TECHNICAL MEMORANDUM

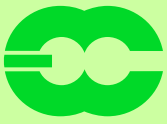
418

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

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

Research Department

September 6, 2003



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

For additional copies please contact

The Library
ECMWF
Shinfield Park
Reading
RG2 9AX
library@ecmwf.int

Series: ECMWF Technical Memoranda

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

Abstract

Since 1997 the European Centre for Medium-Range Weather Forecasts (ECMWF) has made seasonal forecasts 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 a few statistical models, mainly based on past sea surface temperature observations. The relative skill depends strongly on the season. The dynamical models are much 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.

El Niño and La Niña perturb the average weather in many regions and seasons throughout the world. A set of statistical models (STAT) is constructed based on persistence and a lagged regression with an ENSO index over 1901–1986 wherever the correlations are significant. As the number of verification data points is very low (15), the simplest measure of skill is used: the anomaly correlation coefficient of the ensemble mean. To further reduce the sampling uncertainties we restrict ourselves to areas and seasons of known ENSO teleconnections.

The dynamical ECMWF models show better skill in 2-meter temperature forecasts over sea and the tropical land areas than STAT. The modeled ENSO teleconnection pattern to North America is shifted relative to observations, leading to little pointwise skill. Precipitation forecasts of the ECMWF models are very good, better than those of the statistical model, in South-East Asia, the equatorial Pacific and in the Americas in Dec–Feb but in Mar–May the skill is lower there. Overall, S1(S2) show better skill than STAT at lead time 2 months in 29(32) out of 40 regions and seasons of known ENSO teleconnections.

1 Introduction

The use of dynamical models for seasonal forecasting is becoming widespread. In principle, numerical models that represent the dynamics of the atmosphere and ocean 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 et al., 2003), 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; Anderson et al., 1999).

In this paper we compare the seasonal forecasting performance of two state-of-the-art 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 and next for seasonal forecasts of global fields of surface air temperature, mean sea level pressure and precipitation.

To keep the length of the paper manageable, most of the assessment is based on comparison of maps of temporal correlation between ensemble-mean forecast and analyzed values. The use of correlation as a measure of forecast skill does not tell the full story: it ignores any problems of bias and scaling in the model forecasts and problems with the estimated probabilities (e.g., under/over confidence), although any such problems can in principle be dealt with by appropriate post-processing of the model output. It provides a first-order estimate of the forecast skill by ignoring more refined methods of signal detection based on probability thresholds adjusted to cost/loss properties of a specific application (Palmer, 2002). This is particularly relevant in mid-latitudes, where predictability based on the ensemble mean is often low, and skill is better assessed by looking at the

Relative Operative Characteristics (ROC) (Wilks, 1995; Mason, 2003). However, a shift in the ensemble mean is still a useful indicator of possible skill, and it is a robust measure with relatively small uncertainties on the limited number of verification years available.

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 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 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. Another problem is that investigating a large number of independent regions and seasons implies that the chance of ascribing 'good forecast skill' to a random fluctuation is also high.

These are generic problems with verifying seasonal forecasts, particularly in regions where the ratio of predictable 'signal' to unpredictable 'noise' is low. We will give 95% confidence intervals on the correlation coefficients wherever practicable to indicate the uncertainty. We also restrict our comparisons to areas and seasons of known ENSO teleconnections in order to minimize the chances of being misled by random fluctuations. Finally, it is possible that there is true low-frequency variability in teleconnections and seasonal predictability, beyond the apparent variability inherent in a noisy system. This will be investigated using a Monte Carlo approach. We try to be moderate in our discussion of the results of this paper, but the reader is advised to be cautious in interpreting apparent differences in the plotted forecast performance of the various systems, as the differences are only significant when many cases are considered together.

All maps in the paper were obtained from the KNMI Climate Explorer web site (climexp.knmi.nl), which allows the reader to investigate regions and seasons not covered in this paper.

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. This is followed by the verification of the seasonal forecasts themselves: predictions of 2 metre temperature fields in section 4.1, sea level pressure in 4.2 and precipitation fields in 4.3. This is mostly in the form of global maps, although precipitation is discussed regionally and seasonally because of the smaller scales of variability and the strong dependence of skill on the season. We consider both the apparent correlation skill of the models, and their ability to reproduce observed teleconnections of ENSO, on which most of the skill is based. Section 5 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 Balmaseda et al. (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. 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- or 27-member ensemble for every month of the years 1991-1996. In January, April, July and October 27 members were started. The other months had an 11-member ensemble. In contrast to the forecast phase (1997–2002), all hindcasts for the calibration period started from the first of the month.

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_{2m}) 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 used to produce the operational web products.

The data can be sorted by lead month (+0, +1, ...) or by nominal starting date. A '1 Jan' forecast is 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 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 TL95 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 (Vialard et al., 2003).

The set of ocean analyses is augmented with SST perturbations to create an ensemble of ocean initial conditions to use in the coupled forecasts. In addition, stochastic physics is used to perturb the coupled integrations throughout the forecast period. This 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. For further description of S1 and S2, including an assessment of their different characteristics see Anderson et al. (2003).

In S2 the daily mean, maximum and minimum temperatures are stored. The other variables are averages of the 00Z and 12Z values. As in S1, the accumulated variables, precipitation and solar radiation, are daily averages. S2 forecasts are available around the 15th of the month (the ocean analysis runs 11 days behind real time and then a few days are needed to run the 40-member ensemble of forecasts out to 6 months).

2.2 The statistical models

A set of 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 of the predictand 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 ENSO, the second one decadal ENSO. 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

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

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 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.

predictand				predictors	
variable	dataset	resolution	*m*	p	*N*
SST	Kaplan	$5 \times 5^\circ$	1,3	p	20
T2m	HadCRUT	$5 \times 5^\circ$	1,3	p	1
SLP	UKMO	$5 \times 5^\circ$	1,3	p	5
prcp	Hulme	$3.75 \times 2.5^\circ$	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.

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 Niño index forecasts, 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 took the ENSO-CLIPER of Landsea and Knaff (2000) as a reference of a proven, more elaborate, statistical model which includes multivariate predictors. This model 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.

2.3 Data for verification

Most forecasts are compared with the NCEP/NCAR reanalysis (Kalnay et al., 1996), which contains the NCEP OI SST analysis (Reynolds and Smith, 1994). For precipitation the GPCP datset is used (Huffman et al., 1995).

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, in particular Niño3. The Niño3 index is created by averaging SST anomalies over the box 5°S–5°N, 150°–90°W. 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 statistical model STAT are shown in panel c. The forecasts from STAT are very similar to those of the CLIPER and ENSO-CLIPER models which are not shown.

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. Towards the end of the record there is a warm bias in the forecasts of the dynamical models. The statistical models do not capture the onset of the El Niño and La Niña events.

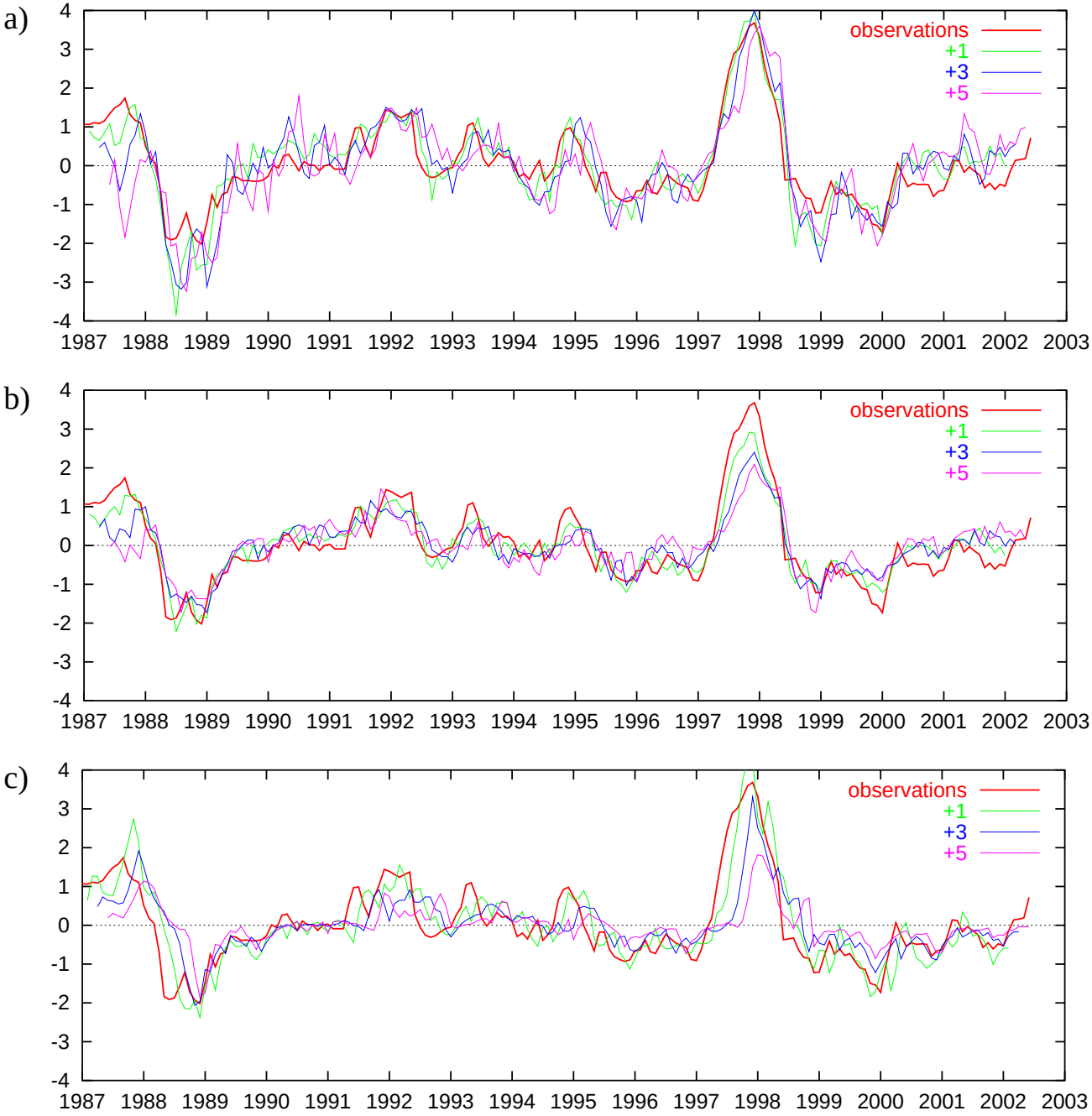


Figure 1: The observed values of the Niño3 index compared with a) the S1 forecasts, b) the S2 forecasts and c) STAT forecasts at +1, +3 and +5 months.

	monthly Niño3			3-monthly Niño3	
lead	+1	+3	+5	+0	+3
S1	0.91^{+2}_{-3}	0.84^{+5}_{-6}	0.73^{+9}_{-11}	0.93^{+2}_{-2}	0.81^{+6}_{-8}
S2	0.94^{+2}_{-3}	0.87^{+4}_{-5}	0.76^{+7}_{-10}	0.95^{+2}_{-3}	0.84^{+4}_{-7}
STAT	0.88^{+4}_{-6}	0.75^{+8}_{-10}	0.61^{+11}_{-13}	0.97^{+1}_{-1}	0.79^{+6}_{-10}
CLIPER	0.88^{+4}_{-6}	0.72^{+10}_{-12}	0.56^{+12}_{-16}	0.95^{+2}_{-2}	0.75^{+8}_{-11}
ENSO-CLIPER				0.84^{+5}_{-5}	0.73^{+7}_{-9}

Table 2: Anomaly correlation coefficients of the S1 and S2 forecasts of monthly and 3-monthly Niño3 compared with STAT, CLIPER and ENSO-CLIPER models. Results from the multivariate ENSO-CLIPER model are only available for a lead time of +0 and +3 months and for 3-month averages. The errors denote the 95% confidence interval computed with a bootstrap method.

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 on the correlations in which 800 time series of 180 (model, observed) pairs 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. If these pairs are independent this is a good estimate of the uncertainty in the correlation coefficient. One can argue to what extent the condition of independence is satisfied. For observations, they are not independent since El Niño has a decorrelation scale of a few months. However, the model results for different months are from independent forecasts, so in that sense the pairs are independent.

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 statistics and the observed scatter. 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.

For predicting the monthly Niño3 index, both dynamical models are significantly better than the statistical models at +3 and +5 months lead time. For 3-month averages, S2 is significantly better than the CLIPER models. S2 is an improvement over S1 in the anomaly correlation coefficient, but S2 is damped relative to S1 and underpredicts the strength of the anomalies. This is discussed further in Anderson et al. (2003). Although the differences in skill between the statistical models are well within the error margins, one sees that the non-local STAT model performs slightly better than the local CLIPER model.

