Did the ECMWF seasonal forecast model outperform statistical ENSO forecast models over the last 15 years?

Geert Jan van Oldenborgh* **Magdalena A. Balmaseda†**
Laura Ferranti† **Timothy N. Stockdale†**
David L. T. Anderson†

April 20, 2005

Abstract
The European Centre for Medium-Range Weather Forecasts (ECMWF) has made seasonal forecasts since 1997 with ensembles of a coupled ocean-atmosphere model (S1). In January 2002, a new version (S2) was introduced. For the calibration of these models, hindcasts have been performed starting in 1987, so that 15 years of hindcasts and forecasts are now available for verification.

Seasonal predictability is to a large extent due to the El Niño — Southern Oscillation (ENSO) climate oscillations. ENSO predictions of the ECMWF models are compared with those of statistical models, some of which are used operationally. The relative skill depends strongly on the season. The dynamical models are better at forecasting the onset of El Niño or La Niña in boreal spring to summer. The statistical models are comparable at predicting the evolution of an event in boreal fall and winter.

1 Introduction

The use of dynamical models for seasonal forecasting is becoming widespread. In principle, numerical models that represent the dynamics of the atmosphere, ocean and land should be able to give better seasonal forecasts than purely statistical approaches, because of their ability to handle a wide range of linear and non-linear interactions and their potential resilience against a changing climate. In practice, model errors are still a substantial source of problems (Latif et al., 2001; Palmer

* KNMI, P. O. Box 201, NL-3730 AE De Bilt, The Netherlands
† ECMWF, Shinfield Park, Reading, RG2 9AX, U.K.

et al., 2004), and it remains unclear to what extent the present generation of numerical forecast models is able to challenge existing empirical methods for seasonal forecasting. Barnston et al. (1999) concluded that dynamical models did not forecast the 1997/98 El Niño and following La Niña better than statistical models. Anderson et al. (1999), Sardeshmukh et al. (2000) and Peng et al. (2000) compared how various models propagate the effect of (prescribed) SST anomalies to make seasonal forecasts, finding comparable or better skill in the statistical models. Our results from coupled models do not support these findings.

In this set of papers we compare the seasonal forecasting performance of two state-of-the-art coupled numerical systems (both from ECWMF) with a statistical forecasting scheme based on lagged regression with SST patterns. As most seasonal predictability is due to El Niño — Southern Oscillations (ENSO) variability, the performance of the three schemes is compared first for ENSO forecasts. For this comparison three more sophisticated statistical forecast models that are used operationally are also included. The companion paper considers seasonal forecasts of global fields of surface air temperature, mean sea level pressure and precipitation.

The period over which the forecasting schemes are verified and compared is 1987–2001. This is primarily due to the numerical forecast results being restricted by lack of ocean observations in the equatorial Pacific before this period. For the simple statistical model we take advantage of the limited verification period by restricting the training period to dates prior to 1987, ensuring a relatively clean statistical forecast for the verification. The operational statistical models have been cross-validated. The fact that we have only 15 years of verification inevitably limits the power of the comparisons we can make: the fluctuations in skill due to the small sample size will often be as large as the differences between the models.

In section 2 of the paper, we describe the numerical and statistical models used, and the observations against which they are verified. Next, in section 3, we discuss the skill in predicting indices of ENSO variability. Section 4 summarizes and concludes the paper.

2 Model, data

2.1 Brief description of the ECMWF models.

2.1.1 System-1

At ECMWF two coupled ocean-atmosphere models have been developed for seasonal forecasting. The first (called System-1, and denoted S1) was introduced in 1997 (Stockdale et al., 1998). At the time that this coupled model was being developed the version of the ECMWF atmospheric model that was used for weather

forecasting was IFS cy15r8. This version was used in S1, though at a lower resolution than the weather forecast model. The coupled model resolution was T63, with 31 vertical levels. The initial conditions for the atmospheric component of the coupled model were obtained from the atmospheric analyses used for weather forecasting but truncated to the lower resolution used in the coupled model. In addition to upper air values, the atmospheric analyses provide initial conditions for soil moisture and snow and ice cover.

The ocean model was a global version of HOPE with a resolution of $0.5^\circ \times 2.8^\circ$ near the equator but lower meridional resolution in the extratropics. The ocean initial conditions were obtained from an ocean analysis system in which all available in-situ thermal data were assimilated. For further details of the ocean data assimilation procedure see Alves et al. (2004) and Balmaseda (2003).

The strategy for creating a forecast ensemble was to run the coupled model out to 6 months ahead starting on successive days within a month, giving 28- to 31-member ensembles. Such forecasts exist from the start of 1997. (In fact, during 1997, computer restraints made it possible to run only 3 times a week but we have recently backfilled these to daily.)

The coupled model is not perfect; in common with all fully-coupled models, it drifts and so the model climatology does not match that of nature. To overcome this, anomalies are calculated with respect to the model climatology which is obtained by running an 11-member ensemble for every month of the years 1991-1996. In January, April, July and October the ensemble was increased to 27 members. In contrast to the forecast phase (1997–2002), all hindcasts for the calibration period started from the first of the month. For these hindcasts, the ensemble was generated by using very small SST perturbations in the initial conditions. The perturbations are negligible in themselves, but because of the chaotic nature of the atmosphere model create substantial spread as the forecast progresses. In order to allow some comparison with System-2 (see later) a further set of integrations was recently performed for the years 1987-1990 when a 5-member ensemble was run for each month of the year. For this set, forecasts were started from 2 days before the start of the month to two days after.

Sea surface temperature (SST), 2-meter temperature ($T_{2\mathrm{m}}$) and sea-level pressure (MSL) have been archived as instantaneous values only at 00Z. The accumulated variables, precipitation (P) and solar radiation (SSR), are daily averages.

