TECHNICAL MEMORANDUM

426

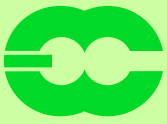
Skill of coupled model seasonal forecasts: A Bayesian assessment of ECMWF ENSO forecasts

C. A. S. Coelho [1], S. Pezzulli [1], M. Balmaseda, F. J. Doblas-Reyes and D. B. Stephenson [1]

Research Department

[1] Department of Meteorology, University of Reading

December 2003



European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

Abstract

This study assesses the skill of December Niño-3.4 forecasts produced by coupled model forecasting systems at ECMWF. A comparison has been made of the skill of several different ECMWF forecasting systems: the old operational system (System 1), the new operational system (System 2), and two other experiments (DEMETER assimilation and DEMETER control) produced as part of the DEMETER project. All forecasts started from initial conditions in August (5-month lead).

The skill is assessed relative to reference empirical persistence forecasts obtained by linear regression of December on the preceding July Niño-3.4 index values over the period 1950-2001. Rather than assess skill by comparing the coupled model forecasts to the empirical forecasts, this study assesses the additional skill obtained by combining the coupled and empirical forecasts. A simple Bayesian approach is used to combine and calibrate the coupled model forecasts using past information contained in the reference empirical forecasts.

Bayesian calibration is found to increase skill by around 11% compared to empirical persistence forecasts and around 3% compared to bias-corrected coupled model forecasts. Bayesian combined forecasts are better calibrated and provide more reliable estimates of forecast uncertainty than raw coupled model forecasts and empirical persistence reference forecasts. The inclusion of ensemble spread information into the calibration model improves neither forecast skill nor estimates of prediction uncertainty.

1 Introduction

Coupled model climate forecasts often contain substantial biases (e.g. Stockdale, 1997; Stockdale *et al.*, 1998). The simplest way to correct forecast biases is to subtract from the coupled model forecasts the mean forecast error. The mean forecast error is a constant that can be estimated from the difference in means of past forecasts and past observations. Such *bias-corrected* forecasts are routinely provided by climate prediction centres such as ECMWF. However, this kind of bias-correction does not correct biases in the ensemble variance nor does it fully exploit all the information contained in past observations.

Uncorrected biases in the coupled model forecasts can easily lead to overly pessimistic conclusions about the skill of the forecasting system. Forecast skill is traditionally assessed by comparing a chosen verification score S_M of the model forecasts to the score S_0 obtained for a reference forecast (Wilks, 1995; Jolliffe and Stephenson, 2003). The reference forecast is often chosen to be either the climatological mean or simple persistence forecasts based on preceeding observations. Reference forecasts are well-calibrated when based on unbiased estimates of past observations and so are generally well-calibrated (reliable). However, the lack of *reliability* (presence of conditional bias) in coupled model forecasts leads to an underestimate of the true skill (*resolution*) of the model forecasts. An alternative approach adopted in this study is to use the simple Bayesian method described in Coelho *et al.* (2004) to first calibrate and combine ECMWF coupled model forecasts with empirical persistence (reference) forecasts. The *additional skill* provided by the coupled model forecasting system is then assessed by comparing the verification score S_{M+0} of the combined forecasts with the score S_0 of the empirical reference forecasts. This cumulative approach has the advantage that skill is obtained by comparing the verification scores of two well-calibrated (unbiased) forecasting systems.

The study focuses on interval forecasts of the Niño-3.4 El Niño-Southern Oscillation (ENSO) index. The Niño-3.4 index is assumed to be normally distributed and therefore fully described by two parameters: the mean and the variance. The predicted mean and variance can be used to construct a *Prediction Interval* (P.I.) in which the future observed index is expected to be found with a given probability. We focus on 5-month lead forecasts of December mean Niño-3.4 index produced with initial conditions on the first of August. This lead time has been chosen for two reasons: a) the peak of Niño-3.4 index SST during ENSO is usually observed in December (Rasmusson and Carpenter, 1982); and b) August is after the spring barrier (Webster and Yang,

1992), when coupled and empirical (statistical) models have comparable level of skill (Oldenborgh *et al.*, 2003). If other months, such as April and August, had been chosen as predictor and predictand, respectively, different results would be obtained (Oldenborgh *et al.*, 2003). An assessment and comparison has been made of different versions of the ECMWF seasonal forecasting systems.

Sections 2 and 3 introduce the coupled model ensemble and empirical forecasts of the Niño-3.4 index used in this study. Section 4 briefly describes the Bayesian method used to combine and calibrate these forecasts, and section 5 presents and compares the results of the combined forecasts with bias-corrected and empirical (reference) forecasts. Section 6 concludes the article with a summary of our findings.

2 Coupled model ensemble forecasts of ENSO

ECMWF numerical model Niño-3.4 index forecasts were obtained from System 1 (SYS1), System 2 (SYS2) and two other experiments, DEMETER assimilation (DEMA) and DEMETER control (DEMC), as part of the Development of a European Multi-model Ensemble system for seasonal to inTERannual prediction (DEMETER) project[1] (Palmer *et al.* 2004). DEMA and DEMC experiments were both performed using the System 2 ECMWF coupled model with initial conditions from the ERA-40 project. SYS1 and SYS2 use initial conditions from the ERA-15 project for the period 1987-1993 and NWP operational data for the period 1993-2001.

Forecast	description	period	members
SYS1	old operational system	1987-2001	5
SYS2	new operational system	1987-2002	5 (40 in 2001 and in 2002)
DEMA	with ocean assimilation	1987-1999	9
DEMC	without ocean assimilation	1958-2001	9

Table 1: Coupled model forecasting systems/experiments investigated in this study.

Figure 1 shows the ECMWF coupled model ensemble forecasts for the seasonal forecast systems/experiments listed in Table 1. The forecasts have been bias corrected. In general, the interannual variability of the bias-corrected forecasts follows that of the observations. However, in several cases the observations lie outside the prediction interval given by the ensemble spread. In section 5 we will discuss quantitative comparisons of the skill and prediction uncertainty of the uncorrected coupled model, bias-corrected and empirical (reference) forecasts.

3 The reference forecast: July-December empirical prediction of Niño-3.4

Figure 2a shows the historical (1950-2001) July and December Niño-3.4 index time series obtained from Reynolds optimum interpolation version 2 SST dataset[2] (Reynolds *et al.*, 2002). The two time series are positively correlated ($r = 0.87$), illustrating the importance of persistence in predictability of the Niño-3.4 index.