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 anomaly correlation coefficient of the monthly and 3-monthly Niño3 index at lead time +3 as a function of the *target* season for S1, S2 and the statistical models. The latter show the spring barrier as the very low predictability of Jun–Aug Niño3. The CLIPER one-month forecast has skill of only $r = 0.2$ for the June prediction from May (not shown). During the rest of the year, the same model has skill as high as $r = 0.6$ at nine months' lead time, forecasting April from July. The ENSO-CLIPER model has a weaker spring barrier than CLIPER, possibly due to the use of atmospheric information that precedes warming. The predictability minimum shown so clearly in the statistical models is not so evident in the ECMWF models. This gain in predictability 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 (Balmaseda et al., 1994, 1995;

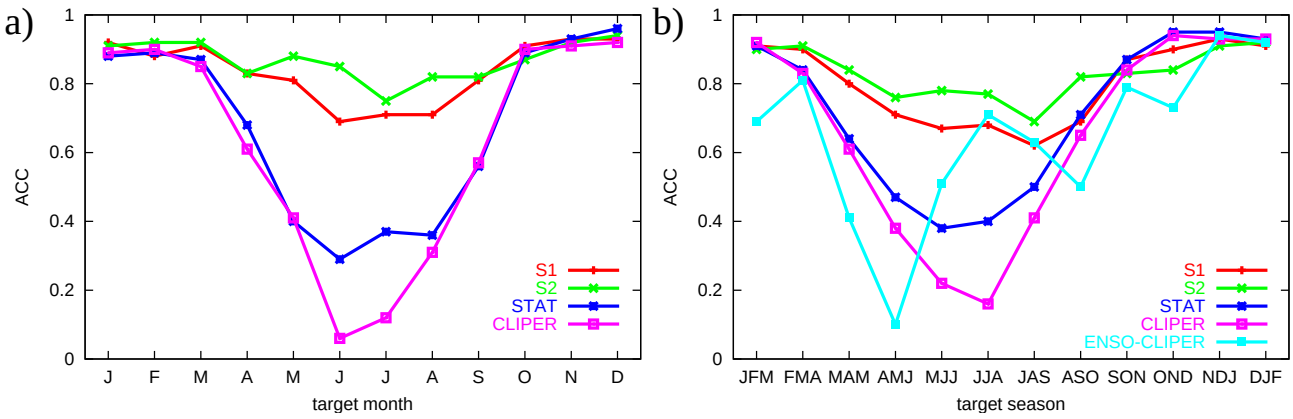


Figure 2: The skill in predicting the a) monthly and b) 3-monthly Niño3 index at a lead time of +3 months.

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.

4 Verification of seasonal forecasts

Next we turn to the quality of the seasonal forecasts: deviations from climatology of the weather averaged over two to four months. Again, the anomaly correlation coefficient will be used as a measure of skill. This skill strongly depends on the location and the season. With only 15 years of data, each combination of location, season and lead time only has 15 independent verifications. To give an idea of the uncertainty in the correlation coefficients, note that for a correlation of zero and gaussian errors, the 95% confidence level is at $r = 0.44$ for the one-sided t-test appropriate for skill estimates. This means that 5% of the area of global maps of correlation coefficients will show anomaly correlations higher than $r = 0.44$ even in the absence of any skill. On the other hand, low values of the correlation in a sample of this size could correspond to relatively high skill

To reduce the number of areas where skill might arise by chance we look specifically at well-known El Niño teleconnections. El Niño is known to be the largest single source of predictable interannual variability. We measure the teleconnections by looking at the simultaneous correlation between Niño3 SST anomalies and the forecast anomalies around the world. This measurement can be made for both the observations and the coupled forecast models. Much of the seasonal climate variability in mid-latitudes is unpredictable, and so the correlation between Niño3 SST and local weather is relatively low. In the case of the model integrations, it would be possible to obviate this by taking the ensemble mean and correlating with the ensemble mean Niño3 SST. In this case the unpredictable 'noise' would be largely removed, and the correlations would be high for those regions where El Niño dominated the predictable part of the variability. Although this is an interesting statistic, it is not one that can be easily compared with observations and it is therefore not used here.

Instead, we calculate the simultaneous correlation for all ensemble members, each with its own model Niño3 index. This is done by constructing two extended timeseries by the juxtaposition of the individual ensemble members. The fact that the model correlations are estimated over many realizations reduces the noise level, but does not change the expected outcome for a single model run. For both S1 and S2, the number of ensemble members available for each year varies, because of the way the ensembles were constructed (see Section 2). We augmented the ensemble sizes to the maximum size (29, 30 or 31 for S1, 41 for S2) by repeatedly and equally sampling the available ensemble members. Thus for S1 in 1987, when only 5 integrations are available and

31 are needed, the first integration is used as a member of the 1st, 6th, 11th, 16th, 21st and 26th ensembles. Where additional members are needed beyond a simple multiple of the available integrations, then randomly chosen integrations are used. In the example given here, the 31st ensemble member would be randomly chosen from the 5 available integrations. Each grid point therefore has 435, 450 or 465 (S1) or 615 (S2) entries. These are, however, not all independent. The observed teleconnections are estimated using the the verification dataset over the 15 years of the verification period.

In the case of large sample sizes and assuming stationarity of ENSO teleconnection, the skill of any statistical model based only on linear correlations with Niño3 SST will be equal to the the absolute value of the observed teleconnections scaled down by the skill in predicting Niño3 SST. The skill of the STAT model is thus expected to bear some similarity to the absolute value of the Niño3 teleconnection pattern. The similarity, however, may not be too close as the STAT model may not be dominated by Niño3, 15 independent realizations is not a large sample size (the 95% confidence level for a two-sided t-test that should be used to asses the significance of the teleconnections, is at $r = 0.51$) and the ENSO teleconnections may not be stationary (Diaz et al., 2001).

4.1 Global 2m temperature fields

First we consider the prediction of temperature outside the El Niño area, which was discussed in section 3, considering land areas as well as ocean. Although a measure of land surface temperature is possible, a more commonly used variable is the air temperature at a height of 2m, denoted by T2m. Over the oceans this temperature is strongly correlated with SST.

Figure 3 shows the skill of the Dec–Feb averaged ensemble-mean temperature prediction from forecasts started on 1st October, measured as the correlation with the T2m for Dec–Feb from the NCEP/NCAR reanalysis. Panel a) shows the results from S1, panel c) from S2 and panel e) from the STAT model. Again, one should keep in mind that every point of the maps of correlation coefficients with observations has data from only 15 years and so the uncertainty in the correlation coefficient is quite large.

In all three models, the skill in the central east Pacific is quite high although parts of the area are blanked out for the STAT model because of a shortage of data. In the north Pacific, too, there is a suggestion of useful skill in the coupled models but this is less apparent in the statistical model. In the Indian ocean the skill of the two coupled models is quite high, but lower in the case of STAT. In the tropical Atlantic the skill of the two coupled models is lower than in the Pacific but still significantly higher than that of STAT. All models have skill in predicting SST in the northern subtropical Atlantic Ocean. They also seem to have skill south of Greenland, most evident in STAT; further analysis shows that this is largely due to persistence. Over land, predictability of T2m is generally lower than over the ocean, reflecting in part the lower heat capacity of land which results in less persistence. Nonetheless there appears to be some skill in the coupled models over parts of South America and of Africa.

In the northern Pacific Ocean and Indian Ocean the skill of both S1 and S2 is higher than expected on the basis of ENSO teleconnections and persistence, which are included in the statistical model. This points to other physics contributing to the skill. The teleconnection in the area extending from the southern Caribbean north-eastwards to Europe (Enfield and Mayer, 1997; Czaja et al., 2002) is simulated well by all three models.

There are sizable ENSO teleconnections to temperature in the Americas at the peak of El Niño in boreal winter (van Loon and Madden, 1981). The South American teleconnections are reproduced quite well by the numerical models (Fig. 3b,d). The statistical model suffers from lack of data in northern South America.

In North America, long-term observations show on average milder winter weather during El Niño along the west coast and the Canada-U.S. border, and warmer weather during La Niña along the Gulf Coast (note the

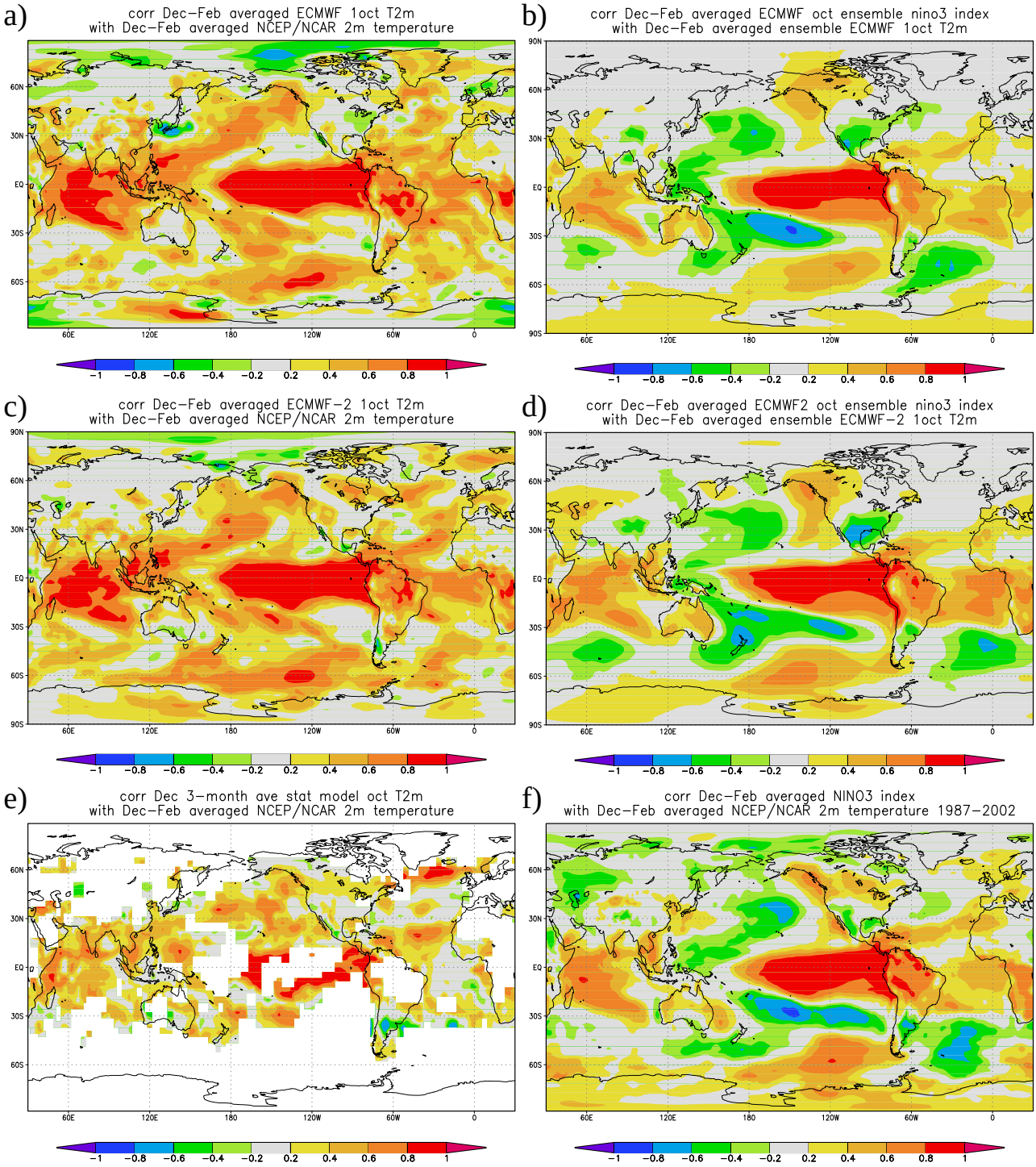


Figure 3: Verification of Dec–Feb T2m forecasts from 1 Oct against the NCEP/NCAR reanalysis. a) Skill of S1, b) S1 Niño3 teleconnection patterns, c), d) same for S2, e) skill of STAT (white points did not have enough data or strong enough teleconnections to merit a prediction), and f) Niño3 teleconnection patterns over 1987–2002.

asymmetry). *Linear* anomaly correlation coefficients are around 0.3 for these teleconnections. They also occurred during the verification period (Fig. 3f). The linear statistical model only shows skill along the coast (Fig. 3e), probably due to the large non-linearity of the teleconnection. Both ECMWF models have correct warm teleconnections in South Alaska (Fig. 3b,d). However, instead of the extension eastwards along the border (Fig. 3f), they show a northwards extension of the teleconnection pattern. The cooling pattern along the Gulf Coast is also situated too far north and is too strong in S2. Due to these teleconnection errors forecast skill is limited to the west coast (Fig. 3a,c). These errors could possibly be corrected for by a downscaling technique, although how much they are due to sampling error is unclear.