Because different strategies were used to generate ensembles and the number of members in an ensemble varies over the period 1987–2001, validation of the system is not trivial. For the 'real-time' forecasts, ensemble-mean monthly-means have been created by averaging the ensemble of daily forecasts starting from the 16th of the previous month to the 15th of the current month. First, monthly mean values were created. A sliding window was used: for each forecast the first

28/30/31 days were averaged to create a monthly mean. For ensemble member 1 this would represent the period from the 16th of one month to the 15th of the next. For ensemble member 2 the averaging period is 17th of one month to the 16th of the next and so on. Then all the monthly mean values are averaged to create a monthly-mean ensemble-mean with a nominal start date of the first of the month. For the calibration period all ensemble members start on the first of the month so monthly means are from the start of the month to the end of the month. For the period 1988–1990, a sliding window is again used but with only 5 members in the ensemble mean. This is a slightly different averaging technique to that which was used to produce the operational web products for S1.

The data can be sorted by lead month (+0, +1, ...) or by nominal starting date. A '1 Jan' forecast was not available until the 26th of January, as the ocean data assimilation system ran 11 days behind real time and the last ensemble member was started from conditions of the 15th Jan. We define the January forecast as month +0, the February forecast as +1, etc. This definition of lead time is consistent with the then-functional ECMWF web site: the number of months between the nominal start date of the forecast and the verification period.

In summary, S1 was a prototype system and subject to different ensemble generation strategies at different periods in its development. This makes the forecasts from S1 quite difficult to use.

2.1.2 System-2

System-2 (S2) was introduced into operational use at the beginning of 2002. It differs from S1 in a number of ways. The atmospheric component is cycle 23r4 of the IFS with a horizontal resolution of T95 and 40 levels in the vertical. The ocean model resolution was increased to $0.3^\circ \times 1.4^\circ$ near the equator and to $1.4^\circ \times 1.4^\circ$ at higher latitudes and the vertical resolution increased from 20 to 29 levels. Changes were also made to the ocean model physics and to the ocean assimilation system. In order to sample some of the uncertainty in the ocean initial conditions, not one, but 5 ocean analyses are performed, from 1987 to present. The different ocean analyses differ in the wind fields used to produce them: perturbations representative of the perceived uncertainty in the wind stress have been added to the ECMWF wind stress.

Patterns of wind stress perturbations were constructed from differences between interannual monthly anomalies of the ERA-15 reanalysis and Southampton Oceanography Centre (SOC) monthly mean wind stresses (Josey et al., 2002), for the period 1980–1997. These differences between two state-of-the-art estimations of wind stress using observations must be representative of the typical uncertainties on the knowledge of the wind stress field. Only the low frequency uncertainty in the wind is important so monthly means are adequate for our purposes. The wind stress perturbations are stratified by calendar month. By linearly interpo-

lating two randomly picked wind stress patterns representative of consecutive months (the full pattern being applied to the middle of each month), daily perturbations can be obtained. These are then used to randomly perturb the daily wind stress that forces the ocean model.

The ocean analysis system consists of an ensemble of five independent ocean analyses, making use of the wind perturbations described above. Member 0 has no wind perturbations applied, members 1 and 2 have the same patterns but of opposite sign, and likewise for members 3 and 4. This method of ensemble generation means that the ensemble-mean winds are not biased relative to the unperturbed member: only a spread is introduced.

The set of ocean analyses is augmented with SST perturbations to create an ensemble of ocean initial conditions to use in the coupled forecasts. We do this in a smilar way to what we did with wind stress: we estimate SST patterns that should be representative of the typical errors in SST products. One set of perturbation patterns has been constructed by taking the difference between 2 different weekly-mean SST analyses (Reynolds OI and Reynolds 2DVAR) from 1985 to 1999 (Reynolds et al., 2002). A second set of SST perturbations has been constructed by taking the difference between Reynolds 2DVAR SSTs and its 1-week persistence. The first set of SST perturbations samples the uncertainties in the SST analysis, whereas the second difference samples the uncertainties due to the fact that the SSTs from NCEP are a weekly-mean product. For each starting date, 2 combinations from these 2 different sets of perturbations are randomly selected and are added to the SSTs produced by the operational ocean analyses with a $+$ and $-$ sign, creating 4 perturbed initial states. The perturbation has full value at the surface but is ramped down to zero at 40m depth. The SST perturbations are not present during the analysis phase, but are added to the initial conditions at the start of a forecast.

Since a burst mode is used for ensemble generation rather than the lagged-average approach, all ensemble members start on the same day. To sample well the effect of different atmospheric forcing on the SST, we need to ensure that the different ensemble members follow a different sequence of synoptic variability after a few days. One way of doing this is to use the so-called 'stochastic physics' (Buizza et al., 1999). The use of stochastic physics is also to represent uncertainties in the parameterisation of subgridscale processes. These parameterisations are meant to represent the average effect of subgrid scale processes on the large-scale flow, but there is also a random component to this effect (e.g., for the same value of the average cloud cover in one model cell, there are many possible vertical and horizontal distributions of the clouds, and thus a range of radiative forcing of the flow). This kind of stochastic forcing is an attempt to take into account these uncertainties in the physical parameterisations by randomly perturbing the atmospheric parameterised physical tendencies at each time step of the model

integration. This introduces a random component in the atmosphere which results in a divergence of synoptic systems in the early range of the forecast. This approach is used in the ECMWF medium range weather ensemble prediction system (Buizza et al., 1999). For further details of the ensemble generation strategy see (Vialard et al., 2003)

The above strategy of creating the ensemble makes it possible to start all forecasts from the 1st of the month. The calibration period is 1987 to 2001 when 5 forecasts, with ocean initial conditions taken from each of the ocean analyses, are made for each month for each year. In May and November these are augmented to 41 by also perturbing SST. The actual forecasts consist of 40 members created for the first of each month, available in real time around the 15th of the month. For further description of S1 and S2, including an assessment of their different characteristics see Anderson et al. (2003).

