performance was obtained using combination methods based on individual model skill compared to methods based on the complete joint behavior of the models. This finding is attributed to the comparatively large expected sampling error in the estimation of the relations between model errors based on the short history. A practical conclusion is that, unless some models have grossly low skill relative to the others, and until the history is much longer than 2 to 3 decades, equal, independent or constrained joint weighting are reasonable courses.

## 1. Introduction

The El Niño-Southern Oscillation (ENSO) phenomenon has global impacts on climate and society (Ropelewski and Halpert 1996; Glantz 2001; Mason and Goddard 2001), and the ability to predict it has provided a conceptual and practical basis for seasonal climate forecasting. Despite improvements in prediction models and observing systems, ENSO forecasts remain uncertain. Forecast uncertainty is due to model and observation error, as well as the intrinsic chaotic nature of the atmosphere. Therefore, forecast information is necessarily probabilistic and the most complete description of a forecast is its probability distribution. Ideally, an ENSO forecast distribution is the probability of the future ENSO state given the current observed state; that is, a conditional probability (DelSole and Tippett 2007). In practice, there are differences between such an ideal forecast distribution and the forecast distribution estimated from a numerical prediction model.

Past model performance data can sometimes be used to improve model output. For instance, model output statistics (MOS; in its simplest form, a regression between forecast and verifying observations) has long been used to correct biases and systematic errors in deterministic forecasts (Glahn and Lowry 1972). The same types of error affect ensemble forecasts and are manifest as errors in the central tendency of the forecast distribution. Additionally, there can be errors in the distribution of the forecast about the central tendency. For instance, not accounting for all sources of uncertainty results in an ensemble that has too little spread. Recently "ensemble MOS" methods have been developed to correct the entire forecast distribution (Wilks 2006; Wilks and Hamill 2007).

Another important development in forecasting is the use of multiple prediction models (Krishnamurti et al. 1999). Dynamical monthly climate predictions using state-of-the art coupled

ocean-atmosphere general circulation models (CGCMs) are currently produced at a number of meteorological global prediction centers worldwide. Many of the coupled models differ in their representation of physical processes, in their numerical schemes, or in their use of observations to construct initial conditions. Therefore, different models have differing biases in their reproduction of interannual variability, as well as in their forecasts of oceanic and atmospheric climate (Shukla et al. 2000). Numerous studies have demonstrated that multi-model forecasts are generally more skillful than single-model forecasts (Krishnamurti et al. 1999; Kharin and Zwiers 2002; Palmer et al. 2004; Kug et al. 2007). Conceptually, three reasons for the enhanced skill of multi-model forecasts are: (i) differing biases may cancel, (ii) the ensemble size is increased so that sampling error is reduced, and (iii) while the multi-model ensemble may have less skill than the "best" model for a particular forecast, it may not be possible to identify the "best" model *a priori* (Hagedorn et al. 2005). The simplest multi-model forecast procedure pools the models and ensembles and estimates the forecast distribution from the multi-model ensemble. More sophisticated methods weight models according to their performance. Using multiple linear regression, Krishnamurti et al. (1999) combined seasonal atmospheric climate (and also weather) forecasts from different models into an optimal multi-model ensemble forecast.

This paper describes the use of ensemble MOS and multi-model combination methods for producing probabilistic ENSO forecasts. We focus on SST anomaly predictions for the NINO3.4 region in the tropical Pacific in view of its demonstrated representativeness of the ENSO phenomenon (Barnston et al. 1997). Our goal is to determine, and explain to the extent possible, the effects of different multi-model prediction weighting techniques (including equal weighting) on the skill of probabilistic monthly ENSO forecasts. The forecasts take the form of probabilities of whether the monthly NINO3.4 index will lie in the upper, lower or middle two quartiles, corre-

sponding to warm, cold or neutral conditions, respectively. Ensembles of retrospective coupled ocean-atmosphere hindcasts using 7 CGCMs are used. Descriptions of the retrospective hindcast data are given in section 2. The two probabilistic hindcast verification measures, one global and the other local in probability space, are described in section 3, and the different ensemble MOS and multi-model techniques used to estimate the forecast probabilities are discussed in section 4. Skill results are presented in section 5, and a summary and some conclusions are given in section 6.

## 2. Data

Forecasts are taken from the DEMETER project (Palmer et al. 2004). The DEMETER project consists of global coupled model seasonal hindcasts from seven coupled models developed by: European Centre for Research and Advanced Training in Scientific Computation (CERFACS), Istituto Nazionale di Geofisica e Vulcanologia (INGV), European Centre for Medium-Range Weather Forecasts (ECMWF), Laboratoire d'Océanographie Dynamique et de Climatologie (LODYC), Max-Planck-Institut (MPI) and Météo-France, the UK Met Office (UKMO). Our analyses are based on the DEMETER forecasts that were started on 1 February, 1 May, 1 August and 1 November, over the common period of 1980-2001, and extend to six months after their start. We consider monthly averages. We refer to the monthly average that contains the start date as the zero-lead forecast, and thus the longest lead forecast is the 5-month lead forecast. We consider forecasts with leads 1 through 5. All models are represented by 9-member ensembles for a total of 63 forecasts for each of the four start times per year for five lead times.

We use as the observations the "extended" NINO3.4 index computed from Kaplan et al. (1998)

until Oct 1981 and from NCEP OIv2 (Reynolds et al. 2002) projected onto the EOFs of Kaplan from Nov 1981 on. Warm, cold and neutral ENSO months are defined as months in which the NINO3.4 index falls in the upper quartile, lower quartile, or middle two quartiles, respectively, of our 22-year base period. Quartile boundaries are calculated on a monthly basis so that the ENSO definitions vary through the calendar year as shown in Fig. 1a. Thus, larger anomalies are required to satisfy the ENSO definition during the later parts of the calendar year when the year-to-year variability is larger. This definition of ENSO events differs from ones such as that of NOAA, in which there is a constant anomaly threshold throughout the seasonal cycle (Kousky and Higgins 2007), or such as both NOAA and IRI where sustained SST anomalies over some multi-month period are required. Figure 1b shows the time series of standardized NINO3.4 anomalies and ENSO category.

## 3. Skill measures

### *a. Ranked probability skill score*

The ranked probability score (RPS) is a measure of the sum of squared differences between the cumulative forecast probabilities and the cumulative observed probabilities for corresponding, progressively increasing category rank. RPS measures forecast reliability and resolution (Murphy 1973), and is a global score in probability space in the sense that it depends on the entire forecast distribution. The RPS of the 3-category probability forecast [$P$(cold), $P$(neutral), $P$(warm)] is

$$\begin{aligned} \mathrm{RPS} &= (P(\text{cold}) - O(\text{cold}))^2 + (P(\text{cold}) + P(\text{neutral}) - O(\text{cold}) - O(\text{neutral}))^2 \\ &= (P(\text{cold}) - O(\text{cold}))^2 + (P(\text{warm}) - O(\text{warm}))^2 \,, \end{aligned} \tag{1}$$

where the observation "probability" $O(\text{cold})$ is one when the observation is in the lowest quartile category and zero otherwise; likewise, $O(\text{warm})$ and $O(\text{neutral})$ are when one when the observation is in the corresponding category and zero otherwise. RPS is oriented such that low scores indicate high forecast quality. The ranked probability skill score (RPSS; Epstein 1969) is

$$\text{RPSS} = 1 - \frac{\text{RPS}}{\text{RPS}_{\text{ref}}} \tag{2}$$

where $\text{RPS}_{\text{ref}}$ is the RPS of a simple reference forecast such as the equal-odds climatological forecast $C = [0.25, 0.50, 0.25]$ or a persistence-based forecast. Positive RPSS indicates a forecast with greater skill than the climatology. In evaluating forecast quality, we average the RPSS of individual predictions using the climatological frequency forecast $C$ as the reference forecast.