4.2 Surface pressure

In the extratropics, the shift of circulation patterns due to El Niño and La Niña is a major source of predictability. The ECMWF models reproduce the overall sense of the winter North American ENSO teleconnection pattern reasonably well (Fig. 4b,d). However, the teleconnections over land appear to be too weak: the center of the Aleution low in the North Pacific is shifted 20–30° too far west, resulting in the poor teleconnections over land that were noted in the T2m forecasts (section 4.1). As in observations, the models have no projection of the ENSO teleconnection patterns onto the NAO: the ensemble correlation for 1 Nov-starts between the Dec–Mar averaged Azores-Iceland NAO and the Niño3 index is $r = -0.10^{+9}_{-9}$ in S1, $r = -0.15^{+8}_{-8}$ in S2. This is in good agreement with the observed value over 1865–1996, $r = -0.05^{+17}_{-18}$.

Dec–Feb sea-level pressure is also predicted well in the area of the Southern Oscillation. Most southern hemisphere teleconnection patterns are also captured.

In contrast, in Jun–Aug the forecast skill is not quite so good (Fig. 5). This may be because the ENSO teleconnections are weaker as panel f) indicates. The predictability is less dominated by ENSO at this time which in turn may be related to the weaker amplitude of El Niño at this time. It is also harder to predict ENSO from April starting dates. Still, the different shapes of modeled and observed Southern Oscillation patterns (Figs 5b,d,f) suggest that model error plays a large role.

Apart from ENSO teleconnections there seems to be some skill in the predictions for eastern North-America and northern and southern Europe in both ECMWF models. The skill in Europe may be connected with the summer predictability discussed in Colman and Davey (1999). There is also good skill ($r > 0.6$) unconnected to ENSO in forecasts of sea-level pressure north-east of New Zealand, especially in S1.

4.3 Precipitation

The spatial decorrelation scales of precipitation are small, making the dangers of 'bump-hunting' correspondingly higher: the chance of finding statistically 'significant' correlations with no physical basis is proportional to the number of independent possible events investigated. As will be shown quantitatively in section 4.3.1, seasonal predictability of precipitation is based mainly on ENSO teleconnections. This is used to restrict the search space: we consider only regions and seasons of known teleconnections (Ropelewski and Halpert, 1987; Kiladis and Diaz, 1989). First the main areas and seasons with ENSO teleconnections are mapped, then for a list of 40 regions and seasons with teleconnections $|r| > 0.4$ over 1901–1986, the skills of STAT, S1 and S2 for 1987–2001 starts are compared.

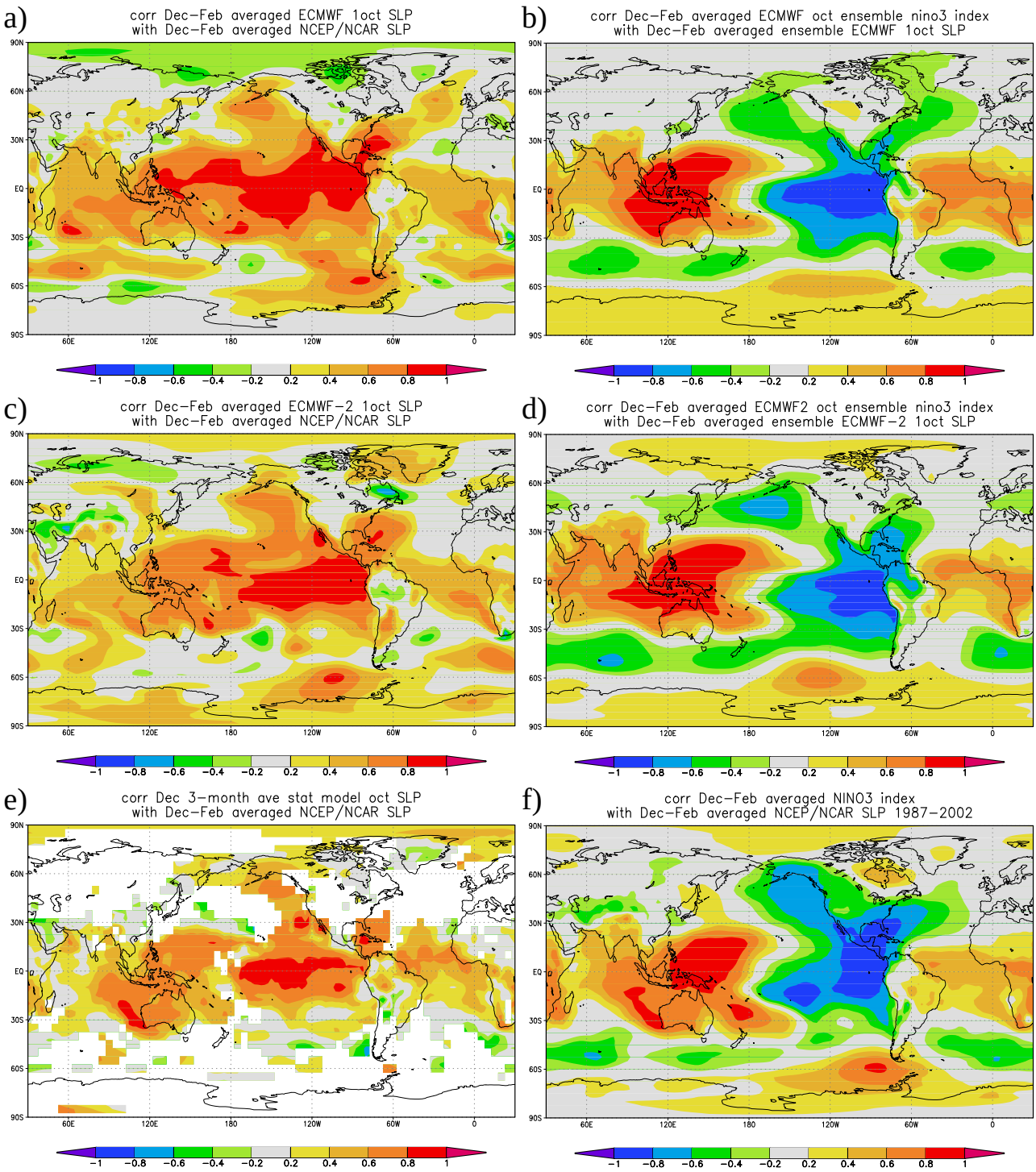


Figure 4: Verification of Dec–Feb SLP forecasts from 1 Oct against the NCEP/NCAR reanalysis. a) Skill of S1, b) S1 Niño3 teleconnection patterns, c), d) same for S2, e) skill of STAT, and f) Niño3 teleconnection patterns over 1987–2002.

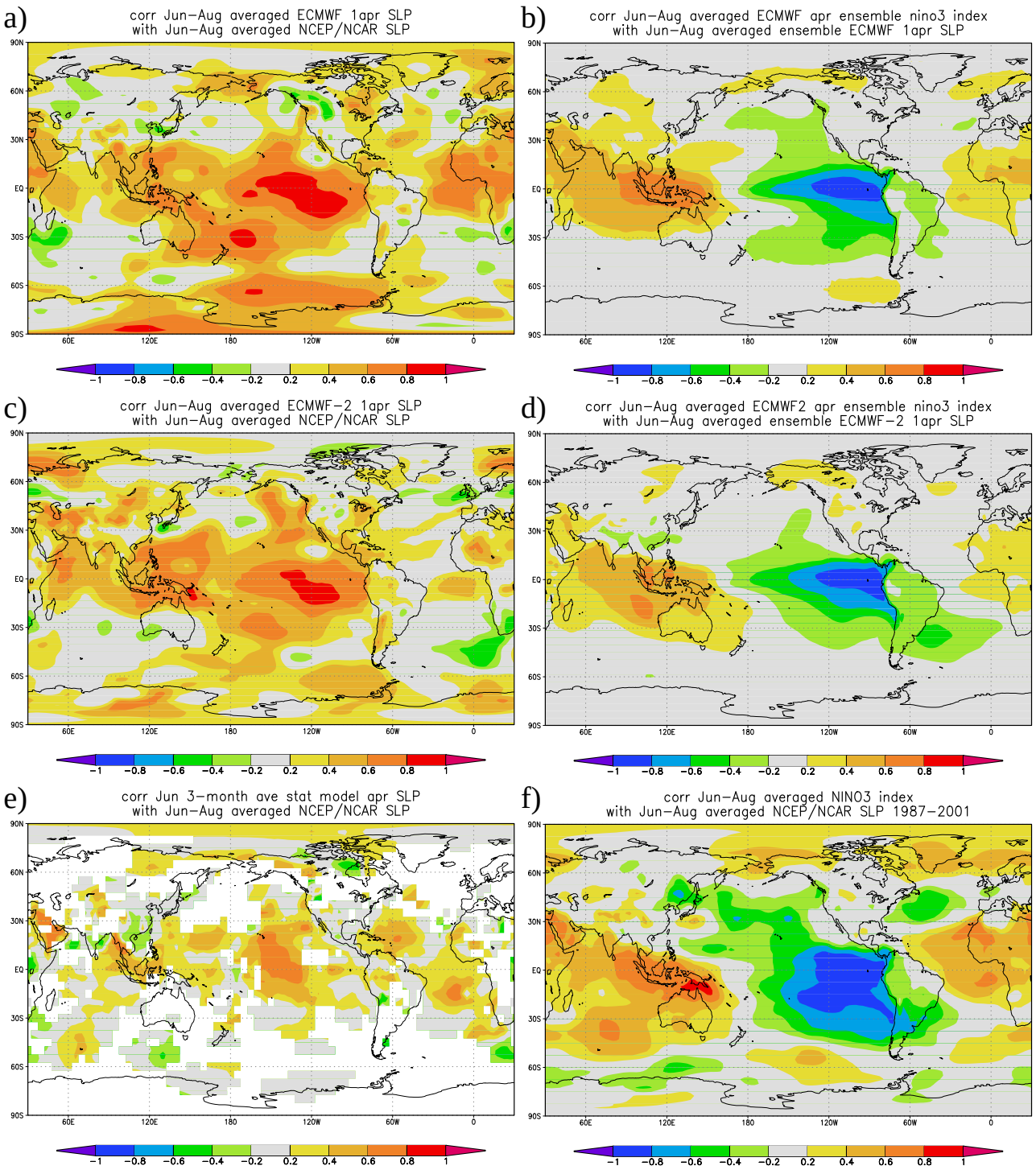


Figure 5: Verification of Jun–Aug SLP forecasts from 1 Apr against the NCEP/NCAR reanalysis. a) Skill of S1, b) S1 Niño3 teleconnection patterns, c), d) same for S2, e) skill of STAT, and f) Niño3 teleconnection patterns over 1987–2002.

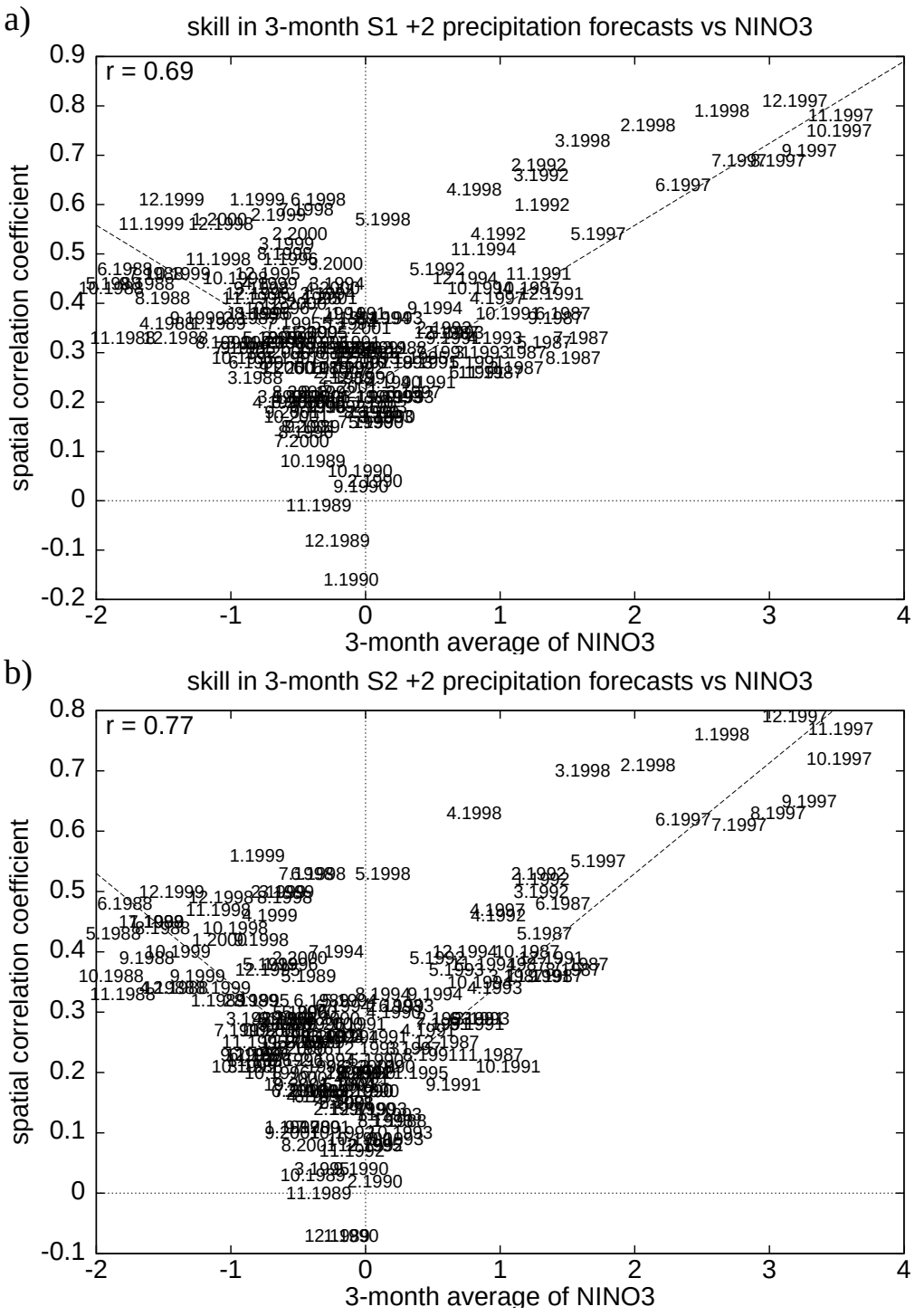


Figure 6: Skill of the +2 month precipitation forecasts (spatial correlation of the global S1 (a) and S2 (b) 3-month average ensemble mean with the GPCP observations) as a function of the value of the Niño3 index. The numbers denote the starting month and year of the 3-month season, the dashed line gives the best fit to the absolute value of Niño3, $r = 0.23 + 0.17|N_3|$ *for S1,* $r = 0.16 + 0.18|N_3|$ *for S2.*

