Forecast calibration and combination:

A simple Bayesian approach for ENSO

C. A. S. Coelho[1], S. Pezzulli, M. Balmaseda (*), F. J. Doblas-Reyes (*)

and D. B. Stephenson

Department of Meteorology, University of Reading

(*) European Centre for Medium-Range Weather Forecasts - ECMWF

Journal of Climate (in press)

January 19, 2004

[1]C. A. S. Coelho, Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading RG6 6BB, U. K., E-mail: c.a.d.s.coelho@reading.ac.uk

ABSTRACT

This study presents a new simple approach for combining empirical with raw (i.e. not bias corrected) coupled model ensemble forecasts in order to make more skillful interval forecasts of ENSO. A Bayesian normal model has been used to combine empirical and raw coupled model December SST Niño-3.4 index forecasts started at the end of the preceding July (5-month lead time). The empirical forecasts were obtained by linear regression between December and the preceding July Niño-3.4 index values over the period 1950-2001. Coupled model ensemble forecasts for the period 1987-99 were provided by ECMWF, as part of the Development of a European Multi-model Ensemble system for seasonal to inTERannual prediction (DEMETER) project. Empirical and raw coupled model ensemble forecasts alone have similar mean absolute error forecast skill score, compared to climatological forecasts, of around 50% over the period 1987-1999. The combined forecast gives an increased skill score of 74% and provides a well-calibrated and reliable estimate of forecast uncertainty.

Key words: ENSO, Bayesian, combination, calibration, empirical, coupled model ensemble, probabilistic interval forecast, skill

1. Introduction

The El Niño-Southern Oscillation (ENSO) is an important large-scale ocean-atmosphere coupled phenomenon that has large impacts on the climate of many regions around the world (Horel and Wallace, 1981; Stoeckenius, 1981; Ropelewski and Halpert, 1986, 1987 and 1989). Since the strong El Niño episode in 1982/3, many efforts have been made to produce routine forecasts of tropical Pacific sea surface temperatures (SST). Long-lead forecasts several months in advance help local governments and industries plan their actions prior to the occurrence of the phenomenon (Patt, 2000).

ENSO forecasts are currently produced using either physically-derived dynamical climate models or empirical (statistical) relationships based on historical data. For a comprehensive review of ENSO forecasting studies developed during the last two decades see Mason and Mimmack (2002). The comparative skill of these two approaches is a subject of much debate (Berliner *et al.*, 2000b). Recent forecast comparisons suggest that empirical models perform at least as well as dynamical coupled models (Barnston *et al.*, 1999; Anderson *et al.*, 1999). Some studies argue that empirical models perform better (e.g. Landsea and Knaff, 2000), while other studies claim that dynamical climate models can give better ENSO forecasts (e.g. Trenberth, 1998).

For both medium-range and seasonal forecasts, it is common practice to use the ensemble technique to cope with the probabilistic nature of the forecasts

(e.g. Stockdale *et al.*, 1998; Taylor and Buizza, 2003 and Palmer *et al.*, 2003). However, using only model produced forecast information ignores all prior (historical) knowledge and is prone to model systematic errors. At this point it is worth stressing the distinction between climate model outputs and observed climate/weather. Climate model outputs should not be treated as observed climate because they contain model structural and parametric errors, which should be corrected by calibration against observations.

Given these two distinct approaches to forecasting, it is natural to ask whether combining them may produce a forecast with more skill than either forecast considered separately. Thompson (1977) was one of the first to show that a simple linear combination of two independent 24 hour weather predictions, obtained by minimising the mean square error of the combined forecast, could reduce the forecast error variance by about 20%. Fraedrich and Leslie (1987) also noted that by linearly combining stochastic short-range forecasts with dynamical model weather predictions it was possible to obtain significantly better prediction skill. Fraedrich and Smith (1989) then extended this approach to seasonal forecasts with lead-times of up to three months. They linearly combined an empirical forecast with a deterministic model forecast for predicting tropical Pacific SST anomalies. It was shown that by minimising the combined forecast mean square error considerable improvement in skill can be obtained. More recently, Metzger *et al.* (2003) have extended the Fraedrich and Smith (1989) combination scheme to predict Niño-3 index anomalies for lead

times up to 24 months. They found that the linear combination of empirical and deterministic forecasts can provide improvement in prediction skill if the predictions of individual schemes are independent and of comparable skill. However, only modest skill improvements were found. Krishnamurti *et al.* (1999, 2000a,b, 2001), Pavan and Doblas-Reyes (2000) and Stefanova and Krishnamurti (2002) have introduced the multi-model method for combining dynamical weather and climate forecasts. The multi-model method linearly combines ensemble forecasts from different models by minimising the mean square error of the combined forecast. It has been demonstrated that the multi-model invariably outperforms any of the individual models.

From this brief review, it is clear that there is still a need for more research into how to produce well-calibrated combined forecasts. The aim of this study is to introduce a simple Bayesian approach and demonstrate it using monthly Niño-3.4 index forecasts at a 5-month lead time. One particular advantage of this method is that it merges valuable past (historical) information with coupled model ensemble forecasts to produce better quality probability estimates of the mean forecast value and its respective uncertainty.

The Bayesian approach has been discussed for decision making in applied meteorology by Epstein (1962) and for statistical inference and prediction in climatology by Epstein (1985). It has also successfully been used in other areas such as hydrology (e.g. Krzysztofowicz 1983; Krzysztofowicz and Herr 2001) and recently in climate studies (e.g. Berliner *et al.* 2000a,b and Rajagopalan *et*

al. 2002). As pointed out by Mason and Mimmack (2002), ENSO forecasts are usually issued in deterministic terms and very little attention has been directed to careful estimation of forecast uncertainty. This study treats ENSO forecasts in probabilistic terms, with particular attention directed to the estimation of prediction uncertainty. For this particular application, Niño-3.4 index interval forecasts are used to summarise the mean and the variance of the predicted normal distribution.

Section 2 introduces the empirical and coupled model ensemble forecasts of the Niño-3.4 index used in this study. Section 3 describes the Bayesian method used to combine the forecasts, and Section 4 presents results of the combined forecasts. Section 5 concludes the article with a summary and a discussion of possible future areas for research.

2. Empirical and coupled model ensemble forecasts of ENSO

These methods here will be demonstrated using 5-month lead forecasts of December mean Niño-3.4 index starting from conditions at the end of the preceding July. Empirical and coupled model ensemble forecasts available over the T=13-year period (1987-99) have been used. This short record is typical of the length of data sets produced by most of the world's climate prediction centres. Details concerning datasets and forecast lead time are given in Appendix A. Figure 1 shows the historical (1950-2001) December Niño-3.4 index time series. The largest El Niño (1972, 1982 and 1997) and La Niña (1970, 1973,

1988 and 1998) events can clearly be seen.

a. Empirical forecast of ENSO

A.1. THE EMPIRICAL MODEL

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. That is, $\theta_t = \beta_o + \beta_1 \psi_t + \varepsilon_t$, 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 can be written more explicitly in probabilistic notation as

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

with the mean given by

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

that is, a linear function of the predictor ψ_t. The standard statistical symbol $\mid$ denotes "given" (conditional upon) and $\sim$ denotes "is distributed as".

Figure 2 shows a scatter plot of the December versus the preceding July Niño-3.4 index for the period 1950-2001 ($N = 52$ observations). The linear regression fit is indicated on Fig. 2 as a solid line. A large amount of the

total variance of December is explained by the preceding July Niño-3.4 index (R^2=0.76). This emphasises the importance of persistence for forecasting the Niño-3.4 index.

A.2. EMPIRICAL MODEL CROSS-VALIDATION

To avoid artificial skill, the empirical model has been evaluated using a cross-validation "leave one 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 3a shows empirical forecasts for the target period 1987-99 (thick line), observed values (thin line) and the December climatological mean of 26.5^oC (short-dashed line). The 95% prediction interval (P.I.) for θ_t "given" ψ_t is also shown (grey area surrounded by long-dashed lines). The 95% prediction interval is defined by

$$\hat{\mu}_{ot} \pm 1.96\,\hat{\sigma}_{ot} \tag{3}$$

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 given by

$$\hat{\sigma}_{ot} = \hat{\sigma}_o \left(1 + \frac{1}{n} + \frac{(\psi_t - \overline{\psi}_t)^2}{n\,S_t^2}\right)^{\frac{1}{2}} \tag{4}$$

where $n = N - 1$ is the total number of years used in the cross-validation, $\overline{\psi}_t = \frac{1}{n}\sum_{i\neq t}\psi_i$ is the long-term climatological mean of the July Niño-3.4 index, $S_t^2 = \frac{1}{n}\sum_{i\neq t}[\psi_i - \overline{\psi}_t]^2$ and $\hat{\sigma}_o = \left[\frac{1}{n-2}\sum_{i\neq t}(\theta_i - \hat{\mu}_{oi})^2\right]^{\frac{1}{2}}$ is the estimated empirical

model standard deviation (see Draper and Smith 1998, Section 3.1).

Eqns. (3) and (4) show that the smallest prediction interval is obtained when the predictor equals its mean value $\psi_t = \overline{\psi}_t$. On the other hand, by moving away from $\overline{\psi}_t$ in either direction the prediction interval increases. The greater distance a particular July Niño-3.4 index (ψ_t) is from the climatological mean value ($\overline{\psi}_t$), the larger is the extrapolation error made when predicting the following December Niño-3.4 index (θ_t). However, the use of Eqn. (4) compared to $\hat{\sigma}_{ot} = \hat{\sigma}_o$ leads to only small changes in practice in the prediction interval, since the S_t^2 term in the denominator is proportional to the sum of n terms of the same magnitude as the term $(\psi_t - \overline{\psi}_t)^2$. The most precise predictions are obtained for July Niño-3.4 index values in the "middle" of the observed range of ψ_t, while for more extreme values further away from the climatological mean, predictions are less precise.