Another simple benchmark forecast that we will evaluate is the conditional climatology, that is, the the historical frequency-based probabilities of each category given the category of the preceding month. This kind of reference forecast was used in Mason and Mimmack (2002), in which it was called the "damped persistence strategy" in order to distinguish it from an outright forecast of persistence (with probability of 100%) of the ENSO state existing at the time of the forecast. For instance, for a February forecast the conditional climatology would be the frequency of each ENSO category conditioned on the state in January. The historical frequencies are computed using the independent period 1900 - 1979.

*b. Ignorance and rate of return*

Ignorance is a likelihood skill score with connections to information theory (Good 1952; Roulston and Smith 2002). Unlike RPSS, ignorance is a local measure and only depends on the forecast probability of observed outcome. Its expected value for a single forecast is the relative entropy be-

tween the forecast and climatological distributions, and its average (over predictions) is the mutual information between forecast and observations under the perfect model assumption (DelSole 2004; DelSole and Tippett 2007).[1] The ignorance $\mathrm{Ig}$ (measured in bits) of an individual forecast is

$$\mathrm{Ig} = -\log_2 P(\text{observed category})\,, \tag{3}$$

where $\log_2$ is the base-2 logarithm and $P(\text{observed category})$ is the forecast probability assigned to the observed category. This skill measure infinitely harshly grades incorrect deterministic forecasts, i.e., forecasts of zero probability of events that occur. The reduction of ignorance compared to a climatological forecast is $\mathrm{Ig} - \mathrm{Ig}_{\text{climo}}$ where

$$\mathrm{Ig}_{\text{climo}} = -\log_2 C(\text{observed category})\,. \tag{4}$$

Because of its connection with relative entropy, ignorance can be converted into a rate of return on an investment or wager. Imagine a gambler with knowledge of a forecast who wagers on the future state of ENSO. When the odds are taken from climatology, the gambler can profit from the forecast information. If the odds are fair, in the sense of the probabilities adding to one, the average rate of return ROR (in %/wager) for the gambler is

$$\text{ROR} = 100 \times (2^{\langle \mathrm{Ig} \rangle - \langle \mathrm{Ig}_{\text{climo}} \rangle} - 1)\,, \tag{5}$$

where the angle brackets denote expectation over a set of forecasts. The skill measure ROR is equivalent to the expected ignorance but is expressed in more familiar units. The compound (geometric mean) rate of return is used to amalgamate ROR values for different forecast starts and leads.

[1] In a Gaussian perfect model setting with a forecast whose correlation with observations is $r$, the mutual information is $-\log(1-r^2)$.

## 4. Estimation of probabilities from ensembles

The methods used here to estimate ENSO probabilities from forecast ensembles can be divided into 3 classes. In the first class are methods that only use forecast model output – observations are not used. We call methods in this class *uncalibrated methods*. Uncalibrated methods assume that the ensemble mean and ensemble statistics are correct. However, past model performance may indicate deficiencies in these quantities as well as strategies for improving forecast probabilities. Methods that use observations and take into account model performance are classified according to whether models are compared with each other. Methods which use equal weighting or individual model performance, independent of the behavior of other models are called *independent calibration methods*, whereas methods in which the contribution of a model to the forecast probabilities depends on the behavior of the other models are called *joint calibration methods*. We now describe the methods in these three classes. A summary list of the methods is given in Table 1.

### *a. Uncalibrated methods*

A simple nonparametric method of estimating ENSO probabilities from a forecast ensemble is to count the number of ensemble members in the cold, neutral and warm categories. Systematic model errors are corrected by using the forecast model's climatology to compute quartile boundaries. Quartile boundaries are computed in a cross-validated fashion so that the current forecast is not included but all other ones are. The category definition varies with start date and lead-time. Further, because of the cross-validation design, the category definition varies slightly even within one start date and lead time, depending on which year is being held out as the target of the forecast.

Probabilities of one or zero occur when all or none of the ensemble members fall into one

category. Such deterministic forecasts are undesirable as they tend to reflect the small ensemble size rather than complete certainty, and can cause problems when likelihood skill measures like ignorance are used. Therefore we use the following *ad hoc* formula to compute the probability of a category:

$$P(\text{category}) = \frac{\text{\# of ensemble members in that category} + 1/3}{\text{total number of ensemble members} + 1}\,. \tag{6}$$

In the case of 63-member ensembles, with which we will often be dealing here, the minimum probability for a category is 0.5% and the maximum probability is 99.0%.

Errors in the climatological mean and variance of a single forecast model are corrected by using quartile definitions based on the model's own forecast history. However, when a multi-model ensemble containing many models is formed, only the systematic errors of the entire ensemble are corrected by this procedure. We refer to this minimally adjusted ensemble as simply the multi-model ensemble (MM). A more refined correction scheme is to form the multi-model ensemble with the anomalies of each model with respect to its own climatology. We refer to this as the multi-model ensemble with bias correction (MM-bc). A further refinement is accomplished by forming the multi-model ensemble with the normalized anomalies of each model, thus removing any systematic differences between the ensemble variances of individual models. We refer to this option as the multi-model with variance and bias correction (MM-vc). In all three correction schemes, the bias and/or variance used to correct a particular forecast are computed without using that forecast–i.e,. cross-validation is always used.

We use the following two schemes to investigate the extent to which forecast probabilities can be parameterized by the multi-model ensemble mean alone. These methods address the question of the relative importance for predictability of changes in forecast mean and of the distribution about

that mean. Kleeman (2002) examined this issue in several simple models including a stochastically forced coupled ocean-atmosphere model used to predict ENSO, finding that changes in the ensemble mean provided most of the prediction utility. Similar results have been seen in other climate problems (Kleeman 2002; Tippett et al. 2004, 2007; Tang et al. 2007). In the context of extended range weather forecasts, Hamill et al. (2004) found ensemble spread not to be a useful predictor in constructing probability forecasts.

In the first scheme (MM-c), the forecast distribution is taken to be a constant distribution about a varying mean. The forecast distribution is formed from the ensemble spreads from all years centered about the current ensemble mean. The constructed ensemble has more members than the actual ensemble by a factor of 22 and so sampling error is reduced. This constructed ensemble has a distribution about the ensemble mean that varies with start month and lead but not with forecast year.

Another way of constructing probabilities that are parameterized by the ensemble mean is to form a generalized linear regression (GLR) between the the ensemble mean and the direct model output probabilities, in particular those from the MM-bc method. We call this method MM-glr. This regression constructs a parametric connection between the ensemble mean and model forecast ENSO probabilities–no observations are used. When the probit model is used in the GLR, as is here the case, the procedure is related to fitting a Gaussian when the distribution is indeed Gaussian but sometimes performs better than Gaussian fitting for data that does not have a Gaussian distribution (Tippett et al. 2007). This method should not be confused with the commonly used method of developing a GLR between the ensemble mean and *observations* which serves to calibrate the model with observations (Hamill et al. 2004). Rather the GLR used in the MM-glr method parameterizes the ensemble probabilities in terms of the ensemble mean but does not correct the model.

Such a regression can serve to reduce the sampling variance of the counting probability estimate due to finite ensemble size (Tippett et al. 2007).

*b. Independent calibration*

The simplest independent calibration method used here is to assume that the bias-corrected multi-model ensemble mean is the best estimate of the forecast mean and then to estimate the forecast distribution about it from past performance rather than from the ensemble distribution. Using a Gaussian distribution to model forecast uncertainty gives the method that we call MM-g. We fit the Gaussian distribution to past performance in a way that accounts for any systematic amplitude errors and bias (see Appendix A). In the MM-g method, the forecast distribution is Gaussian with variance that is constant from one forecast to another. Non-homogeneous Gaussian regression (ngr) offers a more general framework with the forecast variance changing from one forecast to another (Gneiting et al. 2005; Wilks 2006). In the ngr method, the forecast variance is equal to a constant plus a term that is proportional to the ensemble variance. The parameters of the ngr model are found by optimizing the continuous ranked probability score (CRPS) which requires minimizing the absolute forecast error (Gneiting et al. 2005). The constant variance parameters from MM-g are used to initialize the numerical optimization procedure used to minimize the CRPS.