4.3.1 *ENSO-related skill*

The connection between ENSO and the predictability of precipitation can be shown from a plot of the skill of the forecasts against the strength of El Niño. During strong El Niño and La Niña events the skill is higher than during neutral conditions, as shown in Fig. 6. For this plot the skill is the spatial anomaly correlation of the ensemble-averaged forecast with observations. This measure is dominated by high-rainfall regions in the tropics. As expected, the model skill in forecasting precipitation patterns is higher during El Niño and La Niña than during neutral conditions. The excursion to the right represents the large 1997/98 El Niño. Precipitation was not predicted as well during its onset (lower branch) as during its demise (upper branch), as the Niño3 index was underpredicted during the onset and some teleconnections lag El Niño.

A simple model of ENSO-forced precipitation can be constructed by taking precipitation to be proportional to the Niño3 index N_3 plus a noise term ε: $p_{\rm obs} = AN_3 + \varepsilon$. The model prediction should be similar, $p_{\rm mod} = AN_3 + \eta$. In the case of an ensemble average we normally have $\eta^2 \ll (AN_3)^2$: the model noise is reduced by the ensemble averaging. On the other hand, the weather noise is not necessarily small and in places may be larger than the signal. The correlation can be written as

$$r = \frac{\overline{p_{\rm obs} p_{\rm mod}}}{\sqrt{\overline{p_{\rm obs}^2}}\sqrt{\overline{p_{\rm mod}^2}}} = \frac{(AN_3)^2}{\sqrt{(AN_3)^2+\varepsilon^2}\sqrt{(AN_3)^2+\eta^2}} \approx \frac{|AN_3|}{\sqrt{\varepsilon^2}} \tag{2}$$

where the bar denotes spatial averages. If the ENSO signal is smaller than the weather noise, this gives a quasi-linear relationship between correlation and size of El Niño. Motivated by this we attempt a straight line fit to the data, while acknowledging it will not be appropriate for high values of correlation.

Fig. 6 shows that the correlation can be approximated by $r = 0.23 \pm 0.03 + (0.17 \pm 0.03)|N_3|$ for S1 and by $r = 0.16 \pm 0.02 + (0.18 \pm 0.02)|N_3|$ for S2 (the range denotes 2σ symmetrical error estimates). The non-zero constant term shows that there are more factors than ENSO that give rise to predictability that are captured by the ECMWF models.

Limited by the short sample of ENSO events included in the 1987-2001 period it is not possible to infer the existence of asymmetry/symmetry in the skill with respect to El Niño and La Niña. Neither can one exclude the possibility that the seasonal forecast system might present different levels of skill during the positive/negative ENSO phases (Sardeshmukh et al., 2000).

4.3.2 *South-East Asia, Australia and West Pacific*

August–November Globally, the strongest observed effects of ENSO on precipitation occur during the dry season (Aug–Nov) in eastern Indonesia (Berlage, 1957) and the western equatorial Pacific. For example, of 10941 stations in the Global Historical Climate Network (GHCN) v2b with more than 40 years of precipitation data, those with the highest correlation with Niño3 are Banda Island, Indonesia (4.53°S, 129.88°E) with $r = -0.82^{+6}_{-4}$ and Beru, Kiribati (1.40°S, 176.00°E) with $r = +0.82^{+8}_{-13}$ for Aug–Nov rainfall. In eastern Indonesia and the West Pacific this is the dry season, in which drought has large effects. In central and western Indonesia this time of year includes the onset of the monsoon, which starts in September on Sumatra, October on Java and November on Bali. In central Indonesia, the onset is the season most sensitive to ENSO; western Indonesia has no long-term linear ENSO teleconnections in this season.

The ECMWF forecasts from starting dates around 1 Jun (the +2 forecasts) are compared with observations in Fig. 7. One sees that the models are better than both the STAT model and a contemporaneous statistical Niño3 relationship (panel f) in predicting the dry season precipitation anomalies in Indonesia: $r > 0.8$ in the

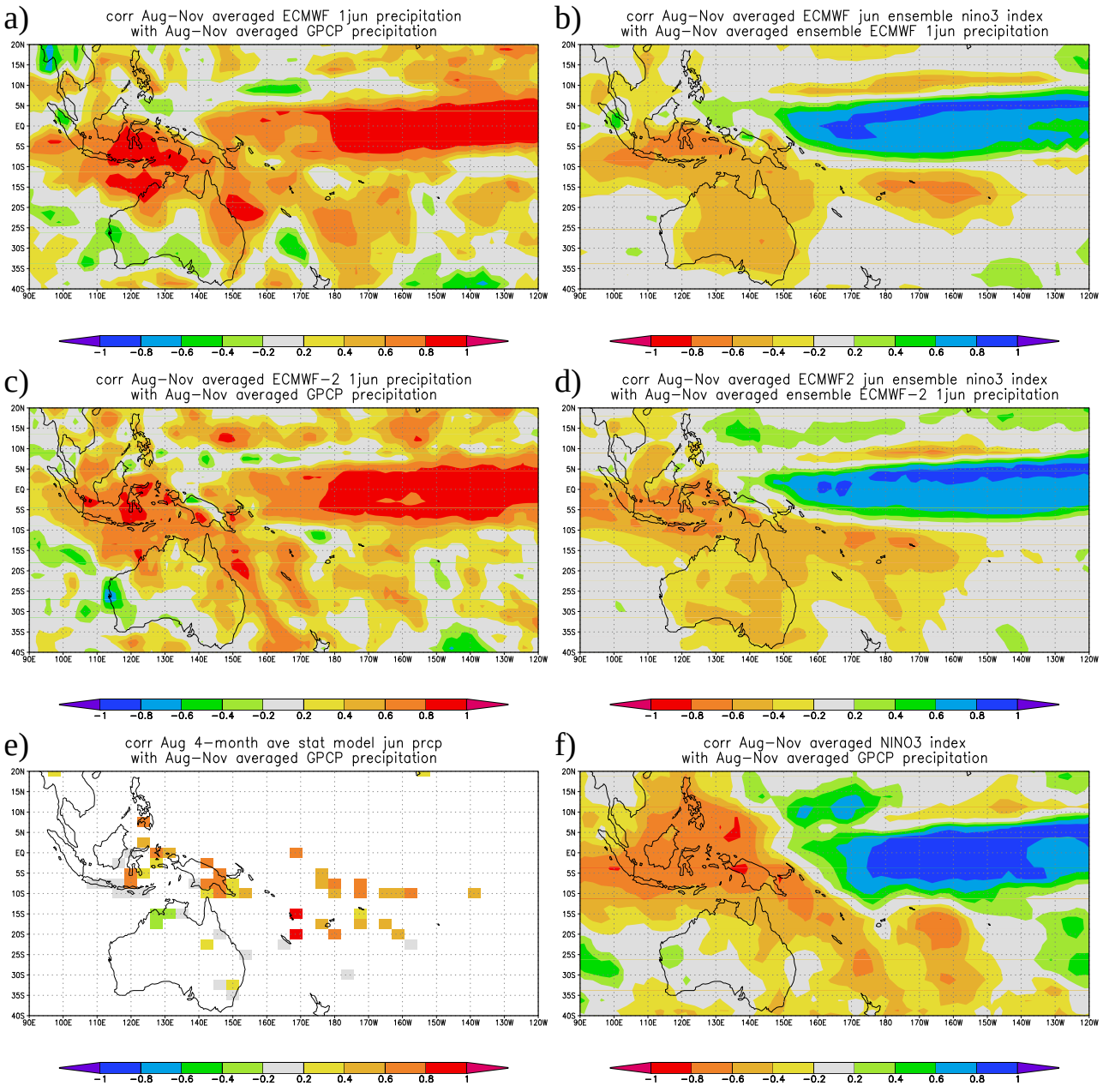


Figure 7: Verification of West Pacific Aug–Nov (dry season) precipitation forecasts from 1 Jun against the GPCP analysis. a) Skill of S1, b) S1 Niño3 teleconnection patterns, c), d) same for S2, e) skill of STAT, and f) Niño3 teleconnection patterns over 1987–2001.

eastern part for S1, in spite of the very rough approximation of the orography and land-sea mask there. The S2 correlations are somewhat lower, possibly due to the limited ensemble size (5) and the smaller amplitude of the predicted ENSO signal. The STAT model has near zero skill in some boxes, for instance in the Java region, in spite of the ENSO teleconnection. This is due to the influence of the second EOF of SST which is included in the STAT model: during the training period this EOF contributed positively to the skill, but this was reversed after 1987.

Although the anomaly correlation of the GCMs is (very) good, the size of the dry-season anomalies was in general underpredicted. For example, in Aug–Nov 1997 the S1 (S2) model ensemble mean forecast was 55% (50%) of the normal rainfall in the region 10°S–EQ, 110°–130°E. Less than 35% was observed.

Eastern Australian rainfall is also forecast rather better by the GCMs than the statistical model, which is mainly based on ENSO ($r(N_3) \approx -0.4$ over the last century). This teleconnection was not as strong in 1987–2001 as in the training period, leading to little skill in the statistical model. Variations in the strength of ENSO teleconnections have occurred before: a 15-year running correlation on NINO3 SST with precipitation at Melbourne (37.82°S, 144.97°E, 1855–1992) is even positive in 1932–1941, in spite of the long-term value of $r = -0.39 \pm 0.13$. A Monte-Carlo study shows that this is not necessarily due to decadal climate variability: most random series with $r = -0.39$ will have 15-year periods with $r > 0$ in 100 years. Apart from the weak observed ENSO teleconnections, the ECMWF models used other predictors to forecast precipitation more succesfully. However, both dynamical models overextend their ENSO teleconnections into central and western Australia, leading to spurious forecasts there.

The strong ENSO teleconnection in the west Pacific is generally handled well by the model giving high skill forecasts. There is a nodal line in the teleconnections which crosses the equator near 150E (panel f). Obviously any forecast skill which is based on ENSO teleconnections will be reduced in the vicinity of the nodal line. The nodal lines in the model (panels b, d) are slightly shifted relative to that in panel f. Perhaps as a consequence of this the skill in Papua New Guinea is reduced. The forecasts in the equatorial central Pacific are very good: r reaches 0.9 there.

North of the equator, in the Philippines, the statistical model only finds strong enough historical teleconnections at one single grid point, where it showed good skill over 1987–2001, $r = 0.68$. The ECMWF models only have skill $r \approx 0.4$ because the ENSO teleconnection observed in this period is not present in the models. The lack of a teleconnection is representatve of the long-term behaviour of other stations in the Philippines, which did not get included in the statistical model. The skill of he statistical model is therefore to a large extent due to chance.

December–February In contrast to the onset date, the intensity of the wet monsoon in central Indonesia is normally not affected by El Niño — once it rains, it pours. In eastern Indonesia the correlations of Dec–Feb rainfall with Niño3 are also much smaller than in the dry season. Becasue of this lack of correlation the statistical model makes few predictions (Fig. 8e). However, the rainfall response in the 1997/98 El Niño was very different to the normal reponse, in that over large parts of Indonesia the monsoon rains were substantially reduced in January. To emphasize the marked reduction in rainfall in January 1988, Fig. 9 shows the relative precipitation anomaly; panel d shows that the rainfall was typically reduced by 50% over Java and estern Indonesia. The S1 model correctly predicted the tendency for a dry January in this specific case (panel a).