Figure 3a shows that the empirical forecast prediction interval does not vary much from year to year, indicating stability of estimates such as $\hat{\sigma}_o$. This simple model provides good forecasts, especially for the 1988 and 1998 La Niña episodes and for the 1997 El Niño episode. Out of the 13 years the model has only once (in 1987) forecast the Niño-3.4 index outside the 95% P.I. Measures of forecast skill and uncertainty will be discussed in more detail in Section 4.

Figure 3b shows the time series of the standardised forecast errors

$$Z_t = \frac{\hat{\mu}_{ot} - \theta_t}{\hat{\sigma}_{ot}} \tag{5}$$

where $\hat{\mu}_{ot}$ is the forecasted mean, θ_t is the observed value and $\hat{\sigma}_{ot}$ is the prediction standard deviation at time *t*. If this empirical model is appropriate, the standardised forecast errors should be distributed as independent normally distributed random variables with zero mean and unit variance. This appears to be the case from Fig. 3b. Although some slight sign of serial correlation may suggest the need of future model extensions, the standardised forecast errors appear to have constant variance and are well centred on zero with no obvious large outliers. The periods 1988-1990 and 1997-1998 have small standardised errors, while 1987, the period 1991-1996 and 1999 have larger standardised errors. The largest standardised forecast error occured in 1987.

b. Coupled model ensemble forecasts of ENSO

Figure 4a shows ECMWF raw (i.e. not bias corrected) coupled model ensemble forecasts for the same period. The ensemble mean of the ensemble of 9 forecasts is shown as a solid thick line. The 95% P. I., 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 thin line shows the observed values of Niño-3.4 and the short-dashed line is the December climatological mean of 26.5^oC. The ensemble system tends to underestimate the Niño-3.4 index and the width of the 95% P. I. is unrealistically smaller than the width of the 95% P. I. of the empirical forecast. Quantitative comparisons of skill and uncertainty of the empirical and raw coupled model forecasts will be discussed in Section

4.

Figure 4b shows the standardised forecast errors for the ECMWF raw coupled model ensemble forecast. Standardised forecast errors (Eqn. **??**) were obtained by dividing the forecast error by the standard deviation of the 9 coupled model forecasts for each year. These forecasts show a clear negative bias towards cooler Niño-3.4 values. Biases are well-known features of coupled model seasonal forecasts (e.g. Stockdale, 1997). The year 1991 produced one of the largest standardised forecast errors due to having a large forecast error and a small ensemble standard deviation.

3. Bayesian method for combining forecasts

The Bayesian method is a consistent probabilistic approach that can be used for combining historical (climatological) information (θ) with dynamical model ensemble mean forecasts ($\overline{X}$). The Bayesian method is firmly based on rigorous probability theory and so can provide well-calibrated probability forecasts.

With no access to a coupled model ensemble mean forecast $\overline{X}$, the only possible probabilistic assessment about the observable variable θ has to be based 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 $p(\theta)$ estimated from historical observations. In Bayesian theory, $p(\theta)$, is known as the *prior distribution* and encapsulates

prior knowledge about likely possible values of θ - from past experience not all values of θ were found to occur equally likely. A more informative prior is the empirical model defined in Section 2a.

However, when a particular ensemble mean forecast $\bar{X} = x$ is known for the future, it is then 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 that the forecast $\bar{X} = x$ is known. Conditioning on forecasts helps to reduce the uncertainty about future values of θ (Jolliffe and Stephenson, 2003; Chapter 9). This procedure is illustrated schematically in Fig. 5. The normal prior probability density (short-dashed line) when combined with a normal likelihood probability density (dashed line) yields a normal posterior probability density (solid line). The posterior distribution p($\theta|\bar{X} = x$) is found from the prior p(θ) by making use of Bayes' theorem

$$\overbrace{p(\theta_t|\bar{X}_t = x)}^{posterior} = \frac{\overbrace{p(\bar{X}_t = x|\theta_t)}^{likelihood}\overbrace{p(\theta_t)}^{prior}}{p(X_t = x)} \tag{6}$$

where θ_t is the observable variable at time t and x is a particular value of 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 a normalising constant (Lee, 1997).

The likelihood p($\bar{X}|\theta$) of obtaining an ensemble mean forecast $\bar{X}$ given observations θ is an essential ingredient in the Bayesian updating procedure that

can be estimated by stratifying 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).

The Bayesian approach has several important advantages over approaches that rely *solely* on sampling ensembles of coupled model forecasts (e.g. Stockdale *et al.*, 1998 and Taylor and Buizza, 2003). Firstly, the Bayesian approach appropriately incorporates prior information about the distribution contained in historical observations (i.e. combination). Secondly, the likelihood estimation provides a natural way of correcting for biases in the model forecasts that often occur in coupled model systems (i.e. calibration). Thirdly, the resulting well-calibrated posterior distribution allows one to generate an arbitrarily large sample (a *mega-ensemble*) of possible climate realizations, of use for example in scenario studies of risk and forecast value (Jolliffe and Stephenson, 2003; Chapter 8). It should be noted that, even for perfect forecasts, ensembles of model forecasts are not realizations of real climate - climate forecasts are variables in model space not in observation space. Climate model forecasts are generally not perfectly calibrated (although some models may produce well-calibrated raw forecasts) and contain uncorrected forecast errors. Ensemble forecast variances, for example, are likely to either underestimate or overestimate posterior uncertainties. In summary, ensemble spread does not generally explain all the forecast uncertainty and ensemble relative frequency does not perfectly estimate the probability of climate.

The Bayesian method has three main steps: a) choice of the prior distribution; b) modelling of the likelihood function; and c) determination of the posterior distribution. For simplicity, it has been assumed in this study of Niño-3.4 that both prior and likelihood distributions are normal (Gaussian). The Niño-3.4 index has already been demonstrated to be well approximated by the normal distribution (e.g. Burgers and Stephenson, 1999 and Hannachi *et al.*, 2003).

a. Choice of the prior distribution

The empirical model based on preceding July values of the Niño-3.4 index defined in Section 2a

$$\theta \sim N(\mu_{ot},\, {\sigma_{ot}}^2) \tag{7}$$

where μ_{ot} is estimated using $\hat{\mu}_{ot} = \hat{\beta}_o + \hat{\beta}_1 \psi_t$ and ${\sigma_{ot}}^2$ is estimated using Eqn. (4), provides an informative and straightforward prior distribution. More sophisticated empirical models could be used in future studies.

b. Modelling of the likelihood function

Figure 6 shows a scatter plot of raw coupled model ensemble forecasts versus the observed December Niño-3.4 index for the period 1987-1999. Ensemble means are depicted using large open circles. The dashed line is what one expects for perfect forecasts in which the forecast values are identical to the observed values. The likelihood $p(\overline{X_t} \mid \theta_t)$ is modelled by performing a weighted linear regression between the ensemble mean forecasts ($\overline{X_t}$) and matching observations

(θ_t):

$$\bar{X}_t \mid \theta_t \sim N(\alpha + \beta\theta_t, \gamma V_t) \tag{8}$$

where α and β are the intercept and slope parameters, respectively. Regression weights are given by $w_t = V_t^{-1}$, where V_t is the sample variance of the ensemble mean estimated from $V = s_X^2/m$, where m is the number of ensemble forecasts ($m = 9$ for our forecasting example). Forecasts with larger ensemble spread have more uncertain ensemble means and so must be given less weight in the regression.

For *independent* ensemble forecasts the variance of the ensemble mean forecast in the likelihood model would be given by V_t (see Clarke and Cooke 1992, Section 10.3). However, if the ensemble members are not independent, the variance differs from V_t. A simple way to ensure consistency is to allow scaling of the ensemble variance V_t by a factor γ in Eqn. (8). Ideally γ should be equal to one but in practice here γ is larger that one. In the case of perfect model, but not independent ensemble members, γ can be interpreted as m/m', where m is the number of ensemble members and m' is the effective number of independent forecasts. The dependency factor γ is obtained as a weighted mean of the square regression residuals

$$\gamma = \frac{1}{n-2} \sum_{t=1}^{n} w_t (\bar{X}_t - \hat{\alpha} - \hat{\beta}\theta_t)^2 \tag{9}$$

where n is the length of the time series and $w_t = V_t^{-1}$. Since the expectation of the ensemble mean is modelled by linear regression ($\alpha + \beta\theta_t$), it follows that the estimated γ will encompass the errors in this linear assumption.

The solid line in Fig. 6 is the best fit linear weighted regression between raw ensemble mean values $\bar{X}_t$ and observations θ_t, corresponding to estimates for the whole period of $\hat{\alpha}$ = 6.24^{o}C, $\hat{\beta}$ = 0.75, and $\hat{\gamma}$ = 7.05. It can be clearly seen that the raw coupled model ensemble forecast is biased. These values and Fig. (6) indicate that: a) the variance in Niño-3.4 explained by the coupled model is underestimated [i.e. $Var(\bar{X}_t) < Var(\hat{\theta}_t)$ since $\beta < 1$]; b) the coupled model generally underestimates the mean SST in the Niño-3.4 region [solid line generally below dashed line in Fig. (6)]; and c) either there are not enough independent ensemble members ($m' = m/\hat{\gamma} = 1.3$) or the error in the coupled model ensemble forecasts cannot be removed by a linear regression.

