Enhancement and Error Estimation of Neural Network Prediction of Niño-3.4 SST Anomalies

YUVAL

Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, British Columbia, Canada

(Manuscript received 12 April 2000, in final form 8 August 2000)

ABSTRACT

A procedure to enhance neural network (NN) predictions of tropical Pacific sea surface temperature anomalies and calculating their estimated errors is presented. A simple linear correction enables more accurate predictions of warm and cold events but can result in introduction of larger errors in other cases. The prediction error estimates aid recognizing erroneously magnified anomalies and are used to sort the predictions into El Niño, La Niña, and neutral states. The error estimation process is based on bootstrap resamplings of the data and construction of a large number of bootstrap prediction replicas. A statistic calculated on the set of bootstrap replicas that corresponds to each of the actual predictions is used to estimate the prediction's errors. The method is demonstrated on NN prediction of the Niño-3.4 index.

1. Introduction

The problem of understanding and predicting the El Niño–Southern Oscillation (ENSO) phenomenon has been in the focus of many research efforts. Barnston et al. (1994) examined the performance of five ENSO prediction systems over a 12-yr period. In a follow-up paper, which focused on the 1997/98 El Niño episode, Barnston et al. (1999) give a summary of 15 statistical and dynamical models that predict the tropical Pacific sea surface temperature (SST) on a regular basis.

Apart from the scientific interest in understanding the causes of the interannual climatic anomalies associated with ENSO, forecasts of the phenomenon are desired due to its important social and economical impact (Enfield 1989; Ropelewski and Halpert 1987). Most important for practical considerations are reliable predictions of the large anomalies associated with El Niño and La Niña conditions. Unfortunately, some prediction schemes (e.g., Yuval 2000; Knaff and Landsea 1997; Penland and Magorian 1993; Keppenne and Ghil 1992) tend to underestimate the amplitude of large anomalies and thus might fail to provide necessary calls for imminent El Niño or La Niña events.

Although the current performance of ENSO predicting models is at best ''moderate'' (Barnston et al. 1994, 1999), relatively small efforts are made to estimate prediction errors. Tangang et al. (1998a,b) calculate only the error bars of their prediction skills. Zebiak et al. (1999), Syu and Neelin (1999), and Saunders et al. (1999) attach to their predictions error bars equal in length to the root-mean-square error (rmse) calculated on hindcast predictions of past data. Obviously this results in underestimation of the errors of difficult to predict future data and overestimation of the errors of easy to predict ones. Penland et al. (1999) provide the more rigorous one-standard-deviation confidence interval of their prediction. Their derivation of the confidence interval relies on some assumptions about the predicting model and the distribution of the errors (Penland 1989) and is specific to their prediction scheme.

This paper presents a simple method to enhance forecasts such that the large events are more accurately predicted. This enhancement can result in magnification of small anomalies and introduction of additional undesired false alarms. Complementing the predictions with estimations of their errors enables evaluation of their uncertainty and forestalls most of the false alarms.

The correction of the predictions is based on a linear relationship between prediction residuals and the observations. The correction enables more accurate prediction of large anomalies while keeping the forecast's correlation skill unchanged and only slightly degrade its overall rmse skill. Another improvement is in the average magnitude of anomalies in the corrected forecast that is more realistic and closer to the observations.

The procedure to estimate the errors is based on bootstrap sampling (Davison and Hinkley 1997; Efron and Tibshirani 1993). The procedure is very general and can

Corresponding author address: Yuval, Dept. of Earth and Ocean Sciences, University of British Columbia, 1461-6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada.
E-mail: yuval@ocgy.ubc.ca

be considered for estimating prediction errors of any model whose development process can be automated and is fast enough to be run a large number of times within economically reasonable time limits. It is mainly suitable for statistical models developed empirically on a given dataset.

Bootstrap estimation of errors is based on the notion of random bootstrap samplings of the original dataset. Bootstrap samples are similar to the original dataset but with possibly some of the data missing and others appearing more than once. Reproductions of the original model can be developed on the bootstrap samples. Prediction capability of each of these model reproductions depends on the similarity between the content of the dataset on which it was developed and the testing data, or future data it will try to predict in operational mode. The bootstrap model reproductions produce a set of bootstrap prediction replicas accompanying each of the actual predictions. A statistic calculated on this bootstrap predictions set is used as the error estimate of the prediction.

In this paper the neural network (NN) training procedure of Yuval (2000) is used to develop the NN predicting model and its bootstrap reproductions. Corrected predictions and corresponding prediction error estimates for lead times of 3, 6, 9, and 12 months are shown and discussed. Integrated in the error estimation process is a scheme to determine whether El Niño or La Niña conditions are to be expected using the definition of El Niño by Trenberth (1997).

The paper begins with a section describing the data and the NN prediction model, and formulating the prediction problem. It is followed by a discussion of the preliminary linear correction that is applied to the predictions. The next section proposes a bootstrap procedure for error estimation appropriate for climatological data and is followed by a section showing the results. The next section presents a scheme that uses the error estimates for sorting the predictions into one of three possible categories: El Niño, La Niña, or neutral state. A discussion summarizes and concludes the paper.

2. Formulation of the prediction problem

a. Data

The data used in this paper came from two sources: the Comprehensive Ocean–Atmosphere Data Set sea level pressure (SLP) data of the equatorial Pacific (Woodruff et al. 1987), and the National Oceanic and Atmospheric Administration SST data of the tropical Pacific (Smith et al. 1996; Reynolds and Smith 1994). Both datasets contained monthly data on a $2^\circ \times 2^\circ$ grid from January 1950 to December 1997. Several processing steps of regriding, removal of monthly means, and smoothing were applied to the original data. An extended empirical orthogonal functions (EEOF) analysis (Tangang et al. 1998b; Weare and Nasstrom 1982) was carried out to compress the data in both space and time. The time series of the first 12 EEOF modes were retained as predictors. The series of average SST anomalies in the Niño-3.4 region is the predictand. The choice of keeping 12 EEOF modes was based on sensitivity analysis to obtain the optimal number. It was found that skills of prediction of the Niño-3.4 index were not sensitive to number of EEOF predictors between 10 and 20. More detailed description of the datasets and the initial processing are given by Tangang et al. (1998a) and Tang et al. (2000).

b. The neural network model

In the last few years NN models were applied in various prediction problems (Tangang et al. 1998a; Goswami and Srividya 1996; Hastenrath et al. 1995; Navone and Ceccatto 1994). Detailed descriptions of NNs and NN model development can be found in the references therein or in textbooks like Bishop (1995) or Cichocki and Unbehauen (1993).

An NN model is a mathematical operator that receives the problem's predictors as an input, manipulates them in a nonlinear fashion, and outputs the problem's predictand. The structure, or architecture, of the operator is specified in advance and the model building process, called the NN training, is a search for the various optimal model parameters. The NN training is an optimization problem in which a set of parameters is found such that a desired value of a cost function is achieved. The cost function is calculated on part of the dataset called the training set. In most cases (e.g., Navone and Ceccatto 1994) another part of the data, called a validation set, is set aside to control the tuning and verify the results of the training. To eliminate the possibility of artificial skill, the final prediction capability must be assessed on a totally independent dataset that is not used in any way during the training.

