A Consolidated CLIPER Model for Improved August-September ENSO Prediction Skill

BENJAMIN LLOYD-HUGHES AND MARK A. SAUNDERS

Benfield Hazard Research Centre, Department of Space and Climate Physics, University College London, Holmbury St Mary, UK

PAUL ROCKETT

Benfield Limited, 55 Bishopsgate, London, UK

Weather and Forecasting (In Press)

Manuscript submitted: 22nd August 2003

Revision submitted: 22nd April 2004

Accepted: 5th May 2004

Corresponding author address: Dr Mark A Saunders, Benfield Hazard Research Centre, Department of Space and Climate Physics, University College London, Holmbury St Mary, Dorking, Surrey, RH5 6NT, United Kingdom. (E-mail: mas@mssl.ucl.ac.uk).

ABSTRACT

A prime challenge for ENSO seasonal forecast models is to predict boreal summer ENSO conditions at lead. August-September ENSO has a strong influence on Atlantic hurricane activity, Northwest Pacific typhoon activity and tropical precipitation. However, summer ENSO skill is low due to the spring predictability barrier between March and May. A 'Consolidated' ENSO-CLIPER seasonal prediction model is presented to address this issue with promising initial results. Consolidated CLIPER comprises the ensemble of 18 model variants of the statistical ENSO-CLIPER (CLImatology and PERsistence) prediction model. Assessing August-September ENSO skill using deterministic and probabilistic skill measures applied to cross-validated hindcasts 1952-2002 and using deterministic skill measures applied to replicated real-time forecasts 1900-1950, shows that the consolidated CLIPER model consistently outperforms the standard CLIPER model at leads from 2 to 6 months for all the main ENSO indices (3, 3.4 and 4). The consolidated CLIPER August-September 1952-2002 hindcast skill is also positive to 97.5% confidence at leads out to 4 months (early April) for all ENSO indices. Optimisation of the consolidated CLIPER model may lead to further skill improvements.

1. Introduction

The predictability of El Niño Southern Oscillation (ENSO) sea surface temperatures (SSTs) has received considerable research over the last two decades. During the 1997-98 strong El Niño and subsequent 1998 moderate La Niña 15 dynamical and statistical ENSO seasonal forecast models were in real-time operation (see Barnston *et al* 1999; Landsea and Knaff 2000 for details and inter-comparisons of model performance). Most ENSO prediction models produce useful forecasts (*i.e*. a correlation skill of at least 0.5) at leads out to 6 months when skill is assessed over all seasons (Kirtman *et al* 2002). However, the predictability of ENSO has a strong seasonal cycle: it is relatively easy to predict boreal winter and spring ENSO conditions from boreal summer but it is difficult to predict boreal summer ENSO conditions from boreal winter and spring. The decrease in forecast skill through the months of March-May is known as the 'spring predictability barrier'. This phenomenon was reported first by Walker and Bliss (1932) who observed that the Southern Oscillation had least persistence across the March-May season. Subsequent studies have documented the ENSO spring predictability barrier in detail (see Torrence and Webster (1998) for a recent review).

Improved seasonal predictions of boreal summer ENSO conditions would bring sound socio-economic benefits. August-September ENSO has a strong influence on Atlantic, US and Caribbean hurricane activity (eg Gray 1984; Bove et al 1998; Saunders et al 2000) which peaks between August and October, Northwest Pacific typhoon activity (Chan 1985; Saunders et al 2000) which peaks between July and October, and patterns of boreal summer tropical precipitation (eg Ropelewski and Halpert 1987; Dai and Wigley 2000). The ability to skillfully predict seasonal hurricane/typhoon activity and seasonal rainfall at longer range would benefit society, business and government by reducing the risk and uncertainty associated with the year-to-year variability in the incidence of such climatic events and conditions.

The statistical ENSO-CLIPER (CLImatology and PERsistence) prediction model is arguably one of the more successful ENSO seasonal forecast models to date (Kerr 2000). ENSO-CLIPER was developed by Knaff and Landsea (1997) as a "no skill" control forecast for comparison with more sophisticated dynamical ENSO prediction models. It is a statistical model based entirely on the linear optimal combination of persistence, month-to-month trend of initial conditions and climatology. The formulation of the ENSO-CLIPER model provides scope for modifying its structure. The sensitivity of this model's summer ENSO skill to changes in the model specification is assessed. The study then examines whether the skill of the standard ENSO-CLIPER model may be improved by combining - or 'consolidating' - hindcasts made with different structural CLIPER variants. This procedure - called ensemble or consensus forecasting - has long been used in numerical weather prediction to improve forecast skill (Thompson 1976) but has been applied only recently to seasonal climate forecasting (Goddard *et al* 2001). For a consensus forecast to achieve skill which is measurably higher than from its individual ensemble members, these members need to show statistical independence (*e.g*. Goerss 2000). Reports indicate that ENSO predictive skill may be improved by combining forecasts made with different predictive models (Unger *et al* 1996; Kirtman *et al* 2002; Mason and Mimmack 2002).

The paper is structured as follows. Section 2 reviews briefly the standard ENSO-CLIPER model, describes how prediction skill and uncertainty are calculated, and details the data sets employed. The results section (3) displays the summer ENSO prediction skill from the standard ENSO-CLIPER model and its temporal stability, and shows the sensitivity of this skill to three factors used in the model's formulation. The factors examined are the predictor significance level test, the teleconnected predictor averaging period and the variance factor used during the optimal combination of predictors. A consolidated CLIPER model is presented comprising the ensemble of 18 model variants of the standard CLIPER model. The August-September ENSO skill of the consolidated model is presented for each ENSO index region (3.4, 3, 4, and 1+2) using deterministic and probabilistic skill measures applied to cross-validated hindcasts 1952-2002 and to replicated real-time forecasts 1900-1950 (deterministic skill only). Section 4 discusses these results and conclusions are drawn in section 5.

2. Methodology

a. Standard ENSO-CLIPER model

A detailed description of the standard ENSO-CLIPER model methodology is provided by Knaff and Landsea (1997) and need not be repeated here. In summary, there are 14 potential predictors available to the model. These predictors are listed by ENSO index region and number in Table 1 and may be categorised as follows:

A) Persistence of predictand SST anomaly (1, 3 and 5-month means). Predictor numbers 1-3.

B) Trend of predictand SST anomaly (1, 3 and 5-month means). Predictor numbers 4-6.

C) Initial condition of teleconnected predictors (3-month mean). Predictor nos 7, 9, 11 and 13.

D) Trend of teleconnected predictors (3-month mean). Predictor numbers 8, 10, 12 and 14.

To illustrate how the 1-, 3- and 5-month means for trend work let us consider a forecast made in early August. This forecast would use the following predictand SST anomalies: July minus June

(1-month trend), May-July minus February-April (3-month trend), and March-July minus October(prior)-February (5-month trend).

Each predictor which correlates with the predictand to the 5% significance level after correction for serial autocorrelation enters a predictor pool from which a leaps-and-bounds (L&B) algorithm (Furnival and Wilson 1974) estimates the optimal combination of N = 1, 2, ...14 predictors. The L&B selection routine works by stepping forward using every possible combination of the predictors until the best multiple regression equations having one, two,….., 14 predictors are found. The final selected model is the one with the largest N that explains at least 2.5% more variance than the N-1 predictor model. This is subject to the caveat that only one of the 1, 3 and 5-month mean predictors in each of the categories (A) and (B) may be selected. If a satisfactory predictor model can be found, multivariate linear regression is applied to produce the forecast; otherwise a zero anomaly (*i.e.* climatology) forecast is recorded.

b. Cross-validated hindcasts 1950-2002

We assess seasonal predictability in two ways: from cross-validated hindcasts for the period 1952-2002 and from replicated real-time forecasts for the independent prior period 1900-1950 (section 2*c*). The standard ENSO-CLIPER model was derived using a fixed 43-year training period 1952-1994. Numerical simulations (Lloyd-Hughes, unpublished research, 2003) indicate that at least 50 forecast-observation pairs are required for a realistic skill estimate. Previous studies of ENSO predictability (*e.g*. Mason and Mimmack 2002; Kirtman *et al* 2002; Latif *et al* 1998) have sought to ameliorate this problem by pooling predictions of different seasons at a given lead. However, this is always at the expense of statistical independence. A cross-validated approach (Wilks 1995) is adopted here to extend the validation period to 51 years (1952-2002). At each step a new model is formulated trained on all data excluding a 5 year block centred on the year of interest (i.e. year blocks of 1952-1956, 1953-1957,….., 1998-2002 are used). This block is tapered at the time series ends. Block elimination is employed to minimise potential skill inflation which might arise from the multi-annual persistence of ENSO conditions. The choice of 5 years follows from the frequency spectrum of the ENSO signal which shows a dominant peak in periodicity at about 4 years (Rasmusson and Carpenter 1982; Trenberth 1997).

Forecast lead time is defined according the convention of the World Meteorological Organization (WMO 2002) where a zero lead forecast is one which employs data up to the end of the month immediately prior to the forecast period starting, *i.e*. predictions issued at the end of July for conditions in August-September are said to be issued at zero lead.

c. Independent (replicated real time) forecasts 1900-1950

Our replicated real-time forecast scheme uses an initial model training period from 1870 to 1899. The training period increases one year at a time as each forecast is made. For example, the August-September ENSO forecast for 1901 is made by training the prediction model on data between 1870 and 1900, and so on. This updating exactly replicates the operation of a real-time forecast. Independent ENSO forecasts for each year between 1900 and 1950 are obtained. 1950 marks the limit of independent forecasts since data after this were used in the structural development of the original CLIPER model.

d. Deterministic skill and uncertainty

August-September ENSO skill is assessed for the cross-validated hindcasts 1952-2002 using deterministic and probabilistic skill measures, and for the replicated real-time forecasts 1900-1950 using deterministic skill measures.

For deterministic seasonal hindcast skill we use the skill metric recommended by the World Meteorological Organisation (WMO 2002). This is the percentage improvement in mean square error over a climatological hindcast, referred to as the mean square skill score, *MSSS*. This skill measure is defined as follows:

$$MSSS = 1 - \frac{MSE_f}{MSE_{cl}} \qquad (1)$$

where $MSE_f = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2$ and $MSE_{cl} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

are respectively the mean squared error of the hindcasts and the mean squared error of climatology hindcasts. Here $\hat{x}_i$ and x_i are respectively the hindcast and observed anomaly values for each of the n = 51 years. The climatologies used here are the 51-year (1952-2002) average for the cross-validation period, and the 51-year (1900-1950) average for the replicated real-time forecast period.

Model skill is compared against ordinary persistence skill for the standard ENSO-CLIPER model and its temporal stability, and for the ENSO-CLIPER model formulated using different values of three sensitivity factors. Persistence is calculated over the same length interval as the predictand period (WMO 2002). For example, the ordinary persistence at a lead of 1 month for the August-September target predictand is calculated as the mean anomaly over the prior two-month period May-June.

Confidence intervals are computed around the *MSSS* skill values using the bootstrap method (Efron and Gong 1983). This involves randomly selecting with replacement, 51 years (in this case) of actual data together with the associated predicted and climatological hindcasts. Upon calculating the *MSSS* skills and repeating many times, a distribution of skill values is obtained from which a 95% two-tailed confidence interval can be readily obtained. This confidence interval means there is a 95% probability that the skill computed over the 51 year period will lie within this uncertainty window. The root mean square skill score (*RMSSS*) is also considered and is calculated in a way identical to equation (1) but with the insertion of the root mean square error in place of the MSE. *RMSSS* places less weight on the correct prediction of extremes and so provides a useful comparison to the *MSSS*.