To avoid introducing artificial skill, both prior and likelihood distribution parameters are estimated using cross-validation by leaving out the year being forecast. The mean cross-validated likelihood estimated parameters are: $\hat{\alpha}$ = 6.27 (1.44) [^{o}C]; $\hat{\beta}$ = 0.75 (0.05); and $\hat{\gamma}$ = 7.05 (0.18), where the values in brackets are the mean of the standard errors obtained for each of the cross-validated estimates.

c. Determination of the posterior distribution

From Bayes' theorem (Eqn. **??**) it can be shown that for a normal prior distribution $\theta \sim N(\mu_{ot}, \sigma^2_{ot})$ and normal likelihood $\bar{X}_t|\theta_t \sim N(\alpha + \beta\theta_t, \gamma V_t)$, the posterior distribution is also normal (Lee, 1997). The resulting normal posterior

distribution is given by

$$\theta_t \mid \bar{X_t} \sim N(\mu_t, \sigma_t^2) \tag{10}$$

with the mean μ_t and the variance σ_t^2 equal to

$$\frac{1}{\sigma_t^2} = \frac{1}{{\sigma_{ot}}^2} + \frac{\beta^2}{\gamma V_t} \tag{11}$$

$$\frac{\mu_t}{\sigma_t^2} = \frac{\mu_{o_t}}{{\sigma_{ot}}^2} + \frac{\beta^2}{\gamma V_t}\left(\frac{\bar{X_t} - \alpha}{\beta}\right) \tag{12}$$

A derivation of Eqns. (11) and (12) is presented in Appendix B. The inverse of the variance is known in statistics as the *precision*. Equation (11) states that the precision of the posterior distribution $\left(\frac{1}{\sigma_t^2}\right)$ is exactly equal to the precision of the prior distribution $\left(\frac{1}{{\sigma_{ot}}^2}\right)$ plus the precision of the ensemble system $\left(\frac{\beta^2}{\gamma V_t}\right)$. Perfectly accurate unbiased forecasts would have precision $1/V_t$. However, forecasts are not perfectly accurate and unbiased and so the precision is instead given by the term $\frac{\beta^2}{\gamma V_t}$.

Equation (12) gives the posterior combined mean (μ_t) as the precision weighted sum of the prior empirical mean (μ_{o_t}) and the raw coupled model ensemble mean $(\bar{X_t})$. Note that the precision of the prior distribution and the precision of the ensemble system are weights for the prior mean and raw ensemble mean, respectively. 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 (term in brackets). Note, however, that the role of the prior diminishes with the increase of the sample size *m* so that the posterior distribution is increasingly

dominated by the likelihood and not very much affected by the prior.

d. Instrumental Calibration and Inverse Regression

Rather than regress the forecasts on the observations, it might at first appear more natural to regress the observations on the forecasts. In other words, one can use the coupled model forecasts as predictors in a regression model to obtain predictions of the observations. However, it should be noted that the (explanatory) forecast values are not deterministic control variables but instead contain large amounts of uncertainty. Furthermore, it can be assumed that climate forecasts are generally more uncertain than are the observed values. For these reasons and what follows, it is better to develop a regression model of the forecasts as a function of the observed values. Least-squares estimation then corresponds to minimising forecast error for fixed values of the observed variable.

The calibration of the forecast $\bar{X}_t$ to the predictand θ_t can be considered as a classical calibration problem for an instrumental device. This is a long standing issue in statistical literature, often referred to as the *inverse regression* problem (Brown, 1994). It is of relevance to probability forecasting and so will be briefly reviewed here.

In the simplest classical calibration setting, a precise instrument gives a measurement θ_t, while a less precise instrument, to be calibrated, produces $\bar{X}_t$

for the same quantity. The calibration database consists of a time series of paired values $\{(\theta_t, \bar{X}_t), t = 1, 2, ..., T\}$. Some classical examples for θ_t and $\bar{X}_t$ are respectively (real) pressures and gauge readings (Seber, 1977), tree-ring counts and (the less precise) carbon dating measure (Draper and Smith, 1998), or a long and costly laboratory method for determining the concentration of a certain enzyme in blood plasma samples and a quick and cheap autoanalyser device (Aitchison and Dunsmore, 1975).

In this study, θ_t is the (more precise) best estimate of the observed Niño-3.4 index, while $\bar{X}_t$ is the (less precise) raw coupled model ensemble-mean forecast of the same index for the same year t. The coupled model forecast can be considered to be an instrument for diagnosing the predictand, and calibrating the forecasts then becomes a standard issue of instrumental calibration (Swets, 1988). The problem of estimating θ_t when a new reading $\bar{X}_t$ becomes available is known as the *inverse regression* problem in statistical literature. This is precisely our problem in calibrating some new forecast $\bar{X}_t$ when an historical database is available.

The established protocol stems at least from Eisenhart (1939) [see also Seber, 1977; Aitchison and Dunsmore, 1975; Draper and Smith, 1998; and Brown, 1982]. Since the errors in θ-values are negligible with respect to the device (forecast) errors, θ_t can be treated as the fixed control values and then one obtains the regression model of $\bar{X}$ versus θ:

$$\bar{X}_t = \alpha + \beta\theta_t + \varepsilon_t \tag{13}$$

where ε_t are independent normally distributed random variables with zero mean and variance σ^2. Then the maximum likelihood (ML) estimate of θ is

$$\hat{\theta}_t = (\bar{X}_t - \hat{\alpha})/\hat{\beta} \tag{14}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the least squares solution of the *calibration equation* (13). To avoid explosive estimates when $\hat{\beta} \approx 0$, truncated forms of Eqn. (14) can be defined.

In summary, the *classical calibration model* considers the conditional distribution of $\bar{X}$ given θ (i.e., $\bar{X}|\theta$), since the calibrating equation (13) describes the stochastic measures conditionally to the true quantities. Whereas Williams (1969) and others advocated using Eqn. (13) to derive the ML estimate (Eqn. 14), one can also think of defining the inverse regression model for $\theta|\bar{X}$ and then use it directly for estimating θ_t. Following this idea, Krutchkoff (1967, 1969), suggested the so-called *inverse estimate*

$$\hat{\theta}_t^K = \hat{a} + \hat{b}\bar{X}_t \tag{15}$$

based on the least squares estimates $\hat{a}$ and $\hat{b}$ obtained from the *inverse regression* model

$$\theta_t = a + b\bar{X}_t + e_t \tag{16}$$

Classical and inverse estimates coincide only when $\bar{X}$ is perfectly correlated with θ in the calibration database. The inverse regression approach is currently the prevalent method for correcting forecast biases in meteorology. The inverse

regression model is the typical regression model used in previous climate forecasting studies (e.g. Kharin and Zwiers, 2002; Pavan and Doblas-Reyes, 2000).

Krutchkoff (1967) used simulations to show that the inverse method can have smaller mean squared error (MSE) than the classical calibration approach (even in the truncated form). This led to a controversy in which the MSE criterion was criticised for this particular case. An alternative criteria was proposed and the conditions of relative superiority of one method over the other were investigated in depth by Williams (1969), Berkson (1969), Halperin (1970) and Hoadley (1970) among others, and later on by Chow and Shao (1990).

The Bayesian approach was useful in clarifying the controversy (Hoadley, 1970; Aitchison and Dunsmore, 1975). Ideally, one would like the conditional distribution of $\theta|\overline{X}$ but of course this cannot be obtained from the conditional distribution of $\overline{X}|\theta$ without also having an estimate of the marginal prior distribution $p(\theta)$. By means of $p(\theta)$ and $p(\overline{X}|\theta)$ the distribution of $p(\theta|\overline{X})$ can be obtained using the Bayes' theorem (Eqn. 6) and the inverse regression problem can be solved. In order to understand the relative merits of classical and inverse estimators, note that both are special cases of the Bayesian estimator but with two different priors (Hoadley, 1970). The classical maximum likelihood estimator corresponds to a diffuse (improper) prior $p(\theta) \propto 1$, which leads to a posterior distribution $p(\theta_t|\overline{X_t})$ that is normal with mean $\hat{\theta}_t$ (Eqn. 14). Hoadley (1970) demonstrated that the inverse estimator $\hat{\theta}_t^K$ corresponds to a Bayesian estimate with the prior for θ centred on the calibration mean $\overline{\theta} = \sum \theta_t/T$. In

other words, by using θ_t values of the calibration dataset $(\theta_t, t = 1, 2, ..., T)$ to estimate a normal prior one finds that the posterior mean is given by $\hat{\theta}_t^K$ (Eqn. 15).

In the current comparison between classical and inverse estimators, the inverse regression will do well if θ_t lies centrally in the set of previous θ-values used in fitting the inverse calibration (Eqn. 16). On the other hand, the truncated classical estimator, corresponding to a proper uniform prior, will be more efficient for more extreme θ_t-values (Brown, 1982). Since the inverse regression prior is centred on the calibration mean $\bar{\theta}$, the comparison of inverse and classical estimates will be unfair to the latter if the calibration database coincides with the verification database.