Regressing the ensemble mean of each model with observations and then taking the averages of the separate regressions is a way of assigning different weights to each forecast model without directly comparing the models. This method is equivalent to multiple linear regression with the assumption that the models errors are uncorrelated. Probabilities are assigned using a Gaussian distribution to model the uncertainty of the averaged regressions. We call this method grsep.

The independent calibration methods above give an entire (Gaussian) forecast distribution from which the probabilities of exceeding any particular threshold can be computed. Two independent calibration methods that give only the categorical probabilities are glro and MM-bow. In glro, generalized regressions are developed separately between the ensemble means of each model and the binary variables for the observed occurrence of each category. Then the resulting probabilities are averaged. Similarly, in MM-bow weights for the ensemble probabilities of each model and the climatological probabilities are found to optimize the log-likelihood of the observations which is proportional to the average ignorance (Rajagopalan et al. 2002). Then the resulting probabilities are averaged similar to Robertson et al. (2004).

*c. Joint calibration*

The most familiar joint calibration method is superensembling where optimal weights are found for the ensemble means by multiple linear regression (Krishnamurti et al. 1999). Superensembling is a special case of the more general method of forecast assimilation (Stephenson et al. 2005). A forecast probability distribution can be computed using a Gaussian distribution to model the uncertainty of the superensemble mean. We call this method gr. Applying methods where the model weights are estimated simultaneously from historical data is potentially difficult in the case of the DEMETER data because the number of models is relatively large compared to the common history period. For instance, the robust estimation of regression coefficients may be difficult. The same potential difficulty applies to the Bayesian optimal weighting (bow) method where optimal weights (relative to climatology) are simultaneously found for the probabilities of each model (Rajagopalan et al. 2002). In the case of Gaussian regression, the number of predictors can be

reduced, and hence the number of parameters to be estimated, using CCA or SVD (Yun et al. 2003). Here we use CCA with two modes (cca).

DelSole (2007) introduced a Bayesian regression framework where prior beliefs about the model weights can be used in the estimation of regression parameters. The ridge regression with multi-model mean constraint (rrmm) method uses as its prior the belief that the multi-model mean is the best solution. This is the same as finding the coefficients that minimize the sum of squared error plus a penalty term that grows as the coefficients become different from 1/(number of models), which is 1/7 in our case. The weight given to the penalty term determines the character of the regression coefficient. When infinite weight is given to the penalty term, the model weights are all the same, and rrmm is identical to MM-g. When no weight is given to the penalty term, the model weights are those given by multiple regression, and rrmm is the same as gr. Another method, ridge regression with multi-model mean regression (rrmmr) uses as its prior the belief that the models should be given approximately the same weight and penalizes coefficients that are unequal but does not penalize their difference from 1/(number of models). Like rrmm, when infinite weight is given to the penalty term, rrmmr is the same as MM-g (the multi-model mean is essentially regressed with observations) and when no weight is given to the penalty term, rrmmr is the same as gr. For both the rrmm and rrmmr methods, the parameter determining the relative weight of the penalty term is computed using a second level of cross-validation.

## 5. Results

### *a. Uncalibrated methods*

A set of skill results arranged by start date and lead is shown in Fig. 2 for uncalibrated schemes, that is, ones that do not use any skill assessment calibration with respect to the observations. Average RPSS and compound average ROR values over all starts and leads are given in Table 1. Removing the mean bias from each model almost always improves skill, and often by substantial margins for both the RPSS and ROR skill measures (MM-bc versus MM). This result is consistent with Peng et al. (2002). Making the inter-ensemble variances of each model identical (MM-vc) tends to slightly further increase RPSS and ROR skills, although the effect is not consistently positive. Two schemes that use only the ensemble means to construct the probability distribution statistically (MM-c and MM-glr) have skill very close to that of MM-vc, but generally do not exceed it. We therefore consider the MM-vc as the benchmark among the ensembling methods to be tested below, as it represents a most general and basic calibration of the models' individual biases in mean and interannual variability, while retaining the individual ensemble distributions with their year-to-year variations of spreads and shapes.

Figure 2 indicates that much of the useful forecast skill can be attributed to variations in the ensemble mean rather than variations in higher order statistics, as MM-c and MM-glr skill tends to be only very slightly lower than MM-vc. This timing is consistent with the conclusions of other recent studies of other forecast variables (Kharin and Zwiers 2003; Hamill et al. 2004; Tippett et al. 2007; Tang et al. 2007). Examination of individual forecasts reveals that the slight shortcoming in MM-c for November starts is due mainly to the single year 1983, when, initialized with cold conditions, all the forecasts were for cold, but the observation was in the neutral category. The

MM-vc forecast had slightly weaker probabilities for the cold category and so was penalized less.

The greatest benefit of the dynamical predictions compared to the conditional climatology frequency forecast occurs for forecasts starting in May. This makes sense in view of the fact that May is the time of the northern spring ENSO predictability barrier – a time when warm or cold ENSO episodes are roughly equally likely to be dissipating (having matured several months earlier) versus growing toward a maturity to occur later that calendar year. The frequency forecast incorporates both possibilities indiscriminately from the entirety of the data history, while the dynamical models have opportunity to respond to initial conditions that may help identify the direction of change in the ENSO condition (e.g. from sub-surface tropical Pacific sea temperature structure). On the other extreme, the lead-1 forecasts starting in February have slightly less skill than the frequency forecasts as measured by RPSS. This could relate to the fact that it is a time of year when existing ENSO episodes are in a process of weakening toward neutral, a tendency that can be captured by the simple frequency forecasts.

Figure 3 shows skills for MM-vc compared with skill for each constituent single model. With just a few exceptions, multi-model skill is superior to the skill of a single model. One reasonably might ask whether the higher skill of MM-vc is due merely to the increased ensemble size. This question can only be answered definitively by comparing simple model and multi-model ensembles of the same size. However, theoretical estimates for the improvement of RPSS as ensemble size is increased can be used to judge whether the superior skill of the multi-model ensemble is comparable to that expected by increasing the ensemble size of a single model. The average RPSS for the individual models over all starts and leads is 0.47 (0.50 excluding MPI). Generalizing the results of Richardson (2001) and Tippett et al. (2007), the RPSS of a perfectly reliable forecast

system depends on ensemble size $N$ according to (see Appendix B for details)

$$\langle \mathrm{RPSS}(N) \rangle = \frac{(N+1)\langle \mathrm{RPSS}(\infty) \rangle - 1}{N} \,. \tag{7}$$

Setting $\langle \mathrm{RPSS}(9) \rangle = 0.5$, gives $\langle \mathrm{RPSS}(\infty) \rangle = 0.55$, and $\langle \mathrm{RPSS}(63) \rangle = 0.54$. On the other hand, the average RPSS of MM-vc is 0.60. Therefore the increase in skill of the multi-model ensemble is not consistent with an increase in skill due to increase in ensemble size alone. Hagedorn et al. (2005) demonstrated that while increasing the ensemble size of a single model did increase skill, the multi-model ensemble retained a slight advantage over the single model ensemble.

*b. Independent calibration*

Skill measures arranged by start date and lead are given in Fig. 4 for the independent calibration methods. Average RPSS and compound ROR are given in Table 1. For the most part, the independent calibration methods have comparable skill with grsep having the best overall performance, with its ROR slightly exceeding that of MM-vc. However, the glro method has significantly poorer skill than the other methods. The poor skill obtained when using a generalized linear regression between the ensemble mean and the occurrence of a category is due the method erroneously giving too strong probabilities. This behavior appears to be an effect of the small sample size and the fact that the predictands used to develop the model are zero or one. Moreover, in some cases the numerical method for estimating the glro parameters does not converge. In contrast, the predictands in MM-glr method are the MM-bc probabilities which are not so extreme and the model is "easier" to fit and tends to give milder probability shifts. Because the glro probabilities sometimes approach zero or one, when they are wrong, the ROR harshly scores them. The MM-bow method suffers from similar problems and has somewhat poor skill compared to the other independent calibration

methods.