For the purpose of this paper the methodology of Yuval (2000) is used to train the NN models. This methodology enables an automatic search for the optimal NN parameters with no need to set aside a validation set. The tuning of the training is done automatically by simultaneously minimizing a predefined cost function and the generalized cross validation function (Haber and Oldenburg 2000; Wahba 1990; Golub et al. 1979). Niño-3.4 index prediction skills using this methodology were found in Yuval (2000) to be almost identical to those of Tang et al. (2000).

c. The prediction problem

The predictors at a particular point in time can be considered for forecasting the predictand a few months ahead. The time shift between predictors and predictands is the prediction lead time, defined here as the time from the center of the period of the latest predictors to the center of the predicted period. This definition

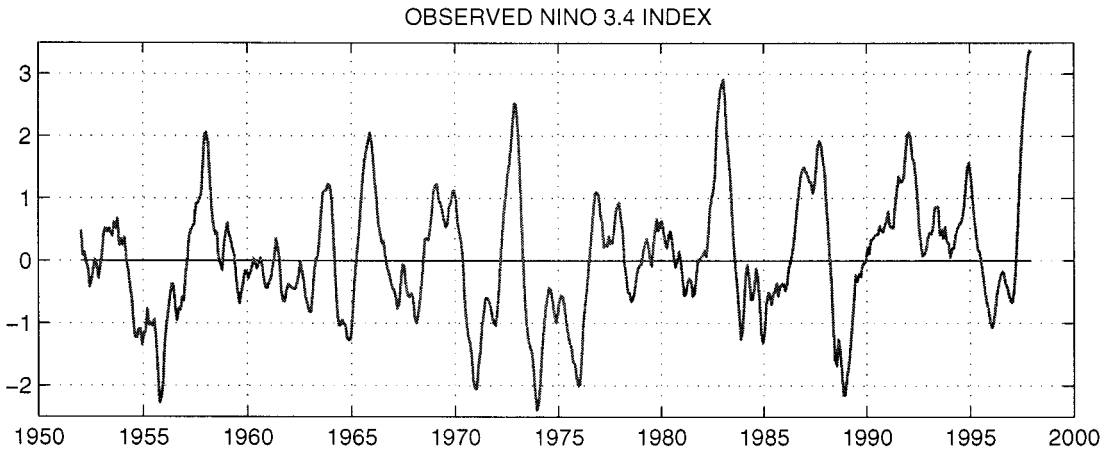

FIG. 1. The Niño-3.4 index time series.

follows Tang et al. (2000) and differs from the definition of Barnston et al. (1994) that defines the lead time as the time from the end of the period of the latest predictors to the center of the predicted period. A set of shifted predictor–predictand pairs can be used to develop a model capable of forecasting future values of the predictands. In order to assess the prediction capability of a model and the usefulness of its estimated errors, both predictions and errors must be calculated on an independent testing set that does not consist any of the data pairs used for training. This prevents unduly adjustment of the predicting models to the training data and artificially inflated prediction skills. As the total length of available SLP and SST data is only about 50 years, a good way to obtain a relatively long testing data record to evaluate the results of this paper is by the following procedure.

For each of the lead times considered (3, 6, 9, and 12 months), a set of predictors–predictand pairs is prepared. This set is partitioned into a few nonoverlapping subsets of data pairs that are used in turn for testing. To each testing subset corresponds a training subset consisting the data pairs that are not included in the testing set. The models producing a testing predictand and its estimated prediction errors are developed using the data pairs in the corresponding training set. The testing segments of predictands and error estimates are then pasted together so the performance can be evaluated on the whole data record.

The choice of number of parts to divide the data is a compromise between the ideal leave-one-out partition scheme (Allen 1974), where each data point is singled out in turn for testing resulting in large training sets, and computation time constraints. It was found experimentally that the final results of this work are not sensitive to division into number of parts between 3 and 10 and partition into five segments was used. A leave-one-out scheme would probably have resulted with slightly improved predictions due to more data available in each training run, but it was not attempted here due to the large number of runs required for the bootstrap error estimation.

The above cross-testing partitioning procedure enables assessing the results for the whole dataset without using any datum for its own prediction or its prediction error estimation. The procedure involves both the hindcast and retroactive predictions discussed in Barnston et al. (1994) and is following the ideas of cross validation discussed in Michaelsen (1987) and Barnston and van den Dool (1993). We emphasize the independence of the testing data because reported results of NN predictions are especially prone to be biased by use of predicted data in their own prediction and only the independence of the testing data ensures that the predictions and error estimates of our final results can be viewed as if they were produced by an operational prediction mode. One exception that was allowed is that the initial processing step of EEOF analysis was not carried out separately for each training set. This exception was allowed also by Tang (2000) and might introduce some amount of artificial skill in the predictions but should not alter the general conclusions regarding the enhancement and error estimation processes.

3. Linear correction of predictions

Figure 1 shows the observed Niño-3.4 time series. The upper panels of Figs. 2–5 show its predictions at 3-, 6-, 9-, and 12-month lead times. Magnitudes of the predictions, especially around extrema in the data series, seem to be smaller on average compared to the observed data. Table 1 gives the average of absolute values of the Niño-3.4 index and that of its predictions at 3-, 6-, 9-, and 12-month lead times. The average magnitude of the observed data is larger than the average magnitude of the predictions and the difference increases with lead time.

The prediction residuals, obtained by subtracting training values predicted by the model from the corresponding observations, are plotted against the ob-

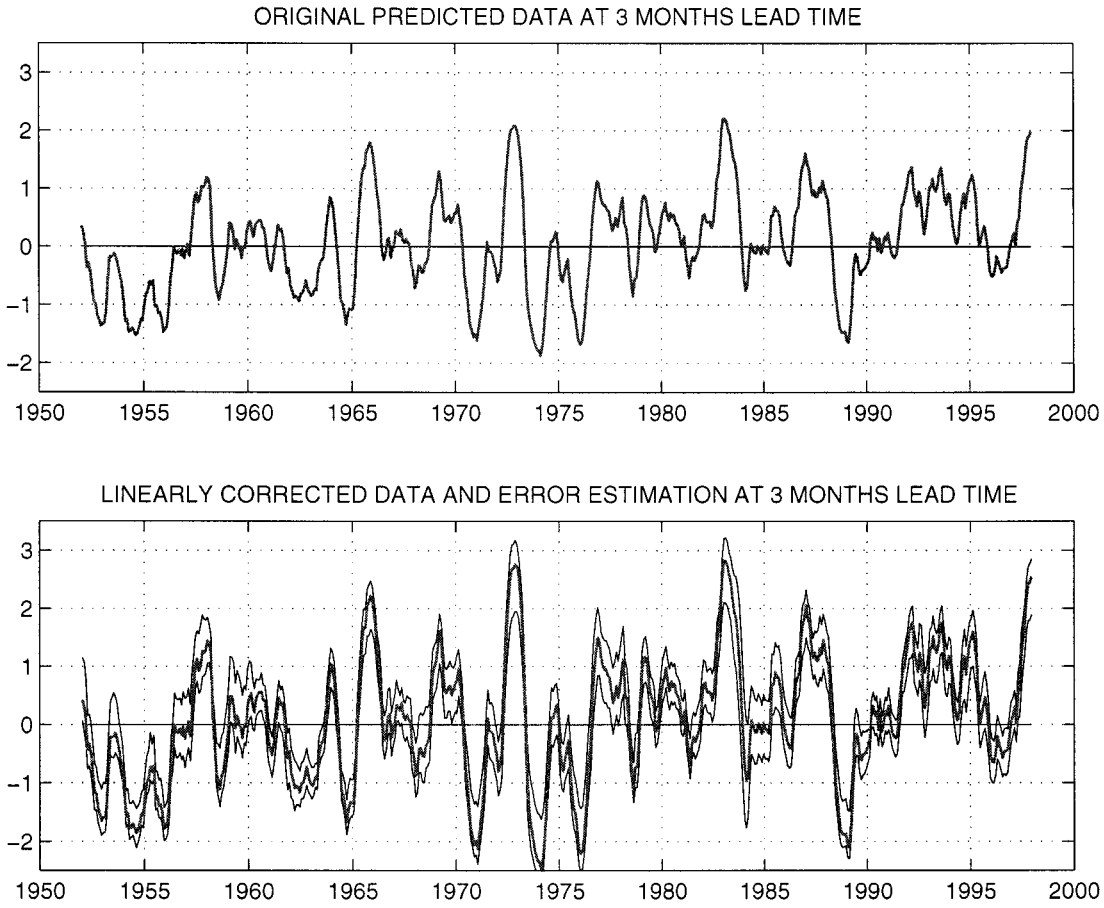

FIG. 2. (upper panel) Original uncorrected predictions at 3-month lead time. (lower panel) The corresponding linearly corrected prediction (thick line) and its error estimates (thin lines).

served values in Fig. 6. The linear least squares fit to the data points in the plots was calculated and is also shown. The reasonable linear fit, noted especially in the plots corresponding to the 6-, 9-, and 12-month lead predictions, implies that the values of the training residuals R_{tr} can be approximated by the linear relation

$$R_{tr} = \alpha' O, \tag{1}$$

where O is an observed value and α' is the proportionality factor calculated using the predictions and observed data of the training set. Similar proportionality was found also between the observations and the cross-tested predictions and thus it can be used to calculate prediction residuals of testing or future data. The problem of course is that at the time of issuing a prediction its corresponding observation is yet unknown and cannot be used to calculate the residual. However, bearing in mind the reasonable correlation between the cross-tested predictions in the upper panels of Figs. 2–5 and the observations in Fig. 1, we can assume that future predictions will be correlated to a certain degree with their corresponding observations. The residuals of testing or operational mode predictions can thus be approximated by

$$R_{te} = \alpha' \tilde{P}, \tag{2}$$

where R_{te} is a residual and $\tilde{P}$ an original prediction. A linearly corrected prediction P can be obtained through the relation

$$P = \tilde{P} + \alpha' \tilde{P} = \alpha \tilde{P}, \tag{3}$$

where $\alpha = 1 + \alpha'$. In order to obtain a testing result that will reflect as closely as possible the expected future performance, α' is calculated for each training set and is used for the linear correction that is carried out on the corresponding testing set.