2.2 The statistical models

A set of simple statistical seasonal forecast models, denoted by STAT, has been developed for comparison with S1 and S2. They were constructed on the basis of observations in the period 1901–1986. The predictors are the persistence and the time series $E_i(t)$ of the first few EOFs of SST in the Kaplan reconstruction (Kaplan et al., 1998). The first EOF describes the main mode of ENSO, the second one has a low-frequency times series and describes the decadal ENSO variability (Zhang et al., 1997). The predictands are fields of SST (Kaplan et al., 1998), T2m (HadCRUT, Jones, 1994; Parker et al., 1994; Jones et al., 1997), SLP (Basnett and Parker, 1997) and precipitation (Hulme et al., 1998). At analysis time t_a the forecast for quantity X at forecast time t_f is simply

$$X(t_f) = p(m_f, m_a)X(t_a) + \sum_{i=1}^{N} a_i(m_f, m_a)E_i(t_a) \qquad (1)$$

where $p(m_f, m_a)$ represents the effect of persistence from calendar month m_a to month m_f and $a_i(m_f, m_a)$ the past effect of EOF i of SST on X, also dependent on the seasonal cycle and lead time.

The model parameters $p(m_f, m_a)$ describing the effect of persistence are obtained from a linear fit to the observations over 1901–1986. The parameters are set to zero at grid points where the fit is not significant at the 2.5% level (one-sided t-test). Next the effect of persistence is subtracted from the observations, and the parameters $a_i(m_f, m_a)$ describing the effect of EOF i are fitted to the resulting fields. For these parameters we demand a (two-sided) significance of $5\%/N$. The number N of EOFs that are taken into account was estimated subjectively by watching from which EOF onwards the teleconnections a_i start to consist of noise far away from the places where the EOF pattern has large amplitudes, again only

predictand				predictors	
variable	dataset	resolution	m	p	N
SST	Kaplan	$5 \times 5°$	1,3	p	20
T2m	HadCRUT	$5 \times 5°$	1,3	p	1
SLP	UKMO	$5 \times 5°$	1,3	p	5
prcp	Hulme	$3.75 \times 2.5°$	1,2,3,4	-	2

Table 1: The parameters of the statistical seasonal forecast model STAT. Persistence is indicated by p, the number of EOFs used by N, the number of months in a season by m. Precipitation models have also been constructed for 2- and 4-month seasons as these often correspond to local wet or dry seasons. After 1994 SLP persistence was taken from the NCEP/NCAR reanalysis.

looking at data prior to 1986. This way, the risk of overfitting is decreased at the expense of reduced skill due to precursors that are not included.

Separate forecast models were made for multi-month seasons instead of summing the monthly forecasts. Both predictor and predictand are summed over the same number of months. The lead time in this case is the number of months between the last month of the predictor season and the first month of the predictand season. The model parameters are given in Table 1. The statistical models are available from the authors.

In the case of forecasts of Niño indices, the forecasts were also compared to a simple damped persistence model CLIPER, which corresponds to the statistical model described above with $N = 0$ and the same Niño index for both predictor and predictand. It is trained on all data up to and including the analysis date.

Finally, we considered a set of proven, more elaborate, statistical models which include multivariate predictors. The first is the ENSO-CLIPER model of Landsea and Knaff (2000), which was downloaded from their web site. Although they discuss results from other regions in their paper, the model is only made available for 3-month averaged Niño3 index forecasts. It is trained on data from 1954 to 1994, so the training period has some overlap with the verification period 1987–2001. As indicated before, the STAT and CLIPER model are totally independent of the verification period.

Also included are the Markov model by Xue et al. (2000) and Constructed Analogue (CA) model by van den Dool (1994); van den Dool and Barnston (1994), which are in operational use at the National Centers for Environmental Prediction, USA. The Markov model projects sea level, SST and zonal wind in the tropical Pacific on three combined EOFs, and describes their evolution by monthly 3×3 matrices fitted over 1980–1995. The CA model fits the observed global SST evolution over the last year as a weighted sum of the years 1956–2001 with posi-

tive and negative weights. Its forecast is the same sum of the observed evolution of these years. The monthly values have been reconstructed from the seasonal forecasts. The forecasts from these two models is available on the 5–8th of the month, earlier than the GCM forecasts. The hindcasts are cross-validated, i.e., for each year a model is constructed on a training period that leaves out that year.

3 Skill in the prediction of ENSO indices

One of the most important tests of any coupled model is to see how well it predicts El Niño and how the skill compares with that of the various statistical models. Widely-used indicators of ENSO conditions are the SST indices for certain regions in the equatorial Pacific. The Niño3 and Niño3.4 indices are created by averaging SST anomalies over the boxes 5°S–5°N, 150°–90°W and 5°S–5°N, 170°–120°W respectively.

The predictions for S1 and S2 are shown for leads of +1, +3 and +5 months in Figs 1a and 1b. The observed SST is also plotted. For comparison, results from the operational statistical Markov and CA models are shown in Figs 1c and 1d. Hindcasts from the simple STAT model are shown in Fig. 1e.

Fig. 1 shows that the 1997 warm event and the following extended cold period were successfully forecast by both the ECMWF models, though S2 underestimated the strength of the warm event. The hindcasts from both coupled models underestimated the 1987 event, possibly due to errors in the ocean initial conditions. The 1988 La Niña was captured. The simple statistical models do not capture the onset of the El Niño and La Niña events. The more sophisticated statistical models have intermediate properties. For instance the CA model fails to capture the onset of El Niño several times, but correctly predicts the following La Niña.