Note, however, that rather than using a different estimation technique for each case, the best method is to choose the best prior for any particular application (the Bayesian approach). To do this, one needs extra information about θ alone. In forecast calibration this is the most common situation, where a short bivariate time series $\{(\theta_t, X_t), t = 1, 2, ..., T\}$ can be used for calibrating and a longer historical climatology can be used to estimate the prior. The utility and flexibility of the Bayesian approach in combining the two sources of information is apparent. The use of more complex prior data including other predictors can further help in adapting the prior to the particular forecasting conditions. A very simple example will be given in this paper by using the previously defined empirical forecast to estimate the prior.

4. Results

Figure 7a shows the mean of the combined forecast (thick line), observations (thin line), the 95% P.I. (grey shaded surrounded by long-dashed line) and the December climatological mean of 26.5^oC (short-dashed line). Comparison of this forecast with the empirical forecast alone (Fig. 3a) and raw coupled model ensemble forecast alone (Fig. 4a) shows that the combined forecasts are in closer agreement with the observations. The 95% P.I.'s are also reduced compared to those of the empirical forecasts indicating a reduction in forecast uncertainty due to combination with raw coupled model forecasts. Unlike the raw coupled model forecasts, only one forecast year (1994) falls outside the 95% P.I., indicating that the forecasts are better calibrated than the raw coupled model forecasts. However, it is worth mentioning that a similar effect could be obtained by crudely removing the mean bias from the raw coupled model forecasts and rescaling the averaged ensemble spread to match the error variance.

Figure 7b shows the combined forecast standardised errors. The smallest errors were found within the period 1987-1993 and in 1995 and 1998. The largest errors were in 1994, 1996, 1997 and 1999. It can be seen that these errors are evenly distributed and centred on zero.

Figure 8 shows plots of the standardised forecast error versus forecast values for the three types of forecasts presented so far. Figure 8b shows that the raw coupled model ensemble forecast is negatively biased. The standardised errors

for the empirical forecast (top panel) and for the combined forecast (bottom panel) are evenly spread around the zero line. Note also that the combined forecast does not show dependency on forecast values. However, this is not the case for the raw coupled model ensemble forecast, in which larger forecast values are associated with larger standardised forecast errors.

Table 1 gives some deterministic verification scores and a measure of forecast uncertainty of seven different forecasts of December Niño-3.4 index for the period 1987-1999. All the forecasts were produced using the cross-validation "leave one out" method and Table 1 summarises the skill of these forecasts in the short 13 year sample period.

• The climatological forecast is given by the historical Niño-3.4 index December mean value ($\bar{\theta}$) of 26.5^oC and the historical December standard deviation (s_θ) of 1.19^oC.

• The empirical forecast is given by $\hat{\mu}_{ot}$ and $\hat{\sigma}_{ot}$, as defined in Section 2a.

• The raw coupled model ensemble forecast is given by $\bar{X}_t$ and s_X, as defined in Section 2b.

• The bias-corrected forecast is given by $\bar{X}_t{}' = \bar{X}_t - \overline{X} + \bar{\theta}$ and s_X, where $\bar{X}_t$ is the raw ensemble mean forecast at time t, and $\overline{X}$ and $\bar{\theta}$ are the time means of the raw ensemble mean forecast and the observed mean values over the forecast period 1987-1999, respectively. This is a special case of a Bayesian forecast

with uniform prior (defined below) and simplified likelihood [$\beta = 1$ and $\gamma = m$ in Eqn. (8)]. The simplified likelihood models the ensemble mean bias as a constant (α) and the sample variance of the ensemble forecast as $mV_t = s_X^2$.

• The combined forecast with uniform prior is given by $\frac{\bar{X}_t - \alpha}{\beta}$ and $\sqrt{\frac{\gamma V_t}{\beta^2}}$. It is obtained by setting ${\sigma_{ot}}^{-2}$ to zero in Eqns. (11) and (12), that is, *all* values of the index 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 raw ensemble mean and it is useful for comparison with the bias-corrected forecast.

• The combined forecast with climatological prior is given by μ_{oc} and σ_{oc}. It is obtained when the December normal climatological distribution [i.e., $N(\bar{\theta}, {s_\theta}^2)$] is used as the prior distribution.

• The combined forecast is given by $\hat{\mu}_t$ and $\hat{\sigma}_t$, as defined in Section 3b.

Mean Squared Error (MSE) and Mean Absolute error (MAE) have been used as verification scores for the forecast means. The MAE skill score given by $SS = 1 - (MAE/MAE_c)$, where MAE_c is the climatological MAE, was used to measure forecast skill. The reason for using this score instead of the MSE skill score is because the MAE skill score provides a more resistant measure for small samples (Jolliffe and Stephenson, 2003). Forecast uncertainty was summarised by the time mean of the predicted forecast standard deviations over the forecast period 1987-1999.

The climatological forecast is the most uncertain and imprecise forecast with the largest MSE and MAE errors and the largest prediction uncertainty (Table 1). The raw coupled model ensemble forecast has (coincidentally) the same MSE as the empirical forecast, and a slightly larger MAE than the empirical forecast. Note that although these two models have similar MSE and MAE their uncertainty estimates are quite different. The width of the 95% P. I. in Fig. 4a, which is proportional to the mean uncertainty shown in Table 1, shows that the coupled model uncertainty is unrealistically underestimated and fails to cover the range of observations. The bias-corrected coupled model forecast has smaller MSE and MAE than the empirical forecast, and a greater skill score than the raw coupled model ensemble forecast. The uniform prior forecast has smaller MSE and MAE than the bias-corrected forecast, a slightly better skill score than the bias-corrected forecast and a much greater skill score than the raw coupled model ensemble forecast. The uniform prior has also smaller errors than the empirical forecast, and a greater skill score than the empirical forecast. These results suggest that the use of prior information helps to improve forecast skill. It also has a larger forecast uncertainty that is between the uncertainty of the raw coupled model ensemble forecast and the empirical forecast. The combined forecast with climatological prior has slightly smaller MSE and MAE than the combined forecast with uniform prior and greater skill scores than the bias-corrected forecast and the raw coupled model ensemble forecast, indicating that the use of climatological prior information helped to improve even more forecast skill. The combined forecast with climatological prior also

has smaller errors than the empirical forecast, and a greater skill score than the empirical forecast. It also has greater forecast uncertainty, which is only slightly smaller than the uniform prior forecast uncertainty. The combined forecast has the smallest values of MSE and MAE of all the forecasts. It also shows an impressive improvement of 23% in skill when compared to the raw coupled model forecasts, indicating that the use of a more informative prior led to additional improvement in forecast skill. Additionally, it provides a much better and more realistic uncertainty estimate compared to the other forecasts.

Table 2 summarises the standardised forecast errors. The mean standardised forecast error shows that the raw coupled model forecast is negatively biased, with the largest mean error of all the forecasts. The climatological forecast, the combined forecast with uniform prior, and the combined forecast with climatological prior have the smallest mean errors, indicating that these forecasts are well-calibrated. The raw coupled model ensemble and the bias-corrected ensemble forecasts have the largest and most unrealistic variances of the standardised forecast errors. All forecasts have variances larger than one suggesting that the prediction uncertainty of the forecasts is being underestimated.

Since these scores are based only on a small sample of forecasts, one might worry that the benefits of using the Bayesian approach are due to chance sampling. However, similar conclusions as here were obtained when the same methodology was applied to three other versions of the ECMWF seasonal forecasting system, one of which had a much longer record of 44 years (Coelho *et*

al., 2003, in preparation). Additional analyses of the robustness of the obtained results have been performed by spliting the 44-year record into 3 samples of 13 forecasts each. It has been found that Bayesian combined forecasts generally provide better and more reliable forecasts than raw coupled model and empirical forecasts.

5. Conclusions

A Bayesian approach for calibrating and combining empirical and raw coupled model ensemble forecasts has been presented. The combined 5-month lead forecast of Niño-3.4 index has been shown to have greater forecast skill than either of the forecasts individually. This indicates that both empirical and raw coupled model ensemble forecasts contain mutually useful information. In other words, neither forecast is sufficient for the other forecast and so increased forecast skill can be obtained by combining both types of forecast. In order to produce improved interval forecasts of the Niño-3.4 index, empirical and coupled model forecasts should be combined together. The combined forecast also provides a more reliable prediction error estimate because it is based on a well-founded calibration approach that incorporates valuable historical information.

Good quality forecasts are expected to have both small prediction errors (good accuracy) and reliable forecast uncertainty estimates. It has been shown that, although the ECMWF raw coupled model ensemble forecast is able to simulate the inter-annual variability of the Niño-3.4 index reasonably well 5

months in advance, it underestimates both the mean SST value in the Niño-3.4 region and forecast uncertainty. The simple empirical model, on the other hand, provides more skillful forecasts compared to the raw coupled model ensemble forecast. These forecasts are less biased and present larger and more reliable uncertainty estimates. When the Bayesian approach was used to combine these two forecasts together, more skillful forecasts were obtained having more accuracy and reliability.

It is important to stress that both the prior and the likelihood model used in this study are simple. More sophisticated regression models could easily produce greater improvements in forecast skill, yet this is not the ultimate aim of this pilot study. It should be noted that some of the forecast errors/uncertainty derive from the modelling assumption used here (e.g. normal-normal model). Our approach does not fully incorporate uncertainty in the likelihood model parameters estimates that could be treated using a hierarchical Bayesian approach (see Berliner *et al.*, 2000b). This methodology also needs to be developed in order to combine ensemble forecasts from different coupled models (multi-model approach).