Clearly the correction in Eq. (3) is not ideal. It is using the predictions instead of the observed data to approximate the residuals. That results in inaccuracies and insufficient correction due to the general small amplitude of the predictions. Also, residuals in the training data are generally slightly smaller than the residuals of independent predictions due to some fit to correlated noise in the training (Yuval 2000). The resultant α is

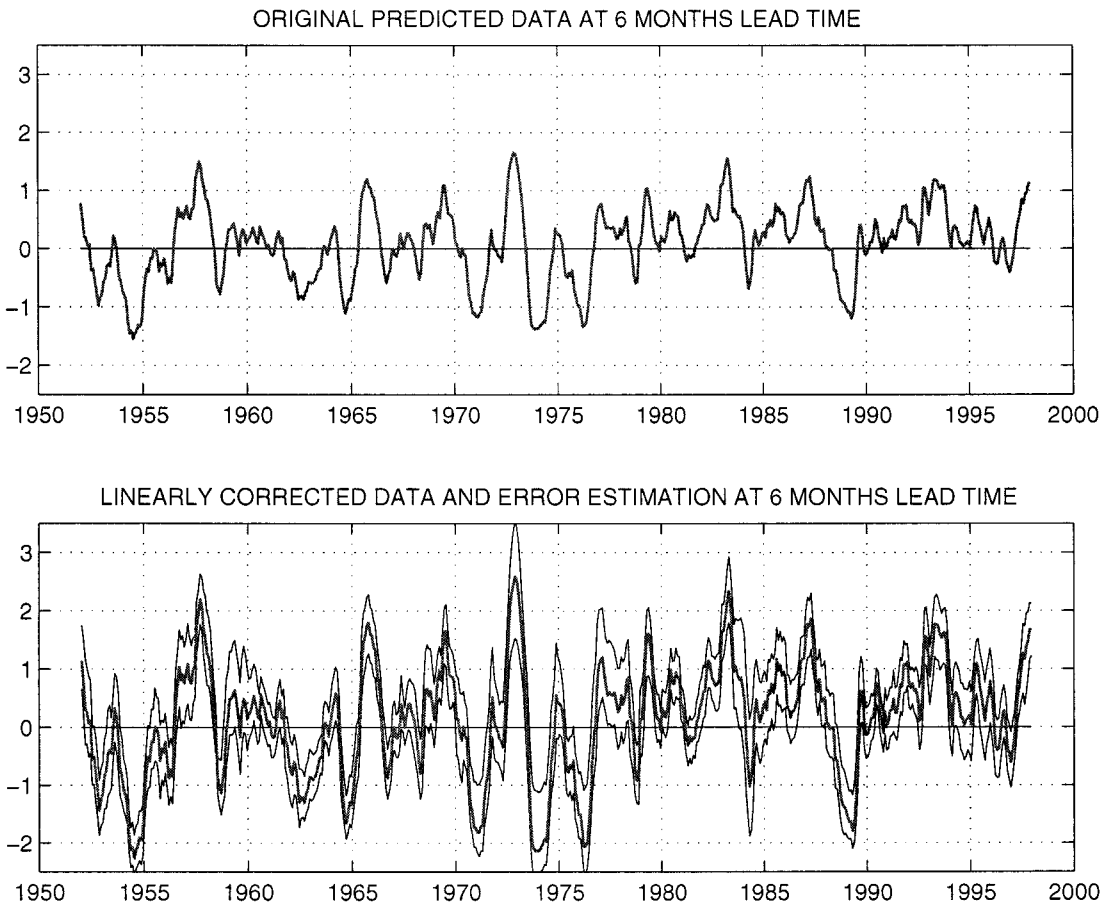

FIG. 3. The same as Fig. 2 but for predictions at 6-month lead time.

too small and this contributes to the insufficiency of the correction. Another point about the correction is that it is mainly appropriate for large anomalies where the linear relationship is more obvious and corrections should be large and more accurate. Corrections become less significant and, for portion of the data, even deleterious as the magnitude of the anomalies decrease.

A more common approach to adjust the dynamical range of testing predictions is to divide their values by the ratio between the standard deviations of predicted and observed training data (Ward and Folland 1991). As noted by Ward and Folland (1991) this method adjusts predictions of different magnitudes by the same factor and thus does not improve the correlation skill while it unduly inflates the rmse error. It is more suited for cases were the quantity of interest is the linear error in the cumulative probability density distribution. The linear correction suggested in this section is tailored to enhance prediction of large anomalies. As will be shown later it clearly improves predictions of large events while keeping the correlation unchanged and only slightly increasing the rmse error. In the following, references will be made only to the linearly corrected predictions and the words ''linearly corrected'' will be omitted for brevity.

4. Bootstrap sampling and error estimation

There are numerous possible sources of prediction errors, for example, inadequate prediction power of the predictors, preliminary processing steps that corrupt the true predictor–predictand relationship, inaccurate predicting model development process, etc. It is impossible to try to estimate those errors directly. A way to estimate the range of possible outcomes of a predicting process is to develop models on many nonoverlapping training datasets and analyze their different predictions of the same test data. Unfortunately, it is not a practical solution for the tropical Pacific datasets (or other climatological data) that are too short and consist usually of no more than few hundred monthly means of the relevant physical properties. In order to evaluate inaccuracy of a prediction process based on models developed on these records we need many realizations of the climate of our planet. It might be that at a certain point in the future GCMs will be able to produce additional