To quantify the performance of the various models, the anomaly correlations for the +1, +3 and +5 month forecasts over the 15 years are given in Table 2. The errors denote the 95% confidence interval computed with a standard bootstrap resampling method (Efron and Tibshirani, 1998) on the correlations in which 800 time series of $180/n$ (model, observed) blocks of n months drawn from the original set with replacement were constructed. The correlation for each of these was calculated. The 95% confidence limit interval is the position of the 20th and 780th value of the sorted correlations. The length of the moving block n was set equal to the decorrelation length of the forecast errors for a given lead time. This is three months for forecasts with lead times +3 and +5, two months for monthly forecasts with smaller lead times, and one month for statistical 3-monthly forecasts at +0. The dynamical 3-monthly +0 forecasts have a decorrelation time

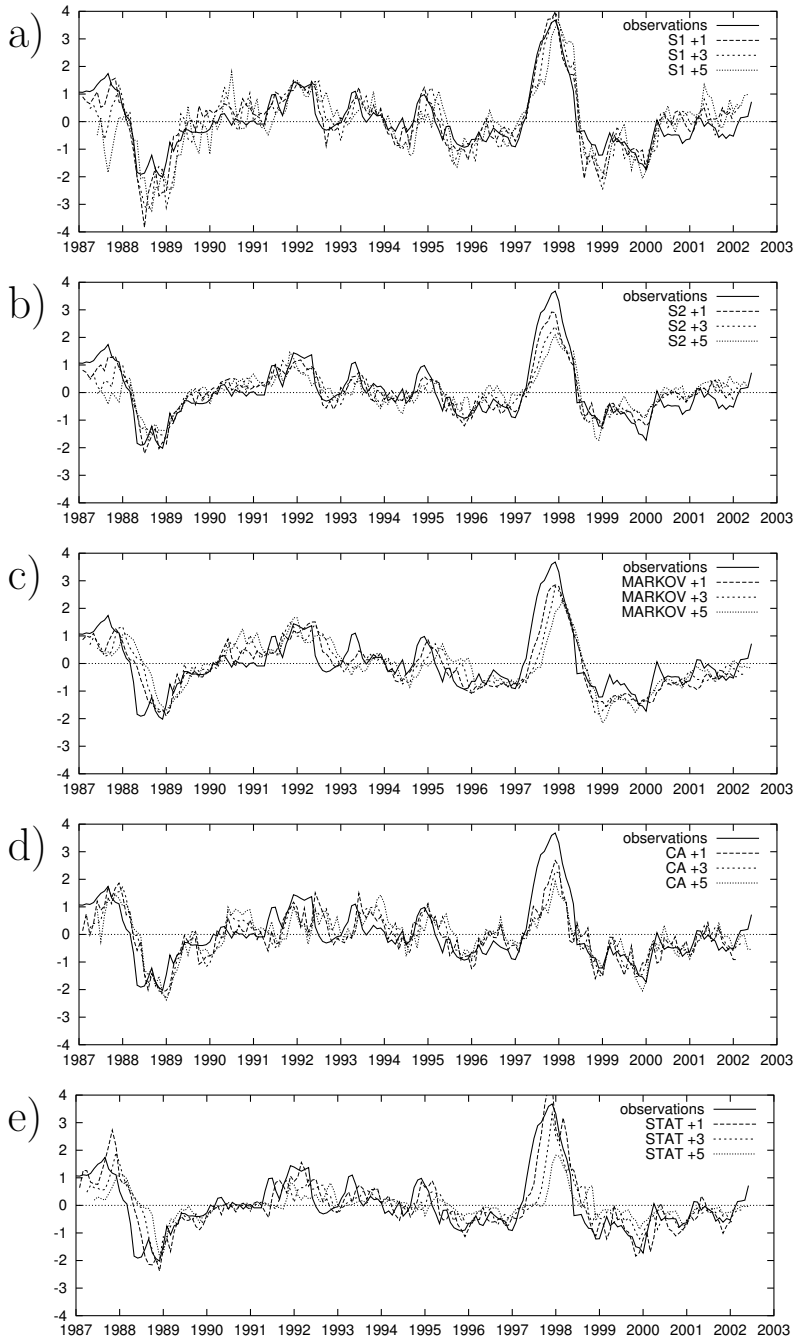

Figure 1: The observed values of the Niño3 index compared with the a) S1, b) S2, c) Markov, d) CA and e) STAT model forecasts at +1, +3 and +5 months.

of 5 months. The error decorrelation is much faster than the decorrelation of ENSO indices.

Confidence limits are included since it is important to be able to judge the difference between the correlation coefficients of the different models against a measure of the uncertainty in these correlation coefficients due to the limited sample sizes. If all the skill were due to just one very big and successful forecast, and the other forecasts were small and unsuccessful, the correlation coefficient could be large, but the uncertainty would also be large since the one good forecast could have been due to sheer luck. This is captured well by the bootstrap method.

The arbitrary verification period 1987–2001 influences the numbers in table 2. In particular, the dynamical systems predicted the 2002/03 El Niño quite well, whereas most statistical models failed to forecast it. The inclusion of this period would have been to the advantage of the dynamical models but this was not done.

Table 2 shows that both dynamical models are better at forecasting the Niño3 index than all the statistical models, including the operational statistical models. The difference is significant at 95% (with a one-sided t test) for monthly forecasts of S2 compared with all statistical models at lead +1 and +3 (probably due to the truncation in EOFs), at lead +5 only compared with damped persistence (CLIPER). In the 3-monthly forecasts the difference between S2 and the statistical models is significant at lead +0, and at +3 with the simple CLIPER and STAT models. At these lead times the chance that one would get a better skill score for S2 by pure chance is less than 5%, for the other dynamical vs statistical comparisons the chance is somewhat higher. In forecasting the more westerly Niño3.4 index the skills of the statistical models are higher than in Niño3, so that at longer lead times the CA model has comparable skill over this period to the dynamical models.

S2 is generally better than S1. At longer leads, S2 has a problem in that the amplitude of its ENSO variability is damped. This is an undesirable feature related to model error, and although its impact on the correlation score is probably modest, it might be related to the lack of advantage of S2 at longer leads. Anderson et al. (2003) give a thorough comparison S1 and S2, and explore the issues involved in the lower amplitude of variability in S2.

A simple multi-model forecast was made by taking the average of the four operational models: S1, S2, CA and Markov. The overall skill is close to the best.