Acknowledgements

We wish to thank Dr. D. L. T. Anderson, head of the seasonal forecast group at ECMWF and Dr. T. N. Palmer, the DEMETER (EVK2-1999-00197) project principal investigator, who kindly provided the ECMWF coupled model hindcasts used in this research. CASC was sponsored by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) precess 200826/00-0. FJDR was supported by DEMETER. We also want to acknowledge two anonymous reviewers for their thoughtful comments and suggestions, which helped to significantly improve this manuscript.

Appendix

A. Datasets and lead time

Historical (1950-2001) Niño-3.4 index data were obtained from Reynolds optimum interpolation version 2 SST dataset (Reynolds *et al.*, 2002). Coupled model Niño-3.4 index ensemble forecasts were available from ECMWF for the period 1987-99, as part of the Development of a European Multi-model Ensemble system for seasonal to inTERannual prediction (DEMETER) project[2] (Palmer *et al.*, 2003). In the DEMETER project, several coupled models are run four times per year, starting the first day of February, May, August and Novem-

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

ber at 00 GMT. Nine ensemble forecasts are produced for the next six months including the starting month. Wind stress and SST perturbations are used to generate the ensemble. However, as part of this research only the ECMWF coupled model forecasts from the DEMETER assimilation experiment have been used. These forecasts were produced using initial conditions from the ERA-40 project and also assimilate subsurface ocean data. Only forecasts started in August to forecast the next December (5-month lead time) have been used. 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 and so gives predictive better skill (Webster and Yang, 1992).

B. Derivation of the posterior distribution

From Eqns. (7) and (8) the prior and the likelihood probability density functions (*pdf*) are respectively:

$$p(\theta_t) = N(\mu_{ot}, \sigma_{ot}^2) = \frac{1}{(2\pi)^{\frac{1}{2}}\,\sigma_{ot}}\; exp\left[-\frac{(\theta_t - \mu_{ot})^2}{2\;{\sigma_{ot}}^2}\right]$$

$$p(\bar{X_t} \mid \theta_t) = N(\alpha + \beta\,\theta_t, \gamma V_t) = \frac{1}{(2\pi)^{\frac{1}{2}}\,(\gamma V_t)^{\frac{1}{2}}}\; exp\left[-\frac{(\bar{X_t} - \alpha - \beta\,\theta_t)^2}{2\;\gamma V_t}\right]$$

Changing the variable to $Y_t = \frac{\bar{X_t} - \alpha}{\beta}$ in the likelihood function then gives

$$p(Y_t \mid \theta_t) = \frac{\beta}{(2\pi)^{\frac{1}{2}}\,(\gamma V_t)^{\frac{1}{2}}}\ exp\left[-\frac{\beta^2(Y_t-\theta_t)^2}{2\ \gamma V_t}\right]$$

which is a normal distribution for the random variable Y_t with mean θ_t and variance $\gamma V_t\ /\ \beta^2$:

$$p(\theta_t) = N(\mu_{ot}, \sigma_{ot}^2)$$
$$p(Y_t \mid \theta_t) = N(\theta_t, \frac{\gamma V_t}{\beta^2})$$

This is the Normal-Normal Bayesian model in standard form. Using Bayes' theorem (Eqn. 6), this can be shown to have a posterior *pdf* which is Normal, with posterior precision (reciprocal variance) given by the sum of prior precision and likelihood precision (Lee, 1997):

$$\frac{1}{\sigma_t^2} = \frac{1}{{\sigma_{ot}}^2} + \frac{\beta^2}{\gamma V_t}$$

while the posterior mean is the weighted average of prior mean and the rescaled forecast Y_t, with weights given by the respective precisions. Substituting Y_t by $\frac{\bar{X_t}-\alpha}{\beta}$ then gives

$$\frac{\mu_t}{\sigma_t^2} = \frac{\mu_{ot}}{{\sigma_{ot}}^2} + \frac{\beta^2}{\gamma V_t}\left(\frac{\bar{X_t}-\alpha}{\beta}\right)$$

REFERENCES

Aitchison J. and I. R. Dunsmore, 1975: Statistical prediction analysis. Cambridge University Press. Cambridge. 273 pp.

Anderson, J., H. van den Dool, A. Barnston, W. Chen, W. Stern and J. Ploshay, 1999: Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bull. Am. Meteorol. Soc.*, **80**(7), 1349-1361.

Barnston, A. G., M. H. Glantz and Y. He, 1999: Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997-98 El Niño episode and the 1998 La Niña onset. *Bull. Am. Meteorol. Soc.*, **80**(2), 217-243.

Berkson, J., 1969: Estimation of a linear function for a calibration line: consideration of a recent proposal. *Technometrics*, **11**, 649-660.

Berliner, L. M., R. A. Levine and D. J. Shea, 2000a: Bayesian climate change assessment. *J. Climate,* **13**, 3805-3820.

Berliner, L. M., C. K. Wikle and N. Cressie, 2000b: Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *J. Climate*, **13**, 3953-3968.

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

Brown, P. J., 1982: Multivariate calibration. *J. R. Statist. Soc. B*, **44**(3), 287-

321.

Brown, P. J., 1994: Measurement, regression and calibration. Oxford Science Publications. Oxford Statistical Science Series, **12**, 210 pp.

Chow, S. and J. Shao, 1990: On the difference between the classical and inverse methods of calibration. *Appl. Statist.*, **39**(2), 219-228.

Clarke, G. M. and D. Cooke, 1992: A basic course in statistics. Edward Arnold. Third edition. 451 pp.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes and D. B. Stephenson, 2003: The skill of Bayesian calibrated ENSO forecasts: A review of ECMWF seasonal forecasting systems. *ECMWF Technical Memorandum*, in preparation.

De Groot, M. H., 1970: Optimal statistical decisions. McGraw-Hill. New York.

Draper, N. R. and H. Smith, 1998: Applied regression analysis. John Wiley and Sons, Inc. Third edition. 706 pp.

Eisenhart, C., 1939: The interpretation of certain regression methods and their use in biological and industrial research. *Ann. Math. Statist.*, **10**, 162-186.

Epstein, E. S., 1962: A Bayesian approach to decision making in applied meteorology. *J. Appl. Meteorol.*, **1**, 169-177.

Epstein, E. S., 1985: Statistical inference and prediction in climatology: A

Bayesian approach. *Meteorological monographs - American Meteorological Society*. **20**(42). 199 pp.

Fraedrich, K. and L. M. Leslie, 1987: Combining predictive schemes in short-term forecasting. *Mon. Wea. Rev.*, **115**, 1640-1644.

Fraedrich, K. and N. R. Smith, 1989: Combining predictive schemes in long-range forecasting, *J. Climate*, **2**, 291-294.

Halperin, M., 1970: On inverse estimation in linear regression. *Technometrics*, **12**, 727-736.

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.

Horel, J. D. and J. M. Wallace, 1981: Planetary-scale atmospheric phenomena associated with the Southern Oscillation. *Mon. Wea. Rev.*, **109**, 813-829.

Jolliffe, I. N. and D. B. Stephenson, 2003: Forecast Verification: A practitioner's guide in atmospheric science. Wiley and Sons. 240 pp.

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

Krishnamurti, T. N., C. M. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548-1550.

Krishnamurti, T. N., D. W. Shin and C. E. Williford, 2000a: Improving tropical precipitation forecasts from a multianalysis superensemble. *J. Climate*, **13**, 4217-4227.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil and S. Surendran, 2000b: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196-4216.

Krishnamurti, T. N., S. Surendran, D. W. Shin, R. J. Correa-Torres, T. S. V. V. Kumar, E. Williford, C. Kummeow, R. F. Adler, J. Simpson, R. Kakar, W. S. Olson and F. J. Turk, 2001: Real-time multianalysis-multimodel superensemble forecasts of precipitation using TRMM and SSM/I products. *Mon. Wea. Rev.*, **129**, 2861-2883.

Krutchkoff, R. G., 1967: Classical and inverse methods of calibration. *Technometrics*, **9**, 525-539.

Krutchkoff, R. G., 1969: Classical and inverse methods of calibration in extrapolation. *Technometrics*, **11**, 605-608.

Krzysztofowicz, R., 1983: Why should a forecaster and a decision maker use Bayes theorem. *Mon. Wea. Rev.*, **19**(2), 327-336.

Krzysztofowicz, R. and H. D. Herr, 2001: Hydrologic uncertainty processor for probabilistic river stage forecasting: precipitation-dependent model. *J. Hydrology*, **249**, 46-68.

Landsea, C. and A. Knaff, 2000: How much skill was there in forecasting the very strong 1997-1998 El Niño? *Bull. Am. Meteorol. Soc.*, **8**(9), 2107-2119.

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

Mason, S. J. and G. M. Mimmack, 2002: Comparison of some statistical methods of probabilistic forecasting of ENSO. *J. Climate*, **15**, 8-29.

Metzger, S., M. Latif, and K. Fraedrich, 2003: Combining ENSO forecasts: A feasibility study. Submitted to Monthly Weather Review.

Palmer, T. N. and co-authors, 2003: Development of a European Ensemble System for Seasonal to Inter-annual Prediction (DEMETER). *Bull. Am. Meteorol. Soc.*. Submitted.

Patt, A., 2000: Communicating probabilistic forecasts to decision makers: A case study of Zimbabwe. Belfer Center for Science and International Affairs (BCSIA), Environment and Natural Resources Program, Kennedy School of Government, Harvard University. Discussion paper 2000-19. Available at http://environment.harvard.edu/gea.