[1] http://www.ecmwf.int/research/demeter/

[2] http://www.cpc.noaa.gov/data/indices/index.html

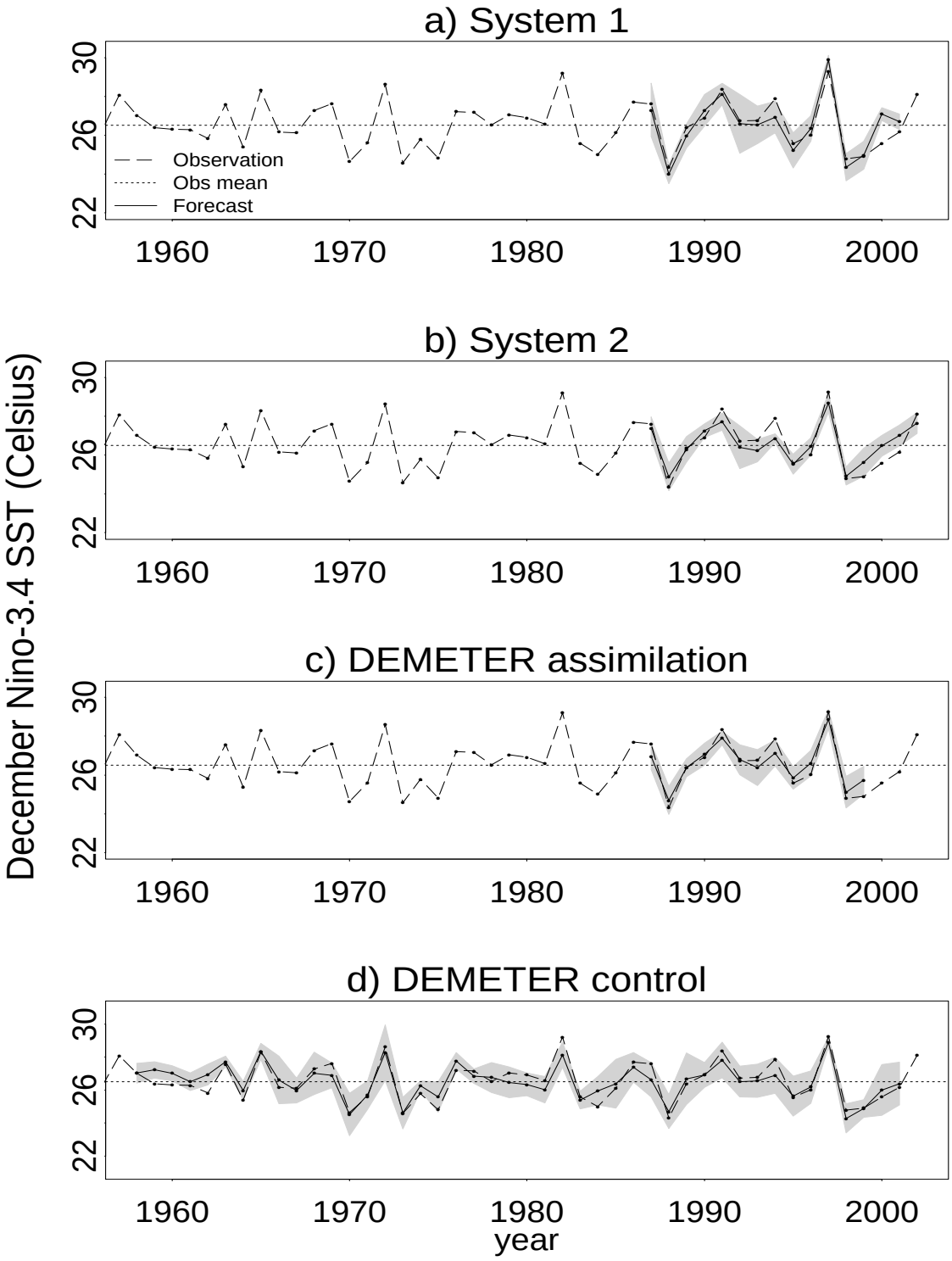


Figure 1: December Niño-3.4 index bias-corrected coupled model ensemble forecasts (^oC) from a) SYS1, b) SYS2, c) DEMA d) DEMC. Observed values (dashed line), forecasts (solid line) and the 95% prediction interval, given by the ensemble mean plus or minus 1.96 the standard deviation of the ensemble forecasts (s_X), is represented by the grey shading. The short-dashed line is the December 1950-2001 climatological mean (26.5 ^oC). Forecasts are given for the periods indicated in Table 1.

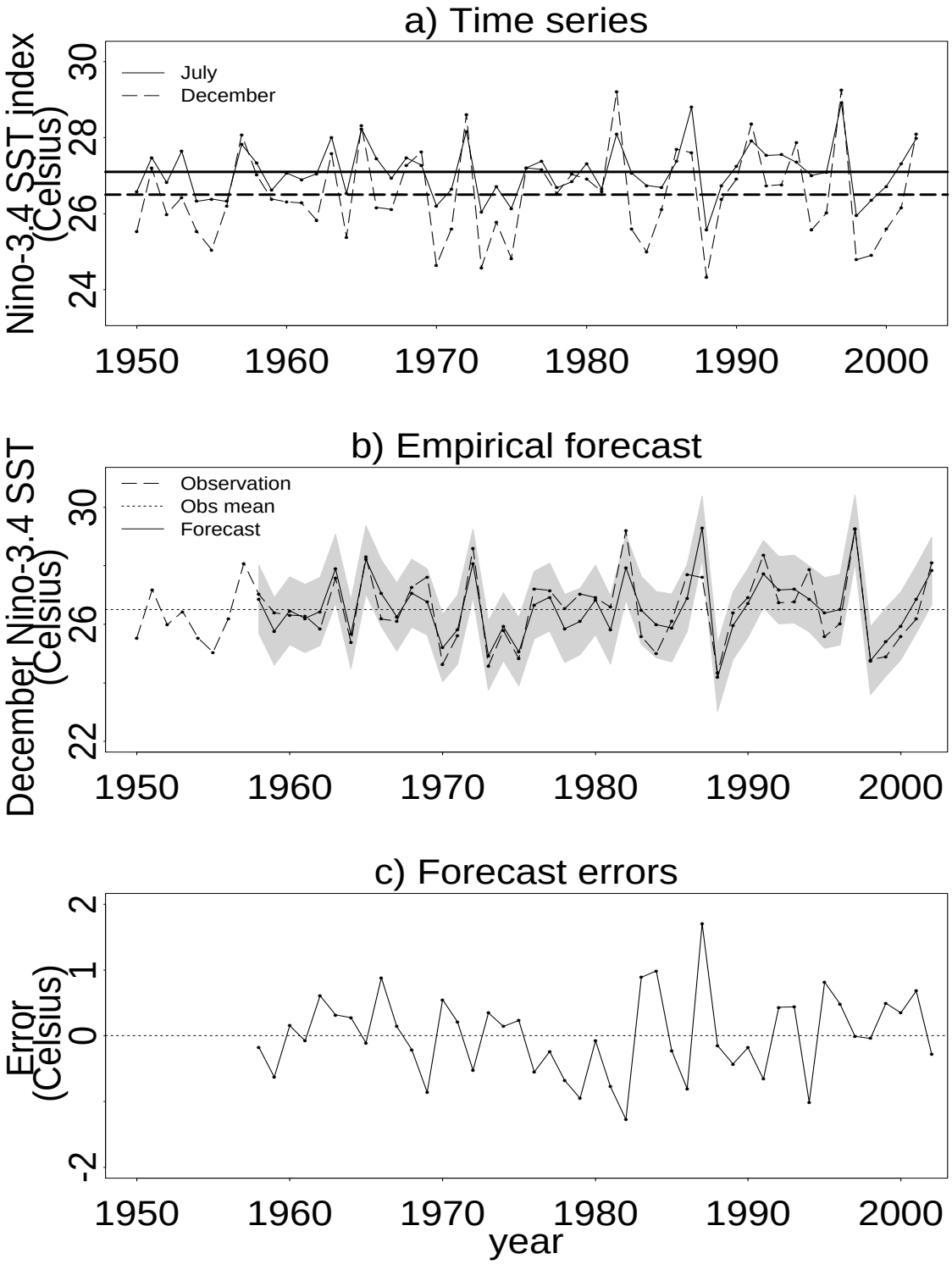


Figure 2: a) Observed July (solid line) and December (dashed line) Niño-3.4 time series (1950-2001) in ^oC. The horizontal thick solid line is the July climatological mean of 27.1^oC and the horizontal thick dashed line is the December climatological mean of 26.5 ^oC for this period; b) December 1958-2002 Niño-3.4 index empirical forecasts (^oC). Observed values (dashed line), forecasts (solid line) and the 95% prediction interval (grey shading). The short-dashed line is the December 1950-2001 climatological mean (26.5 ^oC). c) Forecast errors given by the difference between empirical forecast and observed values.