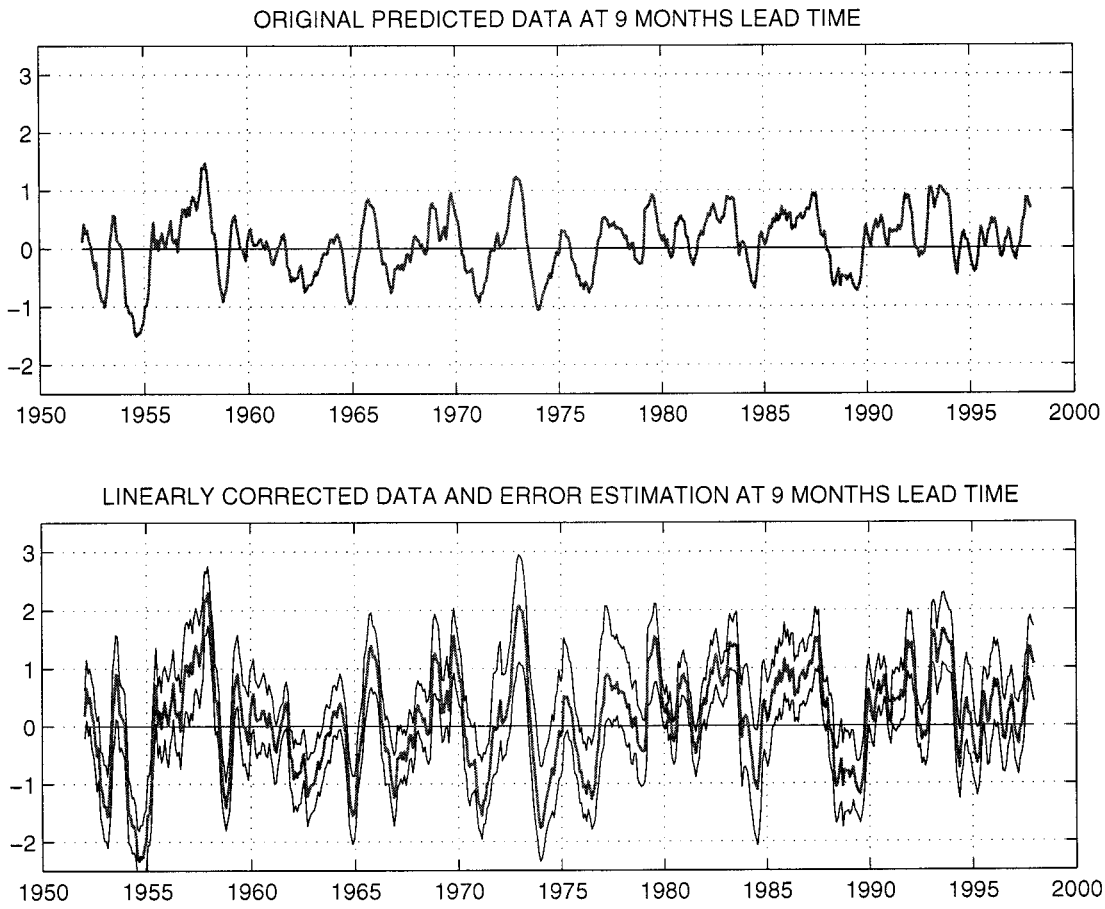

FIG. 4. The same as Fig. 2 but for predictions at 9-month lead time.

reliable realizations of the climate system. Currently this is not feasible and we have to our disposal only the one realization of past real data. A way to simulate a large set of realizations from the available data is described in this section.

The method is based on the idea of creating bootstrap resampling of the data as expounded by Davison and Hinkley (1997) and Efron and Tibshirani (1993). Each of the training sets used to develop the original NN models is resampled B number of times in a way described in the next section. An NN model is developed on each of the new training sets and is used to predict replicas $\tilde{P}_i^*$, $i = 1, 2, \ldots, B$, of each of the the corresponding testing data (or of future data in an operational mode). A linear correction as described in the last section is applied to these prediction replicas and results in a set P_i^*, $i = 1, 2, \ldots, B$, of corrected prediction replicas corresponding to each corrected prediction P. The error of this prediction is estimated by a statistic calculated using the values of the prediction and its replicas P_i^*.

Unfortunately, much of the theoretical and practical work in the area of bootstrap sampling and errors estimation is based on the assumptions of uncorrelated data and certain distributions of the noise. Climatological data have strong serial correlation and they contain correlated noise. The noise is probably correlated also with the data and its distribution is unknown. The section below proposes an alternative bootstrap sampling that might be more suitable for climatological data.

a. Bootstrap sampling and development of bootstrap prediction replicas

Every set of training data pairs is divided into equal-in-length blocks each containing data corresponding to a certain period of time. Many additional training samples are constructed using the moving blocks bootstrap sampling (Efron and Tibshirani 1993). This is a random selection of blocks, with replacement, from the original training set and a matching of the blocks into new sets similar in length to the original one. To ensure that results do not depend on any specific division into blocks, the division is randomly altered for each set by changing the division starting point. Model reproductions are trained on the bootstrap data samples following exactly the same training process as for the original model.

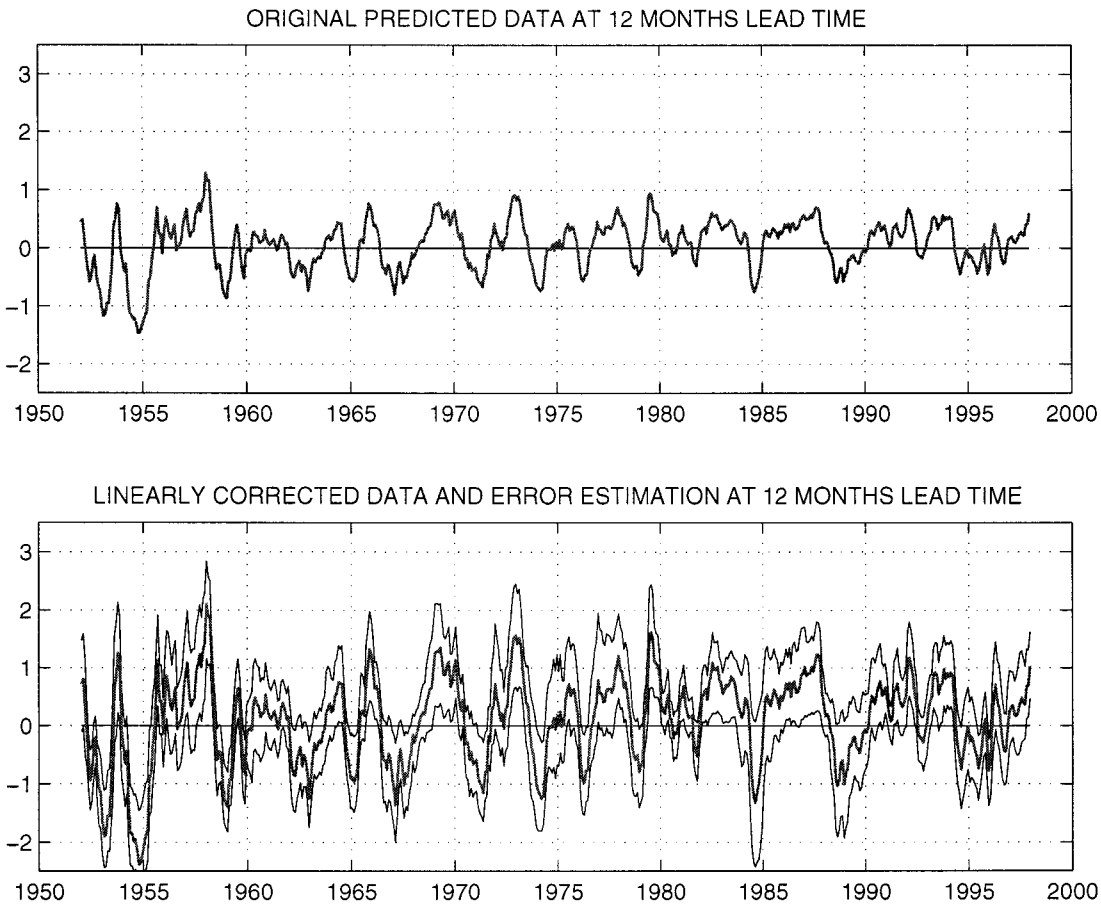

FIG. 5. The same as Fig. 2 but for predictions at 12-month lead time.

An important question regards the length of the data blocks. It seems most reasonable for it to be about the length of the longest autocorrelation found in the predictor and predictand series. This results in nearly independent samples but retains the correlation in the data that we would like to conserve. Experimentation with different block lengths is necessary in order to find a satisfactory one. It was found that for the prediction problem dealt with in this paper the outcomes of using data block lengths in the range of 18–36 months were similar and a block length of 24 months, considered a characteristic timescale of ENSO (Barnston et al. 1994), was chosen to be used for all the results shown in the next sections.

There is no way to know in advance how large the number of bootstrap samples should be. Efron and Tibshirani (1993) suggests that 50 samples are usually enough, and rarely more than 200 are needed, for standard error estimation. An assessment of a sufficient number of bootstrap samples can be made by experimentation with increasingly larger number until the statistical properties of the results are stabilized. Results shown in this paper were obtained using 1000 bootstrap samples that are very similar to the results obtained using smaller sets of 200 bootstrap samples.

The motivation to use the bootstrap method is that a rare feature in the predictor–predictand relationship will be missing from many bootstrap sample sets. The model reproductions developed from sample sets that do not include a rare feature will not be good at predicting similar features when they appear in the future. A large prediction variance is expected in this case and thus a large error should be assigned to the prediction. A common feature is likely to be sampled often and to appear in most of the bootstrap samples. Its prediction variance will be low and it should be assigned a small error. For example, in the case of ENSO prediction, a typical El Niño should be well predicted by most of the model reproductions and the prediction variability and errors are expected to be small. On the other hand, most of the model reproductions will not be good at predicting an event that has little or no precedent. Prediction variability is expected to be high in such cases resulting in large estimated errors.

In this paper the bootstrap method is used to estimate errors. It must be born in mind that the error estimation

TABLE 1. Comparison between the averaged magnitudes of the Niño-3.4 index, the uncorrected predictions, and the linearly corrected predictions. Values are in degrees °C. Also given are the correlation and normalized rmse skills of the original and the linearly corrected predictions.