*c. Joint calibration*

We next explore the potential to improve skill scores further by joint calibration. Skill measures arranged by start date and lead are given in Fig. 5 for the joint calibration methods. Average RPSS and compound ROR are given in Table 1. The joint calibration methods have mostly comparable skills with only occasional improvement on that of MM-vc especially in the November starts. The methods gr and bow have poorer overall skill than the other methods, especially when measured by the ROR. The multiple regression method gr has substantially poorer skill outside of the November starts. The result that gr skill is poor when the skill level is low and is good when the skill level is high (November starts) is is consistent with the usual estimate of the variance of the regression coefficient that depends on sample size and skill level. Apparently, the period of data (22 years) is not long enough to fit the regression coefficients reliably when 7 predictors are used. When forecasts are made on cases within the training sample (i.e., when cross-validation is not used), results (not shown) using multiple linear regression are excellent, and usually exceed those of MM-vc. This indicates overfitting to the short sample data. Similar behavior was seen by Kang and Yoo (2006) who, using an idealized climate models and deterministic skill scores, noted that the superensemble method lead to overfitting that was more severe when skill is low.

One way to avoid overfitting is to reduce the number of predictors and hence the number of parameters that need to be estimated. In the cca method we reduced the number of predictors by using two CCA modes as predictors. The cca method had the best overall performance of the jointly calibrated methods. Comparable performance was achieved by the Bayesian regression

methods which also are designed to avoid overfitting. Penalizing the regression for deviating from the MM-g reduces the tendency to overfit. The choice of the weight given to the penalty is chosen in a second level of cross-validation. This parameter does not seem to be robustly estimated since, if it were, rrmm would always be better than MM-g (it is not) as MM-g is a special case of rrmm when the weight given to the penalty term is infinite. In fact, when the weight is determined using all the data, rrmm is better than MM-g. The Bayesian regression method that penalizes unequal weights, rrmmr, is slightly better over all than MM-g. It is consistently better than MM-g if the penalty weight is chosen using all the data. These problems suggest a fundamental problem with the rrmm and rrmmr methodologies since they are not able to use the data to determine when the differences between the models in the historical record are insufficient to weight one more than another (DelSole 2007).

Thus, while skill results for some of the methods are relatively resistant to overfitting (grsep, rrmm, rrmmr, cca) and can be considered representative of expected skill on truly independent (e.g. future) data, a disappointment is that they are generally not generally superior to the skill of the optimum version of equal weighting for the models used in this study despite apparent skill differences among the models (Fig. 3).

## 6. Summary and conclusions

This study assesses the skill of probabilistic ENSO forecasts and examines the benefits of ensemble MOS and multi-model combination methods using recent coupled model SST forecast histories of monthly NINO3.4 tropical Pacific SST categories. Warm, neutral, and cold categories are defined using the upper, middle two, and lower, quartile categories, respectively. Forecast

data from the seven coupled models of the DEMETER project are used, with start dates spanning the period 1980-2001. Forecast skill is accessed using ranked probability skill score (RPSS) and rate of return (ROR); ROR is a skill measure which attaches an investment value to the forecast information and has the same ranking as the ignorance or log-likelihood skill score (Roulston and Smith 2002).

Three classes of multi-model ensemble MOS methods for producing forecast categorical probabilities are examined: uncalibrated, independent calibration and joint calibration. Uncalibrated methods only use model output, and do not use observations or model performance. Independent calibration methods use observations and model performance to calibrate each model separately. Joint calibration methods use each model's performance as well as relations between the performance of models.

In uncalibrated methods, two model data pre-processing steps are performed: (1) removing the climatology of each model; and (2) normalizing the anomalies to have the same variance. These calibrations to the individual models are shown to increase skills when multi-model ensembling with equal model weighting is used as a baseline condition. Such multi-model ensembles are found to have higher skills than those from the individual models given the same calibrations, with only occasional exceptions. The dependence of RPSS on ensemble size is known for a perfectly reliable model (Tippett et al. 2007). This dependence shows that the advantage in skill of the multi-model ensembles over the single-model average skill is larger than that expected due to the larger size of the multi-model ensemble. Other studies have directly shown that the skill improvement of the multi-model ensemble is greater than that achieved by increasing the ensemble size of a single model (Hagedorn et al. 2005). Therefore the advantage of the multi-model ensemble over small single model ensemble is consistent with both increased ensemble size and reduced model

error. Forecast distributions constructed from the multi-model ensemble mean alone had nearly as much skill as ones constructed from the forecast ensemble, indicating that much of the useful forecast information is contained in the mean of the forecast distribution. Interestingly, while ENSO predictive skill is generally low during boreal summer compared to other times of the year, the advantage of the model forecasts over forecasts based on historical frequencies is greatest at this time.

Overall, independent calibration performed better than joint calibration for both the RPSS and for the ROR score unless steps were take to avoid overfitting. The reason for this seems to be the shortness of the length of the historical record (22) relative to the number of models (7). Joint calibration estimates seven parameters simultaneously from the data while independent calibration estimates a single parameter at a time from the data. Even in the case of univariate linear regression, skill can be degraded when the sample size is small, particularly when the skill level is low (Tippett et al. 2005). Reducing the number of predictors or incorporating prior information about the weights were effective ways to prevent overfitting. On the other hand, joint calibration methods that did not restrict the weights had poorer skill than independent methods.

The conclusion and recommendation is that little or nothing is gained, and something could be lost, in attempting to use the most general joint calibration schemes for 7 models contributing to multi-model ensembles based on 20 to 25 years of model history. Rather, independent calibration, as well as joint calibration with overfitting prevention measures, can be used without skill sacrifice when forecasting outside of the training sample. It is assumed that when one or more models have skills that are clearly out of line with the others in a general and obvious ways, they should simply be removed from the ensembling exercise. However, unless the difference in skill is sufficiently large, removing the "worse" model may have the unintended consequence of degrading

skill (Kharin and Zwiers 2002; Kug et al. 2007). While some of the models used in this study appeared to have generally somewhat higher skills than others, none had unusually low skill.

*Acknowledgments.*

The authors thank Simon Mason and Andreas Weigel for their comments and suggestions. The authors are supported by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NA05OAR4311004). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies.

APPENDIX A

## A Gaussian model for forecast probabilities

To model forecast uncertainty using a Gaussian distribution, we first compute the correlation $r$ between the mean forecast and observations. The correlation corresponds to a signal-to-noise ratio $S$ given by

$$S = \frac{r^2}{1 - r^2} \,,$$

which is the ratio of the signal variance $\sigma_S^2$ to the noise variance $\sigma_N^2$. Therefore the noise variance is given by $\sigma_N^2 = \sigma_S^2 / S$ where we use the variance of previous mean forecasts as the signal variance. The mean of the forecast distribution is the mean forecast and its variance is $\sigma_N^2$. Category boundaries are computed using the climatological distribution whose mean is the long-term average forecast and whose variance is the sum of the signal and noise variances.

## APPENDIX B

### Dependence of RPSS on ensemble size

Richardson (2001) showed that the Brier score (BS) of a reliable forecast system depends on ensemble size $N$ according to

$$\langle \mathrm{BS}(N) \rangle = \left(1 + \frac{1}{N}\right) \langle \mathrm{BS}(\infty) \rangle \,, \tag{B1}$$

where $\langle \cdot \rangle$ denotes expectation over realizations of the observations. Tippett et al. (2007) generalized this result to RPS of reliable forecasts of tercile category probabilities. Remarkably, we show below that the same relation holds independent of the number of categories.