S1 tends to predict reduced rainfall in most of Australia during El Niño (Fig. 8b). Over the whole century observations show that only the north-eastern coast had significant ENSO teleconnections ($r = -0.3 \ldots -0.4$) in this season. During the last 15 years these teleconnections were absent, however, leading to very little skill in both GCM and statistical forecasts for eastern Australian summer precipitation. In contrast, the GCM predictions for western Australia had some skill.

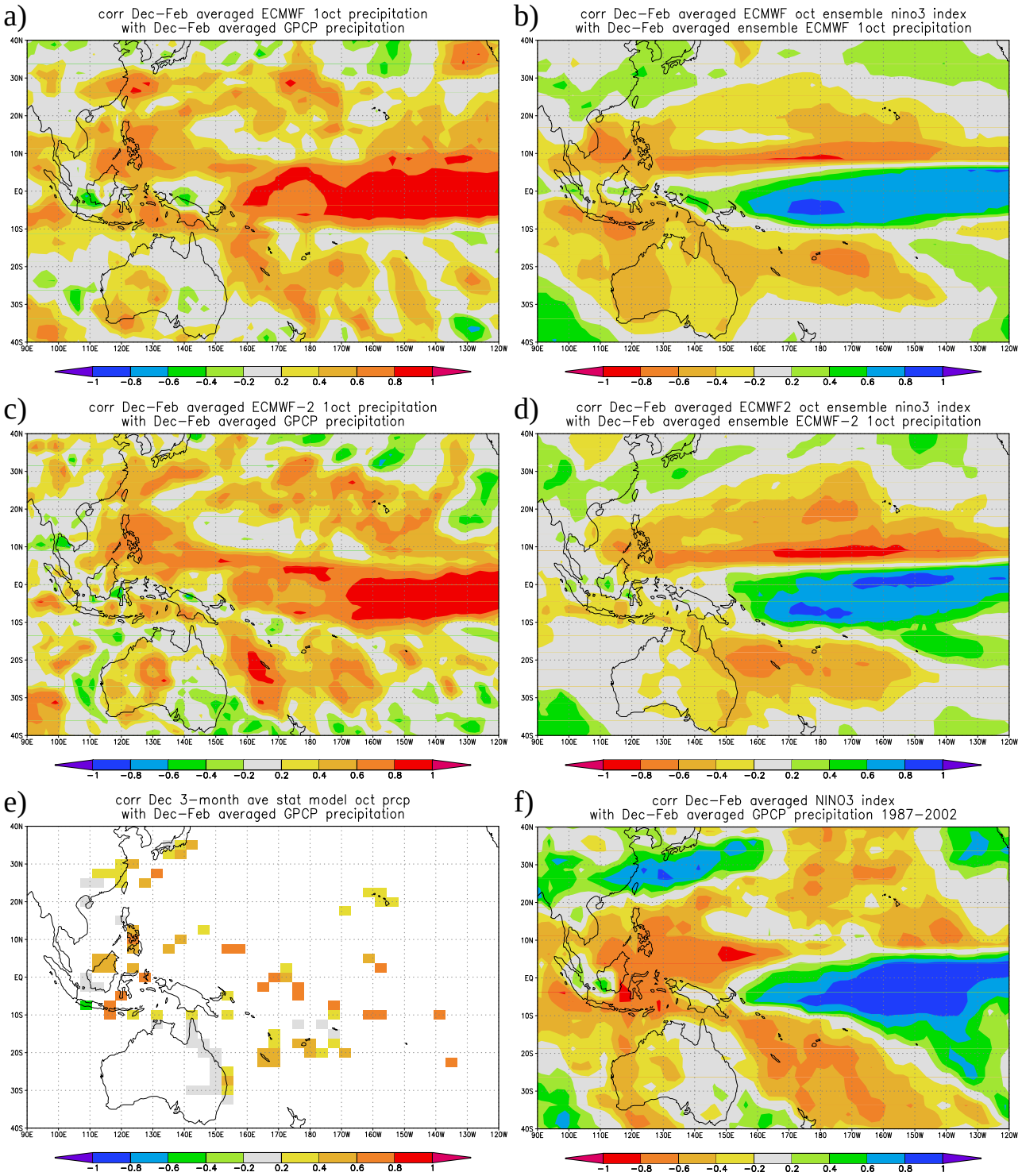


Figure 8: Verification of West Pacific Dec–Feb (rainy season) precipitation forecasts from 1 Oct against the GPCP analysis. a) Skill of S1, b) S1 Niño3 teleconnection patterns, c), d) same for S2, e) skill of STAT, and f) Niño3 teleconnection patterns over 1987–2002.

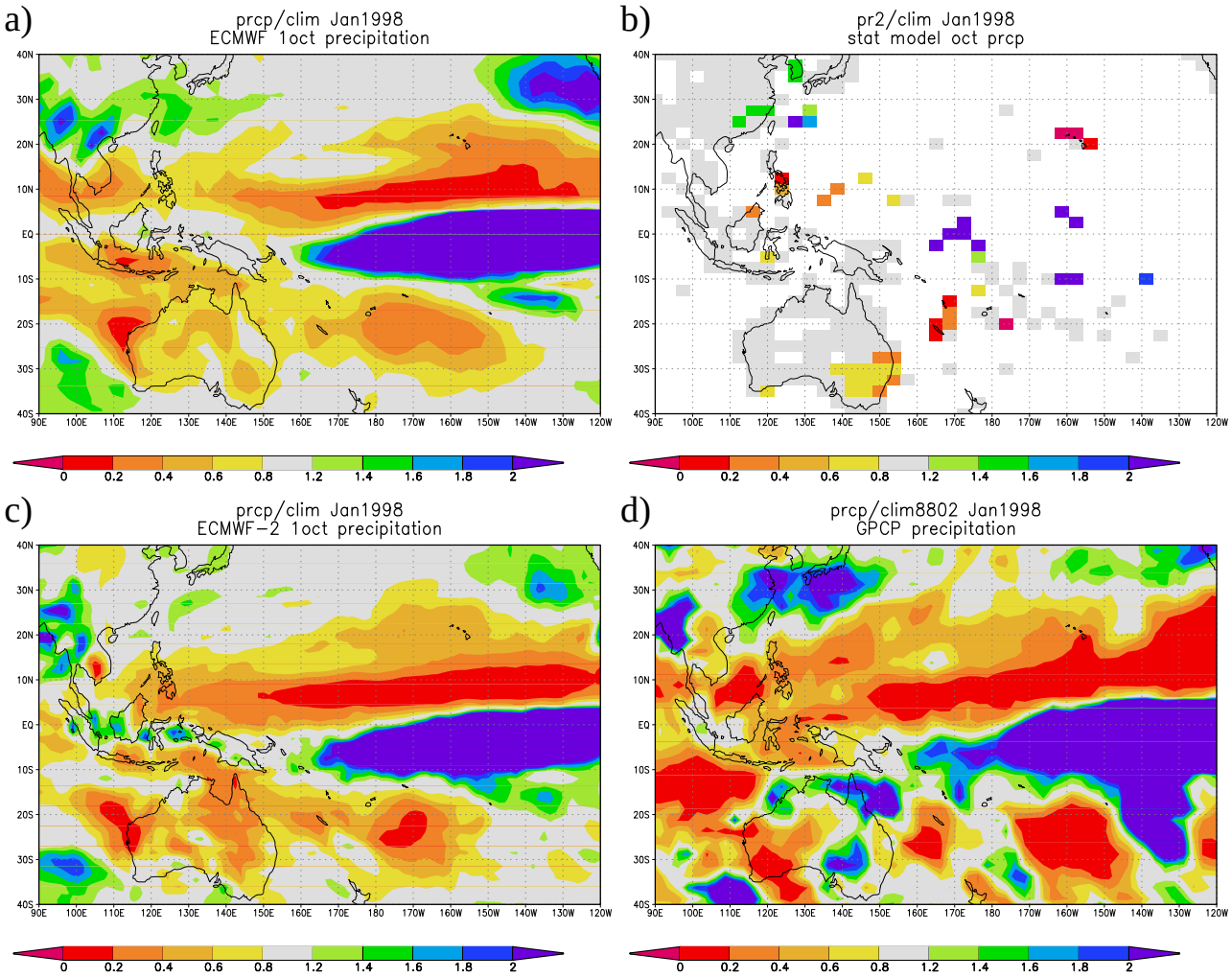


Figure 9: a) The S1 and c) S2 ensemble mean forecasts for the relative rain anomaly in January 1998 from starts around 1 Oct 1997, b) the statistical model forecast and d) observed relative precipitation anomalies.

On the other side of the equator, rainfall anomalies in the Philippines, a standard El Niño teleconnection in this season, were predicted well by the ECMWF models. The wet teleconnection to the Chinese East coast (Tang et al., 1997) is also present. Its modeled strength is similar to the long-term record, which is weaker than Fig. 8f shows.

March–May Rainfall anomalies during the onset of the Asian Monsoon in Mar–May are forecast very well (see Fig. 10) in a zone extending from the Philippines westward to the Andaman Islands (12°N, 93°E). The correlations are higher than can be explained by the model ENSO teleconnections. A lag-correlation study of observational data shows that SST in the region east of the Philippines also influences rain in this zone.

June–July In the remaining months, Jun–Jul, ENSO teleconnections are much weaker in South-East Asia and the West Pacific (not shown). The STAT model performs very badly due to the spring barrier in ENSO predictability. The ECMWF models still have skill $r > 0.8$ in parts of eastern Indonesia and $r \approx 0.6$ in the northern Philippines, in both cases much higher than can be explained by correlation with model Niño3.

4.3.3 The Americas

December–February In the Americas the strongest teleconnections are at the peak of El Niño, Dec–Feb (Ropelewski and Halpert, 1987; Kiladis and Diaz, 1989). These are: rain along the equator in the Pacific just reaching parts of the Peruvian and Ecuadorean coasts, wetter than normal weather in southern Brazil and Uruguay, drought in northern South America and more rain in Mexico and the southern U.S., especially Florida. The forecast skill of the dynamical models was at least as good as that of the statistical in these areas (see Figs 11a,c,e), although the model teleconnections were weaker than observed (Figs. 11b,d,f). There was also skill in north-eastern Brazil, where Dec–Feb is the onset of the rainy season. This skill is not simply related to ENSO and it is absent in our ENSO-based statistical model.

The weak ECMWF model teleconnections to the Californian coast resemble those of the average historical record, which the statistical model is based on. However, in this area there are large decadal variations. A 20-yr running-window correlation analysis of the GHCN precipitation data for Santa Barbara (34°N, 120°E, 1867–2001) shows that the correlation coefficient was around -0.4 from 1890 to 1930, $+0.3$ up to 1990 and $+0.6$ since then. A Monte-Carlo study shows that these decadal variations in the strength of the teleconnection are unlikely to be due to sampling effects ($P < 5\%$). They led to very little skill in the statistical model forecasts on the West Coast in 1987–2001. As the dynamical models are closer to the long-term average than the situation of the last ten years their skill is also low. Other factors that may play a role in lowering skill are the fairly strong nonlinear effects present in this area (Mason and Goddard, 2001) or the wrong position of the Aleutian pole of the PNA-like teleconnection pattern in the ECMWF models.

March–May In Mar–May there are teleconnections in the observations to North-East Brazil and Southwest U.S./North Mexico (Figs 12e,f). Neither of these is present to the same extent in either ECMWF model. This is not due to the spring barrier. The ECMWF models have good skill in Mar–May in South-East Asia and the Pacific Ocean. Also, the statistical model, with a much larger spring barrier, obtains decent skill scores. There are known deficiencies in the climatology of the dynamical models in this season: a too intense ITCZ and a large dry bias over Brazil (Anderson et al., 2003). These probably affect the ENSO teleconnections. (See also http://www.ecmwf.int/products/forecasts/d/charts/seasonal/verification.)