The largest El Niño (1972, 1982 and 1997) and La Niña (1970, 1973, 1988 and 1998) events can clearly be seen.

The simplest 5-month lead empirical model for forecasting December mean Niño-3.4 index uses linear regression with the preceding July mean Niño-3.4 index historical time series as the linear predictor. Mathematically, $\theta_t = \beta_o + \beta_1 \psi_t + \varepsilon_t$, where θ_t and ψ_t are the December and July Niño-3.4 monthly-mean values, respectively, β_o and β_1 are the intercept and slope parameters, respectively, ε_t is a "Normal (Gaussian)" random variable with zero mean and variance σ_o^2 [i.e., $\varepsilon_t \sim N(0, \sigma_o^2)$] and t is the year being forecast. This model, which will be used to produce reference forecasts, can be written explicitly in probabilistic notation as

$$\theta_t \,|\, \psi_t \sim N(\mu_{ot},\, {\sigma_o}^2) \tag{1}$$

with the predicted mean of θ_t given by

$$\mu_{ot} = \beta_o + \beta_1 \psi_t \tag{2}$$

which is a linear function of the predictor ψ_t. The symbol | denotes "given" and $\sim$ denotes "is distributed as". The Niño-3.4 index is known to be well approximated by the normal distribution and so the normal regression model is appropriate (Burgers and Stephenson, 1999; Hannachi *et al.*, 2003).

To avoid artificial skill, the empirical model parameters have been estimated using the cross-validation ("leave one year out") method (Wilks 1995, Section 6.3.6). To produce a forecast for time t, only data at other times (years) different than t have been used to estimate model parameters and errors.

Figure 2b shows empirical persistence forecasts for the period 1958-2002. The 95% P.I. is calculated using $\hat{\mu}_{ot} \pm 1.96\,\hat{\sigma}_{ot}$, where $\hat{\mu}_{ot} = \hat{\beta}_o + \hat{\beta}_1 \psi_t$ is the Niño-3.4 index predicted mean for a particular December and $\hat{\sigma}_{ot}$ is the predicted standard deviation estimated using Eqn. (4) of Coelho *et al.* (2004). This simple model shows surprisingly accurate results, especially for the 1988 and 1998 La Niña episodes and for the recent 1997 and 2002 El Niño episodes (Fig. 2c). Within the 45 years of December hindcasts the model has only forecast the Niño-3.4 index outside the 95% P.I. in 1982 and 1987. This is in agreement with the expected error rate of 5%. Persistence works well for the July to December forecasts that occur after the spring barrier (Oldenborgh *et al.*, 2003).

4 Bayesian calibration and combining of forecasts

If one had no access to an ensemble mean forecast $\bar{X}$, the only possible probabilistic assessment about the observable variable θ would have to be based solely on the assumption that future values of θ will behave like they did in the past. For example, the probability distribution of θ can be estimated by using the climatological probability density function p(θ) based on historical observations. In Bayesian theory p(θ) is known as the *prior distribution* and encapsulates *prior knowledge* about likely possible values of θ - from past experience not all values of θ are equally likely to occur.

However, when a particular ensemble mean forecast $\bar{X} = x$ is available, it is possible to update the prior p(θ) to obtain the (conditional) *posterior distribution* p($\theta|\bar{X} = x$). In other words, this is the probability distribution of θ given the forecast $\bar{X} = x$. Conditioning on forecasts helps to focus the uncertainty about future values of θ. The posterior distribution p($\theta|\bar{X} = x$) is found from the prior p(θ) by making use of Bayes' theorem

$$\overbrace{p(\theta_t|\bar{X}_t = x)}^{posterior} = \frac{\overbrace{p(\bar{X}_t = x|\theta_t)}^{likelihood}\overbrace{p(\theta_t)}^{prior}}{p(X_t = x)} \tag{3}$$

where θ_t is the observable variable at time t and x is the predicted ensemble mean forecast at time t. Note that both the posterior distribution and the likelihood function are considered to be functions of θ_t. Finally, $p(\bar{X}_t = x)$ does not depend on θ_t and therefore only plays the role of normalising constant.

In this study, we will use the empirical model forecasts as the prior. In other words, we assume the normal prior distribution $\theta_t \sim N(\mu_{o_t}, {\sigma_o}^2)$, where μ_{o_t} and ${\sigma_o}^2$ are provided by the empirical regression on preceeding July values (Eqns. 1 and 2).

The likelihood $p(\bar{X}|\theta)$ of obtaining an ensemble mean forecast $\bar{X}$ given an observed value θ is an essential ingredient in this updating procedure that can be estimated by regression of past ensemble-mean forecasts (hindcasts) on past observations. The likelihood provides a convenient summary of the calibration and resolution of past forecasts (Jolliffe and Stephenson, 2003). For more discussion of this method of calibration of forecasts see Coelho *et al.* (2004).

In this study, the likelihood is modelled by linear regression between ensemble mean forecasts ($\bar{X}_t$) and matching observations (θ_t). To demonstrate the method of calibration, instead of *bias − corrected* forecasts issued by ECMWF, *uncorrected* (raw) coupled model outputs will be used.

Two different likelihood models have been examined. One assumes constant variance for the ensemble mean and the other includes ensemble-spread information into the regression model. The likelihood model that assumes constant variance for the ensemble-mean is given by

$$\bar{X}_t \mid \theta_t \sim N(\alpha + \beta\theta_t, \delta) \tag{4}$$

where α and β are the intercept and slope parameters, respectively, and δ is the constant variance parameter. Figure 3 shows scatter plots between the coupled model ensemble forecasts and observed December Niño-3.4 index for the four ECMWF forecasting systems. The solid lines are the best fit linear regression between ensemble-mean values $\bar{X}_t$ and observations θ_t, from which α, β and δ are estimated. This figure reveals that all the forecasting systems are biased ($\hat{\alpha} \neq 0$ and $\hat{\beta} \neq 1$).

The likelihood model that incorporates ensemble spread information into the variance of the ensemble mean is given by

$$\bar{X}_t \mid \theta_t \sim N(\alpha + \beta\theta_t, \delta + \gamma V_t) \tag{5}$$

The estimation of the parameters α, β, δ and γ is done in three steps. In the first step the linear regression between coupled model forecasts and observed December Niño-3.4 index is performed. In the second step the intercept δ and the slope γ are obtained from the regression between the residuals of the regression of the first step and the sample variance of the ensemble mean V_t. The variance V_t is estimated using $V_t = s_X^2/m$, where s_X^2 is the variance of the ensemble (X) and m is the ensemble size given in Table 1. Finally, in the third step, the intercept α and the slope β are obtained from the weighted linear regression between ensemble mean forecasts ($\bar{X}_t$) and matching observations (θ_t), with regression weights given by $w_t = (\delta + \gamma V_t)^{-1}$.