Pavan, V. and F. J. Doblas-Reyes, 2000: Multi-model seasonal hindcasts over

the Euro-Atlantic: skill scores and dynamic features. *Clim. Dynamics*, **16**, 611-625.

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.

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.

Ropelewski, C. F. and M. S. Halpert, 1986: North American precipitation and temperature associated with the El Niño Southern Oscillation (ENSO). *J. Climate*, **114**, 2352-2362.

Ropelewski, C. F. and M. S. Halpert, 1987: Global and regional scale precipitation patterns associated with El Niño/Southern Oscillation. *Mon. Wea. Rev.*, **115**, 1606-1626.

Ropelewski, C. F. and M. S. Halpert, 1989: Precipitation patterns associated with high index phase of Southern Oscillation. *J. Climate*, **2**, 268-284.

Seber, G. A. F, 1977: Linear regression analysis. John Wiley and Sons. New-

York. 465 pp.

Stefanova, L. and T. N. Krishnamurti, 2002: Interpretation of seasonal climate forecast using Brier Skill Score, the Florida State University superensemble, and the AMIP-I dataset. *J. Climate*, **15**, 537-544.

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.

Stoeckenius, T., 1981: Interannual variations of tropical precipitation patterns. *Mon. Wea. Rev.*, **109**, 1233-1247.

Swets, J. A., 1988: Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285-1293.

Taylor, J. W. and R. Buizza, 2003: Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, **19**, 57-70.

Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.*, **105**, 228-229.

Trenberth, K. E., 1998: Development and forecasts of the 1997/98 El Niño: CLIVAR scientific issues. International CLIVAR Project Office. Exchanges;

Newsletter of the Climate Variability and Predictability Program (CLIVAR).

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.

Williams, E. J., 1969: A note on regression methods in calibration. *Technometrics*, **11**, 189-192.

TABLE CAPTIONS

Table 1: Forecast symbols, verification scores, skill score and mean forecast uncertainty. MSE and MAE are the mean squared error and mean absolute error of the mean forecast, respectively. The skill is measured by the MAE skill score (see text for more details)- values in brackets indicate the percentage improvement compared to the ensemble system skill score. Forecast uncertainty is given by the mean predicted forecast standard deviation over the period 1987-1999.

Table 2: The mean and variance of standardised forecast errors.

TABLES

Table 1: Forecast symbols, verification scores, skill score and mean forecast uncertainty. MSE and MAE are the mean squared error and mean absolute error of the mean forecast, respectively. The skill is measured by the MAE skill score (see text for more details)- values in brackets indicate the percentage improvement compared to the ensemble system skill score. Forecast uncertainty is given by the mean predicted forecast standard deviation over the period 1987-1999.

Forecast	Mean	Std. dev.	MSE	MAE	Skill Score	Uncertainty
	μ	σ	$[^oC]^2$	[^{o}C]	[%]	[^{o}C]
Climatology	$\bar{\theta}$	s_θ	1.99	1.16	0	1.19
Empirical	$\hat{\mu}_{ot}$	$\hat{\sigma}_{ot}$	0.47	0.53	55 (+4)	0.61
Raw ensemble	$\bar{X}_t$	s_X	0.47	0.57	51	0.33
Bias-corrected	$\bar{X}_t - \overline{X} + \bar{\theta}$	s_X	0.22	0.40	65 (+14)	0.33
Uniform Prior	$\frac{\bar{X}_t - \alpha}{\beta}$	$\sqrt{\frac{\gamma V_t}{\beta^2}}$	0.18	0.37	68 (+17)	0.39
Climatological Prior	μ_{oc}	σ_{oc}	0.17	0.32	72 (+21)	0.37
Combined	$\hat{\mu}_t$	$\hat{\sigma}_t$	0.13	0.31	74 (+23)	0.32
Perfect forecast	-	-	0	0	100	0

Table 2: The mean and variance of standardised forecast errors.

Forecast	Mean	Variance
Climatology	-0.09	1.52
Empirical	0.24	1.40
Raw ensemble	-1.73	3.32
Bias-corrected	-0.12	2.43
Uniform Prior	0.01	1.41
Climatological Prior	0.04	1.41
Combined	0.20	1.46
Perfect forecast	0	1

FIGURE CAPTIONS

Figure 1: Reynolds Optimum Interpolated December 1950-2001 Niño-3.4 SST index time series in ^oC. The short-dashed line is the climatological mean for this period, 26.5^oC.

Figure 2: Scatter plot of July versus December Niño-3.4 index (^oC). The solid line is the 1950-2001 linear regression model ($\hat{\beta}_o$ = -14.14^oC, $\hat{\beta}_1$ = 1.50, R^2 = 0.76).

Figure 3: (a) December 1987-1999 Niño-3.4 index empirical forecast (^oC). Observed values (thin solid line), forecast (thick solid line) and the 95% Prediction Interval (dashed lines). The short-dashed line is the December 1950-2001 climatological mean (26.5^oC). (b) Standardised forecast error.

Figure 4: (a) December 1987-1999 Niño-3.4 index raw coupled model ensemble forecast (^oC). Observed values (thin solid line), forecast (thick solid line) and the 95% Prediction Interval (dashed lines). The short-dashed line is the 1950-2001 December climatological mean (26.5^oC). (b) Standardised forecast error.

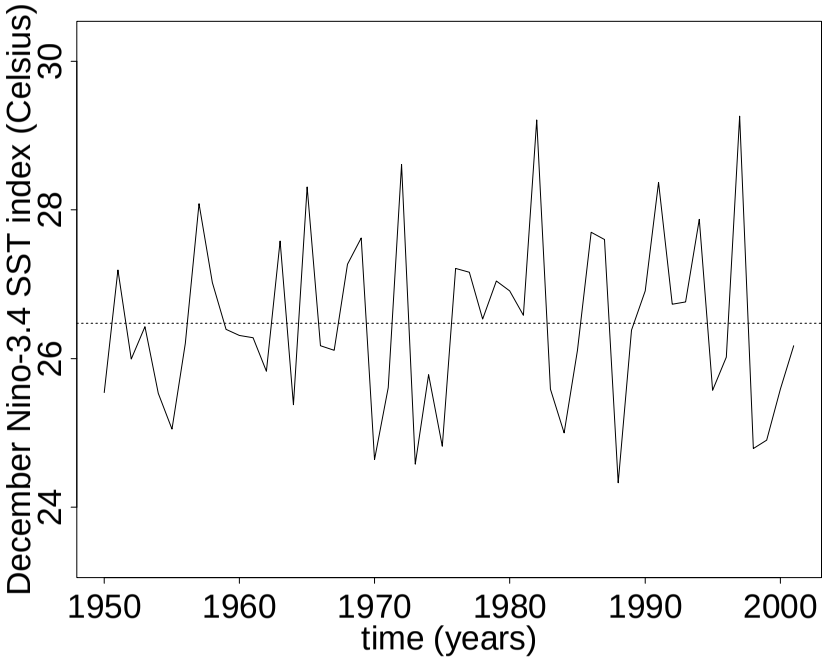
Figure 5: Prior distribution (short-dashed line), likelihood (dashed line) and posterior distribution (solid line).

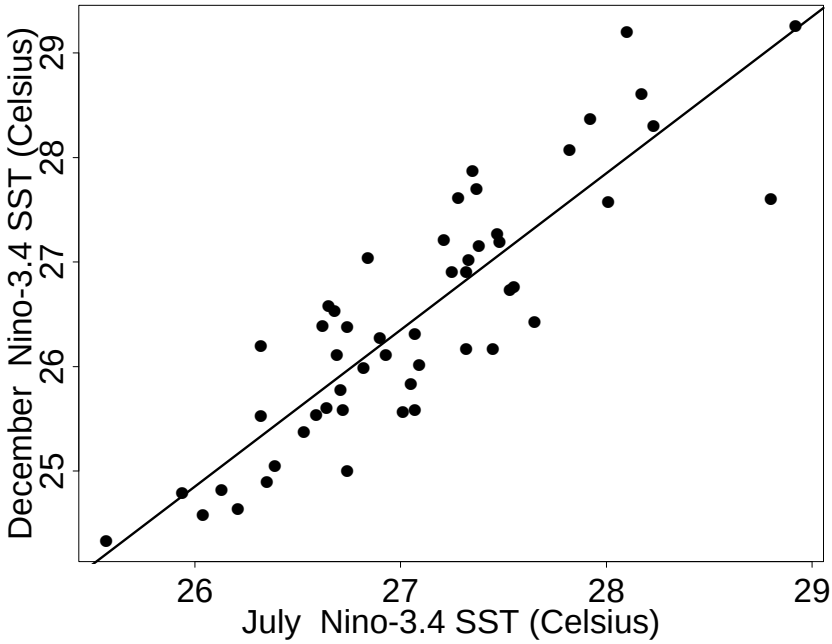
Figure 6: December 1987-1999 Niño-3.4 index likelihood model (^oC). Each black dot is one ensemble member. Big open circles are ensemble means. The