Fully cross-validated *MSSS* with one year at a time withheld may be decomposed (Murphy 1988) into temporal, amplitude and bias errors as follows:

$$MSSS = \left\{ 2\frac{s_{\hat{x}}}{s_x} r_{\hat{x},x} - \left(\frac{s_{\hat{x}}}{s_x}\right)^2 - \left(\frac{\left[E\langle\hat{x}\rangle - E\langle x\rangle\right]}{s_x}\right)^2 + \frac{2n-1}{(n-1)^2} \right\} \Bigg/ \left\{ 1 + \frac{2n-1}{(n-1)^2} \right\} \qquad (2)$$

Here $s_{\hat{x}}$ and s_x are respectively the sample standard deviations of the hindcast and observed values, $r_{\hat{x},x}$ is the product moment correlation of the hindcasts and observations, and $E\langle\cdots\rangle$ represents the expectation or mean value. Although equation (2) is not exact when block elimination is employed, the basic decomposition result will hold. The first three terms in the expansion relate to phase errors (through the correlation), amplitude errors (through the ratio of the hindcast to the observed variances) and the overall bias error. The contribution from each of these terms to the skill improvement afforded by the consolidated ENSO-CLIPER model is considered in section 4.

e. Probabilistic skill

The probabilistic skill measure employed is the rank probability skill score (*RPSS*) (Epstein 1969; Wilks 1995; Goddard *et al* 2003). The computation of *RPSS* begins with the rank probability score (*RPS*) which is defined as:

$$RPS = \sum_{m=1}^{15} \left(CP_{F_m} - CP_{O_m} \right)^2 \tag{3}$$

where the vector CP_{Fm} represents the cumulative probability of the forecast up to bin m, CP_{Om} is the cumulative 'observed' probability up to bin m, and there are 15 equi-sized bins of ENSO sea surface temperature anomaly. CP_{Om} is a step function from zero to one at the bin in which the actual observed value falls. For a perfect forecast $RPS = 0$. The *RPS* is referenced to climatology to give the *RPSS* which, for n forecasts, is defined as:

$$RPSS = 1 - \frac{\sum_{1}^{n} RPS_f}{\sum_{1}^{n} RPS_{cl}} \tag{4}$$

where RPS_f is the *RPS* of the forecast and RPS_{cl} is the *RPS* of the climatology (i.e. zero anomaly) forecast. The maximum *RPSS* is 1; a negative *RPSS* indicates skill worse than climatology. *RPSS* is regarded as a 'harsh' seasonal forecast skill measure with values of 0.10 being considered respectable and values of 0.20 as very good (Barnston, personal communication 2003).

The *RPSS* is computed and compared for three different hindcast formulations 1952-2002, two of which are probabilistic. The first probabilistic hindcast formulation (termed 'normal') fits a normal distribution to the cross-validated hindcast errors 1952-2002. This normal distribution gives a prediction interval around the deterministic hindcast value thereby providing the cumulative probability distribution. The second probabilistic formulation (termed 'ensemble') bins individual ensemble members according to size to obtain directly the cumulative probability distribution. The third formulation (termed 'deterministic') employs the deterministic hindcast values and is included for reference. The climatology cumulative probability distribution is obtained in each case by populating bins with the observed values 1952-2002.

f. Data

The monthly ENSO indices and Southern Oscillation Index (SOI) data employed in the cross-validated hindcasts 1952-2002 are supplied by the US Climate Prediction Center (CPC). The ENSO indices are obtained from a weekly 1° spatial resolution optimum interpolation SST analysis (Reynolds et al. 2002). Although the CPC data begin in 1950 our first cross-validated hindcast is for August-September 1952. The data in 1950 and 1951 are reserved to compute the 5 month trends in predictor categories (A) and (B) at the longest leads. The independent (replicated real-time) ENSO forecasts for 1900-1950 employ the Kaplan *et al*. (1998) reconstructed sea surface temperatures 1870-1950 and historical SOI values from the Climatic Research Unit at the University of East Anglia 1870-1950 (compiled using the method given in Ropelewski and Jones (1987)).

3. Results

a. Standard ENSO-CLIPER cross-validated hindcasts

The standard ENSO-CLIPER model cross-validated hindcast skills for predicting the August-September (henceforth AS) Niño 1+2, 3, 3.4 and 4 indices 1952-2002 are shown in Figure 1. Skills are shown as a function of monthly lead out to 10 months (prior October). *MSSS* decays gradually for all indices from ~90% at zero lead to ~20% at a lead of 4 months. Skill attributable to persistence, whilst initially similar to that of the standard ENSO-CLIPER model, decays more rapidly and (with the exception of Niño 1+2) is always negative at 4 months lead.

The standard CLIPER model provides the largest (~20%) absolute improvement in *MSSS* over persistence at leads of 3 and 4 months. At leads of 5 months and greater the standard ENSO-CLIPER model skill is zero. This is a direct consequence of the model formulation since when no predictors are found (as tends to be the case at the longer leads) no hindcast is made resulting in a zero *MSSS*. The same is not true for persistence which is free to yield wildly inaccurate hindcasts. The slight improvement in persistence skill at the longest leads is noteworthy. This is an artefact of the *MSSS* decomposition, which as shown in Equation (2), contains a term penalising bias. Hindcast bias will be coupled to the annual cycle and is expected to be minimised at 12 months lead.

Confidence in the skill estimates for the standard ENSO-CLIPER model varies with lead. The 95% confidence interval grows from ~10% absolute width at zero lead to 30-60% width at leads of 3-6 months before settling back to ~20% width at longer leads. Thus there is confidence of high skill at short lead and of no skill at long lead. Overall, AS Niño 4 is the best predicted index with model hindcast *MSSS* skill positive to 97.5% confidence at leads out to 4 months or early April and better than persistence at all leads. These findings concur with Barnston and Ropelewski (1992) who reported an increase in ENSO forecast skill from east to west across the Pacific Ocean.

b. Temporal stability

Analyses were performed on the sub-periods 1952-1975 and 1976-2002 to assess the temporal stability of the standard ENSO-CLIPER model AS hindcast skill. These results are displayed by ENSO region in Figure 2 with the early period in the left hand column and the later period on the right. The results for the AS Niño 3.4, 3, and 4 indices appear stable for both CLIPER and persistence. The variation of skill with lead is similar for both time periods and the skill traces for each period generally fit within the other period's 95% confidence intervals. That said, the hindcast skill for the AS Niño 3 index is higher in the first (1952-1975) split while the hindcast skill for the AS Niño 4 index is higher in the second (1976-2002) split. This shift towards higher (lower) AS ENSO skill in the west (east) in recent times is reflected most by the Niño 1+2 index. The latter shows a 60% reduction in absolute skill and a 40% reduction in persistence at leads of 3-5 months between 1952-1975 and 1976-2002.

Kirtman and Schopf (1998) found ENSO skill to be higher in periods where the predictand variance is greatest. Standard deviations of the AS Niño 1+2 index for the first and second splits are 1.0 °C and 1.2 °C respectively. Thus, a change in variance can not explain the change in skill. Examination of the hindcast time series (not shown) reveals that the reduction in the Niño 1+2 skill may arise from the poor prediction of the 1997 El Niño event and from an errant prediction of positive conditions for the summer of 1992 when in reality neutral conditions prevailed. With these years eliminated, the skills in the second split show a closer resemblance to those in the first. A further plausible explanation for the drop in Niño 1+2 skill during the period 1976-2002 relative to 1952-1975 is that in the earlier period El Niño tended to start from Peruvian waters and spread westward. In the more recent period it has tended to start from the central equatorial Pacific and spread eastward. This delay in reaching the South American coast could mean that the Niño 1+2 SST anomalies were less well developed in August-September in the more recent period and thus harder to predict.

The temporal splits in Figure 2 show that the 95% skill confidence intervals for the Niño 3 and Niño 1+2 indices are far wider in the second split than the first. Wang *et al*. (2000) found greater sensitivity in skill for splits of Niño 3 than Niño 4. This was attributed to the increase in SST variance as the equatorial Pacific is traversed from west to east. A similar explanation combined with the poor prediction of the 1997 El Niño may account for the wider confidence intervals in the later split. However, caution must be applied in interpreting skill estimates based on a sample of just 25 years.

c. Sensitivity to significance level

The sensitivity of the standard CLIPER model to the 5% significance level used to screen potential predictors was assessed in terms of *MSSS*. Comparisons were made between models screened at significance levels of 1%, 5% and 10% (all other restrictions being left unchanged). Results for each ENSO region are shown in Figure 3. For completeness each panel includes the standard persistence skill from Figure 1 and the *MSSS* from a 'consensus' model defined as the skill from the average of the hindcasts made with the three individual significance levels. It is clear that the predictor screening significance level has little effect upon the 1951-2002 model performance, changing it at best by ~10%. This result might be expected since poor predictors will be rejected at the subsequent leaps-and-bounds (L&B) predictor optimisation stage. The main advantage of predictor screening is to increase computation efficiency. Each reduction in the number of potential predictors passed to the L&B algorithm yields a saving of at least 6 floating point operations (Furnival and Wilson 1974). Figure 3 also shows that, in general, the consensus model outperforms the individual significance level models.

d. Sensitivity to percentage of variance explained improvement factor

Changes in the *MSSS* 1952-2002 resulting from variation of the PVE (percentage of variance explained) improvement factor passed to the L&B algorithm in the standard CLIPER model were investigated for PVE factors of 1%, 2.5% and 5%. These are shown in Figure 4. Once again, the remaining restrictions were left unchanged. With the exception of the Niño 3.4

index at leads of 2-4 months where *MSSS* differences of 20% are seen, the model skill is found to be insensitive to the PVE improvement factor. Higher values of the improvement factor were also investigated. In general these resulted in a single predictor model since a further predictor could not be found to provide the required leap in PVE.

e. Sensitivity to averaging period

The final CLIPER sensitivity restriction investigated was the averaging period for the teleconnected ENSO initial condition and trend predictors (predictor categories (C) and (D) in section 2*a*). Figure 5 shows skill plots for each region constructed using models built separately using 1, 3 and 6 month averages of the teleconnected predictors. Again other sensitivity factors were left unchanged. The results display a similar pattern to Figure 4 with sensitivity limited to Niño 3.4 at leads of 2-3 months where *MSSS* differences approaching 30% are seen. As with Figure 3, the consensus model generally outperforms the models built with an individual averaging period.

f. A consolidated model

In the absence of any clear physical justification for the level of predictor screening, L&B improvement factor or teleconnected predictor averaging period, it seems reasonable to consolidate the hindcasts from each model into a single aggregate hindcast. A 'consolidated' ENSO-CLIPER model is defined as the mean of 18 ensemble model hindcasts formulated with PVE improvement factors of 1%, 2.5% and 5% and averaging periods of 1-6 months and no predictor screening.

The 'consolidated' CLIPER model 51-year cross-validated skill for the prediction of AS ENSO for all ENSO regions is displayed in Figure 6. Skills from the standard ENSO-CLIPER model are included for comparison (filled circles). For all regions and at all leads it is clear that the consolidated model outperforms (or at worst matches) the *MSSS* skill of the standard CLIPER model. The skill difference between the two models is quantified in Table 2 and discussed below. Confidence intervals for the estimation of *MSSS* are similar overall for both models but narrower for the consolidated model at leads of 0-4 months. The consolidated model *MSSS* skill is positive to 97.5% confidence at leads out to 4 months or early April for all ENSO indices (for Niño 4 and Niño 1+2 it is to leads of 5 months or early March); in comparison, the standard CLIPER *MSSS* skill is positive to 97.5% confidence at leads out to only 1 month for Niño 3.4 and 2 months for Niño 1+2. The consolidated model shows similar temporal stability (not shown) to that seen for the standard CLIPER model but with correspondingly higher skills.