In North-East Brazil there is predictability at this time, the rainy season, due in part to SST in the Atlantic

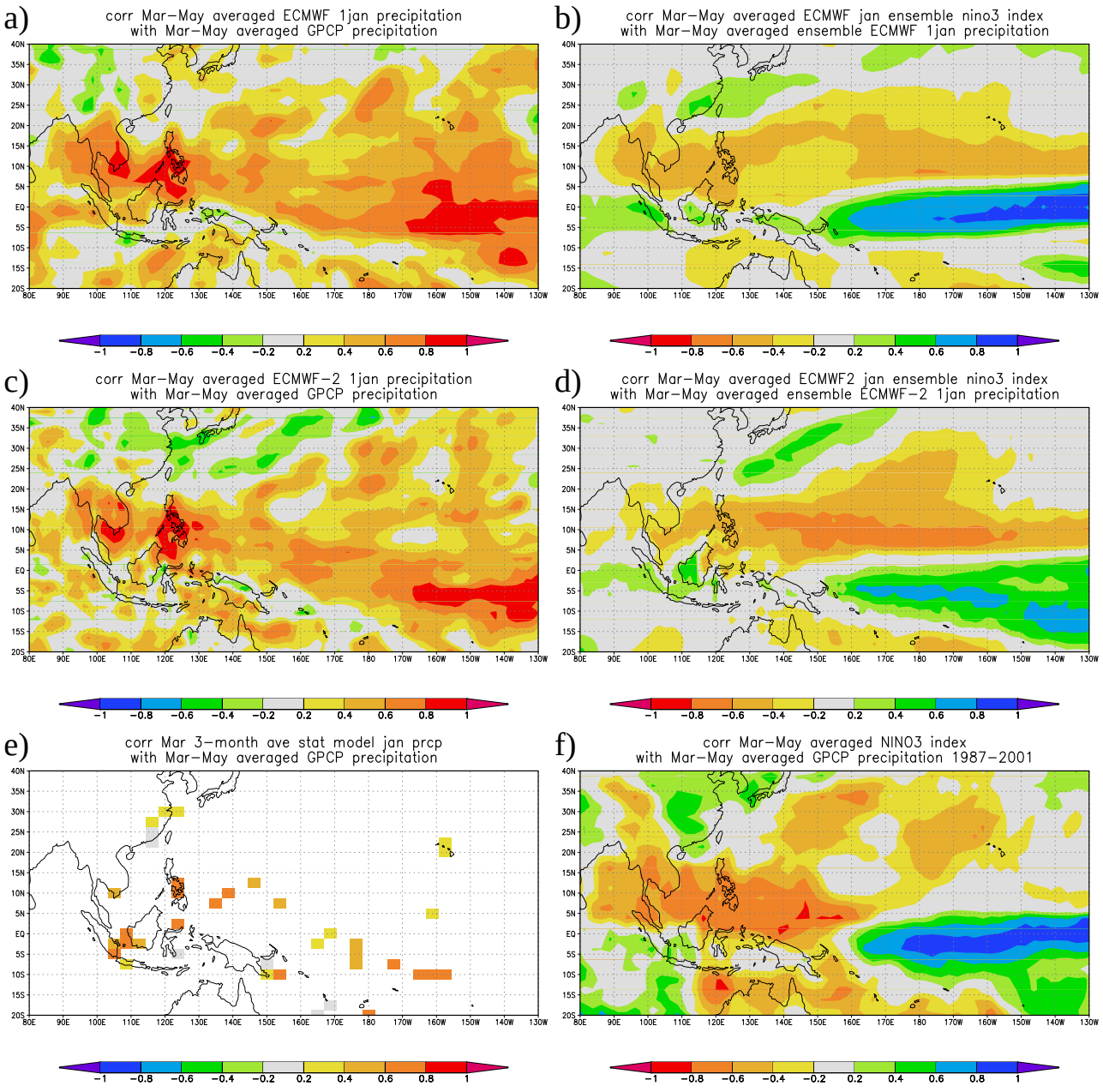


Figure 10: Verification of West Pacific Mar–May (onset Asian monsoon) precipitation forecasts from 1 Jan against the GPCP analysis. a) Skill of S1, b) S1 Niño3 teleconnection patterns, c), d) same for S2, e) skill of STAT, and f) Niño3 teleconnection patterns over 1987–2001.

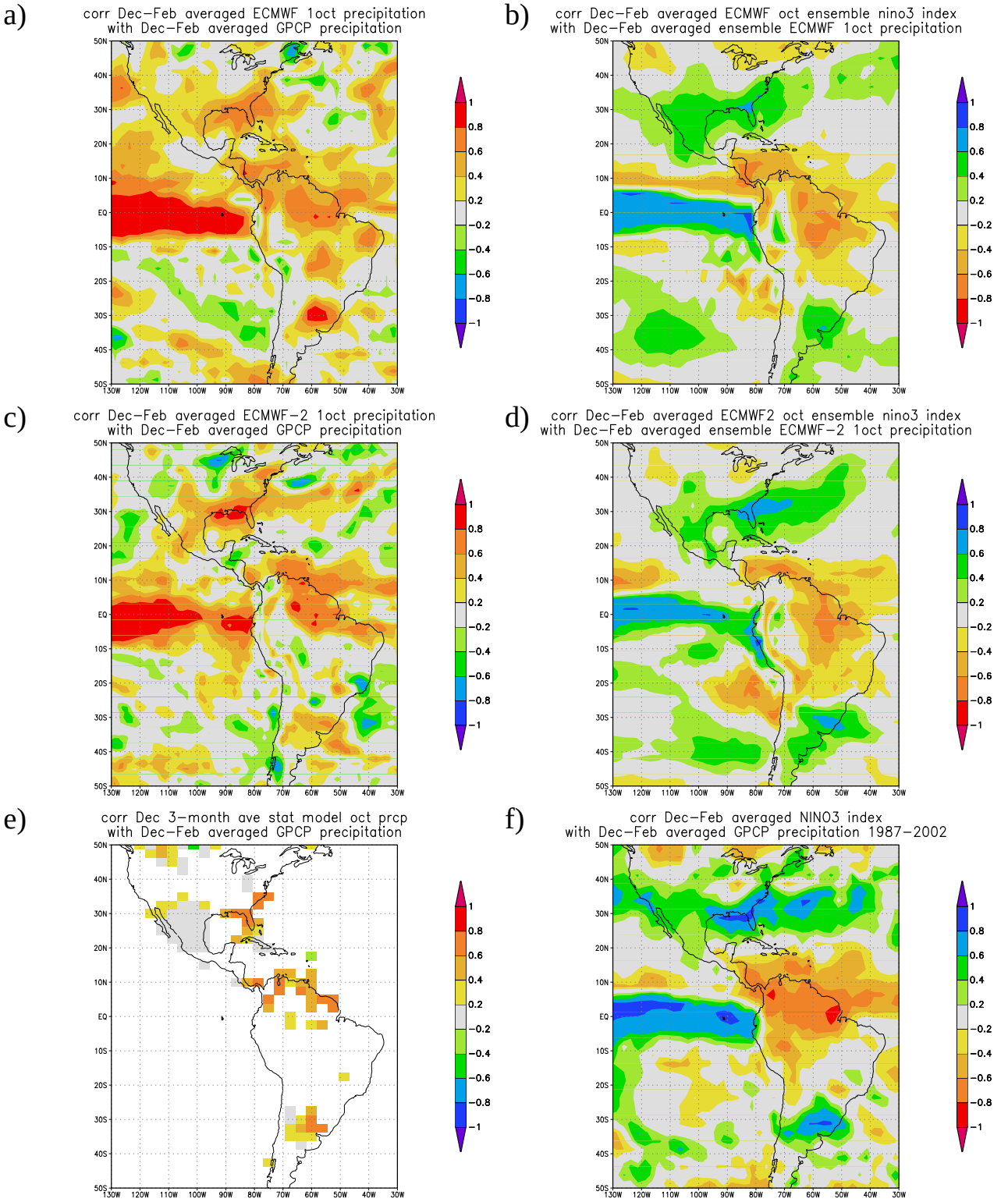


Figure 11: Verification of Dec–Feb precipitation forecasts from 1 Oct in the Americas against the GPCP analysis. a) Skill of S1, b) S1 Niño3 teleconnection patterns, c), d) same for S2, e) skill of STAT, and f) Niño3 teleconnection patterns over 1987–2002.

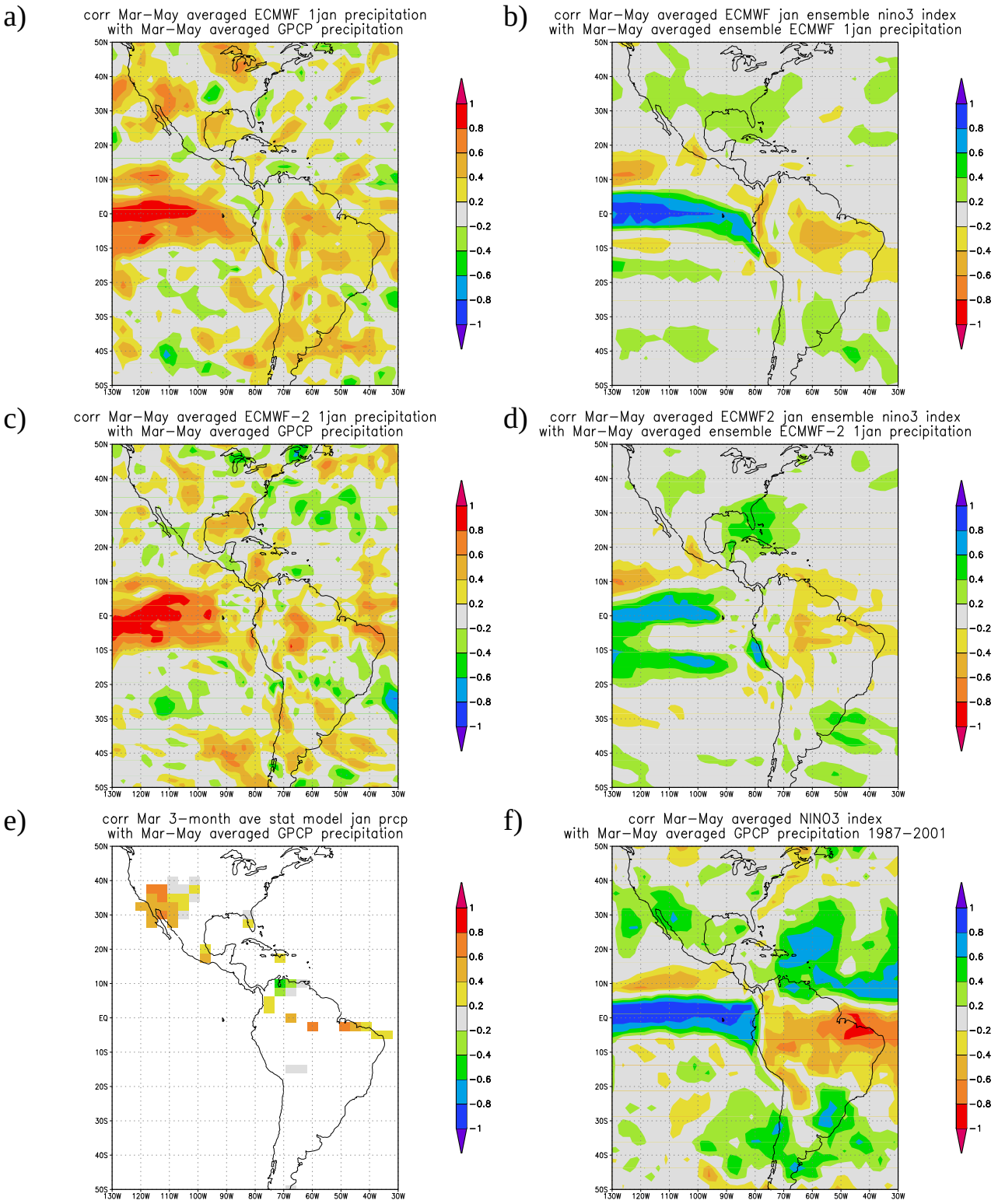


Figure 12: Verification of Mar–May precipitation forecasts from 1 Jan in the Americas against the GPCP analysis. a) Skill of S1, b) S1 Niño3 teleconnection patterns, c), d) same for S2, e) skill of STAT, and f) Niño3 teleconnection patterns over 1987–2001.

Ocean (Hastenrath and Greischar, 1993). S2 shows quite good predictability in this region apparently beyond that due to ENSO. The skill ($r \approx 0.7$) is comparable to the skill of the UK Met Office forecasts for Fortaleza/Quixeramobim (4°S, 39°W) over 1987–1998 (1996 missing) in quintiles, $r = 0.73^{+23}_{-42}$ (Folland et al., 2001).

June–November During the rest of the year the ENSO teleconnections to precipitation over the Americas are much weaker.

4.3.4 Africa

The rainfall over Africa has a strong regional and geographical variation. Rather than show plots for all seasons we will condense the rainfall information neatly into one figure. In Fig. 13 we plot rainfall and correlations with Niño3 in three sectors, corresponding to different times of year with ENSO teleconnections (Hastenrath, 1995). The top-left sector, showing West Africa, has a monsoon climate with rainfall maximum in Jul–Sep. Area-averaged Sahel rainfall in this season has a weak link to El Niño (Palmer, 1986), $r = -0.41$ over 1901–2000. At individual stations the teleconnection is much weaker, so that the statistical model only makes predictions at a few grid boxes.

For the top right sector, we use the period Oct–Nov to capture the short rains. These have a well-established relationship with ENSO (Ogallo, 1988), although the relationship is fairly weak, $r < 0.4$. Rainfall is more strongly connected to Indian Ocean SST (Latif et al., 1999), so predictability could be larger than the ENSO teleconnection suggests. The relationship between the East African long rains and ENSO is less clear. For example Rowell et al. (1994); Philipps and McIntyre (2000), based on different temporal and spatial scales, did not show any significant correlations between the East Africa long rainy season and either the atmospheric or oceanic component of ENSO.