While the seasonally-averaged skill is a useful overall indication, it is quite likely that skill is a function of the time of year and this seasonal variability may differ from model to model. To illustrate this and to intercompare the seasonal dependence of the skill of the different models, we show in Fig. 2 the correlation coefficient of the monthly and 3-monthly Niño3 indices at lead time +3 as a function of the *target* season. In these figures we plot the dynamical models and the operational statistical models, CA, MARKOV and, for the 3-month averaging,

	monthly Niño3			3-monthly Niño3	
lead	+1	+3	+5	+0	+3
S1	0.91^{+3}_{-5}	0.84^{+6}_{-11}	0.73^{+12}_{-18}	0.93^{+3}_{-5}	0.81^{+8}_{-15}
S2	0.94^{+2}_{-4}	0.87^{+5}_{-9}	0.76^{+9}_{-16}	0.95^{+2}_{-4}	0.84^{+6}_{-12}
MARKOV	0.87^{+5}_{-7}	0.75^{+11}_{-14}	0.63^{+13}_{-19}	0.89^{+3}_{-3}	0.72^{+10}_{-16}
CA	0.83^{+5}_{-8}	0.76^{+8}_{-11}	0.70^{+9}_{-12}	0.86^{+4}_{-4}	0.75^{+8}_{-10}
STAT	0.88^{+5}_{-8}	0.75^{+12}_{-18}	0.61^{+16}_{-22}	0.85^{+5}_{-5}	0.63^{+16}_{-20}
CLIPER	0.88^{+5}_{-9}	0.72^{+13}_{-20}	0.56^{+17}_{-24}	0.83^{+6}_{-7}	0.57^{+18}_{-24}
ENSO-CLIPER				0.85^{+4}_{-6}	0.73^{+9}_{-15}
MULTI-MODEL	0.93^{+3}_{-3}	0.87^{+5}_{-9}	0.79^{+8}_{-12}	0.95^{+1}_{-2}	0.85^{+5}_{-10}
	monthly Niño3.4			3-monthly Niño3.4	
lead	+1	+3	+5	+0	+3
S1	0.93^{+2}_{-4}	0.87^{+5}_{-8}	0.79^{+9}_{-15}	0.94^{+3}_{-3}	0.84^{+7}_{-10}
S2	0.95^{+1}_{-2}	0.88^{+4}_{-8}	0.77^{+9}_{-14}	0.96^{+2}_{-2}	0.84^{+6}_{-11}
MARKOV	0.91^{+3}_{-4}	0.81^{+7}_{-8}	0.73^{+9}_{-14}	0.91^{+2}_{-2}	0.78^{+8}_{-9}
CA	0.91^{+3}_{-4}	0.86^{+5}_{-7}	0.78^{+7}_{-10}	0.93^{+2}_{-3}	0.84^{+5}_{-8}
STAT	0.91^{+3}_{-5}	0.79^{+8}_{-12}	0.65^{+12}_{-16}	0.88^{+3}_{-4}	0.67^{+12}_{-15}
CLIPER	0.91^{+3}_{-5}	0.77^{+9}_{-13}	0.61^{+14}_{-16}	0.87^{+5}_{-5}	0.63^{+13}_{-17}
ENSO-CLIPER				0.90^{+3}_{-4}	0.81^{+6}_{-8}
MULTI-MODEL	0.96^{+1}_{-2}	0.90^{+4}_{-5}	0.84^{+6}_{-10}	0.96^{+1}_{-1}	0.88^{+5}_{-6}

Table 2: Anomaly correlation coefficients of the S1 and S2 forecasts of monthly and 3-monthly Niño3 and Niño3.4 compared with MARKOV, CA, STAT, CLIPER, and ENSO-CLIPER models. MULTI-MODEL denotes a simple average of the operational models S1, S2, CA and MARKOV. Results from the multivariate ENSO-CLIPER model are only available for a lead time of +0 and +3 months and for 3-month averages. 0.91^{+3}_{-5} denotes that the 95% confidence interval is 0.86 to 0.94, this interval has been computed with a bootstrap method with moving block lengths equal to the forecast error decorrelation length.

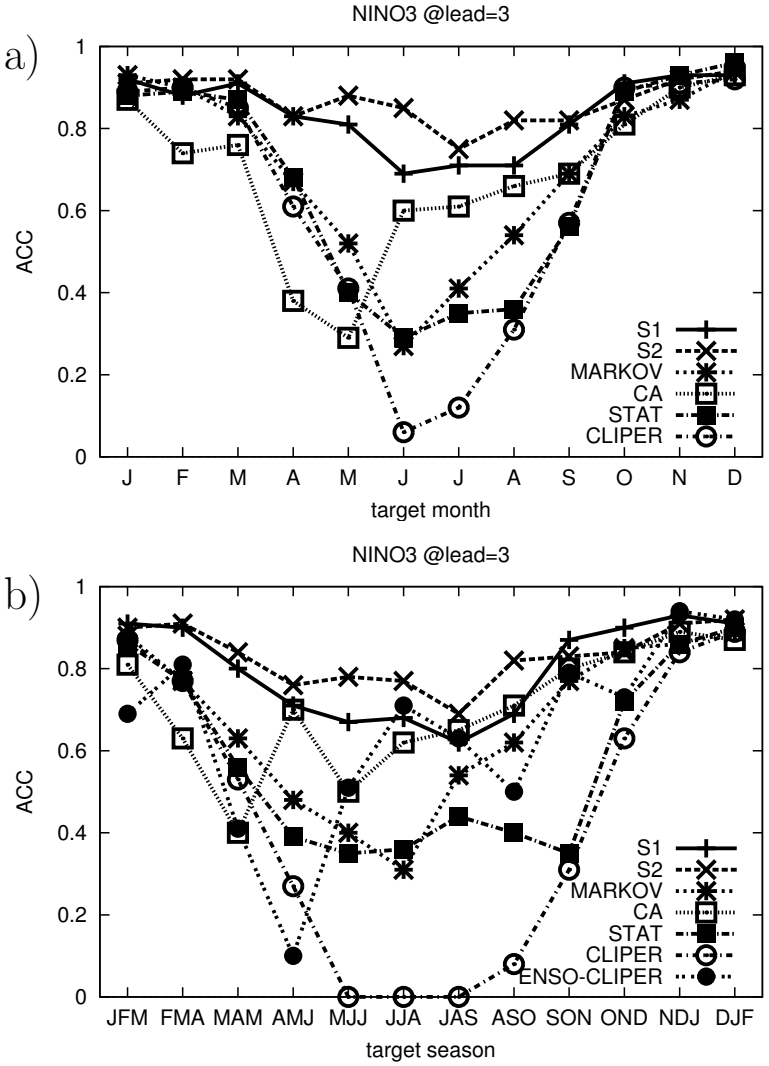

Figure 2: The skill in predicting the a) monthly and b) 3-monthly Niño3 index at a lead time of +3 months.