Absolute percentage improvements in *MSSS* and *RMSSS* of the consolidated model over persistence and of the consolidated model over the standard model are presented in Table 2. The consolidated model outperforms persistence at all leads. Hindcasts from the consolidated and standard models are nearly identical at zero and 1 month leads since all formulations tend to favour simple persistence of the predictand. Similarly, at very long leads when predictors become scarce, all formulations tend to a zero hindcast. It is at leads from 2 to 6 months where the consolidated CLIPER model offers the greatest improvement over the standard CLIPER model for predicting August-September ENSO. Assessed over the 51-year period 1952-2002 the consolidated model provides a 10-20% absolute improvement in *MSSS* at all leads from 2 to 6 months for all the main ENSO index regions 3.4, 3, and 4; for the 1+2 index region the improvement is ~5%. The largest 51-year improvement in *MSSS* is 31% for the AS Niño 3.4 region at 2 months lead. Table 2 also shows that the skill values for improvements in root mean square error are smaller than for *MSSS*. This indicates that a proportion of the consolidated model skill comes from the successful prediction of ENSO extremes.

To aid the further comparison of the consolidated and standard CLIPER cross-validated model skill we include (Table 3) the reduction in root mean square error (RMSE) and mean absolute error (MAE) afforded by the consolidated model over the standard model for each ENSO index and lead. Values for the standard deviation (s.d.), $RMSE_{cl}$ and MAE_{cl} of each August-September ENSO index are also included in Table 3 to help evaluation of these data. Table 3 shows for the Niño 3.4, 3 and 4 index regions at leads between 2 and 6 months that the consolidated model gives a mean improvement of 0.06 to 0.08 °C in RMSE and of 0.05 to 0.06 °C in MAE over the standard model. These improvements may be slightly less than the natural uncertainty associated with the measurement of AS SST in the ENSO regions but the consistency in sign in Table 3 (which would not be expected if the improvements were due to chance) shows that the consolidated model provides a real benefit.

g. Probabilistic skill 1952-2002

The consolidated and standard CLIPER models are compared in terms of their rank probability skill in Table 4. The Table shows the *RPSS* 1952-2002 from the three different hindcast formulations termed ‘deterministic’, ‘normal’ and ‘ensemble’ described in section 2*e*. The ‘deterministic’ formulation is applied to the two CLIPER models and to persistence, the ‘normal’ probabilistic formulation is applied to the two CLIPER models, and the ‘ensemble’ probabilistic formulation is applied perforce only to the consolidated CLIPER model. Table 3 shows, as expected, that the consolidated model *RPSSs* are generally higher than those from the standard model. The consolidated CLIPER ‘normal’ model outperforms the consolidated ‘ensemble’ model which in turn outperforms the ‘deterministic’ model. The consolidated ‘normal’ scores model provides positive *RPSS* at all leads out to 6 months for all ENSO indices in agreement with the *MSSS* deterministic results which also showed skill to 6 months lead. Taking *RPSS* values of 0.10 as being respectable (Barnston, personal communication 2003), the consolidated CLIPER model is seen to provide respectable probabilistic predictive skill for all AS ENSO indices at leads out to 4 or 5 months.

The improvement in *RPSS* of the consolidated model over the standard CLIPER follows directly from the former’s better deterministic skill and narrower error distribution. It is interesting to note that the consolidated 'ensemble’ scores are higher than the ‘deterministic’ ones. This implies that additional information may be contained within the ensemble hindcasts and that simply averaging these together may not yield the best hindcast.

h. Independent forecasts 1900-1950

The replicated real-time forecast skill 1900-1950 of the consolidated ENSO-CLIPER model for predicting the August-September Niño 3.4, 3, 4 and 1+2 indices at monthly leads out to 6 months is compared against persistence in Figure 7 and against the standard ENSO-CLIPER model in Figure 8. The skill measure used is the mean square skill score (MSSS). For all regions and at all leads it is clear that the consolidated model outperforms persistence and outperforms (or at worst matches) the *MSSS* skill of the standard CLIPER model. These skill differences are quantified in Table 5 and discussed below. Confidence intervals for the estimation of *MSSS* are narrower for the consolidated model at all leads and for all ENSO indices; the only exception being Niño 1+2 for the comparison between the consolidated and standard CLIPER models. The consolidated model *MSSS* skill is positive to 97.5% confidence at leads out to 4 months or early April for AS Niño 4 and out to leads of 2 months or early June for the other AS ENSO indices.

Absolute percentage improvements in *MSSS* and *RMSSS* of the consolidated model over persistence and of the consolidated model over the standard model are presented in Table 5 for independent (replicated real-time) forecasts 1900-1950. As with the cross-validated hindcasts 1952-2002 (Table 2) the consolidated model outperforms persistence at all leads and outperforms the standard CLIPER model at leads from 1 to 6 months for the main ENSO index regions 3.4, 3 and 4. Assessed over the 51-year period 1900-1950 the consolidated model provides a 5-10% absolute improvement in *MSSS* at all leads from 1 to 5 months for the Niño 3.4, 3, and 4 regions; for the 1+2 index region there is little improvement. The largest improvement in *MSSS* is 15% for the AS Niño 4 region at leads of 2 and 4 months. The skill values for improvements in *RMSSS* are smaller than for *MSSS*.

4. Discussion

Figures 4 and 5 show that the standard ENSO-CLIPER predictions of Niño 3.4 at leads of 2-3 months are sensitive to both the L&B improvement factor and to the intrinsic averaging procedure imposed upon predictor categories C and D. Figure 9 displays histograms of the number of times that each of the 14 predictors are used in predicting Niño 3.4 1952-2002 at a lead 3 months for averaging periods of 1-6 months. There is considerable variation in the model formulation as the averaging period is changed. As the latter increases there is shift from models reliant upon predictors 6 and 7 to those using predictors 3, 4 and 5. Reference to Table 1 reveals that the dominant predictors under 1 month averaging are the 5 month trend in Niño 3.4 and the persisted 3-month value of Niño 1+2. When the averaging period of the teleconnected SSTs is extended to 6 months these are rejected in favour of shorter period trends and initial conditions of the predictand itself. It appears that teleconnected SSTs (predictors 7 through 14) only become useful when they are computed for a period similar to that of the predictand itself. It is notable that predictors 11-14 are never selected in any model formulation. This is a likely result of the inter-correlation between the predictors and the order in which they are presented to the leaps and bound algorithm. In the situation where the predictor pool is inter-correlated the likelihood of each successive predictor explaining additional variance will decrease with each additional predictor.

The consolidated model is seen to outperform the standard ENSO-CLIPER model for all the indices studied. The greatest improvements are found at leads of 2-6 months which are precisely the leads at which model instability is identified. Averaging the separate models has the effect of reinforcing the consensus of the individual members. Thus, when the models are in agreement a sharp hindcast is issued. Conversely, if there is no consensus the individual predictions will tend to cancel each other out and the hindcast value will tend to zero.

Decomposition of the *MSSS* into temporal, amplitude and bias errors allows an assessment of how each error term contributes to the skill improvement. Plots of correlation (not shown) follow the same pattern as found for *MSSS* (see Figure 6). The consolidated model yields higher and less volatile correlations with the largest improvements seen for Niño 4. The effect of consolidation on the amplitude ratio is neutral. The amplitude ratios for both models are always less than one, *i.e.* they under predict the observed variance in SST. This is apparent particularly at long leads where the hindcasts tend to the climatological value. Bias errors are negligible for both models and are always less than 0.1°C. Thus the skill improvement afforded by the consolidated model must arise from a reduction in the temporal error, *i.e.* through improved prediction of the timing of events.

A simple method for correcting biases in the mean and variance of a hindcast is to perform the linear regression (Déqué 2003)

$$\hat{x}' = E\langle x|\hat{x}\rangle = \beta_0 + \beta_1\hat{x} \tag{5}$$

where β_0 and β_1 are respectively the bias in the mean and variance of the hindcasts. Following the cross-validation procedure, the consolidated hindcasts were recalibrated using parameters estimated from data excluding a 5 year block about the target year. The revised *MSSS* values show little improvement over the raw hindcasts. Since the recalibration amounts to a linear transformation of the hindcast values it cannot change, $r_{\hat{x},x}$, the product moment correlation between the hindcast/observation pairs. Further as noted above, the hindcast bias is negligible. Thus, the only scope for improvement in *MSSS* arises from adjustment of the hindcast variance. Given the minimal improvement in *MSSS* post recalibration, it is concluded that there is no significant bias in the consolidated hindcast variance, and thus the remaining unexplained variance must be attributable to factors outside of the model and/or to non-linear interactions.

Neither the standard nor the consolidated ENSO-CLIPER model is found to be skillful prior to March (lead of 5 months), this corresponding to the onset of the 'spring predictability barrier' (Torrence and Webster 1998). The likely failing of the models results from their heavy reliance (by design) on persistence which often breaks down during this time of the year. The inclusion of long-term trends is insufficient to predict phase changes from winter into summer.

Optimisation of the consolidated CLIPER model may lead to further skill improvements. The model presented here (defined as the mean of an ensemble of 18 models built using 6 teleconnected predictor averaging periods and 3 PVE improvement factors) was selected from the visual inspection of Figures 3-5 and for computational expediency. Improved hindcast skill may be obtained from an optimised multi-ensemble consolidated ENSO-CLIPER model which includes the capability to select ensemble models built (a) using predictors in categories (A) and (B) computed over non 1, 3- and 5-month means, (b) using different predictor significance level screening factors and (c) using more than 18 ensembles. Additional skill may also be obtainable through the deployment of phase dependent models. Previous studies (*e.g.* Mason and Mimmack 2002) have found that ENSO is more predictable when in its positive phase.

5. Conclusions

A 'Consolidated' ENSO-CLIPER seasonal prediction model has been presented to address the issue of improving summer ENSO predictive skill due to the spring predictability barrier between March and May. Consolidated CLIPER comprises the ensemble of 18 model variants of the statistical ENSO-CLIPER (CLImatology and PERsistence) prediction model. Assessing August-September ENSO skill using deterministic and probabilistic skill measures applied to cross-validated hindcasts 1952-2002 and deterministic skill measures applied to replicated real-time forecasts 1900-1950, shows that the consolidated CLIPER model consistently outperforms the standard CLIPER model at all leads from 2 to 6 months for all the main ENSO indices (3, 3.4 and 4). The new model provides up to a 30% (15%) reduction in mean square error 1952-2002 (1900-1950). However, it must be noted that the formulation of the consolidation remains arbitrary, representing a small subset of all the possible CLIPER formulations and thus may be far from optimal. Decomposition of the *MSSS* into correlation, variance ratio and bias shows that the consolidated model also provides superior predictions of the timing and amplitude of ENSO events compared to the standard CLIPER model.

This investigation has focused on the predictability of summer ENSO conditions. Ongoing research will extend the consolidated ENSO-CLIPER results to other seasons and will compare hindcast skill performance and model versatility (ie range of predictand periods, range of forecast lead times and speed of forecast/hindcast computation) to that achieved by leading dynamical ENSO prediction models.

Acknowledgements

Benjamin Lloyd-Hughes is and Paul Rockett was supported by the TSR (Tropical Storm Risk) venture. We thank three referees for helpful comments. We kindly acknowledge the U. S. Climate Prediction Center for providing data.

References

Barnston, A. G., and C. F. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate*, **5,** 1316-1345.

——, M. H. Glantz, and Y. He, 1999: Predictive skill of statistical and dynamic climate models in forecasts of SST during the 1997-98 El Niño episode and the 1998 La Niña onset. *Bull. Amer. Meteor. Soc.*, **80**, 217-243.

Bove, M. C., J. B. Elsner, C. W. Landsea, X. Niu, and J. J. O'Brien, 1998: Effect of El Niño on U.S. landfalling hurricanes revisited. *Bull. Amer. Meteor. Soc.*, **79**, 2477-2482.