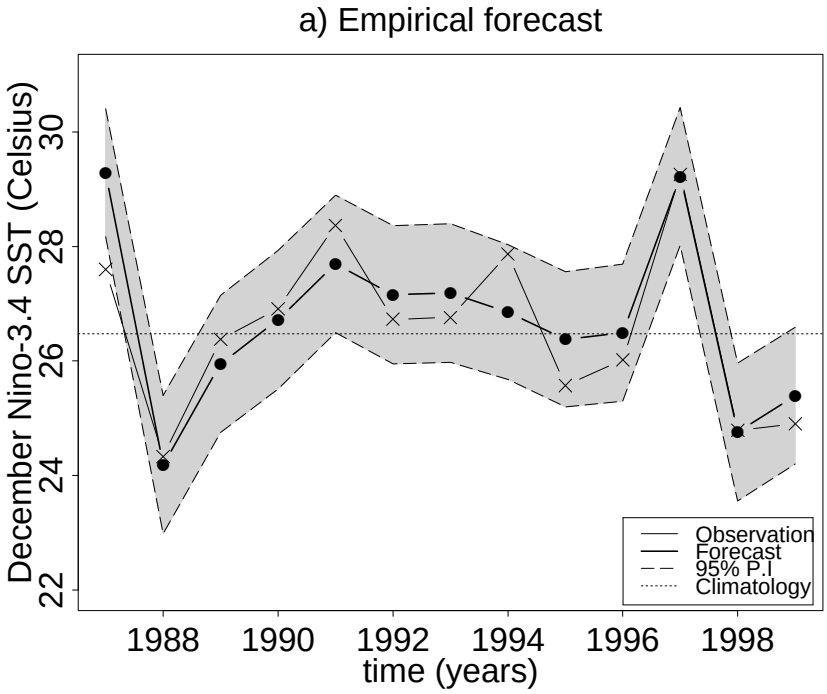
solid line is the regression between raw ensemble means and observations ($\hat{\alpha}$ = 6.24^oC, $\hat{\beta}$ = 0.75, R^2 = 0.95). The dashed line is what would be obtained for perfect forecasts.

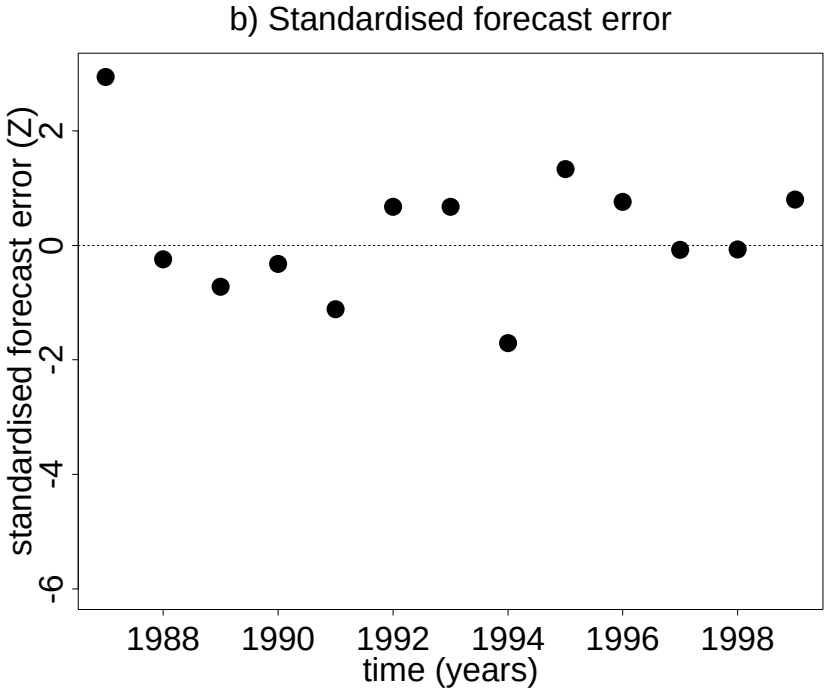
Figure 7: (a) December 1987-1999 Niño-3.4 index combined forecast (^oC). Observed values (thin solid line), forecast (thick solid line) and the 95% Prediction Interval (dashed lines). The short-dashed line is the 1950-2001 December climatological mean (26.5^oC). (b) Standardised forecast error.

Figure 8: Standardised forecast error versus forecast in ^oC for (a) the empirical forecast, (b) the raw coupled model ensemble forecast and (c) the combined forecast.

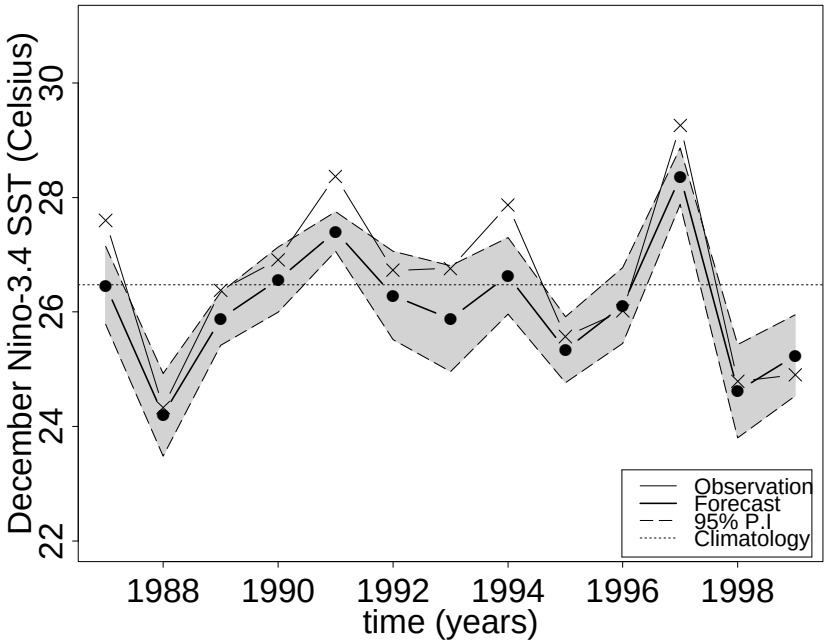

December Nino-3.4 SST index (Celsius)
24
26
28
30
1950
1960
1970
1980
1990
2000
time (years)


December Nino-3.4 SST (Celsius)
25
26
27
28
29
26
27
28
29
July Nino-3.4 SST (Celsius)

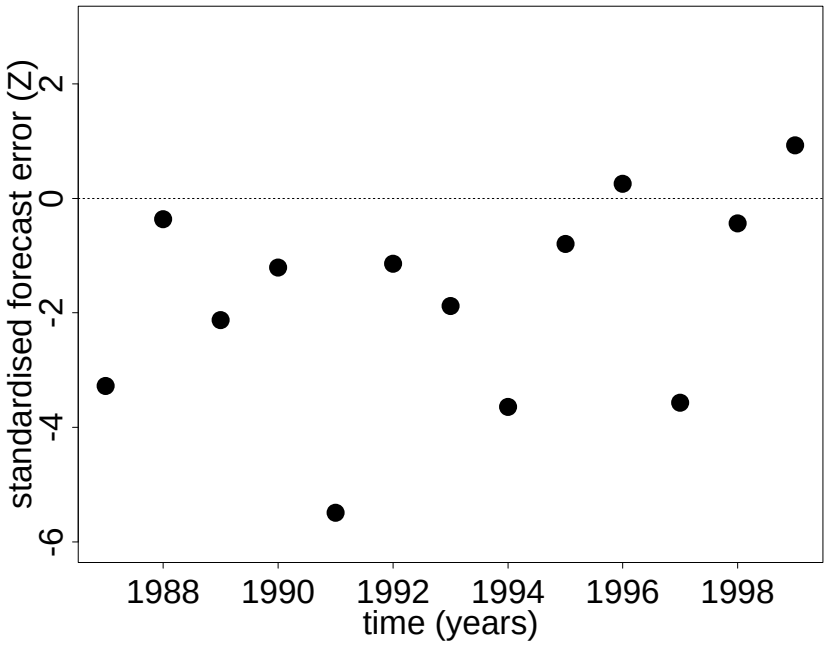

a) Empirical forecast
December Nino-3.4 SST (Celsius)
22
24
26
28
30
1988
1990
1992
1994
1996
1998
time (years)
Observation
Forecast
95% P.I
Climatology

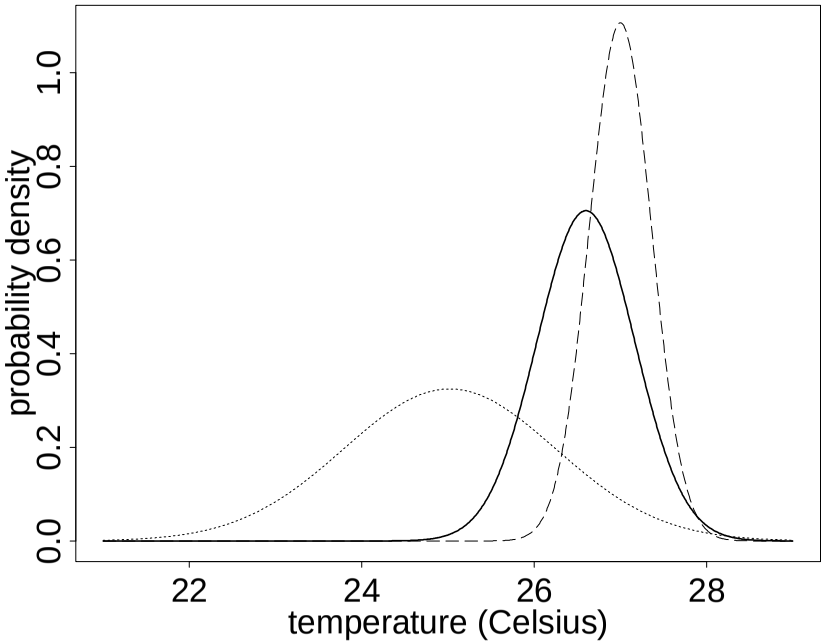

b) Standardised forecast error
standardised forecast error (Z)
-6
-4
-2
0
2
1988
1990
1992
1994
1996
1998
time (years)