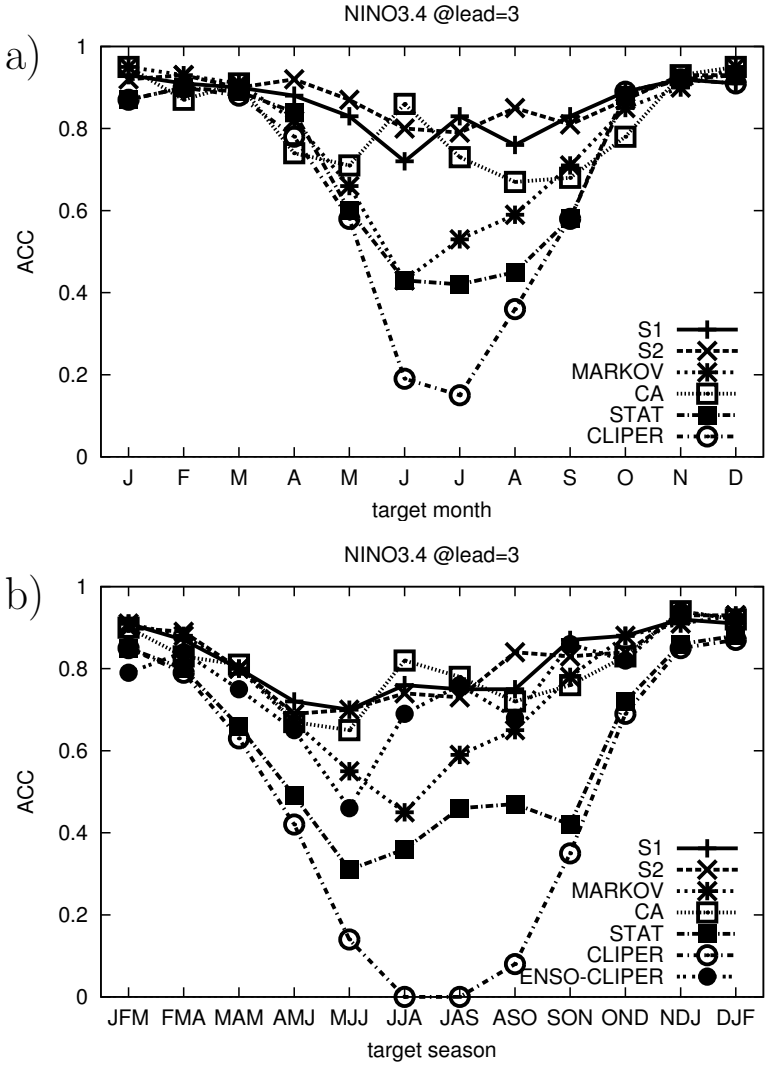

Figure 3: The skill in predicting the a) monthly and b) 3-monthly Niño3.4 index at a lead time of +3 months.

the ENSO-CLIPER (this is the only range for which results are available).

Two things stand-out immediately in these figures. All the statistical models have a marked seasonality in skill, which is much less evident in the dynamical models. The simple statistical models show the spring persistence barrier as the very low predictability of Jun–Aug Niño3 (Webster, 1995; Balmaseda et al., 1995); for the SOI this was already noted by Walker (1924). The CLIPER three-month forecast has skill of less than zero for the June prediction from March, and the skill is even below 0.2 in forecasting June from May (not shown). The MARKOV, STAT and CA models also have a strong seasonality in skill, though not as marked as CLIPER. However, the predictions from the statistical models are good for the winter months and CLIPER, the worst for boreal summer, is better than CA and MARKOV and ranks with the best for winter forecasts (Fig. 2a). The other, perhaps surprising feature is that the 'spring predictability barrier' does not occur at the same time in the different statistical models: the minimum is in June for CLIPER, MARKOV and STAT but in May for CA.

Similar comments apply to Fig. 2b showing 3-monthly forecasts, except that one can now include the ENSO-CLIPER model. This has poor performance in AMJ whereas the other statistical models have poor performance in JJA. The variability in skill from one start date to the next shown by ENSO-CLIPER is surely an undesirable feature. Whether this indicates some over-fitting of the model parameters has not been determined.

Although the NINO3 region has been widely used as an indicator for ENSO, there was a move in the 90's to introduce another region Niño3.4 (Barnston et al., 1997). This was motivated in part by the perception that forecasts were more skillful for this region. We show the results for the various models for Niño3.4 in Fig 3. This should be compared with Fig. 2. As expected, almost all models are more skillful in Niño3.4 than in Niño3, but this especially true for the CA, which over this period is comparable to the dynamical systems in 3-monthly forecasts. The seasonality of the ENSO-CLIPER model is reduced in Niño3.4 but it is still different to that of the other statistical models.

The relatively small seasonality in the skill of the dynamical models is probably in part due to the assimilation and propagation of subsurface oceanic information: a Kelvin wave takes about two months to cross the Pacific Ocean, and slower oceanic processes give some skill beyond that time. A statistical model using subsurface information also has better skill in crossing the spring barrier than the simple SST-based statistical models such as STAT and CLIPER (Balmaseda et al., 1994, 1995; Xue et al., 2000; McPhaden, 2003). Subsurface observations are not used as predictors in STAT because of the poor quality before the TAO array was deployed in the early 1990s. The CA model captures some of this information by fitting to the SST evolution of the past year.