Chan, J. C. L., 1985: Tropical cyclone activity in the northwest Pacific in relation to the El Niño/Southern Oscillation phenomenon. *Mon. Wea. Rev.*, **113**, 599-606.

Dai, A., and T. M. L. Wigley, 2000: Global patterns of ENSO-induced precipitation. *Geophys. Res. Lett.*, **27**, 1283-1286.

Déqué, M., 2003: Continuous variables. *Forecast Verification – A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, Chichester, UK, 97-119.

Efron, B., and G. Gong, 1983: A leisurely look at the bootstrap, the jackknife and cross validation. *American Statistician*, **37,** 36-48.

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

Furnival, G. M., and R. W. Wilson, 1974: Regression by leaps and bounds. *Technometrics*, **16**, 499-511.

Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane, 2001: Current approaches to seasonal-to-interannual climate predictions. *Int. J. Climatol.*, **21**, 1111-1152.

——, A. G. Barnston, and S. J. Mason, 2003: Evaluation of the IRI's "net assessment" seasonal climate forecasts. *Bull. Amer. Meteor. Soc.*, **84**, 1761-1781.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*, **128**, 1187-1193.

Gray, W. M., 1984: Atlantic seasonal hurricane frequency: Part I: El Niño and 30mb quasi-biennial oscillation influences. *Mon. Wea. Rev.*, **112**, 1649-1668.

Kaplan, A., M. A. Cane, Y. Kushnir, A. Clement, M.Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856-1991. *J. Geophys. Res.*, **103**, 18567-18589.

Kerr, R. A., 2000: Second thoughts on skill of El Niño predictions. *Science*, **290**, 257-258.

Kirtman, B. P., and P. S. Schopf, 1998: Decadal variability in ENSO predictability and prediction. *J. Climate*, **11,** 2804-2822.

——, J. Shukla, M. Balmaseda, N. Graham, C. Penland, Y. Xue, and S. Zebiak, cited 2002: Current status of ENSO forecast skill. A report to the Climate Variability and Predictability (CLIVAR) Numerical Experimentation Group (NEG), CLIVAR Working Group on Seasonal to Interannual Prediction. [Available online from http://www.clivar.org/publications/wg_reports/wgsip/nino3/report.htm.]

Knaff, J. A., and C. W. Landsea, 1997: An El-Niño Southern Oscillation climatology and persistence (CLIPER) forecasting scheme. *Wea. Forecasting*, **12,** 633-652.

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

Latif, D., D. Anderson, T. Barnett, M. Cane, R. Kleeman, A. Leetmaa, J. O'Brien, A. Rosati, and E. Schneider, 1998: A review of the predictability and prediction of ENSO. *J. Geophys. Res.,* **103**, 14**,**375-14,393.

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

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to a correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417-2424.

Rasmusson, E. M., and T. H. Carpenter, 1982: Variations in tropical sea surface temperature and surface wind fields associated with the Southern Oscillation/El Niño. *Mon. Wea. Rev.*, **110**, 354-384.

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

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.

——, and P. D. Jones, 1987: An extension of the Tahiti-Darwin Southern Oscillation Index. *Mon Wea. Rev.*, **115**, 2161-2165.

Saunders, M. A., R. E. Chandler, C. J. Merchant, and F. P. Roberts, 2000: Atlantic hurricanes and NW Pacific typhoons: ENSO spatial impacts on occurrence and landfall. *Geophys. Res. Lett.*, **27**, 1147-1150.

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

Torrence, C., and P. J. Webster, 1998: The annual cycle of persistence in the El Niño-Southern Oscillation. *Quart. J. Roy. Meteor. Soc.,* **124,** 1985-2004.

Trenberth, K. E., 1997: The definition of El Niño. *Bull. Amer. Meteor. Soc.*, **78**, 2771-2778.

Unger, D., A. Barnston, H. Van den Dool, and V. Kousky, 1996: Consolidated forecasts of tropical Pacific SST in Niño 3.4 using two dynamical models and two statistical models. *Experimental Long-Lead Forecast Bulletin*, **5**, No. 1, 50-52.

Walker, G. T., and E. W. Bliss, 1932: World Weather V. *Mem. Roy. Meteor. Soc.*, **4**, No. 36, 53-84.

Wang, G., R. Kleeman, N. Smith, and F. Tseitkin, 2002: The BMRC coupled general circulation model ENSO forecast system. *Mon. Wea. Rev.,* **130,** 975-991.

Wilks, D., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

WMO, 2002: *Standardised Verification System (SVS) for Long-Range Forecasts (LRF).* New Attachment II-9 to the Manual on the GDPS (WMO-No. 485), Volume I, WMO, Geneva, 21 pp.

Figure Captions

Figure 1. Cross-validated hindcast skill from the standard ENSO-CLIPER model for predicting the August-September Niño 3.4, 3, 4 and 1+2 indices 1952-2002 at monthly leads out to 10 months. The skill measure used is the mean square skill score (*MSSS*) defined as the percentage improvement in mean square error over a hindcast of zero anomaly; the climatology being 1952-2002. The grey band is a bootstrapped estimate of the 95% confidence interval for the skill measure. The skill and uncertainty from standard persistence are shown by the filled circles and error bars.

Figure 2. As Figure 1 but for the sub periods 1952-1975 (left column) and 1976-2002 (right column).

Figure 3. The sensitivity of the standard ENSO-CLIPER model cross-validated hindcast skill to the significance level imposed during predictor screening for the prediction of August-September Niño 3.4, 3, 4 and 1+2 indices 1952-2002 at monthly leads to 9 months. The 'consensus' skill refers to the average of the three hindcasts obtained using significance levels of 1%, 5% and 10%. The standard persistence skill from Figure 1 is included for reference.

Figure 4. The sensitivity of the standard ENSO-CLIPER model cross-validated hindcast skill to the PVE improvement factor passed to the leaps and bounds algorithm for the prediction of August-September Niño 3.4, 3, 4 and 1+2 indices 1952-2002 at monthly leads to 9 months. The 'consensus' skill refers to the average of the three hindcasts obtained using leaps and bounds improvement factors of 1%, 2.5% and 5%.

Figure 5. The sensitivity of the standard ENSO-CLIPER model cross-validated hindcast skill to the teleconnected predictor averaging period used in the model formulation for the prediction of August-September Niño 3.4, 3, 4 and 1+2 indices 1952-2002 at monthly leads to 9 months. The 'consensus' skill refers to the average of the three hindcasts obtained using averaging periods of 1, 3 and 6 months.

Figure 6. Cross-validated hindcast skill 1952-2002 of the consolidated ENSO-CLIPER model compared against the standard ENSO-CLIPER model for predicting the August-September Niño 3.4, 3, 4 and 1+2 indices at monthly leads out to 9 months. The skill measure used is the mean square skill score (*MSSS*) defined as the percentage improvement in mean square error over a hindcast of zero anomaly; the climatology being 1952-2002. The grey band is a bootstrapped estimate of the 95% confidence interval for the skill measure. The skill and uncertainty from the standard ENSO-CLIPER model are shown by the filled circles and error bars.

Figure 7. Replicated real-time forecast skill 1900-1950 of the consolidated ENSO-CLIPER model compared against persistence for predicting the August-September Niño 3.4, 3, 4 and 1+2 indices at monthly leads out to 6 months. The skill measure used is the mean square skill score (*MSSS*) defined as the percentage improvement in mean square error over a forecast of zero anomaly; the climatology being 1900-1950. The grey band is a bootstrapped estimate of the 95% confidence interval for the skill measure. The skill and uncertainty from persistence are shown respectively by the filled circles and error bars.

Figure 8. Replicated real-time forecast skill 1900-1950 of the consolidated ENSO-CLIPER model compared against the standard ENSO-CLIPER model for predicting the August-September Niño 3.4, 3, 4 and 1+2 indices at monthly leads out to 6 months. The skill measure and presentation format are the same as in Figure 7.

Figure 9. Histograms of the standard ENSO-CLIPER predictors selected for making hindcasts of the August-September Niño 3.4 index 1952-2002 at a lead of 3 months (early May) for models built with teleconnected predictor averaging periods from 1 to 6 months. The predictor numbers (1 to 14) correspond to the classification in Table 1.

Table Captions

TABLE 1. Predictor pools in the standard ENSO-CLIPER model for predicting the Niño 3.4, 3, 4 and 1+2 indices. IC and TR represent respectively initial condition and trend predictors with the numeral designating whether these are 1, 3 or 5 month means as defined by Knaff and Landsea (1997). SOI is the Southern Oscillation Index.

TABLE 2. Improvement afforded by the consolidated ENSO-CLIPER model over (a) persistence and (b) the standard ENSO-CLIPER model for predicting August-September Niño 3.4, 3, 4 and 1+2 as a function of monthly lead from cross-validated hindcasts 1952-2002. Values are given as the absolute difference in *MSSS* and *RMSSS* (in brackets).

TABLE 3. Comparison of the consolidated ENSO-CLIPER and standard ENSO-CLIPER cross-validated hindcast skill 1952-2002 for predicting August-September (AS) Niño 3.4, 3, 4 and 1+2 1952-2002 as a function of monthly lead in terms of root mean square error (RMSE), mean absolute error (MAE) and the improvement in each measure that the consolidated model offers.

TABLE 4. Rank probability skill score (*RPSS*) 1952-2002 of the consolidated ENSO-CLIPER model compared against the standard ENSO-CLIPER model and persistence for predicting August-September (AS) Niño 3.4, 3, 4 and 1+2 as a function of monthly lead out to 6 months. The *RPSS* is compared for different CLIPER probabilistic hindcasts as described in the text.

TABLE 5. Improvement afforded by the consolidated ENSO-CLIPER model over (a) persistence and (b) the standard ENSO-CLIPER model for predicting August-September Niño

3.4, 3, 4 and 1+2 as a function of monthly lead from replicated real-time forecasts 1900-1950. Values are given as the absolute difference in *MSSS* and *RMSSS* (in brackets).

TABLES

TABLE 1. Predictor pools in the standard ENSO-CLIPER model for predicting the Niño 3.4, 3, 4 and 1+2 indices. IC and TR represent respectively initial condition and trend predictors with the numeral designating whether these are 1, 3 or 5 month means as defined by Knaff and Landsea (1997). SOI is the Southern Oscillation Index.

Predictor number	Predictand Niño 3.4	Niño 3	Niño 4	Niño 1+2
1	Niño 3.4 IC-1	Niño 3 IC-1	Niño 4 IC-1	Niño 1+2 IC-1
2	Niño 3.4 IC-3	Niño 3 IC-3	Niño 4 IC-3	Niño 1+2 IC-3
3	Niño 3.4 IC-5	Niño 3 IC-5	Niño 4 IC-5	Niño 1+2 IC-5
4	Niño 3.4 TR-1	Niño 3 TR-1	Niño 4 TR-1	Niño 1+2 TR-1
5	Niño 3.4 TR-3	Niño 3 TR-3	Niño 4 TR-3	Niño 1+2 TR-3
6	Niño 3.4 TR-5	Niño 3 TR-5	Niño 4 TR-5	Niño 1+2 TR-5
7	Niño 1+2 IC-3	Niño 1+2 IC-3	Niño 1+2 IC-3	Niño 3 IC-3
8	Niño 1+2 TR-3	Niño 1+2 TR-3	Niño 1+2 TR-3	Niño 3 TR-3
9	Niño 3 IC-3	Niño 3 IC-3	Niño 3 IC-3	Niño 4 IC-3
10	Niño 3 TR-3	Niño 3 TR-3	Niño 3 TR-3	Niño 4 TR-3
11	Niño 4 IC-3	Niño 3.4 IC-3	Niño 3.4 IC-3	Niño 3.4 IC-3
12	Niño 4 TR-3	Niño 3.4 TR-3	Niño 3.4 TR-3	Niño 3.4 TR-3
13	SOI IC-3	SOI IC-3	SOI IC-3	SOI IC-3
14	SOI TR-3	SOI TR-3	SOI TR-3	SOI TR-3

TABLE 2. Improvement afforded by the consolidated ENSO-CLIPER model over (a) persistence and (b) the standard ENSO-CLIPER model for predicting August-September Niño 3.4, 3, 4 and 1+2 as a function of monthly lead from cross-validated hindcasts 1952-2002. Values are given as the absolute difference in *MSSS* and *RMSSS* (in brackets).