	Observed data	Uncorrected predictions			Corrected predictions		
Lead times	Avg mag	Avg mag	Correlation	rmse	Avg mag	Correlation	rmse
3-month	0.77	0.67	0.83	0.73	0.85	0.83	0.77
6-month	0.77	0.51	0.68	0.94	0.77	0.68	1.01
9-month	0.77	0.43	0.59	1.04	0.70	0.59	1.09
12-month	0.77	0.36	0.52	1.08	0.61	0.52	1.14

process is not accurate and has its own errors. Main reasons for inaccurate bootstrap estimates are insufficient number of bootstrap sampling and severely small number of data in the original set. The number of sampling that was carried out here should be sufficient for the purpose of error estimation but the short length of climatological data is a main impediment in the process of estimating their prediction errors. Considering the many possible outcomes of the complicated climate system it is clear that our dataset forms only a very small, and may be not representative, fraction of a very long time series. The error estimation process of its predictions should be expected to have only limited accuracy. A qualitative assessment of the accuracy of the error estimation will be discussed with the results. More quantitative methods to estimate the inaccuracy of the bootstrap estimations exist (Efron 1992; Beran 1997) but were not attempted in this paper.

b. *Error estimation*

The set of differences between an actual prediction and its bootstrap replicas ($P_i^* - P$), $i = 1, 2, \ldots, B$, is called here a bootstrap deviation set. An estimate of

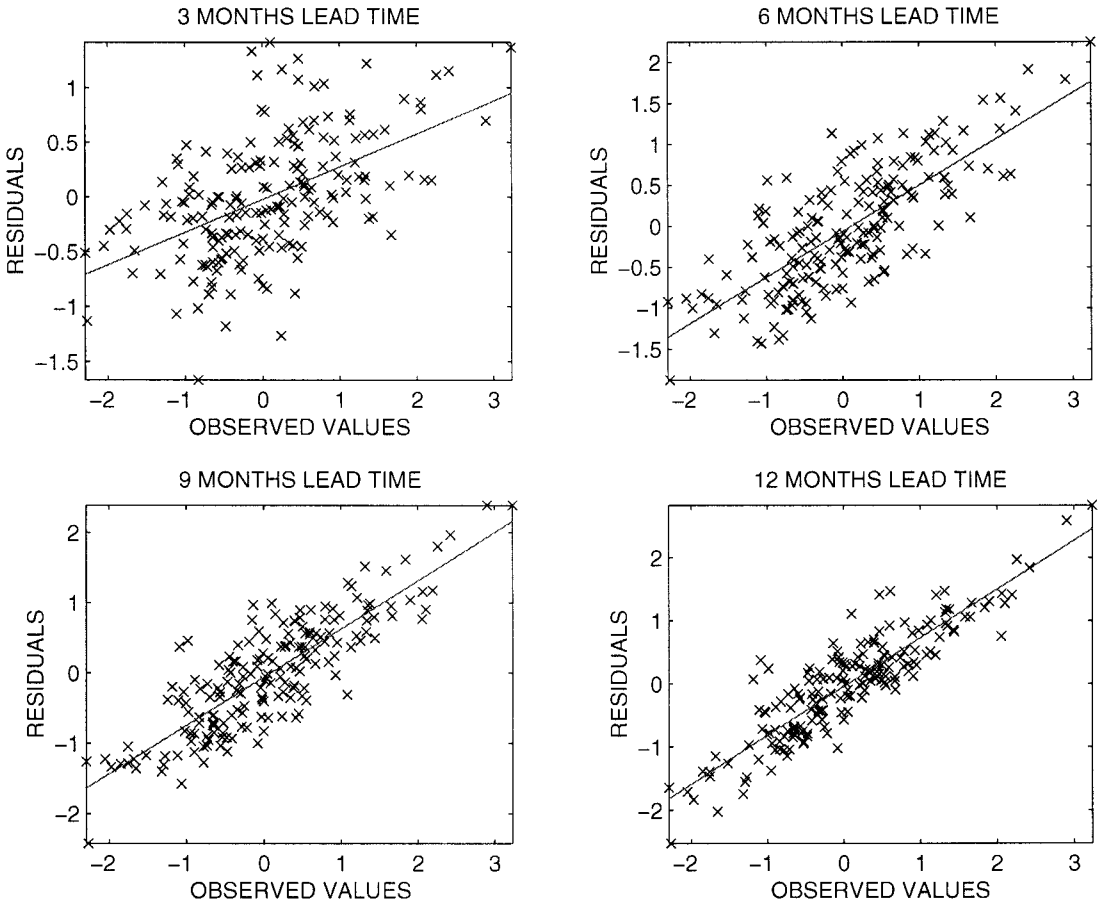

FIG. 6. Residuals, obtained by subtracting the values of uncorrected training predictions from the observed Niño-3.4 data, plotted against the observed values. Only every third data point was plotted for visual purposes.

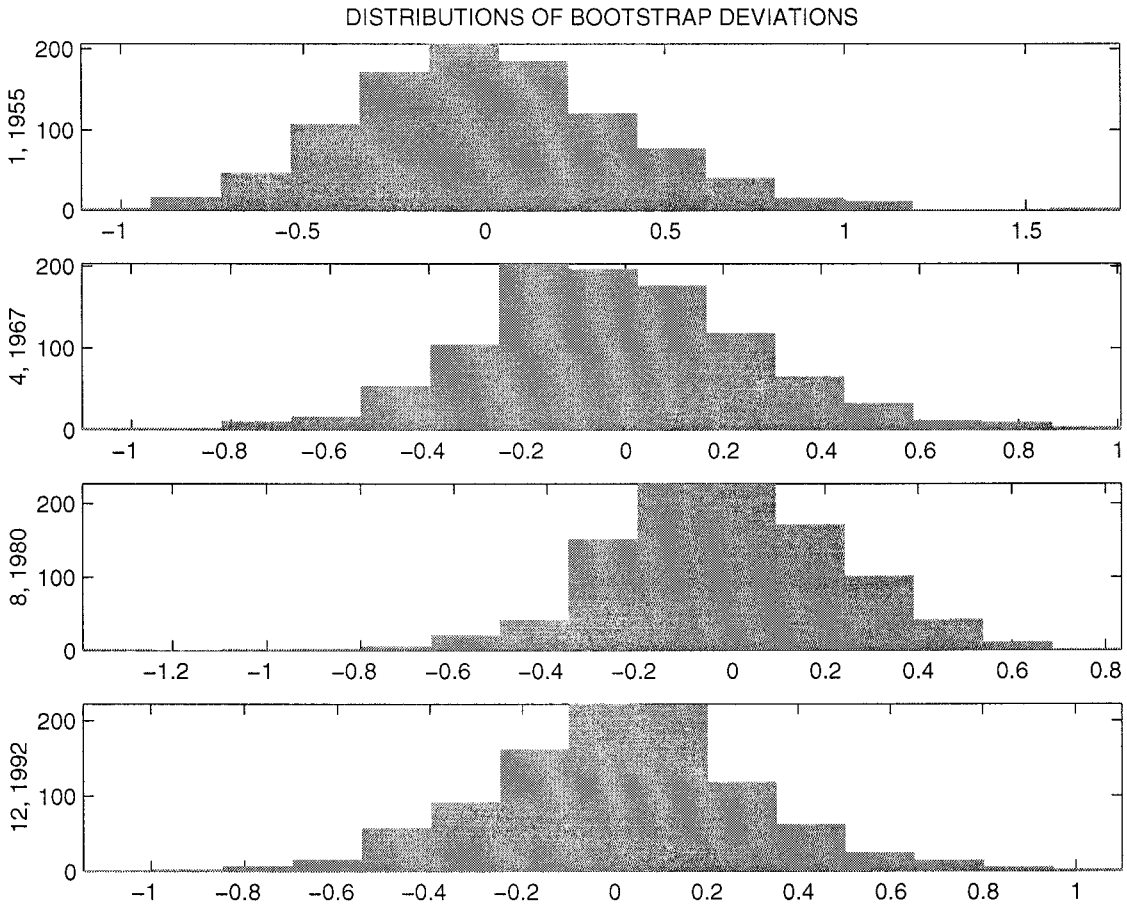

FIG. 7. Histograms of the bootstrap deviations at four representative time points.

the prediction error can be obtained by calculating some statistic of this set. Figure 7 shows the histograms of the bootstrap deviations at four representative time points. As can be seen in the plots the distributions of the deviations are not necessarily symmetric and thus we choose to calculate the statistic for positive and negative deviations separately.