A multi-model average could combine the strong points of the dynamical

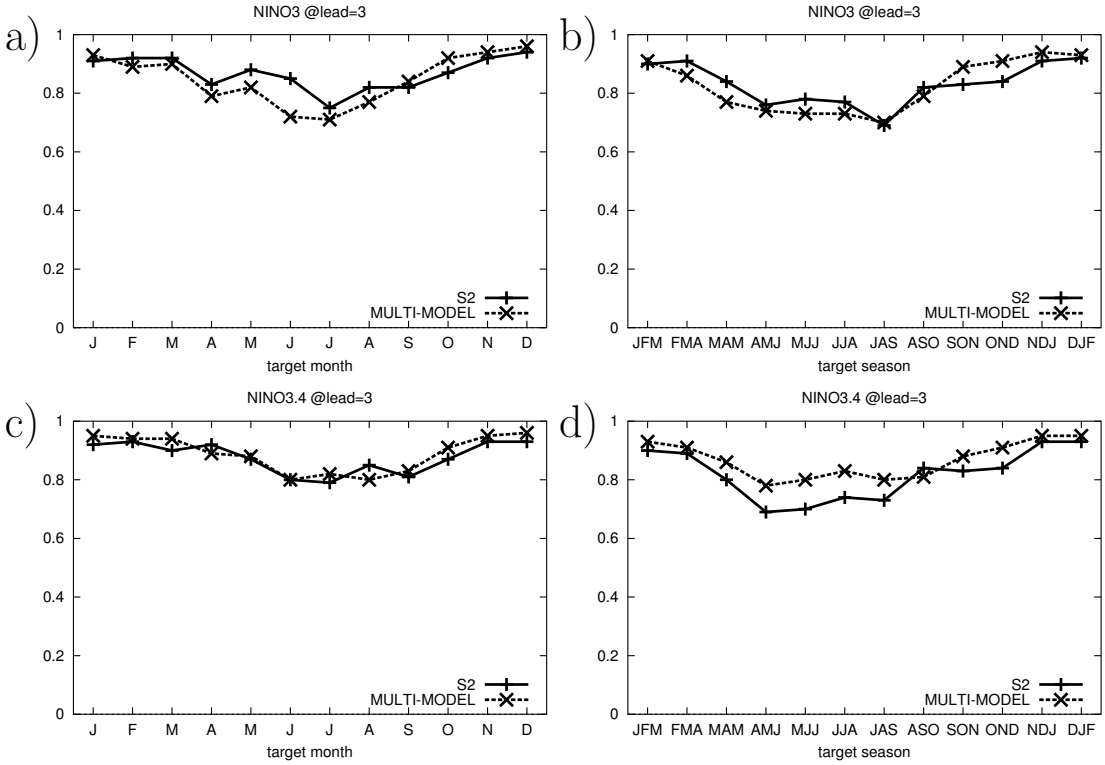

Figure 4: The skill in predicting the a) monthly and b) 3-monthly Niño3, and c) monthly and d) 3-monthly Niño3.4 indices at a lead time of +3 months.

and statistical models. The seasonality of the correlation coefficient for the multi-model and S2 is shown in Fig. 4 for the Niño3 and Niño3.4 indices; this can be compared with Figs 2 and 3. Panels a) and b) show that S2 is quite hard to beat for the Niño3 region, as the statistical models have less skill there. Panel c) shows that the multi-model and S2 are pretty much equivalent. Panel d) shows that for 3-monthly forecasts the multi-model has less seasonality in skill than S2, clearly beating S2 in Niño3.4 at this lead time.

4 Conclusions

We have compared the skill of the ECMWF seasonal forecast models over 1987–2001 starts to that of a set of statistical forecast models in predicting ENSO. The anomaly correlation coefficient is used as the skill measure, since it is not affected by the biases in the mean state and amplitude of the variability. In this article the first-order forecast, the ensemble mean, is used.

The ECMWF seasonal forecast models have proven to be good El Niño prediction systems. Over 1987–2001 starts the yearly-averaged anomaly correlations are significantly higher than those of the simple statistical models considered, and higher than those of operational statistical models. However, this is an average of two very different regimes. The skill of the dynamical models is higher than the skill of the statistical models during the spring barrier: the onset of El Niño or La Niña is forecast better. In fact, S1 predicted the start and amplitude of the 1997-98 event very well most of the time, in real forecast mode, although the model still underestimated the explosive growth during spring 1997. S2 is more damped than S1, but has higher correlation scores. However, once an El Niño event is established in boreal summer, statistical forecasts are already quite good, and model errors prevent the GCMs giving a better forecast than the statistical ones, especially at longer lead times.

Of course, the opposition of dynamical and statistical methods in seasonal forecasting is a false antithesis. While a purely statistical/empirical approach is feasible, despite its limitations, a dynamical model-based forecast needs calibration and interpretation. Further, the move towards multi-model forecasting provides a context for combining the better aspects of both statistical and dynamical models to produce more accurate and reliable forecasts than either class of model alone. A simple multi-model forecast is the average of the four operational models: S1, S2, CA and Markov. The skill scores are indeed the highest, though not convincingly better than S2. More sophisticated methods to combine forecast models with different characteristics would probably yield even better results.

Acknowledgements We would like to thank Huug van de Dool and Yan Xue for providing hindcasts and useful comments.

References

Alves, J. O. S., M. A. Balmaseda, D. L. T. Anderson, and T. N. Stockdale, 2004: Sensitivity of dynamical seasonal forecasts to ocean initial conditions. *Quart. J. Roy. Meteor. Soc.*, **130,** 647–668. See also ECMWF Technical Memorandum 369.

Anderson, D. L. T., T. Stockdale, M. A. Balmaseda, L. Ferranti, F. Vitart, P. Doblas-Reyes, R. Hagedorn, T. Jung, A. Vidard, A. Troccoli, and T. Palmer, 2003: Comparison of the ECMWF seasonal forecast systems 1 and 2, including the relative performance for the 1997/8 El Niño. Technical Memoranda 404, ECMWF, Shinfield Park, Reading, U.K.

Anderson, J., H. van den Dool, A. G. 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. Amer. Met. Soc.*, **80,** 1349–1362.