Niño index	Lead (months) 0	1	2	3	4	5	6
(a) Consolidated CLIPER improvement over persistence							
3.4	5 (7)	14 (12)	22 (14)	55 (30)	110 (49)	179 (70)	233 (84)
3	2 (3)	6 (6)	16 (11)	34 (19)	78 (39)	126 (53)	204 (77)
4	4 (5)	12 (10)	18 (12)	42 (26)	87 (45)	108 (47)	115 (48)
1+2	13 (16)	28 (25)	30 (19)	21 (12)	10 (5)	34 (17)	69 (30)
(b) Consolidated CLIPER improvement over standard CLIPER							
3.4	0 (0)	0 (0)	31 (19)	10 (6)	7 (4)	17 (8)	12 (6)
3	0 (0)	11 (9)	7 (5)	16 (10)	15 (9)	18 (9)	18 (9)
4	0 (0)	6 (5)	18 (12)	15 (10)	26 (16)	23 (12)	7 (4)
1+2	2 (3)	2 (3)	-5 (-4)	19 (11)	7 (4)	7 (4)	1 (1)

TABLE 3. Comparison of the consolidated ENSO-CLIPER and standard ENSO-CLIPER cross-validated hindcast skill 1952-2002 for predicting August-September (AS) Niño 3.4, 3, 4 and 1+2 1952-2002 as a function of monthly lead in terms of each model's root mean square error (RMSE), mean absolute error (MAE) and the improvement offered by the consolidated model over the standard model for each measure.

Lead (months)	RMSE (°C) Consolidated CLIPER	RMSE (°C) Standard CLIPER	RMSE (°C) Difference	MAE (°C) Consolidated CLIPER	MAE (°C) Standard CLIPER	MAE (°C) Difference
(a) AS Niño 3.4 (s.d. = 0.81°C; $RMSE_{cl}$ = 0.81°C; MAE_{cl} = 0.64°C)						
0	0.26	0.26	0.00	0.21	0.21	0.00
1	0.41	0.41	0.00	0.34	0.33	0.01
2	0.57	0.73	-0.15	0.45	0.57	-0.12
3	0.62	0.67	-0.05	0.50	0.56	-0.06
4	0.68	0.72	-0.03	0.56	0.59	-0.03
5	0.74	0.81	-0.07	0.58	0.63	-0.05
6	0.77	0.82	-0.05	0.60	0.64	-0.04
(b) AS Niño 3 (s.d. = 0.89°C; $RMSE_{cl}$ = 0.89°C; MAE_{cl} = 0.69°C)						
0	0.31	0.31	0.00	0.24	0.24	0.00
1	0.48	0.56	-0.08	0.37	0.46	-0.09
2	0.61	0.65	-0.04	0.48	0.49	-0.01
3	0.69	0.78	-0.09	0.52	0.61	-0.09
4	0.71	0.79	-0.08	0.54	0.59	-0.05
5	0.82	0.90	-0.08	0.63	0.69	-0.06
6	0.82	0.91	-0.09	0.65	0.70	-0.05
(c) AS Niño 4 (s.d. = 0.57°C; $RMSE_{cl}$ = 0.58°C; MAE_{cl} = 0.47°C)						
0	0.21	0.21	0.00	0.16	0.16	0.00
1	0.29	0.32	-0.03	0.24	0.27	-0.03
2	0.37	0.44	-0.07	0.31	0.35	-0.05
3	0.38	0.44	-0.06	0.32	0.36	-0.04
4	0.41	0.50	-0.09	0.35	0.43	-0.08
5	0.51	0.57	-0.07	0.39	0.46	-0.07
6	0.53	0.55	-0.02	0.42	0.44	-0.02
(d) AS Niño 1+2 (s.d. = 1.08°C; $RMSE_{cl}$ = 1.08°C; MAE_{cl} = 0.83°C)						
0	0.36	0.40	-0.03	0.29	0.32	-0.03
1	0.48	0.51	-0.03	0.37	0.41	-0.04
2	0.72	0.67	0.04	0.55	0.53	0.02
3	0.87	0.99	-0.12	0.59	0.72	-0.12
4	0.94	0.98	-0.04	0.67	0.71	-0.04
5	0.96	1.00	-0.04	0.71	0.75	-0.04
6	1.05	1.06	-0.01	0.79	0.81	-0.02

TABLE 4. Rank probability skill score (*RPSS*) 1952-2002 of the consolidated ENSO-CLIPER model compared against the standard ENSO-CLIPER model and persistence for predicting August-September (AS) Niño 3.4, 3, 4 and 1+2 as a function of monthly lead out to 6 months. The *RPSS* is compared for different CLIPER probabilistic hindcasts as described in the text.

	Rank probability skill score (*RPSS*)					
Lead (months)	Persistence (deterministic)	Standard CLIPER (deterministic)	Consolidated CLIPER (deterministic)	Consolidated CLIPER (ensemble)	Consolidated CLIPER (normal)	Standard CLIPER (normal)
			(a) AS Niño 3.4			
0	0.49	0.52	0.46	0.53	0.64	0.65
1	0.05	0.23	0.17	0.33	0.45	0.46
2	-0.28	-0.16	-0.04	0.18	0.29	0.09
3	-0.46	-0.31	-0.16	0.03	0.21	0.14
4	-1.00	-0.25	-0.10	-0.01	0.14	0.11
5	-1.38	-0.28	-0.28	-0.06	0.06	-0.03
6	-1.71	-0.40	-0.19	-0.10	0.03	-0.03
			(b) AS Niño 3			
0	0.45	0.48	0.48	0.48	0.60	0.60
1	0.26	0.15	0.26	0.37	0.47	0.36
2	0.15	0.01	-0.10	0.16	0.31	0.27
3	-0.32	-0.19	-0.05	0.13	0.20	0.09
4	-0.63	-0.27	-0.13	0.06	0.18	0.09
5	-0.96	-0.52	-0.41	-0.06	0.07	-0.03
6	-1.26	-0.41	-0.38	-0.10	0.07	-0.03
			(c) AS Niño 4			
0	0.52	0.67	0.67	0.69	0.73	0.73
1	0.11	0.15	0.26	0.41	0.51	0.46
2	-0.11	-0.15	0.00	0.14	0.34	0.20
3	-0.33	-0.26	-0.15	0.19	0.34	0.24
4	-0.78	-0.41	-0.22	0.14	0.28	0.10
5	-1.04	-0.55	-0.18	-0.03	0.13	-0.01
6	-1.07	-0.59	-0.22	-0.08	0.09	0.02
			(d) AS Niño 1+2			
0	0.24	0.46	0.48	0.50	0.60	0.57
1	-0.14	0.32	0.37	0.51	0.52	0.47
2	-0.33	0.10	-0.03	0.15	0.26	0.30
3	-0.22	-0.25	-0.11	0.14	0.18	0.05
4	-0.30	-0.30	-0.17	-0.01	0.11	0.06
5	-0.50	-0.39	-0.33	-0.05	0.08	0.03
6	-0.77	-0.44	-0.41	-0.22	-0.01	-0.02

TABLE 5. Improvement afforded by the consolidated ENSO-CLIPER model over (a) persistence and (b) the standard ENSO-CLIPER model for predicting August-September Niño 3.4, 3, 4 and 1+2 as a function of monthly lead from replicated real-time forecasts 1900-1950. Values are given as the absolute difference in *MSSS* and *RMSSS* (in brackets).

Niño index	Lead (months) 0	1	2	3	4	5	6
(a) Consolidated CLIPER improvement over persistence							
3.4	3 (3)	5 (4)	6 (3)	25 (13)	65 (29)	99 (42)	121 (47)
3	1 (2)	9 (7)	19 (11)	31 (15)	41 (17)	98 (40)	129 (50)
4	3 (3)	14 (11)	21 (13)	24 (15)	36 (20)	44 (22)	68 (31)
1+2	5 (6)	8 (6)	2 (1)	14 (7)	21 (9)	33 (14)	45 (18)
(b) Consolidated CLIPER improvement over standard CLIPER							
3.4	3 (3)	7 (5)	10 (6)	5 (2)	7 (3)	1 (0)	36 (16)
3	0 (1)	2 (2)	11 (7)	0 (0)	4 (2)	7 (3)	27 (12)
4	0 (0)	12 (9)	15 (10)	8 (5)	15 (9)	14 (7)	15 (7)
1+2	0 (0)	-2 (-1)	5 (3)	-5 (-2)	-2 (0)	5 (2)	9 (4)

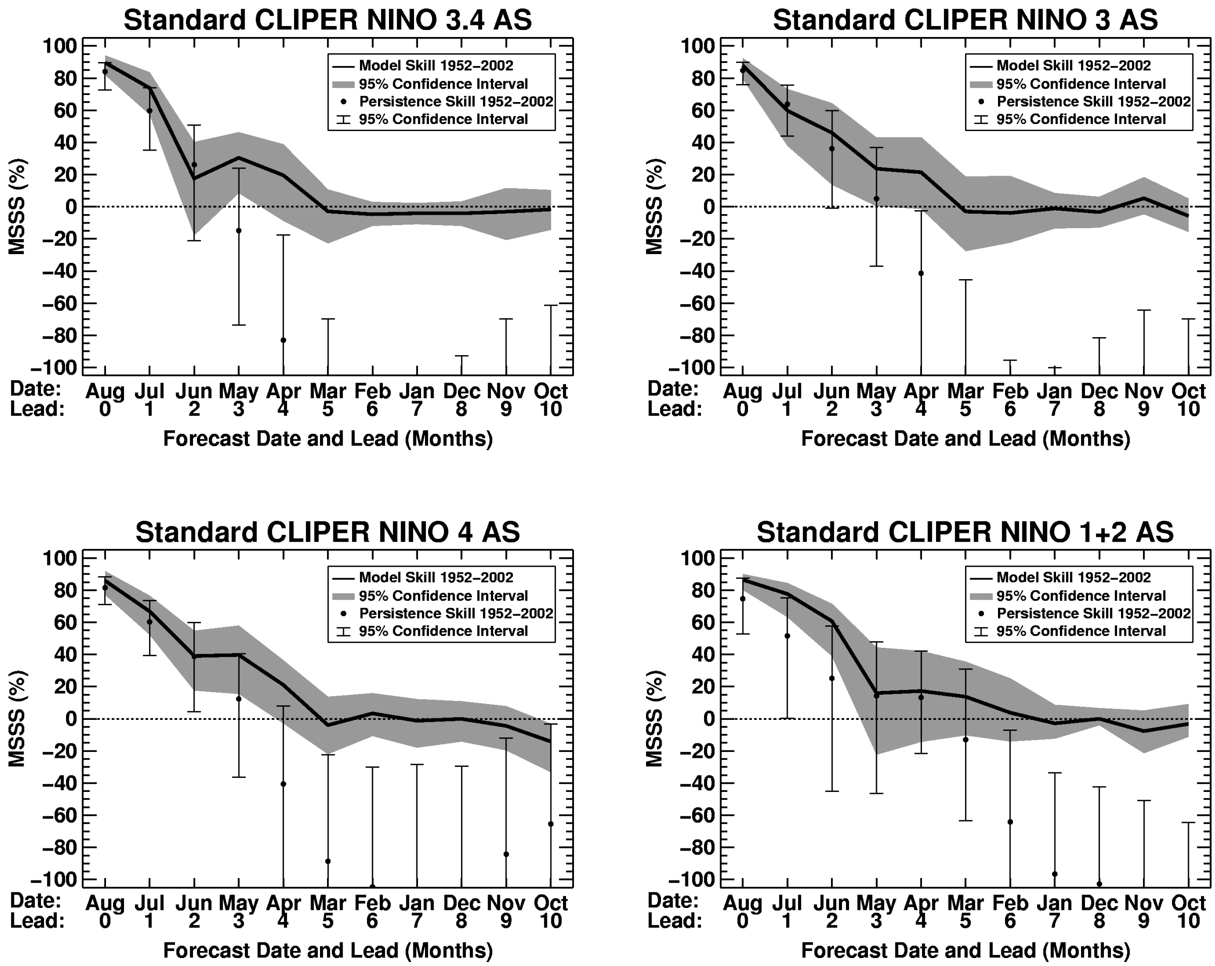

Figure 1. Cross-validated hindcast skill from the standard ENSO-CLIPER model for predicting the August-September (AS) Niño 3.4, 3, 4 and 1+2 indices 1952-2002 at monthly leads out to 10 months. The skill measure used is the mean square skill score (*MSSS*) defined as the percentage improvement in mean square error over a hindcast of zero anomaly; the climatology being 1952-2002. The grey band is a bootstrapped estimate of the 95% confidence interval for the skill measure. The skill and uncertainty from standard persistence are shown respectively by the filled circles and error bars.