The statistic considered for the error estimation is the square root of the means of the sum of squared bootstrap deviations

$$\sigma^+ = \left(\frac{1}{N^+}\sum_{i=1}^{N^+}(\{P_i^*\}^+ - P)^2\right)^{1/2};$$

$$\sigma^- = -\left(\frac{1}{N^-}\sum_{i=1}^{N^-}(\{P_i^*\}^- - P)^2\right)^{1/2}, \qquad (4)$$

where the sets $(\{P_i^*\}^+ - P)$ and $(\{P_i^*\}^- - P)$ are the positive and negative bootstrap deviations sets and N^+ and N^- are the corresponding sizes of the sets such that $N^+ + N^- = B$. Other statistics, for example, mean of the deviations or the maximum values in the distributions could be considered as well, but experimentation showed that use of different statistics results in similar error estimates.

The values of σ^+ and σ^- give a measure of the relative accuracy expected from each prediction. For the purpose of error estimation a multiplication factor can be used to scale the bootstrap deviations to appropriate error bar lengths. This scaling is arbitrary and should be one appropriate to achieve a certain goal. A practical important goal is optimal flagging of predicted anomalies as El Niño or La Niña events. A more heuristic approach to use the bootstrap deviations would have been construction of confidence intervals. Unfortunately, accurate estimates of distributions are needed in order to construct reliable confidence intervals and the level of accuracy achieved in bootstrapping predictions of short climate data records is not adequate for this purpose.

5. Results

Figure 8 shows the positive and negative error estimates of the predictions at 3-, 6-, 9-, and 12-month leads. The error estimates are defined as the values of

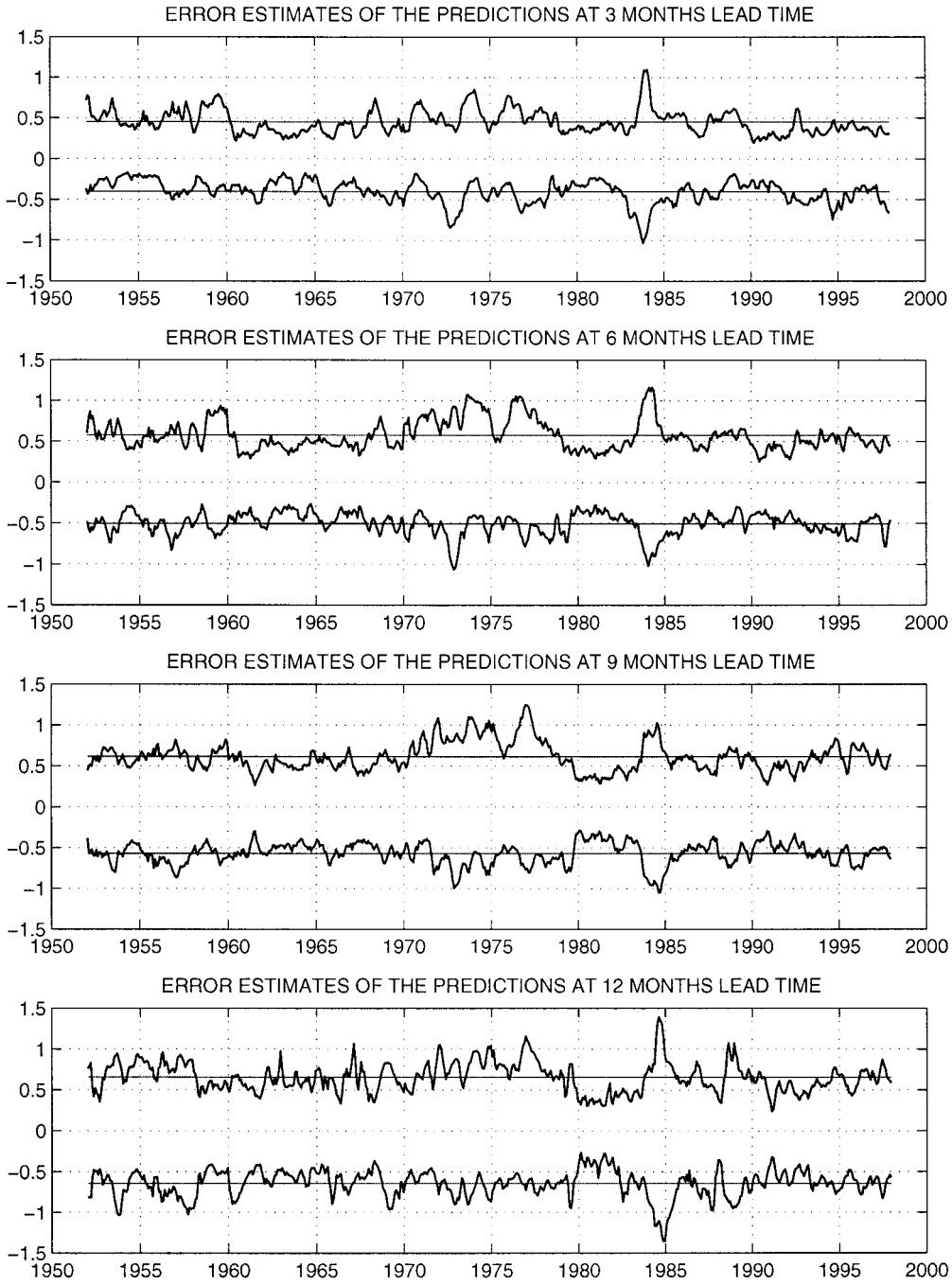

FIG. 8. The positive and negative error estimates of the predictions at (from top to bottom) 3-, 6-, 9-, and 12-month lead times. The straight horizontal lines are the corresponding average values.

σ^+ and σ^- multiplied by the coefficient 1.5. The choice of the coefficient is based on considerations regarding optimal flagging of El Niño and La Niña events and is explained in the next section. The vertical distance between the error estimate curves at a time point defines the length of the error bar and gives a relative measure of confidence we should have in the prediction.

Averages of the error estimates are also plotted in Fig. 8. The average error bar lengths of predictions at 3-, 6-, 9-, and 12-month leads are 0.86°, 1.08°, 1.18°, and 1.30°C, correspondingly, indicating the expected general deterioration of predictability with lead time. The large standard deviations of the error bars, equaling 0.22°, 0.29°, 0.29°, and 0.32°C, indicate considerable variations of the error estimates with time. Periods of large errors are usually, but not always, associated with

large Niño-3.4 anomalies. Positive anomalies are somewhat larger on average. This is probably due to the average magnitude of El Niño events that is larger than that of La Niña events, and the general proportionality of error estimates to prediction magnitudes.

The lower plots of Figs. 2–5 combine the information provided by the linearly corrected forecasts with that provided by their error estimates. These plots should be compared to the corresponding upper plots that show the uncorrected predictions at 3-, 6-, 9-, and 12-month lead times and to the Niño-3.4 index in Fig. 1.

Large peaks and troughs in the Niño-3.4 series are generally better fitted by the corrected predictions than by the corresponding uncorrected ones in the upper panels. As shown in Table 1, the average magnitudes of the corrected predictions are larger and close to those of the observed data. In terms of correlation and rmse skills, also given in Table 1, the two forecasts are very similar. The magnitudes of the corrected predictions are slightly smaller on average than those of the observations because the insufficient correction factor that the training data provide. The decreasing average magnitudes of corrected predictions with lead time is probably due to increasing correlated noise and decreasing sufficient linear correction. The large average magnitude of the 3-month lead corrected prediction is due to the relatively poor linear fit of the prediction residuals to the observation at that lead time.

The error estimates attached to the linearly corrected predictions in Figs. 2–5 can aid in assessing the significance of a predicted anomaly. A positive (negative) anomaly can be confirmed as such if its lower (upper) limit is above (below) zero. A visual evaluation shows that, in general, the 3-, 6-, and 9-month lead forecasts of large events are correctly confirmed by their error estimates—predictions of most of the positive and negative events mentioned in Table 2 in Trenberth (1997) have their corresponding lower and upper limits above or below the zero line. A conspicuous exception is the period between the begining of 1982 to the end of 1985 where the predictions are consistently offset by a few months from the true anomalies. The error estimates at the beginning of the period are not larger than the average and do not indicate a probability for the imminent very strong 1982/83 El Niño. Very large error estimates since mid-1983 do imply some irregular conditions and the low confidence one should have in predictions during this period.