To avoid introducing artificial skill, both prior and likelihood distribution parameters are estimated using the cross-validation method, in which parameters are obtained by leaving out the year being forecast. The mean values of the cross-validated likelihood estimated parameters for the likelihood models of Eqns. (4) and (5) for the common period (1987-99) among the four available forecasts are given in Tables 2 and 3. Values in brackets are standard errors. It can be noted that the coupled model ensemble-mean forecasts are biased ($\hat{\alpha} \neq 0$ and $\hat{\beta} \neq 1$). Note, however, that $\hat{\alpha}$ and $\hat{\beta}$ in Table 2 do not have exactly the same values as the estimates indicated in the legend of Fig. 3 because Table 2 estimates have been obtained using ensemble forecast data

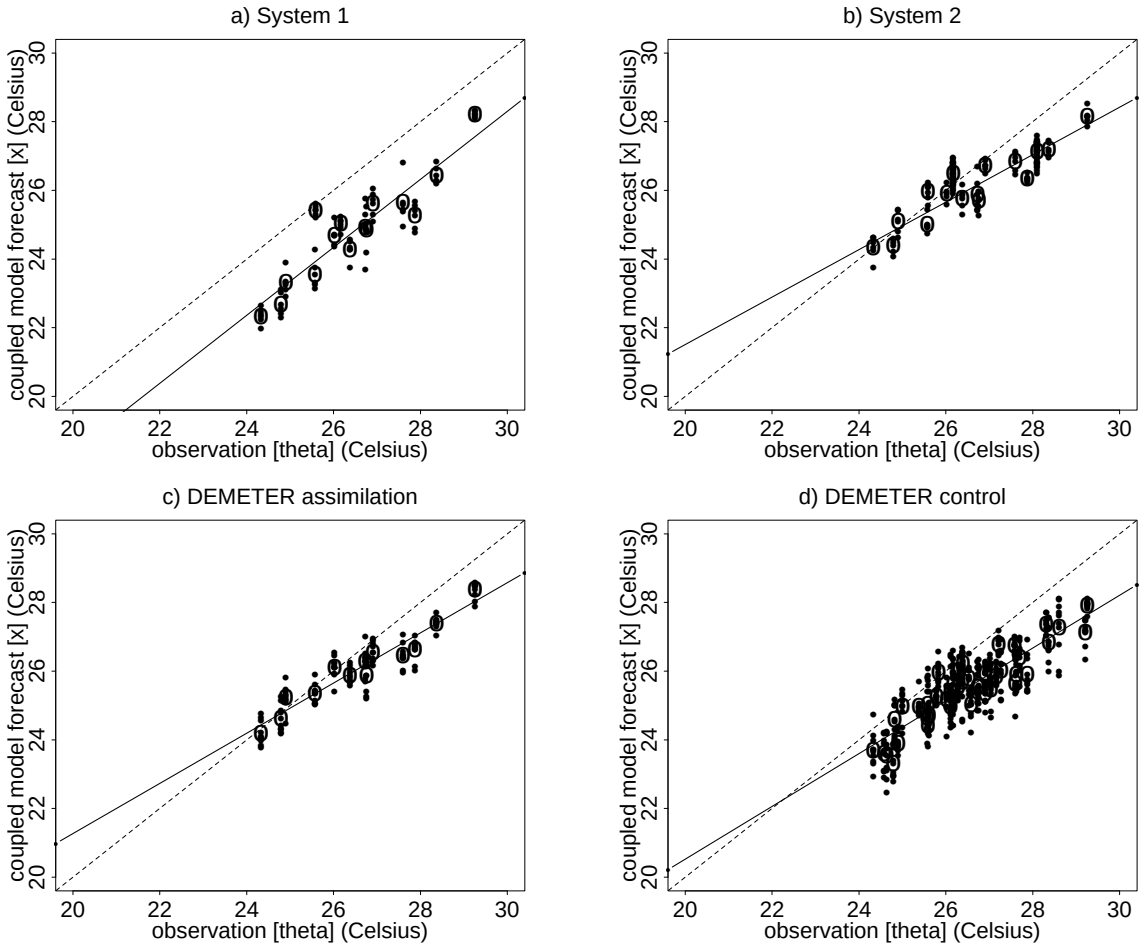


Figure 3: December Niño-3.4 index likelihood models assuming constant (δ) variance for the ensemble mean ($\bar{X}$). a) SYS1 ($\hat{\alpha} = -1.39^oC$, $\hat{\beta} = 0.99$ and $R^2 = 0.84$), b) SYS2 ($\hat{\alpha} = 7.69^oC$, $\hat{\beta} = 0.69$ and $R^2 = 0.86$), c) DEMA ($\hat{\alpha} = 6.65^oC$, $\hat{\beta} = 0.73$ and $R^2 = 0.93$) and d) DEMC ($\hat{\alpha} = 5.14^oC$, $\hat{\beta} = 0.77$ and $R^2 = 0.83$). Each black dot is one ensemble member. Big open circles are ensemble means. The solid lines are regressions between ensemble means and observations. The dashed lines is what would be obtained for perfect forecasts. The four likelihood models were constructed using ensemble forecasts for the periods indicated in Table 1.

for the common period (1987-99) while in Fig. 3 all the available data have been used to estimate these parameters. The estimated values in Tables 2 and 3 indicate that: a) SYS2, DEMA and DEMC underestimate the inter-annual variance of the observed Niño-3.4 index ($\hat{\beta} < 1$); b) all systems generally underestimate the mean SST in the Niño-3.4 region [solid lines generally below dashed lines in Fig. (3)]; and c) inclusion of ensemble spread does not provide additional information when forecasting Niño-3.4 index at 5-month lead ($\hat{\gamma}\overline{V}_t \approx 0$ in Table 3). This is in agreement with recent results by Jewson *et al.* (2003), who proposed an inverse regression calibration model of observed values on coupled model ensemble-mean forecasts.

System/experiment	$\hat{\alpha}$ [^oC]	$\hat{\beta}$	$\hat{\delta}$ [$^oC^2$]	R^2	m
SYS1	-3.22 (2.31)	1.05 (0.09)	0.21 (0.01)	0.94	5
SYS2	6.76 (1.85)	0.72 (0.07)	0.13 (0.01)	0.91	5
DEMA	6.69 (1.65)	0.73 (0.06)	0.11 (0.01)	0.93	9
DEMC	0.60 (2.56)	0.93 (0.10)	0.26 (0.01)	0.90	9

*Table 2: Mean values of cross-validated estimated likelihood parameters for the likelihood model which assumes constant variance for the ensemble mean (Eqn. 4), R^2 and number of members (*m*) for the common period 1987-99. Values in brackets are standard errors.*

System/experiment	$\hat{\alpha}$ [^oC]	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\delta}$ [$^oC^2$]	$\hat{\gamma}\overline{V}_t$ [$^oC^2$]	R^2	m
SYS1	-2.41 (2.33)	1.02 (0.09)	-1.77 (1.91)	0.22 (0.09)	-0.07 (0.32)	0.93	5
SYS2	7.18 (1.82)	0.71 (0.07)	-3.07 (1.94)	0.15 (0.05)	-0.06 (0.39)	0.91	5
DEMA	6.51 (1.56)	0.74 (0.06)	2.95 (3.62)	0.04 (0.05)	0.03 (0.33)	0.94	9
DEMC	0.51 (2.43)	0.93 (0.09)	1.07 (3.47)	0.15 (0.14)	0.03 (0.20)	0.91	9

*Table 3: Mean values of cross-validated estimated likelihood parameters for the likelihood model which incorporates ensemble spread information for the variance of the ensemble mean (Eqn. 5), estimate of the term $\hat{\gamma}\overline{V}_t$, R^2 and number of members (*m*) for the common period 1987-99. The variance $\overline{V}_t$ is the time mean of V_t. Values in brackets are standard errors.*