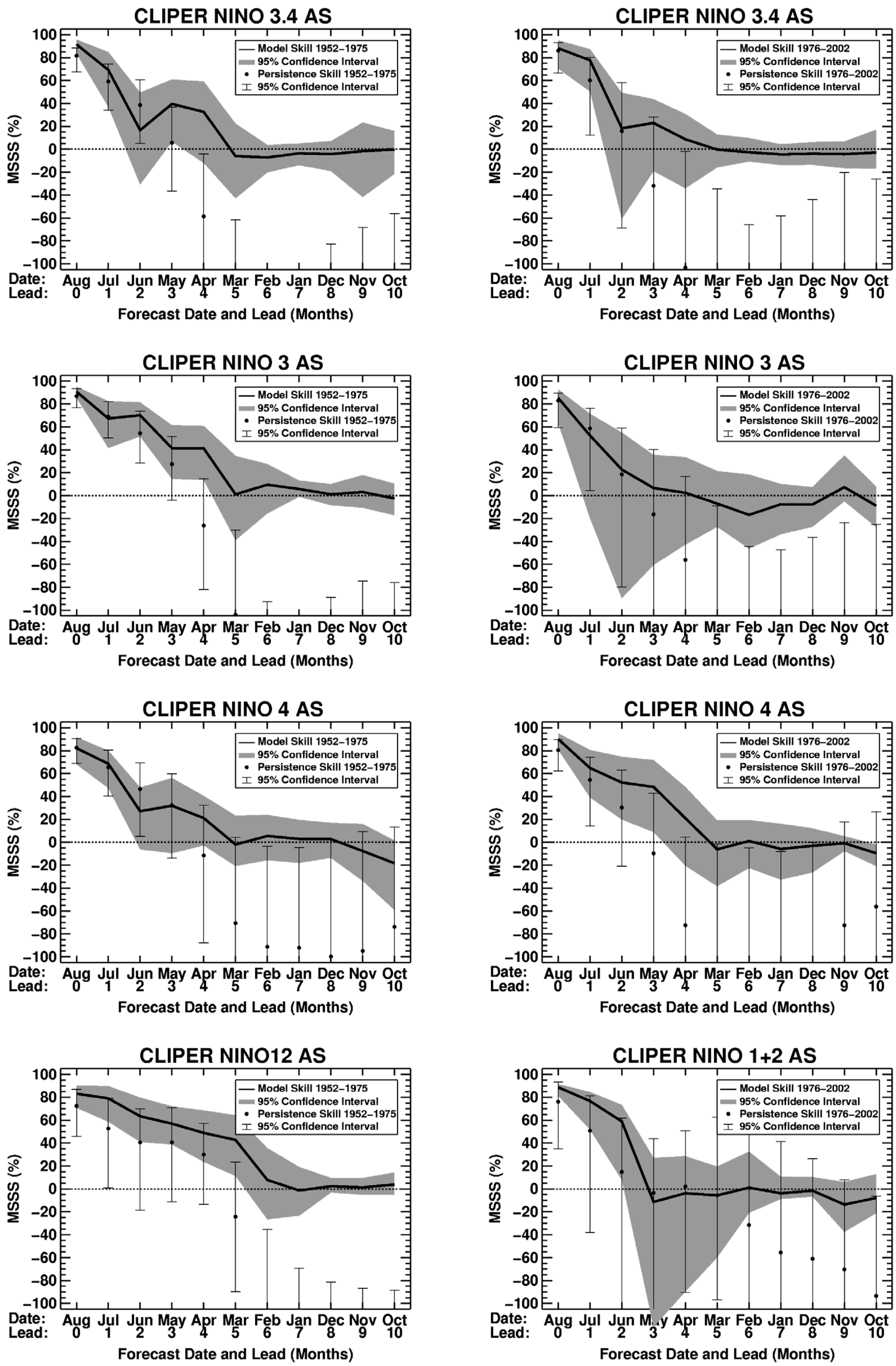

Figure 2. As Figure 1 but for the sub periods 1952-1975 (left column) and 1976-2002 (right column).

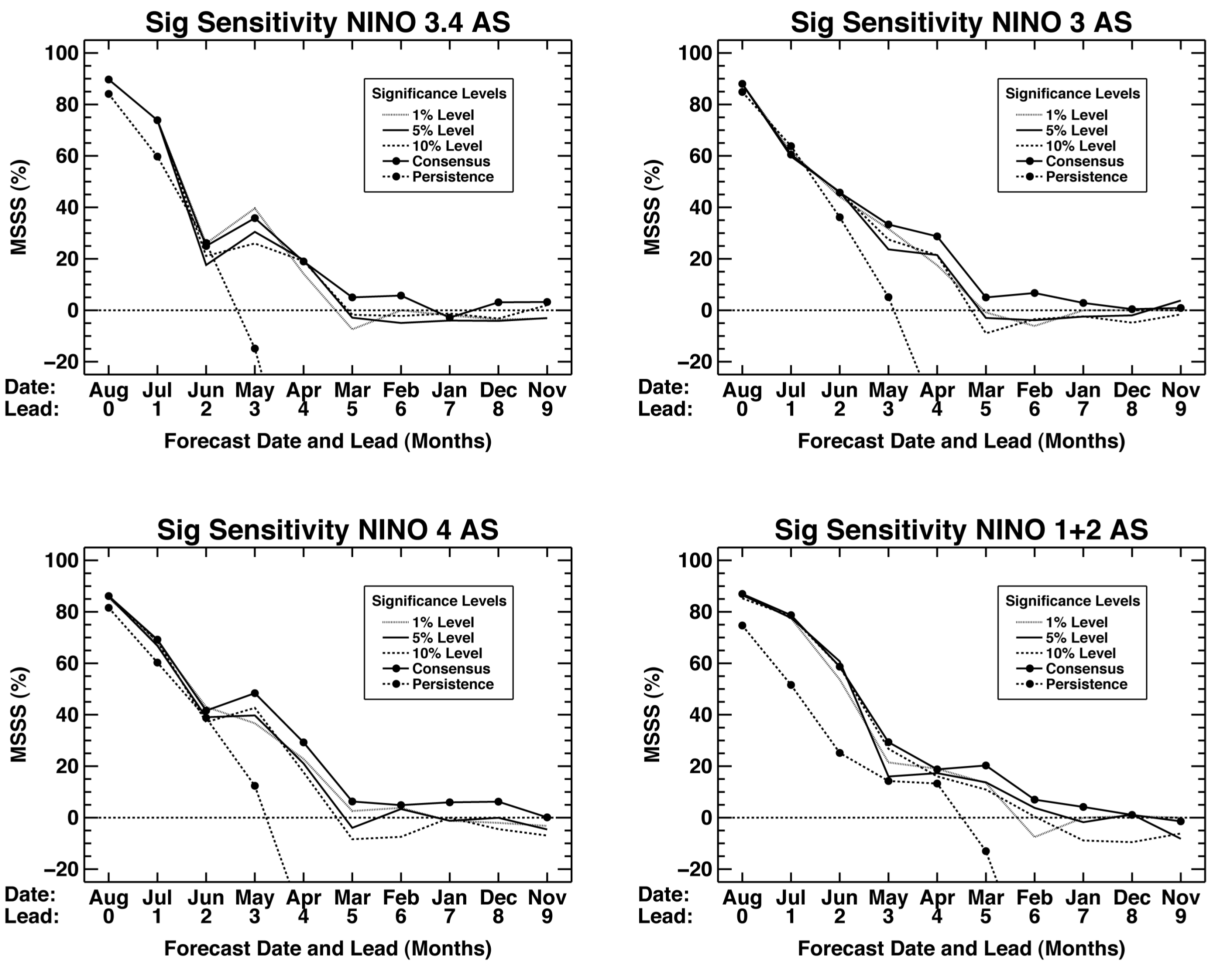

Figure 3. The sensitivity of the standard ENSO-CLIPER model cross-validated hindcast skill to the significance level imposed during predictor screening for the prediction of August-September (AS) Niño 3.4, 3, 4 and 1+2 indices 1952-2002 at monthly leads to 9 months. The 'consensus' skill refers to the average of the three hindcasts obtained using significance levels of 1%, 5% and 10%. The standard persistence skill from Figure 1 is included for reference.