The RPS of a $m$-category forecast is

$$\mathrm{RPS} = \sum_{i=1}^{m} \left[ \sum_{j=1}^{i} P_j - O_j \right]^2 , \tag{B2}$$

where $P_i$ is the probability assigned to the $i$-th category and $O_i$ is one when the observation falls into the $i$-th category and zero otherwise. Assuming that the forecast is reliable means that the forecast probability is indeed the probability that an observation will fall into the $i$-th category. This assumption of reliability allows us to computed the expected value of the RPS as

$$\langle \mathrm{RPS} \rangle = \sum_{l=1}^{m} P_l \sum_{i=1}^{m} \left[ \sum_{j=1}^{i} P_j - \delta_{jl} \right]^2 , \tag{B3}$$

where the Kronecker delta $\delta_{ij}$ is defined to be one when $i = j$ and zero otherwise. Direct manipu-

lation of this expression gives

$$\begin{aligned}
\langle \mathrm{RPS} \rangle &= \sum_{l=1}^{m} P_l \sum_{i=1}^{m} \left[ \sum_{j=1}^{i} P_j - \delta_{jl} \right] \left[ \sum_{k=1}^{i} P_k - \delta_{kl} \right] \\
&= \sum_{l=1}^{m} \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{k=1}^{i} P_l (P_j - \delta_{jl})(P_k - \delta_{kl}) \\
&= \sum_{l=1}^{m} \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{k=1}^{i} P_l (P_j P_k - \delta_{kl} P_j - \delta_{jl} P_k + \delta_{jl}\delta_{kl}) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{k=1}^{i} P_j P_k - P_k P_j - P_j P_k + \delta_{jk} P_j \\
&= \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{k=1}^{i} \delta_{jk} P_j - P_j P_k \\
&= \sum_{i=1}^{m} \sum_{k=1}^{i} P_k - \left[ \sum_{k=1}^{i} P_k \right]^2 = \sum_{i=1}^{m} C_i (1 - C_i)
\end{aligned} \tag{B4}$$

where the cumulative probability is defined by

$$C_i \equiv \sum_{k=1}^{i} P_k \,. \tag{B5}$$

If the forecast probabilities come from an $N$-member ensemble, the forecast probability of the $i$-th category is $P_i + \epsilon_i$ where $\epsilon_i$ is the error in the forecast of the $i$-th category due to sampling variability. The expected RPS of the $N$-member ensemble is

$$\langle \mathrm{RPS}(N) \rangle = \sum_{l=1}^{m} P_l \sum_{i=1}^{m} \left[ \sum_{j=1}^{i} P_j + \epsilon_j - \delta_{jl} \right]^2 \,. \tag{B6}$$

The reliability of the forecast system means that the errors are unbiased, $\langle \epsilon_i \rangle = 0$ and that only quadratic error terms appear in $\langle \mathrm{RPS}(N) \rangle$. In particular, a direct calculation gives

$$\langle \mathrm{RPS}(N) \rangle = \langle \mathrm{RPS} \rangle + \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{k=1}^{i} \langle \epsilon_j \epsilon_k \rangle \,. \tag{B7}$$

Simplification gives

$$\begin{aligned}\langle \mathrm{RPS}(N) \rangle &= \langle \mathrm{RPS} \rangle + \sum_{i=1}^{m} \left\langle \left[ \sum_{j=1}^{i} \epsilon_j \right] \left[ \sum_{k=1}^{i} \epsilon_k \right] \right\rangle \\ &= \langle \mathrm{RPS} \rangle + \frac{1}{N} \sum_{i=1}^{m} C_i (1 - C_i) = \left( 1 + \frac{1}{N} \right) \langle \mathrm{RPS} \rangle \,,\end{aligned} \tag{B8}$$

where we use the fact that the cumulative probabilities estimated from the ensemble are binomially distributed. Equation (7) follows directly.

The result in (B8) can be interpreted as expressing the fact that the RPS of an $N$-member ensemble is the sum of the sampling variability of the observation and the ensemble. Equation (B8) can be used, assuming reliability, to estimate the infinite-ensemble skill given the finite-ensemble skill, similar to the debiased RPSS of Weigel et al. (2007). Although Weigel et al. (2007) use a different set of assumptions, their debiased RPSS agrees to leading order with the result here.

REFERENCES


Barnston, A. G., M. Chelliah, and S. B. Goldenberg, 1997: Documentation of a highly ENSO-related SST region in the equitorial Pacific. *Atmosphere-Ocean*, **35**, 367–383.

DelSole, T., 2004: Predictability and Information Theory Part I: Measures of Predictability. *J. Atmos. Sci.*, **61**, 2425–2440.

DelSole, T., 2007: A Bayesian framework for multimodel regression. *J. Climate*, **20**, 2810–2826.

DelSole, T. and M. K. Tippett, 2007: Predictability: Recent insights from information theory. *Rev. Geophys.*, in press.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987.

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.

Glantz, M. H., 2001: *Currents of Change: Impacts of El Niño and La Niña on Climate and Society*. Cambridge University Press.

Gneiting, T., A. 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.

Good, I. J., 1952: Rational decisions. *Journal of the Royal Statistical Society Ser. B*, **14**, 107–114.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus A*, **57 (3)**, 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

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

Kang, I.-S. and J. Yoo, 2006: Examination of multi-model ensemble seasonal prediction methods using a simple climate system. *Clim. Dyn.*, **26**, 285 – 294.

Kaplan, A., M. A. Cane, Y. Kushnir, A. C. Clement, M. B. Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856-1991. *J Geophys. Res.-Oceans*, **103 (C9)**, 18 567–18 589.

Kharin, V. V. and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate*, **15**, 793–799.

Kharin, V. V. and F. W. Zwiers, 2003: Improved seasonal probability forecasts. *J. Climate*, **16**, 1684–1701.

Kleeman, R., 2002: Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.*, **59**, 2057–2072.

Kousky, V. E. and R. W. Higgins, 2007: An alert classification system for montoring and assessing the enso cycle. *Weather and Forecasting*, **22**, 353–371.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford,

S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **286**, 1548–1550.

Kug, J. S., J. Lee, and I. Kang, 2007: Global sea surface temperature prediction using a multimodel ensemble. *Mon. Wea. Rev.*, **135**, 3239–33 247.

Mason, S. J. and L. Goddard, 2001: Probabilistic precipitation anomalies associated with ENSO. *Bull. Amer. Met. Soc.*, **82**, 619–638.

Mason, S. J. and G. M. Mimmack, 2002: Comparison of some statistical methods of probabilistic forecasting of ENSO. *Mon. Wea. Rev.*, **15**, 8–29.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.

Palmer, T., et al., 2004: Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). *Bull. Am. Meteor. Soc.*, **85**, 853–872.

Peng, P., A. Kumar, H. van den Dool, and A. G. Barnston, 2002: An analysis of multimodel ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.*, 18–1 to 18–12, doi:10.1029/2002JD002712.

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**, 1792–1811.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609–1625.

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.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.*, **132**, 2732–2744, doi: 10.1175/MWR2818.1.

Ropelewski, C. F. and M. S. Halpert, 1996: Quantifying Southern Oscillation-precipitation relationships. *J. Climate*, **9**, 1043–1059.

Roulston, M. S. and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660.

Shukla, J., et al., 2000: Dynamical seasonal prediction. *Bull. Am. Meteor. Soc.*, **81**, 2593–2606.

Stephenson, D. B., C. A. S. Coelho, F. J. Doblas-Reyes, and M. Malmaseda, 2005: Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus A*, **57**, 253–264.

Tang, Y., H. Lin, J. Derome, and M. K. Tippett, 2007: A predictability measure applied to seasonal predictions of the Arctic Oscillation. *J. Climate*, **20**, 4733–4750.

Tippett, M. K., A. G. Barnston, D. G. DeWitt, and R.-H. Zhang, 2005: Statistical correction of tropical Pacific sea surface temperature forecasts. *J. Climate*, **18**, 5141–5162.

Tippett, M. K., A. G. Barnston, and A. W. Robertson, 2007: Estimation of seasonal precipitation tercile-based categorical probabilities from ensembles. *J. Climate*, **20**, 2210–2228.

Tippett, M. K., R. Kleeman, and Y. Tang, 2004: Measuring the potential utility of seasonal climate predictions. *Geophys. Res. Lett.*, **31**, L22 201, doi:10.1029/2004GL021575.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124.