The calibration and combination of ECMWF coupled model forecasts with empirical persistence (reference) forecasts is determined using Bayes' theorem (Eqn. 3). From this theorem, it can be shown that for a normal prior and normal likelihood, the posterior distribution is also normal (e.g. Lee, 1997). The resulting normal posterior distribution in our case is given by

$$\theta_t \mid \bar{X}_t \sim N(\mu_t, \sigma_t^2) \tag{6}$$

with the mean μ_t and the variance σ_t^2 given by

$$\frac{1}{\sigma_t^2} = \frac{1}{{\sigma_{ot}}^2} + \frac{\beta^2}{\sigma_{Lt}^2} \tag{7}$$

$$\frac{\mu_t}{\sigma_t^2} = \frac{\mu_{ot}}{{\sigma_{ot}}^2} + \frac{\beta^2}{\sigma_{Lt}^2}\left(\frac{\bar{X}_t - \alpha}{\beta}\right) \tag{8}$$

where σ_{Lt}^2 is the variance of the likelihood, which can be either equal to δ or $\delta + \gamma V_t$ depending on the likelihood model that is used [Eqns. (4) or (5)]. Equations (6), (7) and (8) define the Bayesian normal-normal model for calibration and combination of forecasts (Coelho *et al.*, 2004). The calibration comes from the term in brackets in Eqn. (8), where the mean bias of the ensemble system is corrected when the difference between $\bar{X}_t$ and α is divided by the re-scaling factor β. The combination is due to the inclusion of the empirical persistence forecast

(μ_{ot}) in Eqn. (8). When the variance ${\sigma_{ot}}^2$ is imposed to be infinite in Eqns. (7) and (8) we have the so-called combined forecast with uniform prior (see definition below). More explanations about the Bayesian method are given in Coelho *et al.* (2004). The skill of calibrated forecasts produced using this approach will be compared to the skill of empirical persistence (reference) and bias-corrected forecasts in section 5.

Four different correction methods were applied to December Niño-3.4 index forecasts as described below:

a) *bias-corrected forecast* given by $\mu_t = \bar{X}_t - \overline{X} + \bar{\theta}$ and $\sigma_t = s_X$, where $\bar{X}_t$ is the ensemble mean forecast at time t, $\overline{X}$ and $\bar{\theta}$ are the time means of the ensemble mean forecast and the observed mean values over the forecast period, respectively, and s_X is the standard deviation of the ensemble forecasts (X). This represents a shift in the forecast values, which is constant in time, and no correction is applied to s_X. Basically the mean forecast bias is removed from the forecasts. This is a special case of a Bayesian forecast with uniform prior (defined below) and simplified likelihood estimate [$\beta = 1$ in Eqns. (4) and (5)]. The simplified likelihood models the bias of the ensemble mean as a constant $\alpha = \overline{X} - \bar{\theta}$ and the sample variance of the ensemble forecast as $\sigma_{Lt}^2 = s_X^2$.

b) combined forecast with *uniform prior* given by $\mu_t = \frac{\bar{X}_t - \alpha}{\beta}$ and $\sigma_t = \frac{\sigma_{Lt}}{\beta}$. It is obtained when ${\sigma_{ot}}^2$ is taken to be infinite in Eqns. (7) and (8), that is, *all* values of the index in the range $[-\infty, \infty]$ are equally likely. This prior characterises a "no-previous-information" reference case. The combined forecast with uniform prior can be seen as a Bayesian bias-correction in the ensemble mean and is useful for comparison with the bias-corrected forecast. Note, however, that $\sigma_t = \frac{\sigma_{Lt}}{\beta}$ is not the same as $\sigma_t = s_X$ of the bias-corrected forecast.

c) combined forecast with *climatological prior* given by $\mu_t = a + b\bar{X}_t$ and $\sigma_t = \lambda^{1/2}$, where a, b and λ are constant parameters estimated from the linear regression between observed values θ and coupled model ensemble-mean forecasts $\bar{X}$.

d) *combined forecast* given by μ_t and σ_t defined by Eqns. (7) and (8).

Forecast	Likelihood	Prior
a) bias-corrected	$\bar{X}_t \mid \theta_t \sim N(\alpha + \theta_t, s_X^2)$	Uniform (i.e. ${\sigma_{ot}}^2 \to \infty$)
b_1) uniform prior 1	$\bar{X}_t \mid \theta_t \sim N(\alpha + \beta\theta_t, \delta)$	Uniform (i.e. ${\sigma_{ot}}^2 \to \infty$)
b_2) uniform prior 2	$\bar{X}_t \mid \theta_t \sim N(\alpha + \beta\theta_t, \delta + \gamma V_t)$	Uniform (i.e. ${\sigma_{ot}}^2 \to \infty$)
c) climatological prior	$\bar{X}_t \mid \theta_t \sim N(\alpha + \beta\theta_t, \delta)$	$\theta_t \sim N(\theta_o, \sigma_o^2)$ (*)
d_1) combined 1	$\bar{X}_t \mid \theta_t \sim N(\alpha + \beta\theta_t, \delta)$	$\theta_t \sim N(\beta_o + \beta_1\psi_t, {\sigma_{ot}}^2)$
d_2) combined 2	$\bar{X}_t \mid \theta_t \sim N(\alpha + \beta\theta_t, \delta + \gamma V_t)$	$\theta_t \sim N(\beta_o + \beta_1\psi_t, {\sigma_{ot}}^2)$

Table 4: Likelihood and prior distributions for the correction methods applied to December Niño-3.4 index forecasts. The numbers 1 and 2 in front of uniform prior and combined forecasts are used to distinguish the likelihood model that has been used, i.e. either the model that assumes constant variance (δ) for the ensemble mean or the model that incorporates spread information into the variance of the ensemble mean ($\delta + \gamma V_t$). (*) θ_o and σ_o^2 are the climatological mean and the climatological variance of θ, respectively, which are obtained with the same dataset used to estimate the likelihood.

These four correction methods can all be seen as combined forecasts with particular likelihood and prior distributions as indicated in Table 4. Each likelihood model provides a different way for the calibration of coupled model forecasts against observations. The incorporation of previous knowledge is obtained by the use of the prior distribution. When the prior is taken to be uniform, no-previous-knowledge is incorporated in the correction method. Note that by using the likelihood model $\bar{X}_t \mid \theta_t \sim N(\alpha + \beta\theta_t, \delta)$ and the normal prior $\theta_t \sim N(\theta_o, \sigma_o^2)$, where θ_o and σ_o^2 are the climatological mean and the climatological variance of θ, respectively, which were obtained with the same dataset used to estimate the likelihood, one gets a posterior distribution that is exactly the same as the linear regression between observed values θ and coupled model ensemble-mean forecasts $\bar{X}$. This posterior distribution is given by the inverse regression model $\theta_t \mid \bar{X}_t \sim N(a + b\bar{X}_t, \lambda)$ that is here

refered to as combined forecast with *climatological prior* (forecast c) of Table 4). The proof of this equivalence can be found in Hoadley (1970).

5 Results

Figure 4 shows the mean of the combined forecast obtained using Eqns. (6), (7) and (8) and the likelihood model that assumes constant variance for the ensemble mean (Eqn. 4), i.e. forecast d_1 of Table 4. Comparisons of these forecasts with the empirical (reference) forecast alone (Fig. 2b) and coupled model ensemble forecasts alone (Fig. 1) show that the ensemble spread of the combined forecast is smaller than the uncertainty of the empirical forecast and larger than the ensemble spread of the coupled systems. Most of the observations now lie inside the 95% prediction interval, indicating that the combined forecast P.I. is more reliable, due both to incorporation of climatological information and calibration of the models. Results in Tables 5 and 6 (discussed below) indicate that the improved reliability of the combined forecast with respect to the coupled model does not come only from the widening of the ensemble spread, but also from a more skillful mean forecast.

Tables 5 and 6 show the root mean squared error (RMSE), the MSE skill score and the mean prediction uncertainty for the empirical persistence forecasts, the uncorrected ensemble forecasts, and the four different correction methods applied to December Niño-3.4 index forecasts as described in section 4 (Table 4). Results are shown separately for each of the four ECMWF ensemble forecasts detailed in Table 1.