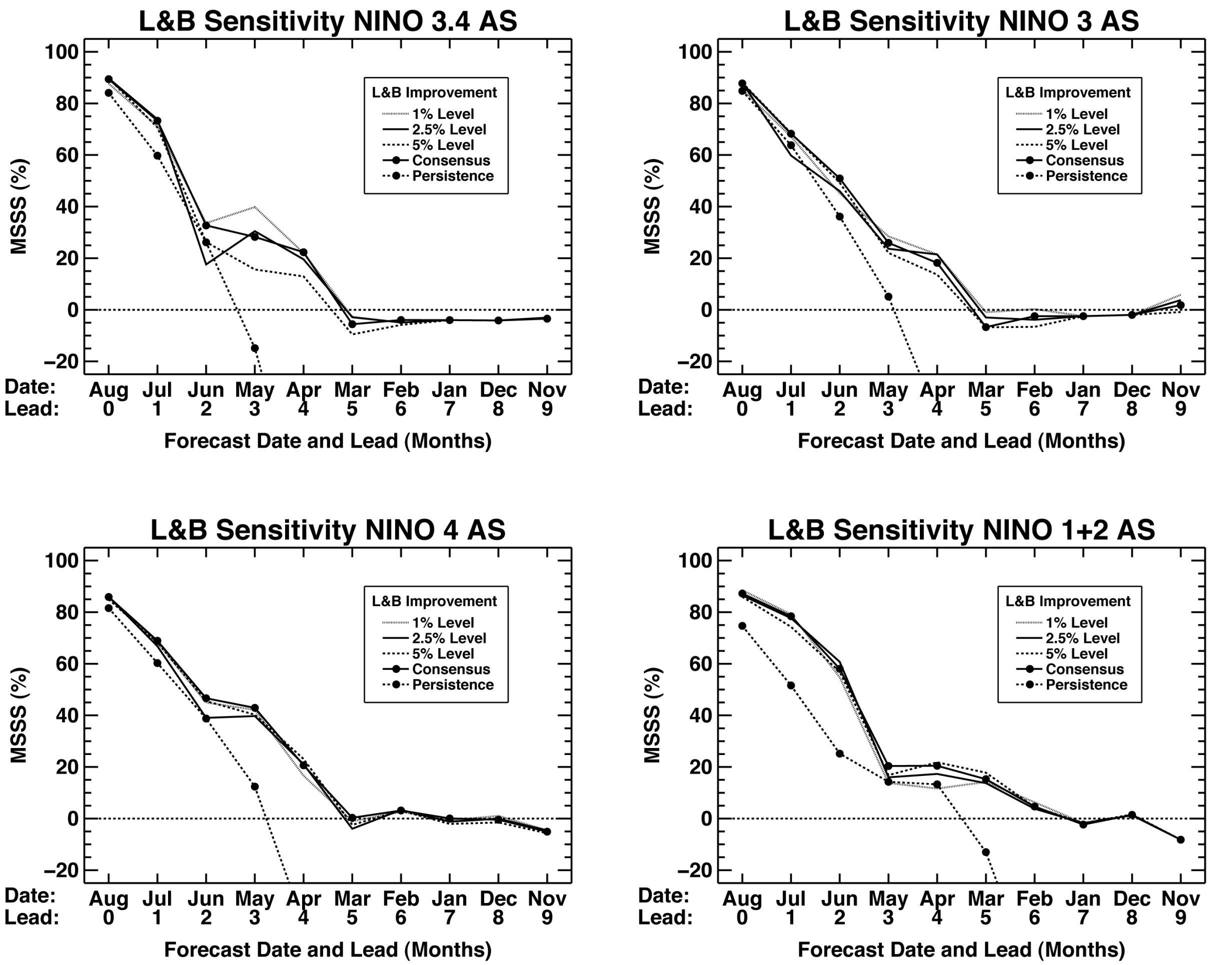

Figure 4. The sensitivity of the standard ENSO-CLIPER model cross-validated hindcast skill to the PVE improvement factor passed to the leaps and bounds algorithm for the prediction of August-September (AS) Niño 3.4, 3, 4 and 1+2 indices 1952-2002 at monthly leads to 9 months. The 'consensus' skill refers to the average of the three hindcasts obtained using leaps and bounds improvement factors of 1%, 2.5% and 5%.

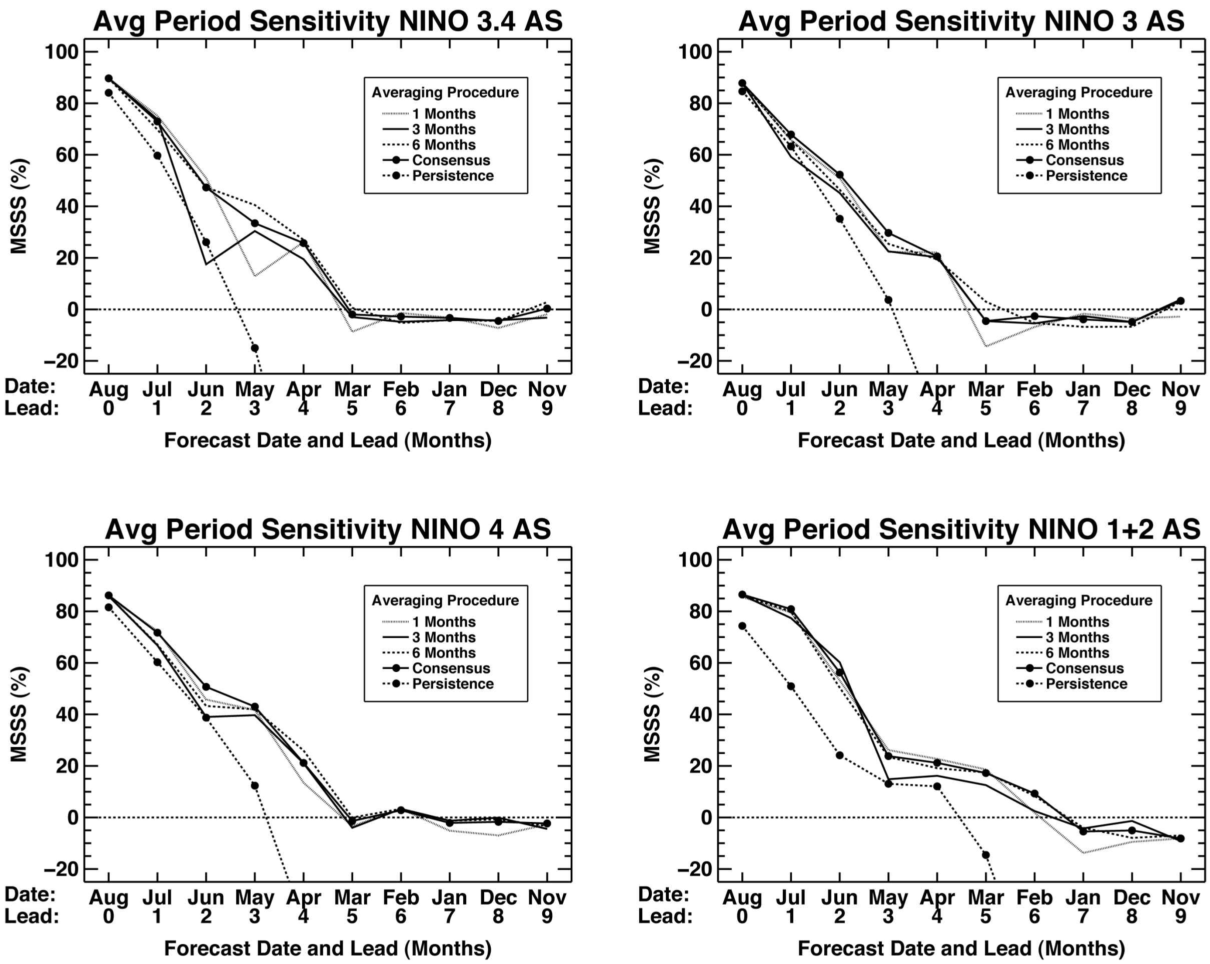

Figure 5. The sensitivity of the standard ENSO-CLIPER model cross-validated hindcast skill to the teleconnected predictor averaging period used in the model formulation for the prediction of August-September (AS) Niño 3.4, 3, 4 and 1+2 indices 1952-2002 at monthly leads to 9 months. The 'consensus' skill refers to the average of the three hindcasts obtained using averaging periods of 1, 3 and 6 months.

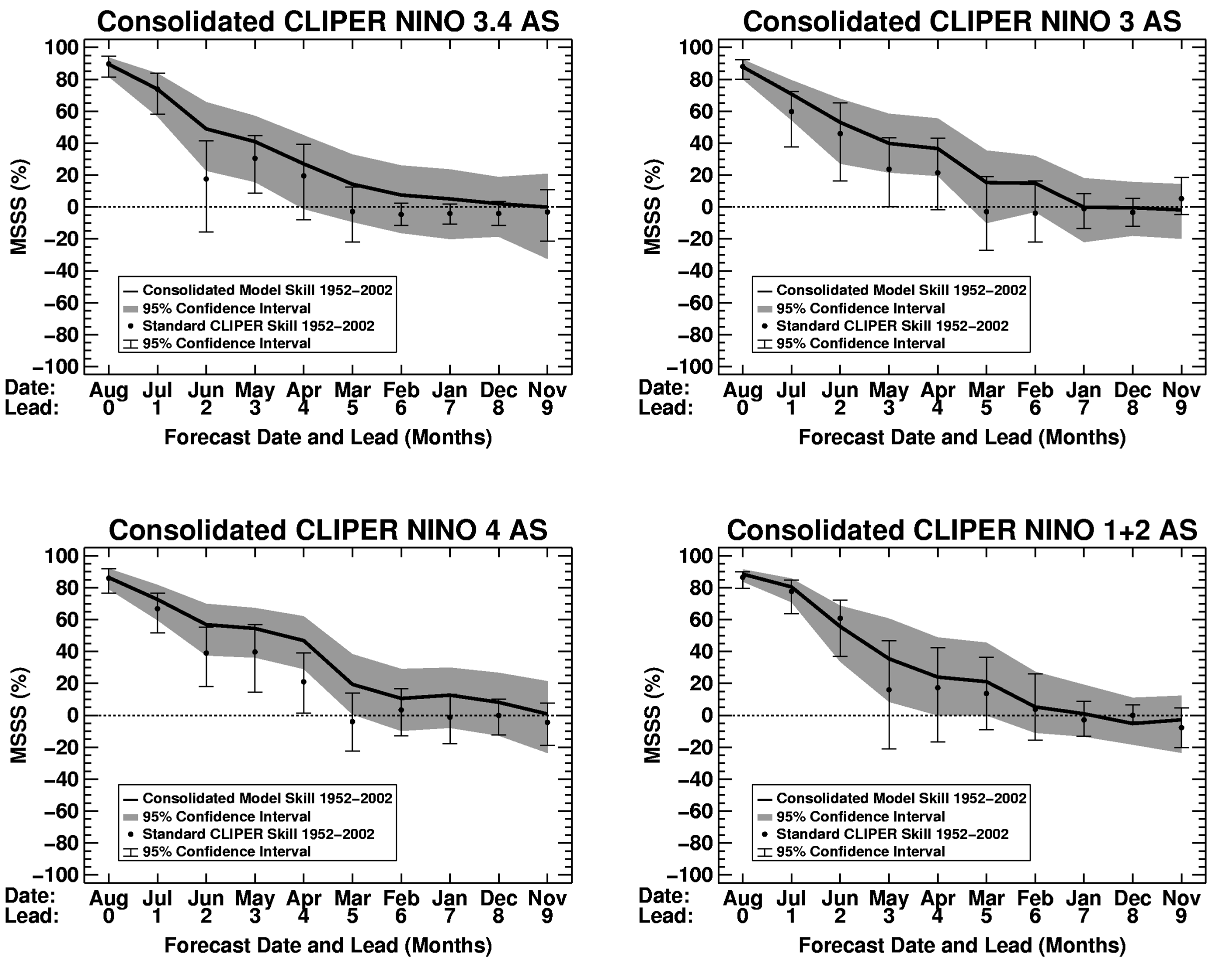

Figure 6. Cross-validated hindcast skill 1952-2002 of the consolidated ENSO-CLIPER model compared against the standard ENSO-CLIPER model for predicting the August-September (AS) Niño 3.4, 3, 4 and 1+2 indices at monthly leads out to 9 months. The skill measure used is the mean square skill score (*MSSS*) defined as the percentage improvement in mean square error over a hindcast of zero anomaly; the climatology being 1952-2002. The grey band is a bootstrapped estimate of the 95% confidence interval for the skill measure. The skill and uncertainty from the standard ENSO-CLIPER model are shown respectively by the filled circles and error bars.