Balmaseda, M. A., 2003: Ocean data assimilation for seasonal forecasts. In *ECMWF seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean*, pages 301–325, ECMWF, Shinfield Park, Reading, U.K.

Balmaseda, M. A., D. L. T. Anderson, and M. Davey, 1994: ENSO prediction using a dynamical model coupled to a statistical atmosphere. *Tellus*, **46A,** 497–511.

Balmaseda, M. A., M. K. Davey, and D. L. T. Anderson, 1995: Seasonal dependence of ENSO prediction skill. *J. Climate*, **8,** 2705–2715.

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

Barnston, A. G., Y. He, and M. H. Glantz, 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. Amer. Met. Soc.*, **80,** 217–244.

Basnett, T. A. and D. E. Parker, 1997: Development of the global mean sea level pressure data set GMSLP2. Climatic Research Technical Note 79, Hadley Centre, Meteorological Office, Bracknell, U. K. 16pp plus Appendices. Data are available from www.cru.uea.ac.uk/cru/data/pressure.htm.

Buizza, T., M. J. Miller, and T. N. Palmer, 1999: Stochastic simulation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125,** 2887–2908.

Efron, B. and R. J. Tibshirani, 1998: *An introduction to the bootstrap*. Chapman and Hall.

Hulme, M., T. J. Osborn, and T. C. Johns, 1998: Precipitation sensitivity to global warming: Comparison of observations with HadCM2 simulations. *Geophys. Res. Lett.*, **25,** 3379–3382. Available from www.cru.uea.ac.uk/cru/data/.

Jones, P. D., 1994: Hemispheric surface air temperature variations: a reanalysis and an update to 1993. *J. Climate*, **7,** 1794–1802. Data are available from www.cru.uea.ac.uk/cru/data/temperature.

Jones, P. D., T. J. Osborn, and K. R. Briffa, 1997: Estimating sampling errors in large-scale temperature averages. *J. Climate*, **10,** 2548–2568.

Josey, S. A., E. C. Kent, and P. K. Taylor, 2002: On the wind stress forcing of the ocean in the SOC climatology : Comparisons with the NCEP/NCAR, ECMWF, UWM/COADS and Hellerman and Rosenstein datasets. *J. Phys. Oceanogr.*, **32,** 1993–2019.

Kaplan, A., M. A. Cane, Y. Kushnir, A. C. Clement, M. B. Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856–

1991. *J. Geophys. Res.*, **103,** 18567–18589. Data are available from ingrid.ldgo.columbia.edu.

Landsea, C. W. and J. A. Knaff, 2000: How much skill was there in forecasting the very strong 1997-98 El Niño? *Bull. Amer. Met. Soc.*, **81,** 2107–2119.

Latif, M., K. Sperber, J. Arblaster, P. Braconnot, D. Chen, A. Colman, U. Cubasch, C. Cooper, P. Delecluse, D. De Witt, L. Fairhead, G. Flato, T. Hogan, M. Ji, M. Kimoto, A. Kitoh, T. Knutson, H. Le Treut, T. Li, S. Manabe, O. Marti, C. Mechoso, G. Meehl, S. Power, E. Roeckner, J. Sirven, L. Terray, A. Vintzileos, R. Voß, B. Wang, W. Washington, I. Yoshikawa, J. Yu, and S. Zebiak, 2001: ENSIP: the El Niño simulation intercomparison project. *Climate Dyn.*, **18,** 255–276.

McPhaden, M., 2003: Tropical Pacific Ocean heat content variations and ENSO persistence barriers. *Geophys. Res. Lett.*, **30,** 1480.

Palmer, T. N., A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. Délécluse, M. Déque, E. Diez, F. J. Doblas-Reyes, H. Feddersen, R. Graham, S. Gualdi, J.-F. Guérémy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A. P. Morse, B. Orfila, P. Rogel, J.-M. Terres, and Thomson M. C., 2004: Development of a european multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). *Bull. Amer. Met. Soc.*, **85,** 853–872.

Parker, D. E., P. D. Jones, A. Bevan, and C. K. Folland, 1994: Interdecadal changes of surface temperature since the late 19th century. *J. Geophys. Res.*, **99,** 14373–14399. Data are available from www.cru.uea.ac.uk/cru/data/temperat.htm.

Peng, P., A. Kumar, A. G. Barnston, and L. Goddard, 2000: Simulation skills of the SST-forced global climate variability of the NCEP-MRF9 and Scripps-MPI ECHAM3 models. *J. Climate*, **13,** 3657–3679.

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

Sardeshmukh, P. D., G. P. Compo, and C. Penland, 2000: Changes of probability associated with El Niño. *J. Climate*, **13,** 4268–4286.

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 den Dool, H. M., 1994: Searching for analogues, how long must one wait? *Tellus*, **46A,** 314–324.

van den Dool, H. M. and A. G. Barnston, 1994: Forecasts of global sea surface temperature out to a year using the Constructed Analogue method. In *Proceedings of the Climate Diagnostics Workshop*, pages 416–419, College Park, MD.

Vialard, J., F. Vitart, M. A. Balmaseda, T. Stockdale, and D. L. T. Anderson, 2003: An ensemble generation method for seasonal forecasting with an ocean-atmosphere coupled model. Technical Memoranda 417, ECMWF, Shinfield Park, Reading, U.K.

Walker, G. T., 1924: Correlation in seasonal variations of weather IX. *Mem. Indian Meteorol. Dep.*, **24,** 275–332.

Webster, P. J., 1995: The annual cycle and the predictability of the tropical coupled ocean-atmosphere system. *Meteorology and Atmospheric Physics*, **56,** 33–55.

Xue, Y., A. Leetmaa, and M. Ji, 2000: ENSO prediction with Markov models: The impact of sea level. *J. Climate*, **13,** 849–871.

Zhang, Y., J. M. Wallace, and B. S. Battisti, 1997: ENSO-like interdecadal variability: 1900–93. *J. Climate*, **10,** 1004–1020.