Forecast	SYS1	SYS2	DEMA	DEMC
empirical	0.67 (75)	0.65 (77)	0.69 (76)	0.62 (74)
uncorrected	1.76 (-73)	0.77 (68)	0.69 (76)	1.11 (16)
bias-corrected	0.57 (82)	0.56 (83)	0.47 (89)	0.51 (82)
uniform prior 1	0.67 (75)	0.60 (80)	0.46 (89)	0.58 (77)
uniform prior 2	0.66 (75)	0.60 (81)	0.45 (90)	0.57 (78)
climatological prior	0.59 (81)	0.56 (83)	0.41 (92)	0.53 (83)
combined 1	0.56 (83)	0.52 (85)	0.39 (92)	0.45 (87)
combined 2	0.49 (86)	0.52 (85)	0.39 (93)	0.44 (87)

Table 5: RMSE in ^oC and MSE skill score in percentage (in brackets) of all forecast systems/experiments investigated for the empirical persistence forecasts, uncorrected coupled model forecasts, and for the four different correction methods applied to December Niño-3.4 index forecasts detailed in Table 4, i.e., bias-correction, Bayesian combined with uniform prior assumption, Bayesian combined with climatological prior and general Bayesian combined with empirical prior (combined). Values were obtained for the periods indicated in Table 1.

The MSE skill score in Table 5 is defined as $SS = 1 - (MSE/MSE_c)$, where MSE_c is the climatological MSE obtained from the historical (1950-2001) Niño-3.4 index December mean value ($\overline{\theta}$) of 26.5^oC. The mean prediction uncertainty in Table 6 is measured by the time mean of the predicted forecast standard deviations over the forecast period of each system/experiment as indicated in Table 1. The MSE skill score is a measure of how much better is a forecast than the climatological forecast. Empirical forecasts RMSE and skill scores (Table 5) do not have exactly the same values for all systems/experiments because forecast periods are different

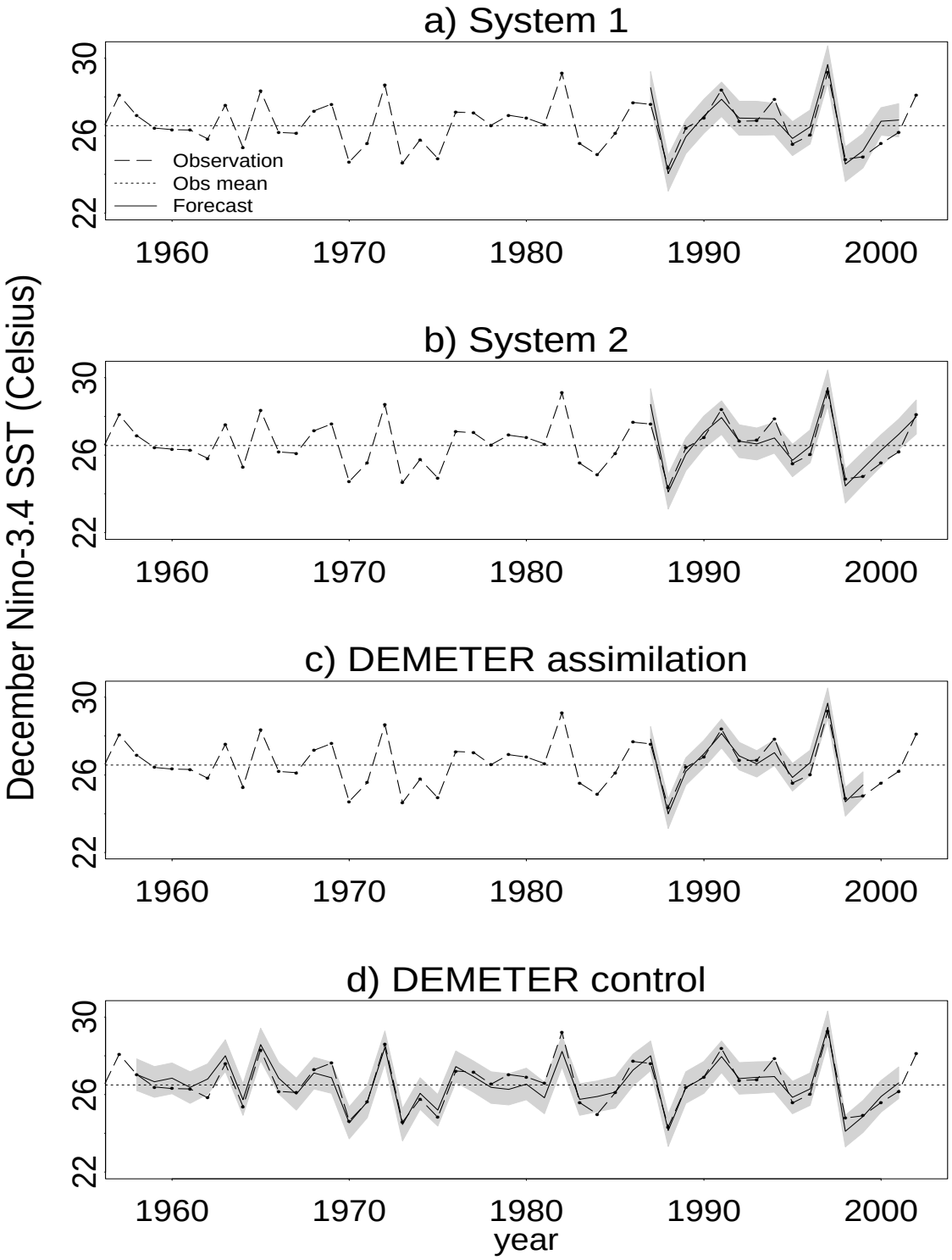


Figure 4: December Niño-3.4 index combined 1 forecast (°C) [d_1) of Table 4] for a) SYS1, b) SYS2, c) DEMA and d) DEMC. Observed values (dashed line), forecasts (solid line) and the 95% prediction interval (grey shading). The short-dashed line is the December 1950-2001 climatological mean (26.5 °C). Forecasts are given for the periods indicated in Table 1.