The accuracy of predicted magnitudes and durations of events deteriorates with lead time. The error estimates of the 12-month lead predictions are relatively very large and straddle in many cases both sides of the zero line. This indicates low predictive power and low confidence that we should have in the predictions at that lead time.

Small erroneous anomalies in the uncorrected prediction are magnified by the linear correction and might become large enough to be mistaken as events. Examples are the negative anomalies during 1994/95 that the 6-, 9-, and 12-month lead time forecasts predict and the the positive anomalies in 1979 appearing in all the lead time predictions. In the first case the additional information provided by the error estimates would probably prevent a wrong interpretation. In the second case the error bars on the corrected predictions are not long enough but considering the lower limits set by the error estimates, the results are comparable to the original uncorrected forecast.

There are large false alarms stemming from bad predictions that we would have liked to be warned about by the error estimates. Examples are the large negative anomalies that the forecasts at all lead times predict during 1952/53 and 1959. In the case of the 1952/53 false alarm, the upper limit is clearly below the zero line indicating, wrongly, a call for a negative event. That is also the case with the 9- and 12-month lead predictions during 1959. The upper limits of the corresponding 3- and 6-month lead predictions are very close to zero implying that the negative prediction might be wrong. The fact that warnings do not come in the cases of the 1952/53 and 1959 anomalies reminds us that the error estimation scheme is certainly not infallible. The failure of both prediction and error estimations in these cases might point to a problem with the predictors during the first decade of the data records. The inaccuracy in predicting the time of the peak and the duration of the 1954/56 La Niña event is consistent with this suggestion.

6. Flagging El Niño and La Niña events

The upper panel of Fig. 9 shows the 5-month running mean of the Niño-3.4 index. Values exceeding ± 0.4°C for at least six consecutive months are stippled. The stippling indicates El Niño and La Niña events according to the definition of ENSO by Trenberth (1997). Notwithstanding the arbitrariness of that definition, it is based on scientific experience and reasoning and is probably a good choice for quantitative comparisons and for practical economical considerations. It is desirable to flag in advance future warm and cold events that will later be declared El Niño or La Niña according to that definition.

The prediction error estimates were defined in a previous section as σ^+ and σ^- multiplied by a coefficient. The choice of the coefficient is arbitrary and depends on the purpose for which the error estimates are to be used. We chose as criterion for scaling σ^+ and σ^- an optimal sorting of predictions of past data into one of the three categories: El Niño, La Niña, or neutral, where a prediction can be defined as El Niño (La Niña) if its lower (upper) limit is above (below) the zero line.

Optimal sorting of predictions to one of the three states can be defined as the one maximizing the total number of hits minus the number of false alarms in the available data, where a hit is defined as a correct flagging

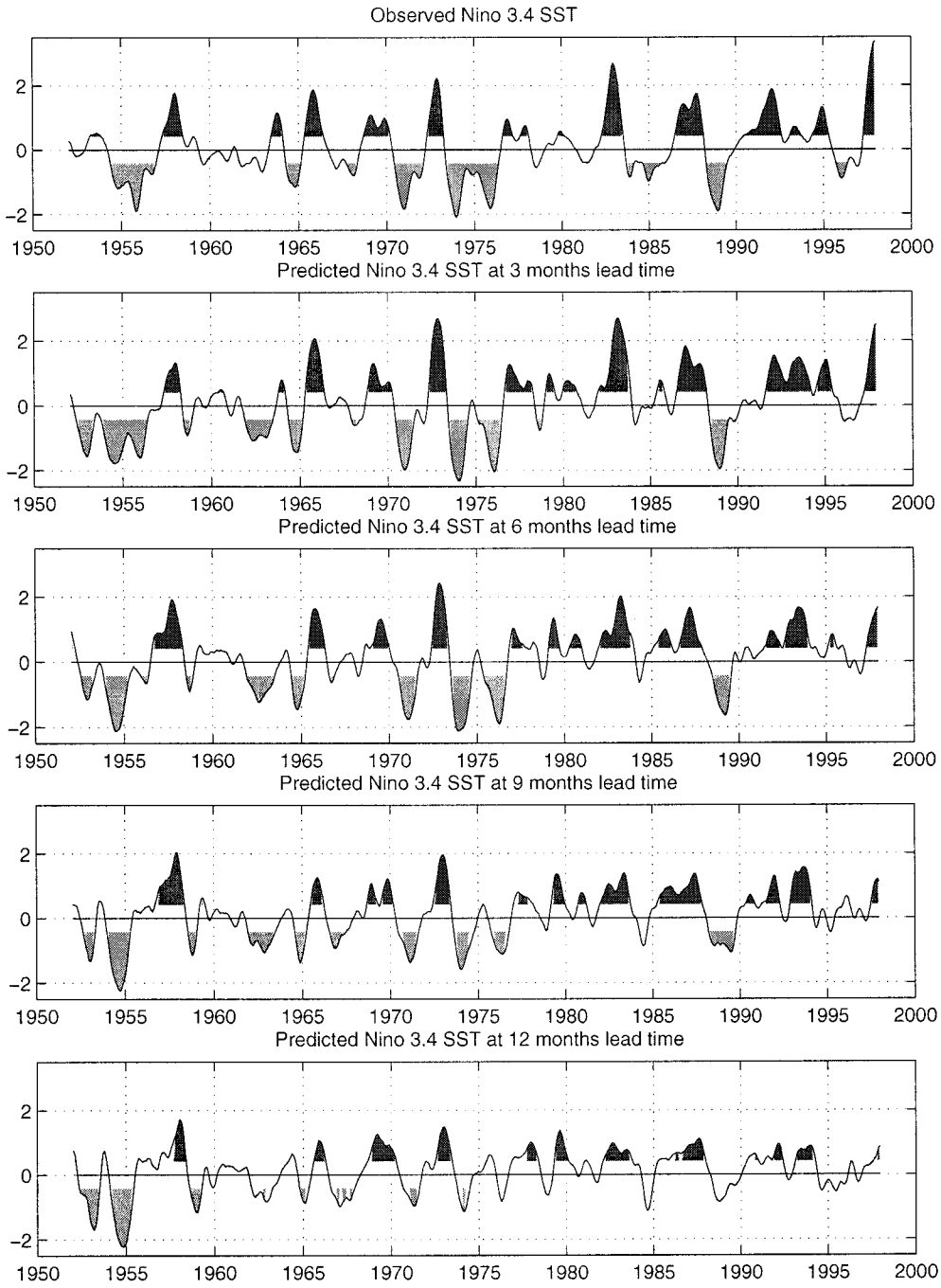

FIG. 9. (upper panel) A 5-month running mean of the Niño-3.4 index. Values exceeding the threshold of ±0.4°C are stippled to indicate ENSO events following the definition of Trenberth (1997). The second, third, fourth, and fifth panels show the 5-month running means of the 3-, 6-, 9-, and 12-month lead predictions with ENSO events stippled according to the flagging scheme discussed in the text.

of an event, and a false alarm is a call for an event that does not transpire. A month-by-month counting of hits and false alarms in the predictions at all lead times together found the optimal scaling coefficient to be 1.5. The sensitivity of the flagging system to the value of the scaling coefficient is low and coefficients in the range 1.2–2.0 result in quite similar sorting.

The 5-month running means of the testing predictions at 3-, 6-, 9-, and 12-month lead times appear in Fig. 9 (in that order) below the Niño-3.4 index. The stippled

areas indicate predictions that were flagged as El Niño or La Niña events. Each of the lower plots of Fig. 9 is a prediction, at the corresponding lead time, of the Niño-3.4 index in the upper panel. How well the plots in the lower panels actually emulate the one on top depends on both the accuracy of the predictions and the quality of the error estimation. The 3- and 6-month lead time plots are quite similar to the one of the observed data indicating good ability of the prediction and error estimation scheme to call for onset and termination of El Niño and La Niña events six months in advance. The 9-month lead plot is reasonable too but a significant deterioration is noted. Duration of many events is too short and few are missing altogether. The 12-month lead plot shows only very few events flagged, reflecting the fact that the error bars of the 12-month lead prediction are very long and straddle in many cases both sides of the zero line. In all the plots there are also cases of false alarms, most notably the false call for a La Niña during 1952/53.