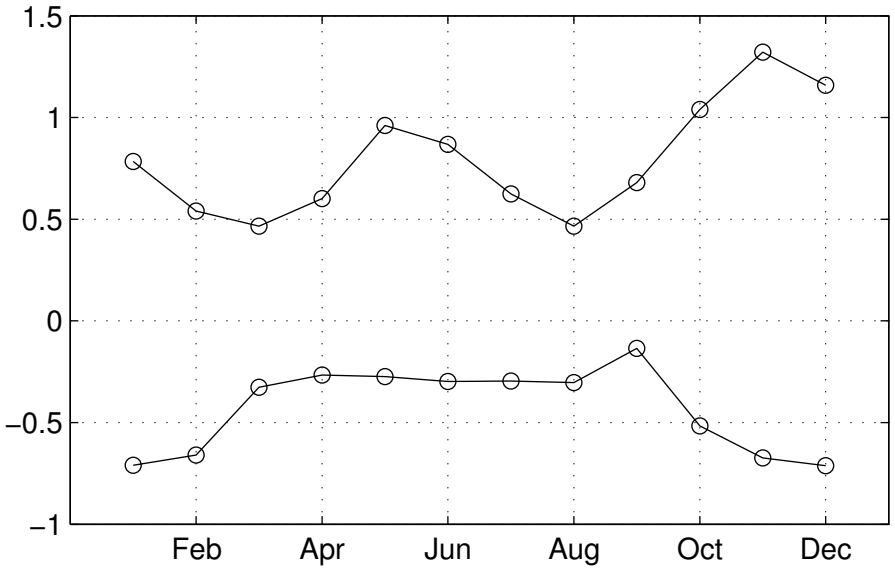
Wilks, D. S., 2006: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorol. Appl.*, **13**, 243–256.

Wilks, D. S. and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390.

Yun, W. T., L. Stefanova, and T. N. Krishnamurti, 2003: Improvement of the multimodel superensemble technique for seasonal forecasts. *J. Climate*, **16**, 3834–3840.

**List of Figures**

(a)




(b)

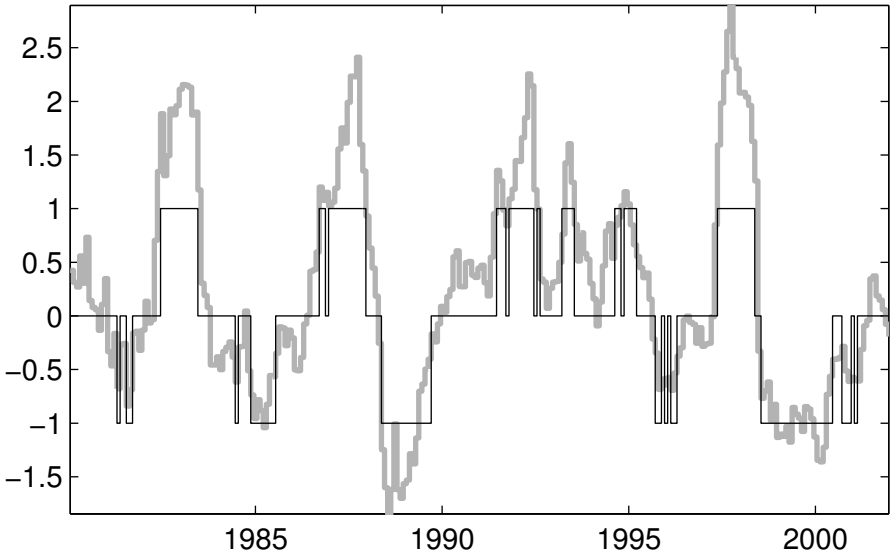
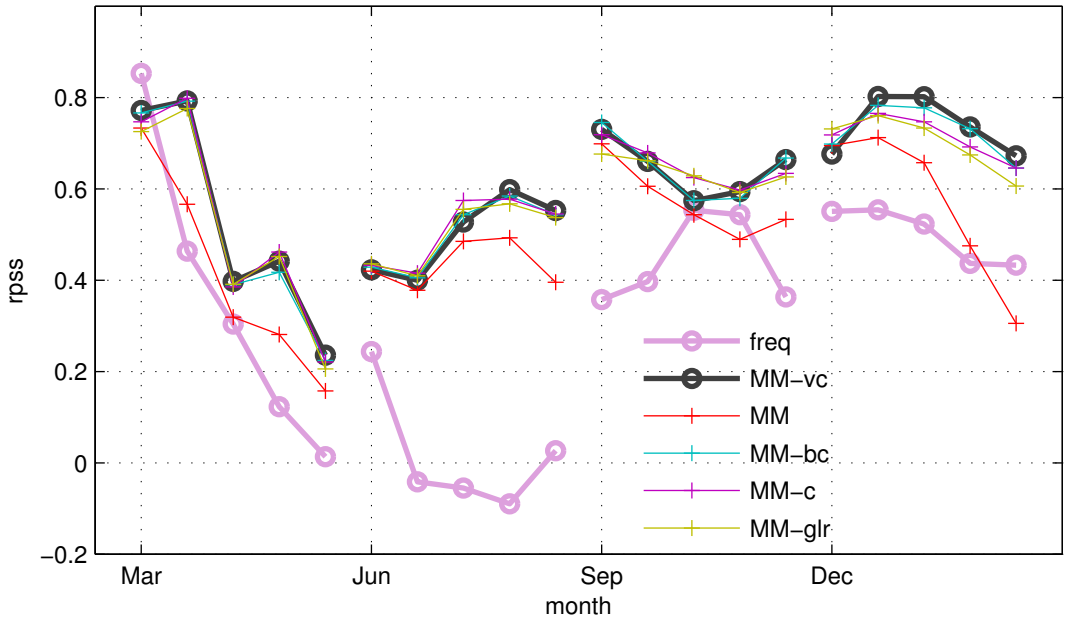



FIG. 1. (a) Upper and lower quartile boundaries in degrees of observed NINO3.4 anomalies as a function of calendar month for the period Jan 1980 - Dec 2001. (b) Time series of standardized NINO3.4 anomalies (gray) and ENSO category (black). Values 1,0, and -1 correspond to warm, neutral and cold conditions, respectively.

(a)




(b)