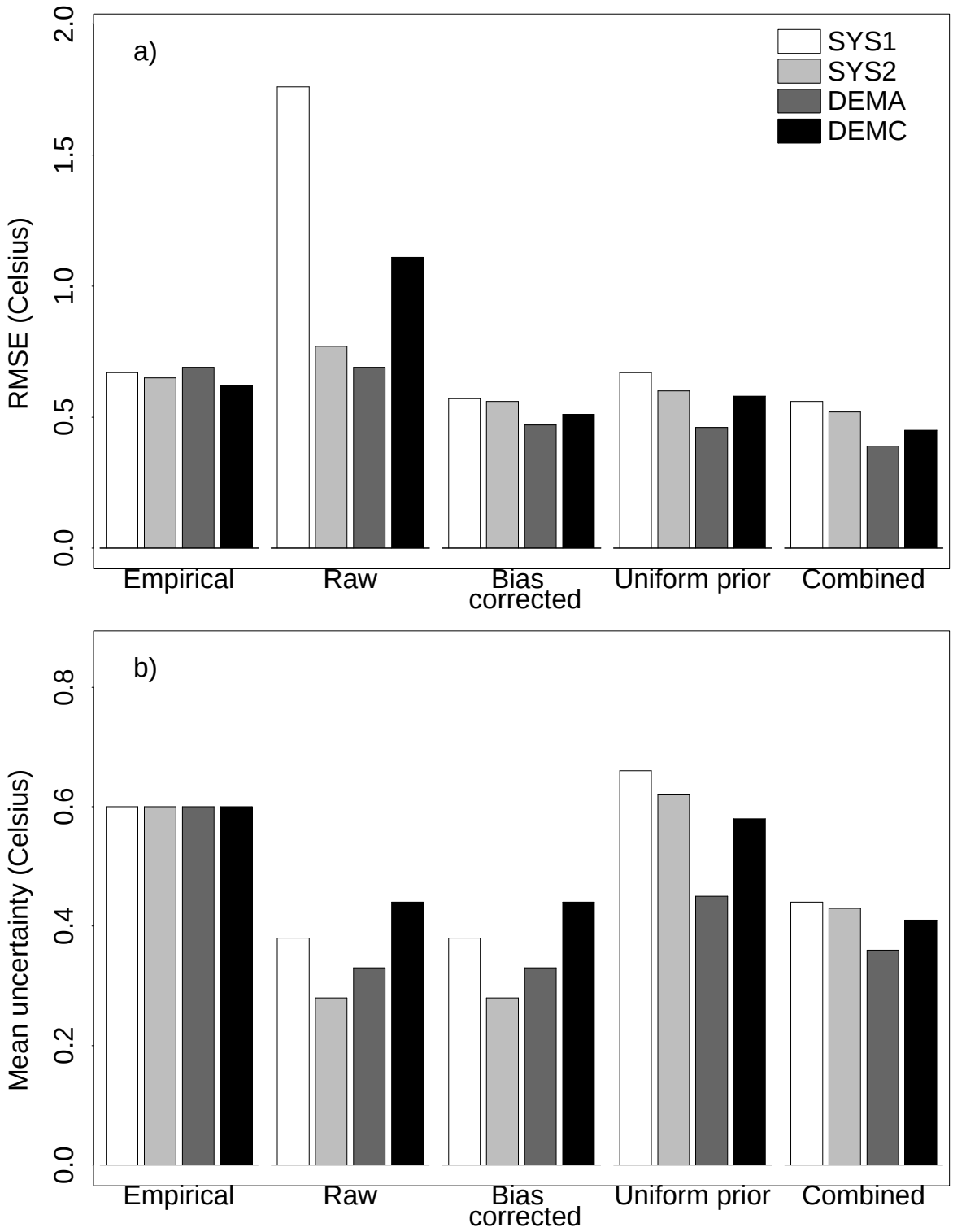


Figure 5: a) RMSE in ^{o}C and b) mean prediction uncertainty in ^{o}C of the 5-month lead December Niño-3.4 index forecasts for the empirical persistence, uncorrected coupled model, bias-corrected forecasts, Bayesian calibrated forecast with uniform prior assumption [b_1) uniform prior 1 of Table 4] and general Bayesian combined with empirical prior [d_1) combined 1], which use the likelihood model of Eqn (4). Note that the different systems/experiments were evaluated over different periods given in Table 1.

Forecast	SYS1	SYS2	DEMA	DEMC
empirical	0.60	0.60	0.60	0.60
uncorrected	0.38	0.28	0.33	0.44
bias-corrected	0.38	0.28	0.33	0.44
uniform prior 1	0.66	0.62	0.45	0.58
uniform prior 2	0.64	0.58	0.38	0.54
climatological prior	0.60	0.56	0.43	0.53
combined 1	0.44	0.43	0.36	0.41
combined 2	0.43	0.42	0.32	0.39

Table 6: Mean prediction uncertainty in ^oC of all forecast systems/experiments investigated for the empirical persistence forecasts, uncorrected coupled model forecasts, and for the four different correction methods applied to December Niño-3.4 index forecasts detailed in Table 4, i.e., bias-correction, Bayesian combined with uniform prior assumption, Bayesian combined with climatological prior and general Bayesian combined with empirical prior (combined). Values were obtained for the periods indicated in Table 1

for the different systems/experiments (Table 1). However, by coincidence the mean uncertainty of the empirical forecasts have the same value, although the periods are different (Table 6). Uniform prior 1 and uniform prior 2 are expected to have the same RMSE and MSE skill score because they use the same prior and the same model for the ensemble-mean in the likelihood (Table 4). However, Table 5 shows that the RMSE and the MSE skill score of uniform prior 1 and uniform prior 2 forecasts are very similar but not exactly the same. These slight differences are due to sampling errors in the estimation of model parameters.

Table 5 and Fig. 5a show that uncorrected coupled model forecasts generally have the largest forecast RMSE and the poorest skill. The empirical (reference) forecasts generally outperform uncorrected forecasts. However, bias-corrected coupled model forecasts generally outperform the empirical, uncorrected, uniform prior 1 and uniform prior 2 forecasts. Bias-corrected, uniform prior 1 and uniform prior 2 forecasts incorporate no-previous-information, i.e. both assume infinite variance for the prior distribution ($\sigma_{ot}{}^2 \rightarrow \infty$ as indicated in Table 4). This means that the resulting forecast improvement is due to the calibration, which is given by the likelihood. The fact that bias-corrected forecasts generally outperform uniform prior forecasts suggests that the calibration used to produce bias-corrected forecasts, which does not have the scaling factor β, is better than the calibration used to produce uniform prior forecasts. Note, however, that for DEMA this is not the case. DEMA uniform prior forecasts, which are calibrated using the scaling factor β, have smaller RMSE than bias-corected forecasts. Combined forecasts with climatological prior [forecast c) of Table 4], which use a similar likelihood model to uniform prior forecasts but a more informative prior given by a normal distribution with mean and variance obtained from the same historical values of θ used to build the likelihood, outperform uniform prior forecasts and have comparable RMSE and MSE skill score to bias-corrected forecasts. This indicates that the use of a better prior helped to reduce forecast error. Bayesian combined 1 and combined 2 forecasts, which use a similar likelihood model to climatological prior and uniform prior forecasts and a more refined prior, generally outperform all other forecasts. This indicates that these forecasts are better calibrated due to the use of a better and more informative prior than those used in the other forecasts. In order to check whether or not the scaling factor β is important for the calibration, combined forecasts using the model $\overline{X_t} \mid \theta_t \sim N(\alpha + \theta_t, \delta)$ for the likelihood and $\theta_t \sim N(\beta_o + \beta_1 \psi_t, \sigma_{ot}{}^2)$ for the prior, have been performed for all the systems here analysed and their results compared to the results of the combined 1 and combined 2 forecasts of Table 5. Very similar results to the combined 1 and combined 2 forecasts of Table 5 have been found, indicating that the major contribution for the combination is from the prior.

Table 5 shows that the Bayesian the combined 1 forecasts improve the skill of bias-corrected ECMWF coupled model forecasts by approximately 3% (on average among all forecasting systems) and the skill of empirical

persistence forecasts by approximately 11%. The incorporation of ensemble spread information [i.e. the inclusion of the term (γV_t) in the likelihood model of combined 2] does not help to substantially increase forecast skill. The little additional skill of 3% for SYS1 and 1% for DEMA obtained with the inclusion of γV_t into the calibration model (i.e. combined 2) compared to the skill of the model that assumes δ equals constant variance for the ensemble mean (i.e. combined 1) was perhaps achieved by the estimation of the additional parameter γ. As shown in Table 3 the term $\hat{\gamma}\overline{V}_t$ is nearly zero, which suggests that this little additional skill may just be an artifact of overfitting the calibration model. Table 6 shows a measure of uncertainty - given by the ensemble spread - for all forecast/calibration methods here investigated. Reliable forecasts are expected to have uncertainty estimates as close to the forecast RMSE as possible. Uncorrected and bias-corrected forecasts have uncertainty estimates too small compared to the forecast RMSE, indicating that the coupled model forecasts are overconfident and not reliable. The calibration achieves forecast reliability by increasing the uncertainty to match the level of RMSE, as can be seen for uniform prior 1, uniform prior 2 and climatological prior forecasts. For non-uniform prior forecasts (i.e., combined 1 and combined 2) both RMSE and uncertainty are reduced as expected from the Bayesian combination. Compared to the uncorrected and bias-corrected forecasts, the Bayesian combination has narrowed the gap between RMSE and uncertainty, i.e. it has improved the forecast reliability. Note, however, that combined 1 and combined 2 forecasts are still overconfident, with mean uncertainty estimates less than the forecast RMSE.