The bottom sector covering southern Africa is for summer (Dec–Feb), when there is a (weak) ENSO teleconnection to less rain during El Niño. In fact the full twentieth century only shows teleconnections to the southwest corner, although shorter records also hint at possible teleconnections to Zimbabwe (Cane et al., 1994).

The last 15 years show stronger ENSO teleconnections to Africa (Fig. 13f). In East Africa this is entirely due to the heavy rains in 1997. This dependence on one event means the uncertainty in these correlations is large. For instance at Mombasa, Kenya (4.03°S, 40.10°E, 1890–1991), the correlation over 1987–2001 is $r = 0.65$ with a 95% confidence interval $-0.60 < r < 0.92$, which easily includes the long-term value $r = 0.30 \pm 0.18$. In the Sahel, a Monte-Carlo study of decadal changes in 1000 random time series with the same long-term regression and correlation with Niño3 as Jul–Sep Sahel rainfall shows that the recent correlation of $r = -0.52 \pm 0.26$ falls well within the range expected due to the sampling uncertainty and decadal changes in ENSO variability. The same holds for the Dec–Feb southern Africa rainfall, which over 1987–2002 had stronger and more northerly teleconnections than earlier in the twentieth century.

Figs 13b,d show that S1 and S2 have only weak ENSO teleconnections and panels a,c show that there is only limited skill with S1 somewhat better than S2. The skill around the Persian Gulf is related to a weak ENSO teleconnection that also shows up in the historical record. The STAT model is very incomplete due to lack of data. In the regions with enough historical observations, its skill (Fig. 13e) is lower than the absolute value of the strength of the historical ENSO teleconnections over 1901–1986. For comparison, the combined statistical/dynamical model Sahel forecasts of the UK Met Office (Folland et al., 1991) had skill only in the Sudan.

Except in South Africa, the skill shown by S1 in fig 13a and to a lesser extent S2 in 13c is not directly attributable

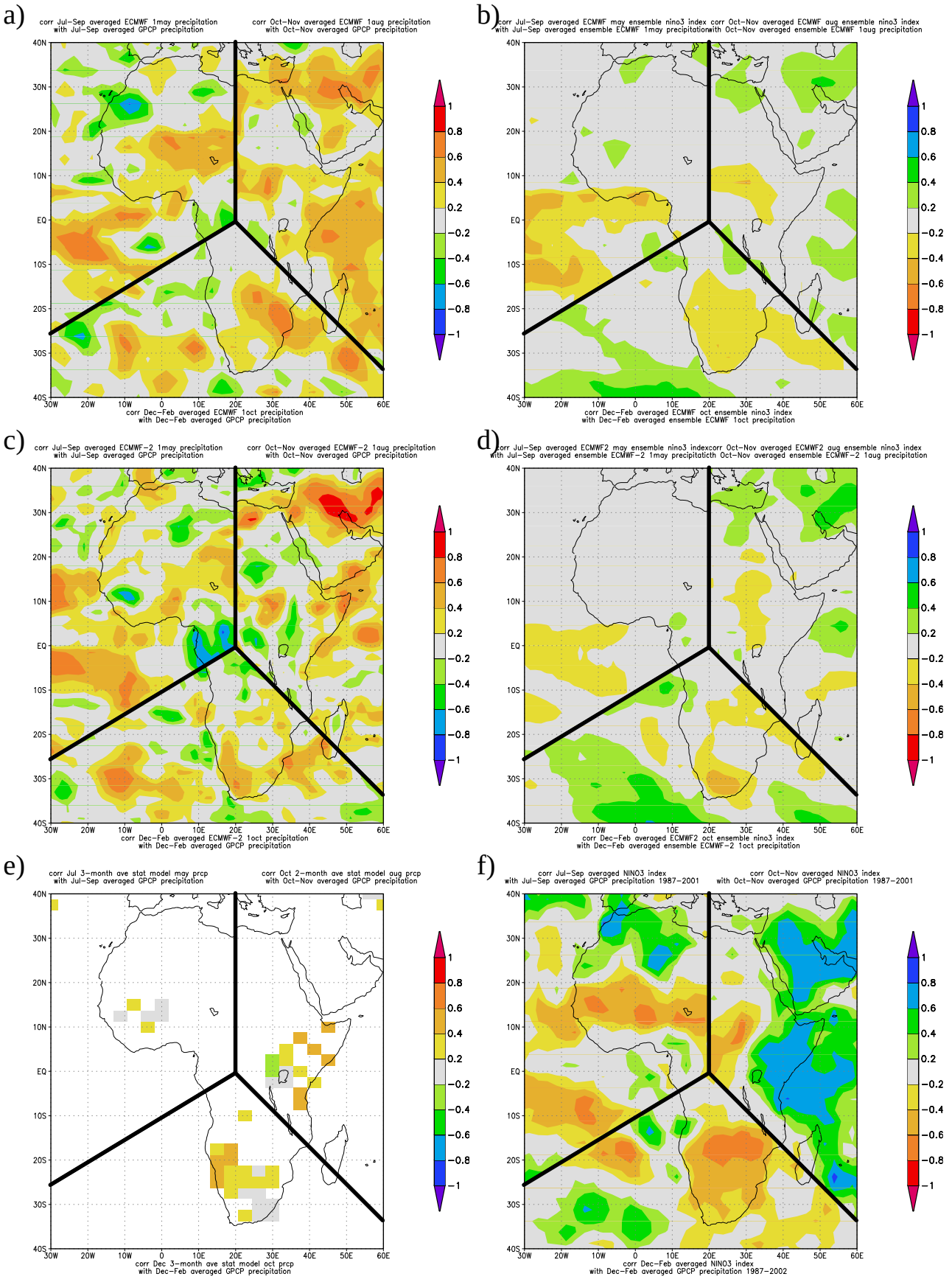


Figure 13: Verification of West-African monsoon (Jul–Sep, top left), East African short rains (Oct–Nov, top right) and South African summer (Dec–Feb, bottom) precipitation forecasts at lead time +2 months against GPCP analysis. a) Skill of S1, b) S1 Niño3 teleconnections, c), d) same for S2, e) skill of STAT, f) observed ENSO teleconnections.

to ENSO as can be seen by comparing with the ENSO (Niño3) teleconnections in Figs 13b,d. These are much weaker than the correlations between Niño3 and observed precipitation over the last 15 years, (fig 13f) and even weaker than those in the historical record.

We are aware of the failure to capture the strong flooding over Kenya/Somalia in late 1997 in the seasonal prediction model. In fact the rainfall was located off the coast and further model studies using the observed SSTs for this period failed to capture the heavy rainfall over land in this region though predicting the excess rainfall over the ocean to the east of Kenya and Somalia was well done. The Latif et al., 1999 study also did poorly at predicting rainfall over land, suggesting either it was not predictable or that the models were in error. In fact it is striking how poorly the S1 and S2 models seem to capture the El Niño teleconnections over Africa (Fig 13b,d) even though they perform quite well in other regions.

4.3.5 Other ENSO precipitation teleconnections

The Jun–Aug South Asian monsoon historical teleconnection was too weak to enter the simple statistical model. The monsoon was also not forecast well by the ECMWF models. The increased strength of the Oct–Dec monsoon rains in southern India and Sri Lanka during El Niño was captured well by the statistical model, but not by the ECMWF models. Curiously, both S1 and S2 predicted the relatively wet Dec–Feb 1997/98 in central India correctly, although this is not an ENSO teleconnection.

The weak ENSO teleconnection to spring precipitation in Europe (van Oldenborgh et al., 2000) lead to some skill in the statistical system in Morocco, Ireland and Scotland, and the Ukraine. However, it was not reproduced in either dynamical model. S2 did have skill in Europe, although based on effects other than on the teleconnection (not shown).

4.3.6 Systematic comparison

Due to the limited number of years for which forecasts (hindcasts) are available, few of the above differences in apparent skill between the statistical and dynamical models are likely to be considered highly significant in a statistical sense when assessed individually. Our ability to compare the skill of the models can be improved by sacrificing geographical and seasonal detail, and considering aggregate statistics. To do this, we selected 40 regions and 3- or 4-month seasons of historical simultaneous ENSO teleconnections with anomaly correlations $r > 0.4$ on the basis of the Hulme, GPCP and GHCN datasets. For the area-averaged precipitation in each of these regions and seasons the skill of the STAT, S1 and S2 models at lead time +2 was computed. The results are shown in Table 3. The error estimates show the 95% confidence interval computed with a bootstrap method. When the lower bound exceeds zero the number is printed in bold, corresponding to $P < 2.5\%$ in a one-sided significance test.

Of the 40 areas and seasons where skill is expected on the basis of ENSO teleconnections, S1 outperforms STAT in 29 and S2 outperforms STAT in 32. As the skill in predicting ENSO is much higher than the skill in the seasonal precipitation over most of the earth, the forecasts are fairly independent even though they are based on the same set of El Niño and La Niña events. Assuming full independence, the null hypothesis that the STAT model has equal skill to the ECMWF models can be rejected at the 0.5% level (S1) and 0.01% level (S2); the dependencies will increase these numbers. In contrast, the difference between S1 (16 better) and S2 (23 better) could easily have arisen by chance if both systems had equal skill. The same conclusions hold for the +1 and +3 forecasts (not shown). Analysis of Table 3 suggests that the lead of the ECMWF models over the statistical model exists at all times of year, but is most dominant in the March–August period, with S1(S2) outperforming STAT in 9(12) out of 12 forecasts. This is the time when the GCMs are producing substantially better forecasts