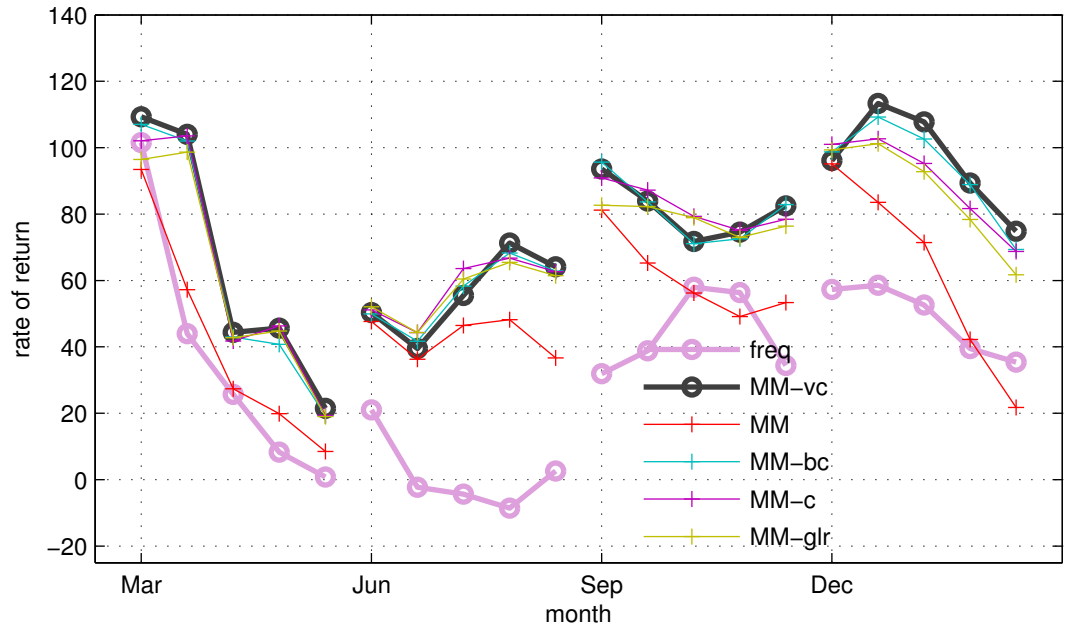
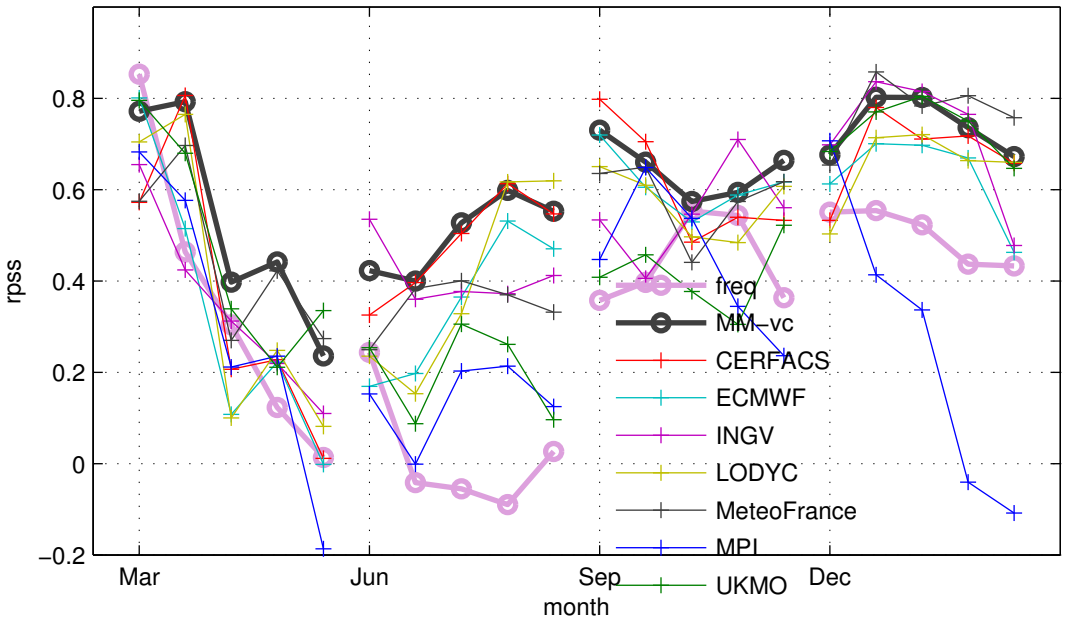



FIG. 2. (a) RPSS and (b) ROR for *uncalibrated* methods, i.e., ones that do not calibrate with respect to the observations. "Freq" refers to the conditional climatology reference forecast.

(a)




(b)

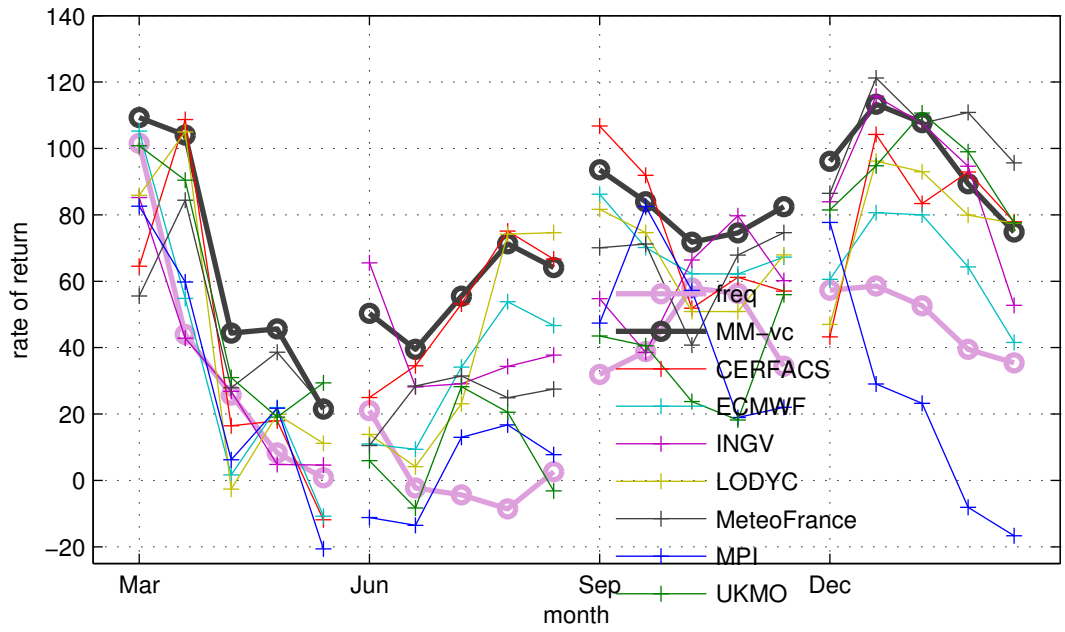
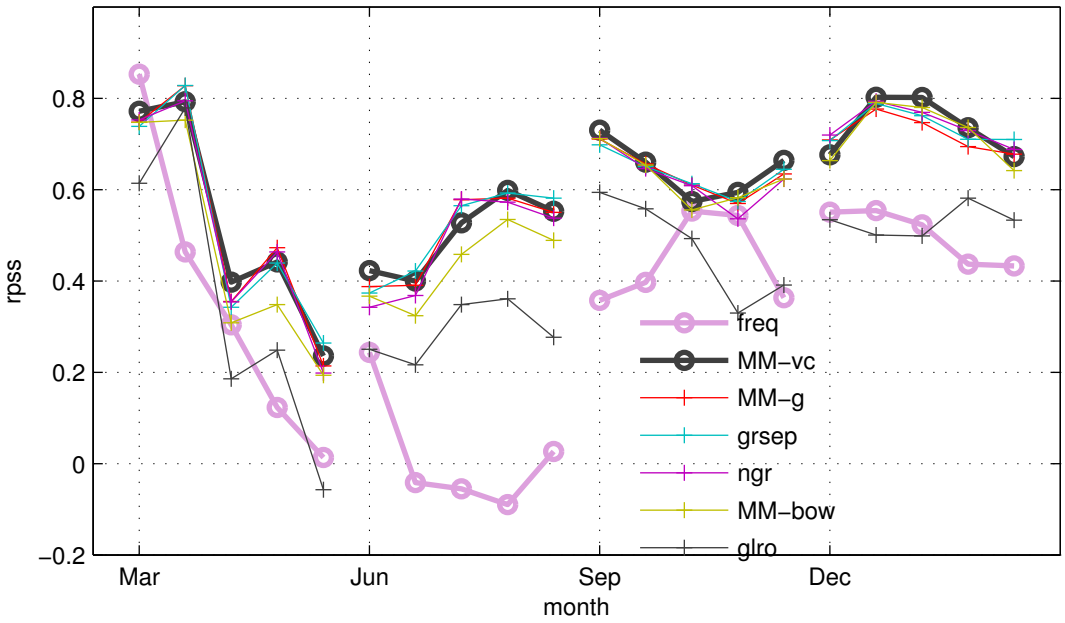



FIG. 3. As in Fig. 2 but for MM-vc, freq. and individual models.