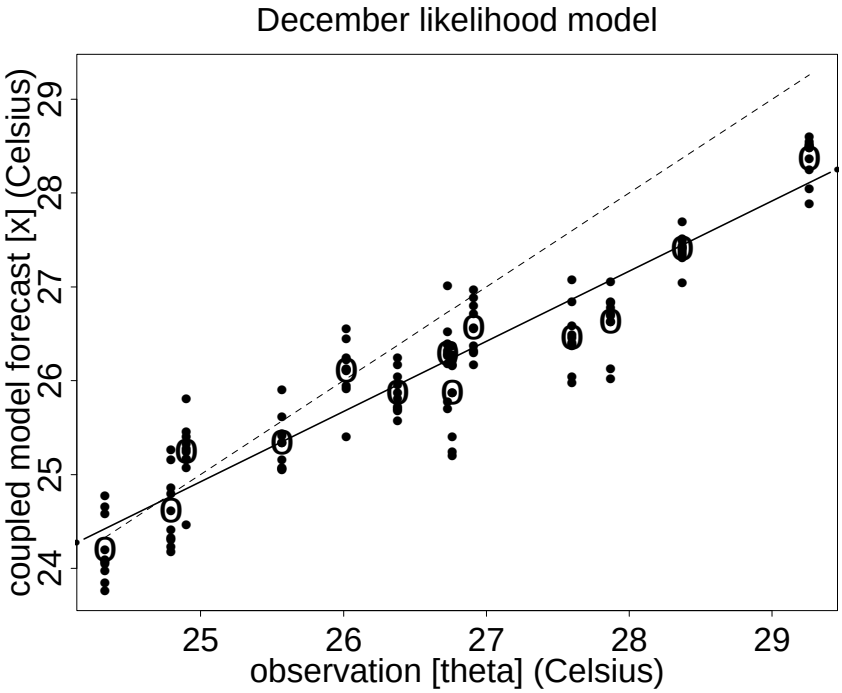
a) Raw coupled model ensemble forecast

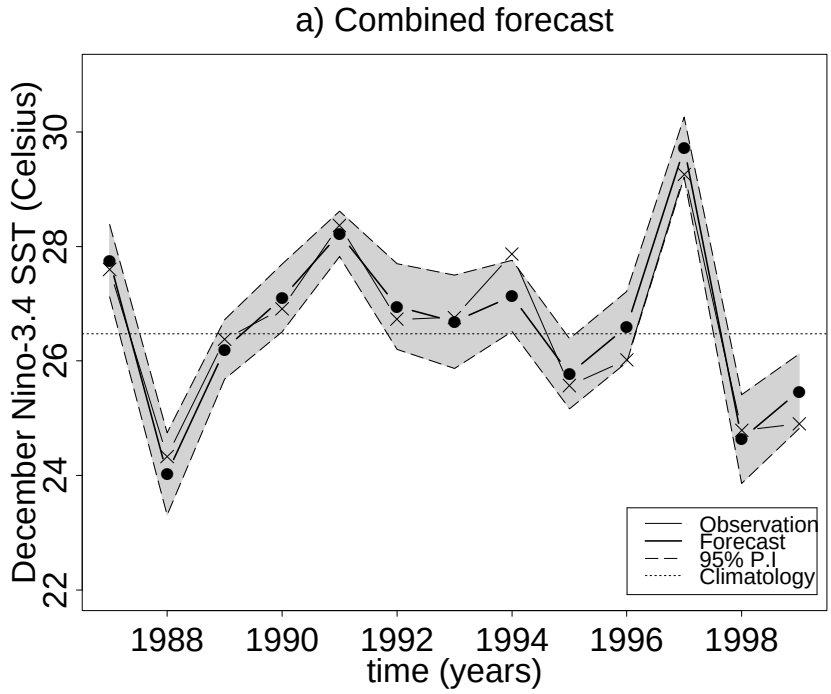


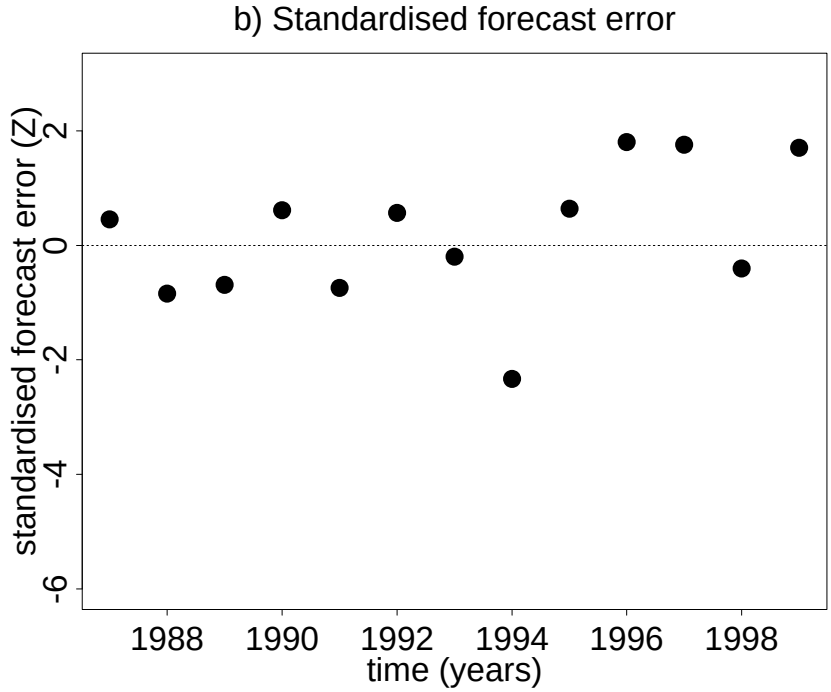
b) Standardised forecast error

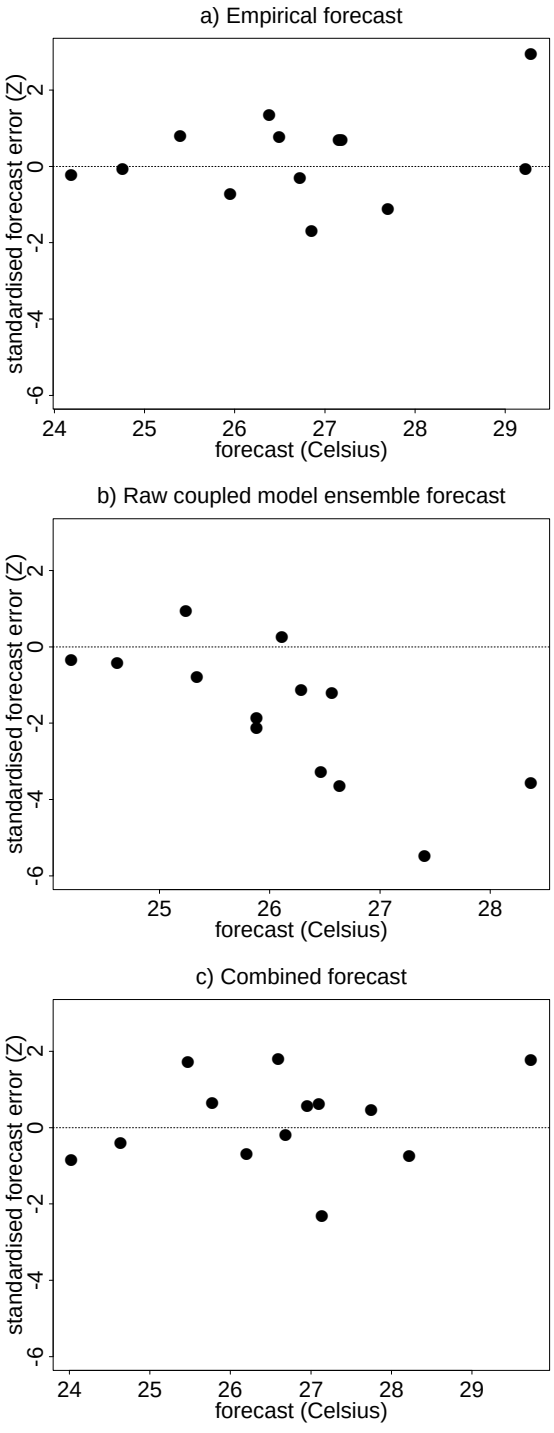



probability density
0.0
0.2
0.4
0.6
0.8
1.0
22
24
26
28
temperature (Celsius)


December likelihood model
coupled model forecast [x] (Celsius)
24
25
26
27
28
29
25
26
27
28
29
observation [theta] (Celsius)


a) Combined forecast
December Nino-3.4 SST (Celsius)
22
24
26
28
30
1988
1990
1992
1994
1996
1998
time (years)
Observation
Forecast
95% P.I
Climatology


b) Standardised forecast error
standardised forecast error (Z)
-6
-4
-2
0
2
1988
1990
1992
1994
1996
1998
time (years)


a) Empirical forecast
standardised forecast error (Z)
2
0
-2
-4
-6
24
25
26
27
28
29
forecast (Celsius)
b) Raw coupled model ensemble forecast
standardised forecast error (Z)
2
0
-2
-4
-6
25
26
27
28
forecast (Celsius)
c) Combined forecast
standardised forecast error (Z)
2
0
-2
-4
-6
24
25
26
27
28
29
forecast (Celsius)