Table 6 and Fig. 5b show that coupled model forecasts give the smallest forecast uncertainty estimates. Both uncorrected and bias-corrected mean prediction uncertainties have exactly the same values because bias-correction does not correct biases in the the ensemble variance. Empirical, uniform prior and combined with climatological prior forecasts give larger uncertainty estimates than uncorrected and bias-corrected forecasts. Combined 1 and combined 2 forecasts generally give intermediate uncertainty estimates between the values of uncorrected and empirical forecasts. However, the inclusion of spread information (V_t) does not substantially alter the mean prediction uncertainty of Bayesian calibrated forecasts. Note the similarity between the mean prediction uncertainty for the uniform prior 1 and uniform prior 2 and also for the combined 1 and combined 2 forecasts when the two likelihood models (assuming the variance of the ensemble mean as δ or $\delta + \gamma V_t$) have been used.

6 Conclusions

The skill has been assessed for 5-month lead December Niño-3.4 forecasts produced by different coupled model systems at ECMWF. All four forecasting systems were found to be able to reproduce inter-annual variations in the December Niño-3.4 index five months in advance. The DEMA system gives the most accurate forecasts with the smallest RMSE (0.69^oC) and the highest MSE skill score (76%). It is worth noticing that the DEMA system uses the same coupled model as SYS2. The differences between the two systems lie a) on the initial conditions (DEMA make use of ERA40 data), and b) the period for which they are validated.

For the case chosen in this study, the empirical forecast (based on a regression model) has comparable skill to that of the coupled models. Empirical regression forecasts are unbiased, present reliable uncertainty estimates and therefore can be used as good reference forecasts. However, it should be remembered that the empirical forecasts of December Niño-3.4 index using the previous July index as predictor had high skill because these two chosen months determine a particularly favourable forecast scenario. July is after the spring barrier when ENSO persists more (spring barrier phenomenon). If other months had been chosen as predictor and predictand, such as April (predictor) and August (predictand), forecasts based on persistence would have been worse (Oldenborgh *et al.*, 2003).

Instead of comparing the skill of coupled model forecasts with the skill of empirical calibrated (reference) forecasts, our strategy here has been to use a simple Bayesian method to first calibrate the forecasts and then

combine these two forecasts. The skill of the resulting forecast can be used to assess how much *additional skill* the calibrated combined forecasts can provide compared to the reference forecasts. The combined forecasts have an average skill (over all the forecasting systems) of 87% which is 11% more than the 76% average skill of the empirical forecasts and 3% more than the 84% average skill of the bias-corrected coupled model forecasts.

In addition to being more skillful, combined forecasts are more reliable than bias-corrected coupled model forecasts. All the coupled models are overconfident because they underpredict forecast uncertainty (i.e., their ensemble spread is too narrow). The empirical forecast is reliable (by construction) at the expense of quite large prediction uncertainty. The Bayesian combination offers a more reliable product than bias-corrected coupled model forecasts, and also a sharper forecast (i.e. with reduced uncertainty) than empirical model forecasts.

Two different Bayesian models have been tested: one that assumes constant variance for the ensemble mean, and another that incorporates ensemble spread information into the variance of the ensemble mean. It has been found that the inclusion of the ensemble spread into the variance of the ensemble mean provides little additional information, in agreement with Jewson *et al.* (2003). Skill and prediction uncertainty estimates are not altered by the inclusion of ensemble spread information into the calibration model.

The Bayesian method is a powerful tool for the calibration of coupled model seasonal forecasts - the ensemble mean and ensemble spread biases can easily be corrected with the Bayesian approach. Furthermore, the method allows one to include useful past (historical) information - not available in the short calibration period. In addition, the Bayesian method facilitates the treatment of flow dependent forecast uncertainty. However, it is too early to fully exploit this capability with state of the art seasonal forecasting sytems because coupled model ensemble spread still provides little useful information.

In this paper, each of the coupled models has been calibrated individually, and each coupled model has been combined with the empirical forecasts separately. This is just the first step towards a more generalized usage of the Bayesian method, where information from different models can be used simultaneously. Work is underway to apply the method to a multi-model forecast system.

Acknowledgements

We wish to thank Dr. D. L. T. Anderson, head of the seasonal forecast section at ECMWF and Dr. T. N. Palmer, the DEMETER (EVK2-1999-00024) project principal investigator, who kindly provided the ECMWF coupled model hindcasts used in this research. Thanks are also due to Matt Sapiano for his help with figures. CASC was sponsored by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) precess 200826/00-0. FJDR was supported by DEMETER.

REFERENCES

Burgers, G. and D. B. Stephenson, 1999: The "Normality" of El Niño. *Geophysical Research Letters*, **26**(8), 1027-1030.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes and D. B. Stephenson, 2004: Forecast calibration and combination: A simple Bayesian approach for ENSO. *J. Climate*, **17**, 1504-1516.

Hannachi, A., D. B. Stephenson and K. R. Sperber, 2003: Probability-based methods for quantifying nonlinearity in the ENSO. *Clim. Dynamics*, **20**, 241-256.

Hoadley, B., 1970: A Bayesian look at inverse linear regression. *J. Amer. Statist. Ass.*, **65**, 356-369.

Jewson, S., F. J. Doblas-Reyes and R. Hagedorn, 2003: The assessment and calibration of ensemble seasonal forecasts of equatorial pacific ocean temperature and the predictability of uncertainty. *Atmos. Sci. Letters*, Submitted.

Jolliffe, I. N. and D. B. Stephenson, 2003: Forecast Verification. A Practioner's Guide in Atmospheric Aciences. Wiley and Sons. 240 pp.

Lee, P. M., 1997: Bayesian statistics: an introduction. Arnold. Second edition. 344 pp.

Palmer, T. N. and co-authors, 2004: Development of a European Ensemble System for Seasonal to Inter-annual Prediction (DEMETER). *Bull. Am. Meteorol. Soc.*, (in press).

Rasmusson, E. M. and T. H. Carpenter, 1982: Variations in tropical sea surface temperature and surface wind fields associated with the Southern Oscillation/El Niño. *Mon. Wea. Rev.*, **109**, 1163-1168.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stockes and W. Wang,2002: An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**(13), 1609-1625.

Stockdale, T. N., 1997: Coupled ocean-atmosphere forecasts in the presence of climate drift. *Mon. Wea. Rev.*, **125**, 809-818.

Stockdale, T. N., D. L. T. Anderson, J. O. S. Alves and M. A. Balmaseda, 1998: Global seasonal rainfall forecasts using a coupled ocean-atmosphere model. *Nature*, **392**, 370-373.

van Oldenborgh, G. J., M. A. Balmaseda, L. Ferranti, T. N. Stockdale and D. L. T. Anderson, 2003: Did the ECMWF seasonal forecast model outperform a statistical model over the last 15 years?, *ECMWF Technical Memorandum*, **418**.

Webster, P. J. and S. Yang, 1992: Monsoon and ENSO: Selectively interactive systems. *Q. J. R. Meteorol. Soc.*, **118**, 877-926.

Wilks, D. S., 1995: Statistical methods in the atmospheric sciences: An introduction. Academic Press. First edition. 467 pp.