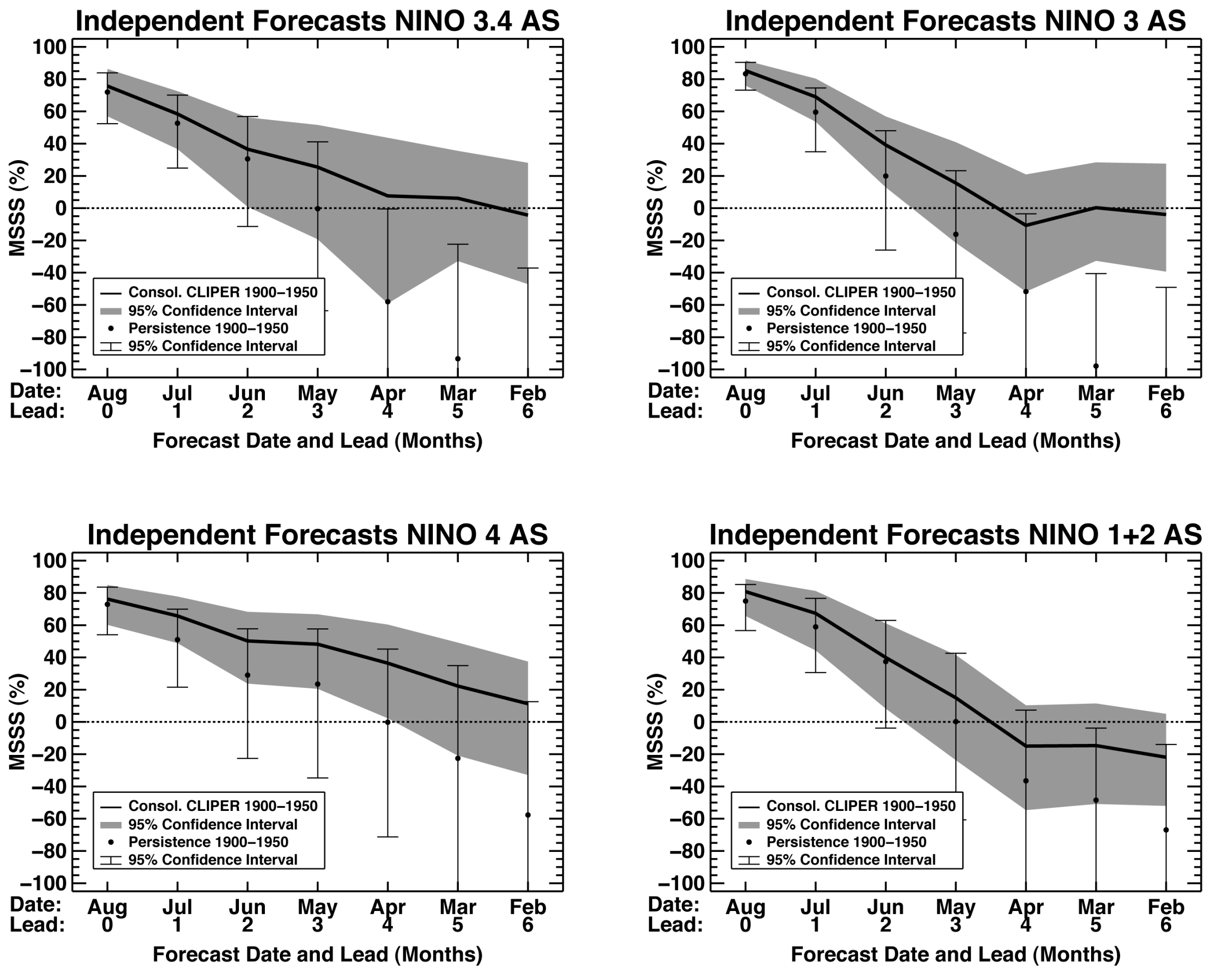

Figure 7. Replicated real-time forecast skill 1900-1950 of the consolidated ENSO-CLIPER model compared against persistence for predicting the August-September (AS) Niño 3.4, 3, 4 and 1+2 indices at monthly leads out to 6 months. The skill measure used is the mean square skill score (*MSSS*) defined as the percentage improvement in mean square error over a forecast of zero anomaly; the climatology being 1900-1950. The grey band is a bootstrapped estimate of the 95% confidence interval for the skill measure. The skill and uncertainty from persistence are shown respectively by the filled circles and error bars.

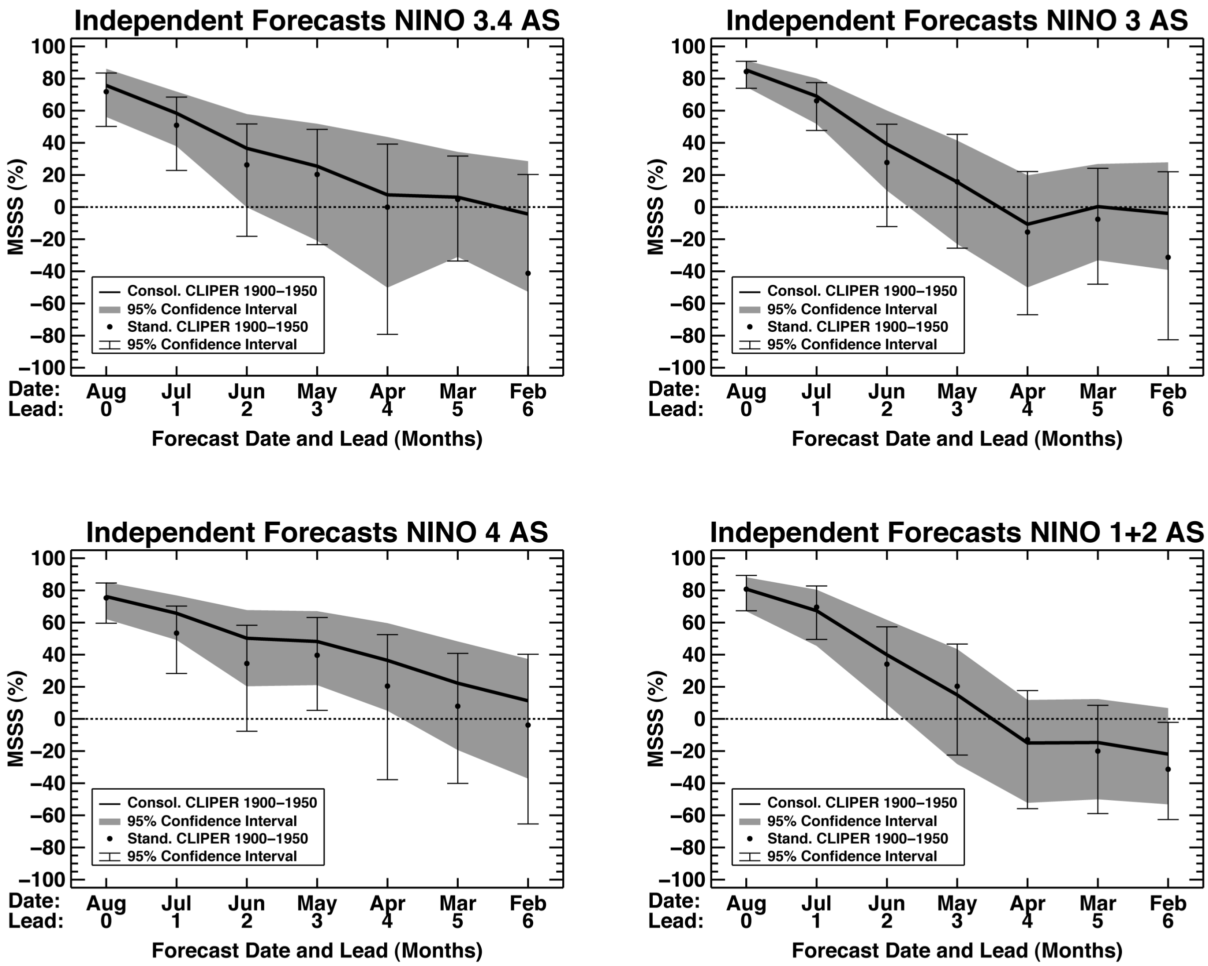

Figure 8. Replicated real-time forecast skill 1900-1950 of the consolidated ENSO-CLIPER model compared against the standard ENSO-CLIPER model for predicting the August-September (AS) Niño 3.4, 3, 4 and 1+2 indices at monthly leads out to 6 months. The skill measure and presentation format are the same as in Figure 7.

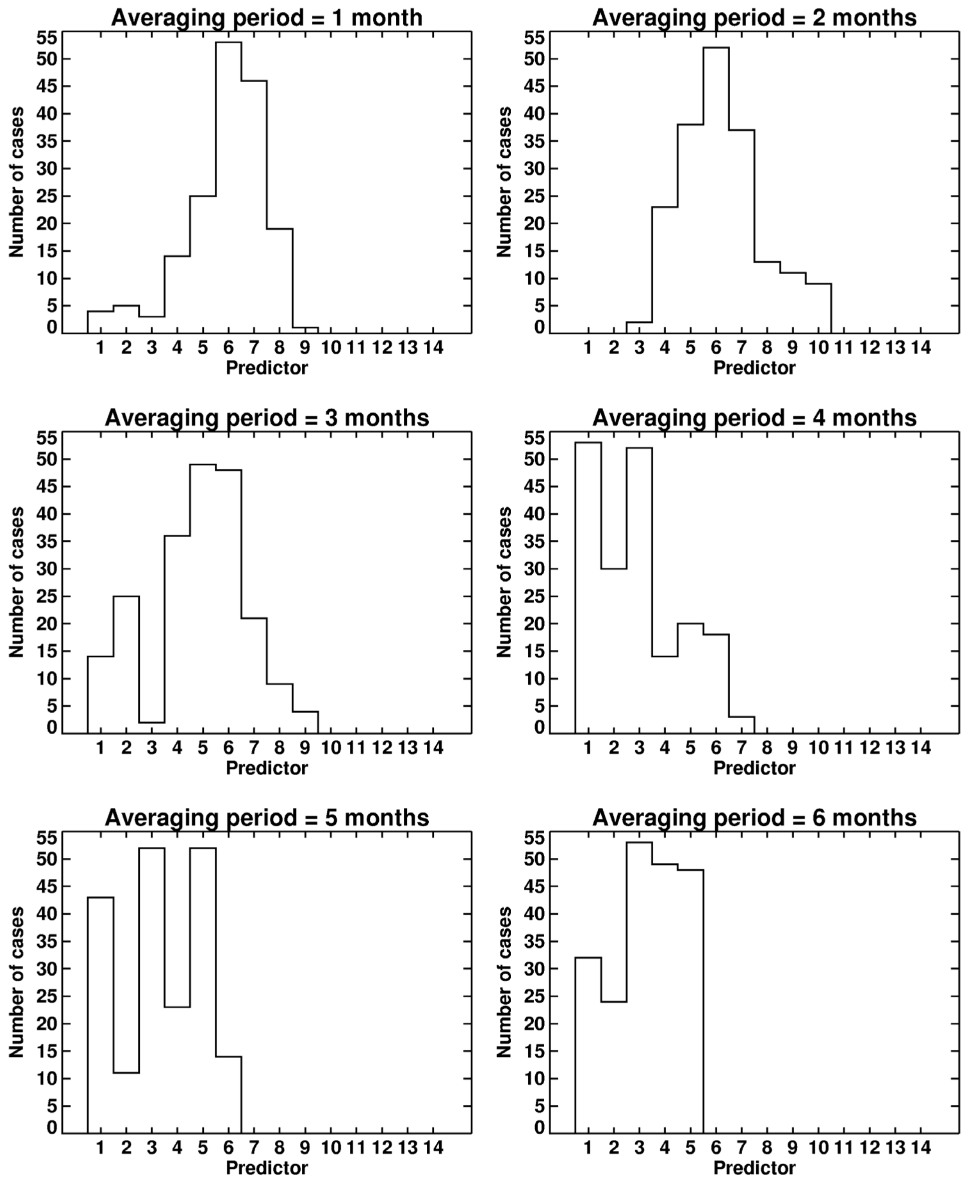

Figure 9. Histograms of the standard ENSO-CLIPER predictors selected for making hindcasts of the August-September Niño 3.4 index 1952-2002 at a lead of 3 months (early May) for models built with teleconnected predictor averaging periods from 1 to 6 months. The predictor numbers (1 to 14) correspond to the classification in Table 1.