Region	latitude	longitude	season	STAT	S1	S2
East Indonesia	10°S–EQ	110°E–130°E	Aug–Nov	0.37^{+38}_{-45}	$\mathbf{0.87}^{+9}_{-19}$	$\mathbf{0.82}^{+13}_{-8}$
	–	–	Dec–Feb	$\mathbf{0.68}^{+23}_{-41}$	$\mathbf{0.63}^{+23}_{-42}$	0.32^{+44}_{-37}
	–	–	May–Jul	0.00^{+0}_{-0}	$\mathbf{0.64}^{+25}_{-36}$	$\mathbf{0.82}^{+11}_{-20}$
Philippines	5°N–10°N	120°E–130°E	Dec–Feb	$\mathbf{0.54}^{+28}_{-39}$	$\mathbf{0.72}^{+20}_{-31}$	$\mathbf{0.72}^{+20}_{-40}$
	–	–	Mar–May	$\mathbf{0.70}^{+17}_{-31}$	$\mathbf{0.82}^{+13}_{-39}$	$\mathbf{0.85}^{+10}_{-21}$
	–	–	Sep–Nov	$\mathbf{0.69}^{+17}_{-23}$	0.18^{+64}_{-56}	$\mathbf{0.56}^{+26}_{-32}$
East Australia	40°S–20°S	145°E–155°E	Jul–Sep	0.00^{+0}_{-0}	0.63^{+29}_{-64}	0.73^{+21}_{-84}
	–	–	Oct–Dec	0.31^{+32}_{-46}	0.37^{+44}_{-55}	0.42^{+34}_{-49}
South Pacific	20°S–10°S	160°E–200°E	Oct–Dec	$\mathbf{0.76}^{+15}_{-20}$	$\mathbf{0.60}^{+24}_{-49}$	$\mathbf{0.57}^{+22}_{-54}$
	–	–	Jan–Apr	-0.30^{+49}_{-44}	0.46^{+38}_{-50}	-0.19^{+54}_{-45}
	–	–	May–Jul	-0.48^{+63}_{-40}	$\mathbf{0.80}^{+12}_{-24}$	$\mathbf{0.70}^{+18}_{-38}$
Equatorial Pacific	5°S–5°N	170°E–240°E	Dec–Feb	$\mathbf{0.78}^{+16}_{-30}$	$\mathbf{0.95}^{+4}_{-19}$	$\mathbf{0.89}^{+8}_{-30}$
	–	–	Mar–May	$\mathbf{0.69}^{+21}_{-33}$	$\mathbf{0.82}^{+14}_{-38}$	$\mathbf{0.78}^{+17}_{-35}$
	–	–	Jun–Aug	0.01^{+60}_{-37}	$\mathbf{0.67}^{+26}_{-22}$	$\mathbf{0.77}^{+15}_{-33}$
	–	–	Sep–Nov	$\mathbf{0.56}^{+26}_{-25}$	$\mathbf{0.94}^{+5}_{-61}$	$\mathbf{0.92}^{+6}_{-67}$
South China	20°N–30°N	110°E–120°E	Nov–Jan	0.24^{+48}_{-76}	0.19^{+38}_{-65}	0.63^{+26}_{-73}
	–	–	Feb–Apr	0.19^{+40}_{-59}	0.13^{+50}_{-72}	$\mathbf{0.49}^{+31}_{-31}$
Hawaii	15°N–25°N	160°W–155°W	Nov–Jan	0.33^{+44}_{-61}	$\mathbf{0.70}^{+21}_{-35}$	0.12^{+50}_{-77}
	–	–	Feb–Apr	0.34^{+42}_{-73}	0.36^{+40}_{-50}	$\mathbf{0.57}^{+25}_{-38}$
New Mexico	30°N–40°N	110°W–95°W	Oct–Jan	0.37^{+44}_{-56}	0.44^{+43}_{-54}	0.40^{+43}_{-56}
	–	–	Feb–May	-0.09^{+55}_{-47}	0.40^{+38}_{-57}	0.41^{+35}_{-48}
Florida	25°N–30°N	85°W–80°W	Dec–Mar	$\mathbf{0.76}^{+18}_{-44}$	$\mathbf{0.57}^{+26}_{-31}$	$\mathbf{0.72}^{+17}_{-36}$
NW US	45°N–55°N	125°W–105°W	Nov–Feb	0.30^{+43}_{-44}	0.41^{+33}_{-55}	0.45^{+43}_{-70}
Mexico	20°N–30°N	110°W–95°W	Oct–Dec	0.41^{+44}_{-67}	0.15^{+51}_{-58}	$\mathbf{0.78}^{+13}_{-21}$
	–	–	Jan–Mar	0.11^{+44}_{-43}	0.47^{+28}_{-50}	0.49^{+35}_{-54}
	–	–	Jul–Sep	0.00^{+0}_{-0}	-0.08^{+51}_{-44}	0.15^{+39}_{-46}
N South America	EQ–15°N	70°W–50°W	Jun–Aug	0.17^{+37}_{-34}	$\mathbf{0.54}^{+32}_{-41}$	$\mathbf{0.57}^{+31}_{-54}$
	–	–	Sep–Nov	0.49^{+31}_{-54}	$\mathbf{0.80}^{+15}_{-39}$	$\mathbf{0.69}^{+18}_{-31}$
	–	–	Dec–Feb	$\mathbf{0.67}^{+23}_{-35}$	$\mathbf{0.78}^{+15}_{-25}$	$\mathbf{0.87}^{+10}_{-18}$
NE Brazil	10°S–EQ	45°W–35°W	Apr–Jul	0.43^{+31}_{-45}	$\mathbf{0.77}^{+17}_{-30}$	$\mathbf{0.73}^{+21}_{-41}$
Uruguay	40°S–30°S	65°W–55°W	Nov–Jan	0.65^{+22}_{-67}	$\mathbf{0.77}^{+15}_{-51}$	$\mathbf{0.80}^{+13}_{-31}$
Coastal Ecuador	10°S–EQ	85°W–80°W	Feb–May	0.00^{+0}_{-0}	0.59^{+28}_{-76}	0.53^{+30}_{-64}
Central Asia	35°N–45°N	60°E–75°E	Oct–Jan	0.05^{+49}_{-46}	$\mathbf{0.45}^{+34}_{-38}$	0.47^{+31}_{-59}
India	15°N–30°N	70°E–90°E	Jun–Aug	0.14^{+55}_{-61}	-0.19^{+55}_{-47}	0.23^{+40}_{-55}
Sri Lanka	5°N–10°N	75°E–85°E	Oct–Dec	$\mathbf{0.55}^{+27}_{-37}$	-0.10^{+46}_{-43}	0.19^{+38}_{-48}
East Africa	5°S–5°N	30°E–40°E	Oct–Dec	0.14^{+51}_{-88}	0.34^{+54}_{-40}	-0.12^{+74}_{-49}
Southern Africa	35°S–25°S	15°E–35°E	Nov–Feb	0.26^{+33}_{-41}	$\mathbf{0.59}^{+21}_{-23}$	0.28^{+46}_{-62}
SW Europe	10°S–10°N	30°E–40°E	Jul–Sep	0.00^{+0}_{-0}	-0.12^{+57}_{-66}	0.17^{+50}_{-60}
	–	–	Oct–Dec	0.11^{+53}_{-88}	$\mathbf{0.45}^{+46}_{-35}$	-0.07^{+83}_{-51}
Central Europe	10°N–40°N	45°E–55°E	Oct–Dec	0.19^{+50}_{-59}	$\mathbf{0.53}^{+27}_{-49}$	$\mathbf{0.83}^{+13}_{-47}$

Table 3: Regions and seasons of known ENSO teleconnections, and the skill of the three models at lead time +2 months over 1987–2001 starts. The error estimates denote the 95% confidence intervals.

than the STAT model of ENSO-related SST anomalies, and the dominance is thus unsurprising. The fact that the GCMs appear still to have the advantage at times of year when their SSTs are not better than those of the STAT model, is an encouraging sign of the potential of GCM-based seasonal forecasting. In Sep–Feb 20(20) of the 28 seasons/areas were forecast better by S1(S2) than STAT, which are both significant at the 2% level if all forecasts are assumed independent.

The dynamical models also perform better than the model ENSO teleconnections alone would indicate. The observed skill at lead time +2 months is larger than the absolute value of the correlation coefficient with modeled Niño3 in 29 cases out of 40 in S1, and in 34 in S2. THis confirms the impression from the previous sections that even in areas of known ENSO teleconnections other mechanisms that give rise to seasonal predictability are modeled correctly by the dynamical models.

5 Conclusions

We have compared the skill of the ECMWF seasonal forecast models over 1987–2001 starts to that of an SST-based statistical forecast model in forecasting ENSO and its teleconnections. The anomaly correlation coefficient is used as the skill measure. This assumes that the biases in the mean state and variability are known and corrected for. 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. The yearly averaged anomaly correlations are significantly higher than those of the three statistical models considered. However, this is an average of two very different regimes. The skill of the dynamical models is much higher than the skill of the statistical models during the spring barrier: the onset of El Niño or La Niña is forecast much better. In fact, S1 predicted the start and amplitude of the 1997-98 event very well most of the time, in real forecast mode. S2 is more damped than S1, but has higher anomaly 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 one.

The dynamical models also forecast SST better than the statistical model in the Atlantic and Indian Oceans. Temperature over land is forecast well in northern South America, but the ENSO teleconnections to North America in winter are sufficiently displaced that the 2m temperature forecasts do not show much pointwise skill away from the west coast.

The skill of precipitation forecasts depends very strongly on the region and season considered. With only 15 years of data and small decorrelation scales the chance of finding accidental skill is very high. We therefore restricted ourselves to regions and seasons of known ENSO teleconnections. During El Niño and La Niña the skill of the dynamical model precipitation forecasts (as measured by a spatial correlation, which emphasizes the tropics) is much higher than during neutral conditions, showing the importance of ENSO for seasonal forecasting of precipitation.

In South-East Asia and the West Pacific, parts of Australia and the Americas in Dec–Feb the dynamical models forecast precipitation better than the statistical model, due to the inclusion of other SST-weather interactions. In contrast, in Mar–May the teleconnections to the Americas are not forecast well by the GCMs. Also, the (fairly weak) teleconnections to the Indian Monsoon, East Africa and Europe are not reproduced well.

Overall, the ECMWF models are on average significantly better than the statistical model in forecasting precipitation at lead times +1, +2 and +3 months in 40 regions and seasons of known ENSO teleconnections, even though the statistical model is based on lagged ENSO regressions. The ECMWF precipitation forecasts are also better than can be accounted for by ENSO teleconnections alone, demonstrating that value is being extracted from the global, comprehensive nature of the forecast system.

And so what do we conclude about the relative merits of dynamical and statistical seasonal forecasts? The first point to make clear is that the statistical models used here, although entirely respectable, may not necessarily be the best that can be constructed, especially on a regional basis – one thinks of the Nord-Este region of Brazil as one example where this may be true. It is very likely to be the case that for some regions, statistical models still have a definite advantage over at least the two dynamical models considered here.

The argument often put forward in favour of GCMs is that they have a (supposedly) clear development path which will lead to improved forecasts in the future. Experience has shown that it is not easy to make big improvements to either the models or the seasonal forecast systems which use them. Nonetheless, models and forecast systems are being improved, and over time this is likely to lead to a transformation in the quality of the GCM-produced forecasts. A key question, which we have addressed in this paper, is whether the 'benefit' of GCMs is something that lies only in a possibly distant future, or is something which is tangible now. We hope we have made clear that although the GCMs are far from uniformly helpful, they have a real contribution to make to practical forecasting today.

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.

References

Anderson, D. L. T., T. Stockdale, M. A. Balmaseda, L. Ferranti, F. Vitart, P. Doblas-Reyas, 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., D. L. T. Anderson, J. O. Alves, T. Stockdale, and J. Vialard, 2003: Diagnostics and validation of the ECMWF-System1 ocean analysis. (in preparation).

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., 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.

Berlage, H. P., 1957: *Fluctuations of the general atmospheric circulation of more than one year, their nature and prognostic value*. Number 69 in Mededelingen en verhandelingen. KNMI.

Cane, M. A., G. Eshel, and R. W. Buckland, 1994: Forecasting Zimbabwean maize yields using eastern equatorial Pacific sea surface temperature. *Nature*, **370,** 204–205.

Colman, A. and M. Davey, 1999: Prediction of summer temperature, rainfall and pressure in Europe from preceding winter North Atlantic Ocean temperature. *Int. J. Climatol.*, **19,** 513–536.

Czaja, A., P. van der Vaart, and J. Marshall, 2002: A diagnostic study of the role of remote forcing in Tropical Atlantic variability. *J. Climate*, **15,** 3280–3290.

Diaz, H. F., M. P. Hoerling, and J. K. Eischeid, 2001: ENSO variability, teleconnections and climate change. *Int. J. Climatol.*, **21,** 1845–1862.

Enfield, D. B. and D. A. Mayer, 1997: Tropical Atlantic sea surface temperature variability and its relation to El Niño-Southern Oscillation. *J. Geophys. Res.*, **102,** 929–945.

Folland, C. K., A. W. Colman, D. P. Rowell, and M. K. Davey, 2001: Predictability of northeast Brazil rainfall and real-time forecast skill, 1987–98. *J. Climate*, **14,** 1937–1958.

Folland, C. K. J., J. Owen, M. N. Ward, and A. Colman, 1991: Prediction of seasonal rainfall in the Sahel region using empirical and dynamical methods. *J. Forecast*, **10,** 21–56.

Hastenrath, S., 1995: Recent advances in tropical climate prediction. *J. Climate*, **8,** 1519–1532.

Hastenrath, S. and L. Greischar, 1993: Further work on the prediction of northeast Brazil rainfall anomalies. *J. Climate*, **6,** 743–758.

Huffman, G. J., R. F. Adler, B. Rudolf, U. Schneider, and P. R. Keehn, 1995: A technique for combining satellite data, raingauge analysis and model precipitation information into global precipitation estimates. *J. Climate*, **8,** 1284–1295.

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.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaver, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetma, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgens, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, and R. Jenne, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Met. Soc.*, **77,** 437–471.

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.

Kiladis, G. N. and H. F. Diaz, 1989: Global climatic anomalies associated with extremes in the Southern Oscillation. *J. Climate*, **2,** 1069–1090.

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., D. Dommenget, M. Dima, and A. Grötzner, 1999: The role of Indian Ocean sea surface temperature in forcing East African rainfall anomalies during December/January 1997/98. *J. Climate*, **12,** 3497–3504.

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.

Mason, I. B., 2003: Signal detection theory and the ROC. In Jolliffe, I. T. and D. B. Stephenson, editors, *Forecast Verification. A Practioner's Guide in Atmospheric Science*, page 66. Wiley & Sons Ltd.

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

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

Ogallo, L., 1988: Relationships between seasonal rainfall in East Africa and the SO. *Int. J. Climatol.*, **8,** 31–43.

Palmer, T. N., 1986: Influence of the Atlantic, Pacific and Indian Oceans on Sahel rainfall. *Nature*, **322,** 251–253.

Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **481,** 747–774.

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., 2003: Development of a european multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). Submitted to BAMS.

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.

Philipps, J. and B. McIntyre, 2000: ENSO and interannual variability in Uganda: Implications for agricultural management. *Int. J. Climatol.*, **20,** 171–182.

Reynolds, R. W. and T. M. Smith, 1994: Improved global sea surface temperature analyses using optimum interpolation. *J. Climate*, **7,** 929–948. Niño indices are available from the Climate Prediction Center at www.cpc.noaa.gov/data/indices/.

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

Rowell, D. P., J. M.Ininda, and M. N. Ward, 1994: The impact of global sea surface temperature patterns on seasonal rainfall in East Africa. In *Proc. Int. Conf. on Monsoon Variability and Prediction*, pages 666–672, Trieste, Italy. WMO.

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.

Tang, Y, W Zhen, and J. Xu, 1997: Relation between drought and flood patterns in the southwestern China and seasonal variation of SST in the Pacific ocean. *Oceanographical Et Limnologia Sinica,* **28,** 88–96. (in Chinese).

van Loon, H. and R. A. Madden, 1981: The Southern Oscillation, Part I. Global associations with pressure and temperature in northern winter. *Mon. Wea. Rev.,* **109,** 1150–1162.

van Oldenborgh, G. J., G. Burgers, and A. Klein Tank, 2000: On the El Niño teleconnection to spring precipitation in Europe. *Int. J. Climatol.*, **20,** 565–574.

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.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: an Introduction.* Academic Press, 464pp.