(a)




(b)

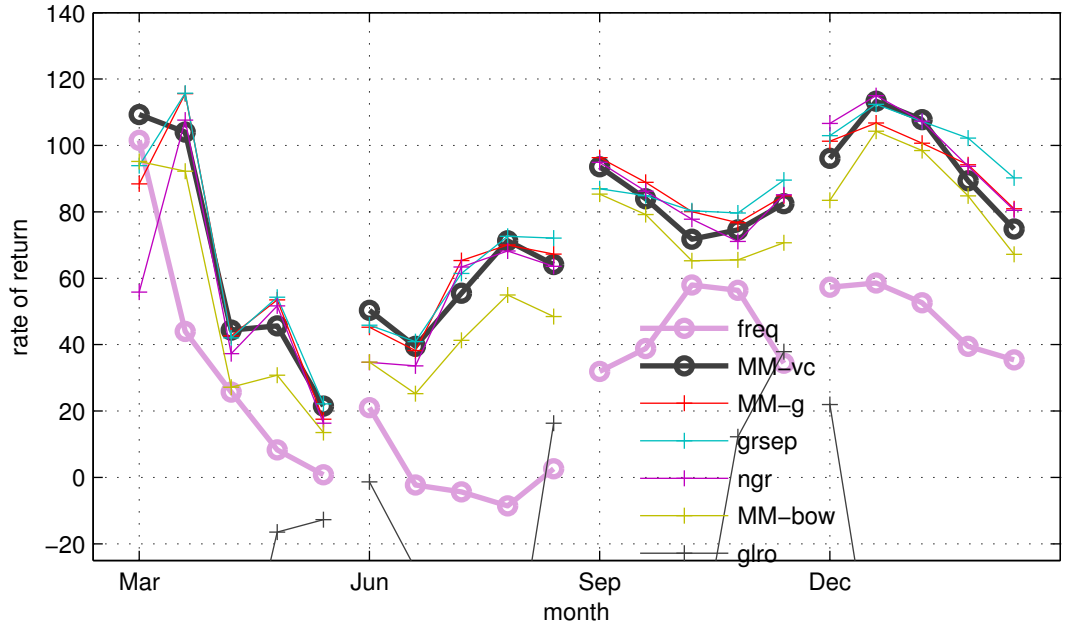
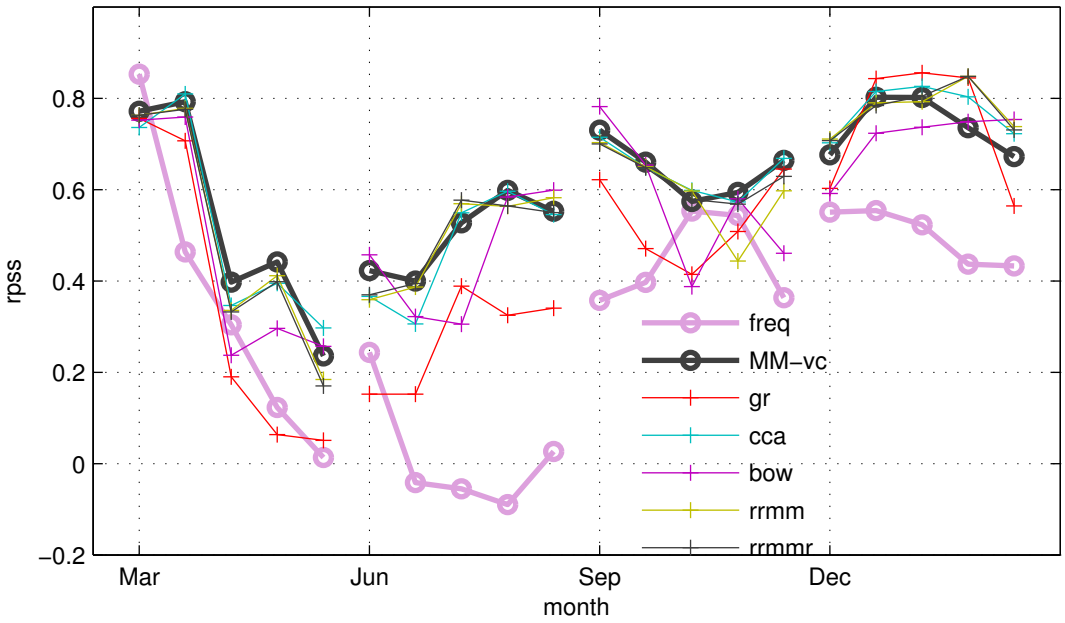



FIG. 4. As in Fig. 2 but for *independently calibrated* methods.

(a)




(b)

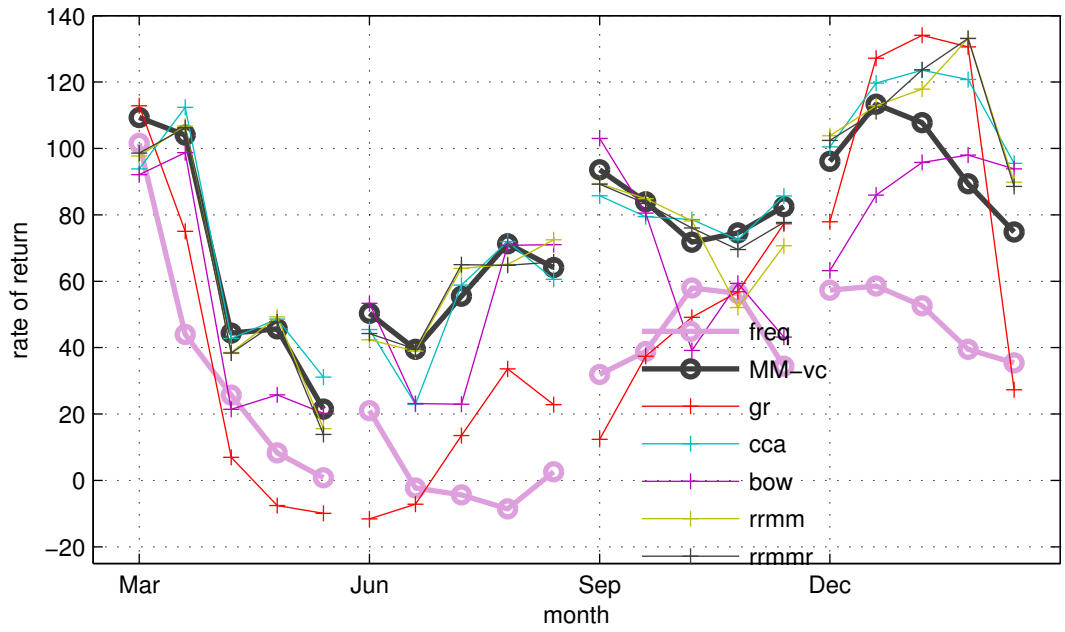



FIG. 5. As in Fig. 2 but for *jointly calibrated* methods.

**List of Tables**

| *Uncalibrated* | Compute probabilities from: | ROR | RPSS |
|---|---|---|---|
| MM-vc | standardized anomalies relative to each model's climatology | **72.8** | **0.603** |
| MM-bc | anomalies relative to each model's climatology | 71.5 | 0.599 |
| MM-c | multi-model mean plus constant historical ensemble | 71.4 | 0.599 |
| MM-glr | GLR using multi-model mean as predictor | 69.1 | 0.587 |
| MM | multi-model ensemble | 50.2 | 0.497 |
| *Independent calibration* | | ROR | RPSS |
| grsep | average separate regressions | **75.9** | **0.6** |
| MM-g | ensemble mean anomalies plus Gaussian | 73.8 | 0.595 |
| ngr | Non-homogeneous Gaussian regression | 70.3 | 0.59 |
| MM-bow | Bayesian optimal weighting of climatology and model means | 61.1 | 0.563 |
| glro | GLR between ensemble mean anomalies and observations | -64.9 | 0.412 |
| *Joint calibration* | | ROR | RPSS |
| cca | Two-mode CCA | **75** | **0.601** |
| rrmmr | Bayesian regression with unequal weights penalized | 74.3 | 0.595 |
| rrmm | Bayesian regression with weights different from 1/7 penalized | 73.5 | 0.59 |
| bow | Bayesian optimal weighting of each model probability | 60.3 | 0.55 |
| gr | Regression with all models | 40.5 | 0.475 |

TABLE 1. List of the methods and their key properties. Compound averaged ROR and average RPSS computed from all starts and leads. Methods in each class are listed in order of compound averaged ROR.