Assuming that flagging future events should be successful at a level similar to that of flagging the testing predictions of past data, the flagging scheme can be considered for declaring future anomalies as El Niño and La Niña events. It must be noted that a simple scheme looking for an optimal threshold to define predictions as ENSO events can be as effective as the flagging scheme presented in this section in sorting anomalies into the three states. Such schemes might require determination of more arbitrary parameters and be more sensitive to their values. Whatever flagging scheme is used it should be interpreted in conjunction with error estimates of the predictions in order to arrive at the best conclusions.

7. Discussion and conclusions

We presented procedures to enhance predictions of large events in the Niño-3.4 index and to estimate the errors in the predictions. Combined together these procedures enable improved evaluation of large anomalies that the model predicts. The presented corrected predictions and their errors estimations are true testing results and assuming stationarity of the physical system can be viewed and evaluated as if they were carried out in an operational mode. The linear correction provides a simple and effective remedy in cases where the original predictions of the Niño-3.4 index tend to underestimate the magnitude of large events.

The bootstrap error estimation scheme that is used is automatic and objective. An arbitrary decision to be made is about the value of the coefficient that multiplies the square roots of the means of the squared bootstrap deviations. It should be noted that one such coefficient must be chosen, otherwise the default of unity, that has no special reason to be used, is the choice. One could think of choosing different coefficients for the different lead times and even different coefficients for positive and negative anomalies. The choice made in this paper is to use the minimum possible number of arbitrary coefficients. The selection of the coefficient was made such that past El Niño and La Niña events are optimally flagged. That choice also provides the means for sorting future predicted Niño-3.4 SST anomalies into El Niño, La Niña, or neutral states.

The bootstrap error estimates have themselves inherent errors. Methods to evaluate the accuracy of bootstrap error estimates exist (Efron 1992; Beran 1997) but were not attempted in this paper. The relative merit of the flagging scheme depends on the quality of both the predictions and the the error estimation method. A quantitative assessment of the flagging scheme requires a definition of a fidelity measure. As such definition is subjective and there is no consensus on a universal one, it is left to the reader to decide about the worthiness of the scheme.

Barnston et al. (1994) noted a moderate level of mean skill in the ENSO prediction models. They also mentioned the need to supplement predictions with expression of uncertainty but speculated that presenting forecasts in a probabilistic way might confuse and frustrate many users. It will probably require many more years and new types of observations to substantially improve the mean skill of ENSO predictions. This paper proposes and provides a way to improve prediction of the more important events. The error estimates accompanying the improved predictions give some measure of their uncertainty and the El Niño and La Niña flagging scheme extracts from the error estimations the categorical forecasts that industry and the general public usually prefer.

Acknowledgments. Thanks to Rich Pawlowicz for suggesting the use of the bootstrap method and for helpful discussions. Benyang Tang carried out the EEOF analysis of the SLP and SST data. William Hsieh and Adam Monahan reviewed a preliminary version of this paper and contributed many useful comments. This work is supported by grants to William Hsieh from Environment Canada and from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Allen, D., 1974: The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16,** 125–127.

Barnston, A. G., and H. M. van den Dool, 1993: A degeneracy in cross-validation skill in regression-based forecasts. *J. Climate*, **6,** 963–977.

——, and Coauthors, 1994: Long-lead seasonal forecasting—Where do we stand? *Bull. Amer. Meteor. Soc.*, **75,** 2097–2114.

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

Beran, R. J., 1997: Diagnosing bootstrap success. *Ann. Inst. Stat. Math.*, **49,** 1–24.

Bishop, C. M., 1995: *Neural Networks for Pattern Recognition.* Clarendon Press, 482 pp.

Cichocki, A., and R. Unbehauen, 1993: *Neural Networks for Optimization and Signal Processing.* John Wiley and Sons, 526 pp.
Davison, A. C., and D. V. Hinkley, 1997: *Bootstrap Methods and their Application.* Cambridge University Press, 582 pp.
Efron, B., 1992: Jackknife-after-bootstrap standard errors and influence functions. *J. Roy. Stat. Soc.,* **B54,** 83–127.
——, and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap.* Chapman & Hall, 436 pp.
Enfield, D. B., 1989: El Niño, past and present. *Rev. Geophys.,* **27,** 159–187.
Golub, G. H., M. Heath, and G. Wahba, 1979: Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics,* **21,** 215–223.
Goswami, P., and Srividya, 1996: A novel neural network design for long range prediction of rainfall pattern. *Curr. Sci.,* **70,** 447–457.
Haber, E., and D. W. Oldenburg, 2000: A GCV based method for nonlinear ill-posed problems. *Comput. Geosci.,* **4,** 41–63.
Hastenrath, S., L. Greischar, and J. van Heerden, 1995: Prediction of the summer rainfall over South Africa. *J. Climate,* **8,** 1511–1518.
Keppenne, C. L., and M. Ghil, 1992: Adaptive filtering and prediction of the Southern Oscillation Index. *J. Geophys. Res.,* **97,** 20 449–20 454.
Knaff, J. A., and C. W. Landsea, 1997: An El Niño–Southern Oscillation climatology and persistence (CLIPER) forecasting scheme. *Wea. Forecasting,* **12,** 633–652.
Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.,* **26,** 1589–1600.
Navone, H. D., and H. A. Ceccatto, 1994: Predicting Indian monsoon rainfall—A neural network approach. *Climate Dyn.,* **10,** 305–312.
Penland, C., 1989: Random forcing and forecasting using Principal Oscillation Pattern analysis. *Mon. Wea. Rev.,* **117,** 2165–2185.
——, and T. Magorian, 1993: Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *J. Climate,* **6,** 1067–1076.
——, L. Matrosova, K. Weickmann, and C. Smith, 1999: Forecast of Tropical SSTs using linear inverse modeling (LIM). *Exp. Long-Lead Forecast Bull.,* **8,** 38–41.
Reynolds, R. W., and T. M. Smith, 1994: Improved global sea surface temperature analysis using optimum interpolation. *J. Climate,* **7,** 929–948.
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.
Saunders, A., M. Ghil, and J. D. Neelin, 1999: Forecasts of Niño 3 SST anomalies and SOI based on singular spectrum analysis combined with the maximum entropy method. *Exp. Long-Lead Forecast Bull.,* **8,** 39–41.
Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes, 1996: Reconstruction of historical sea surface temperatures using orthogonal functions. *J. Climate,* **9,** 1403–1420.
Syu, H., and J. D. Neelin, 1999: Prediction of NINO3 SST anomaly in a hybrid coupled model with a piggy-back data assimilation initialization. *Exp. Long-Lead Forecast Bull.,* **8,** 10–12.
Tang, B., W. W. Hsieh, A. H. Monahan, and F. T. Tangang, 2000: Skill comparisons between neural networks and canonical correlation analysis in predicting the equatorial Pacific sea surface temperatures. *J. Climate,* **13,** 287–293.
Tangang, F. T., W. W. Hsieh, and B. Tang, 1998a: Forecasting the regional sea surface temperature of the tropical Pacific by neural network models, with wind stress and sea level pressure as predictors. *J. Geophys. Res.,* **103,** 7511–7522.
——, B. Tang, A. H. Monahan, and W. W. Hsieh, 1998b: Forecasting ENSO events: A neural network–extended EOF approach. *J. Climate,* **11,** 29–41.
Trenberth, K. E., 1997: The definition of El Niño. *Bull. Amer. Meteor. Soc.,* **78,** 2771–2777.
Wahba, G., 1990: *Spline Models for Observational Data.* SIAM, 169 pp.
Ward, M. N., and C. K. Folland, 1991: Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol.,* **11,** 711–743.
Weare, B. C., and J. S. Nasstrom, 1982: Examples of extended empirical orthogonal function analysis. *Mon. Wea. Rev.,* **110,** 481–485.
Woodruff, S. D., R. J. Slutz, R. L. Jenne, and P. M. Steurer, 1987: A comprehensive ocean–atmosphere data set. *Bull. Amer. Meteor. Soc.,* **68,** 1239–1250.
Yuval, 2000: Neural network training for prediction of climatological time series, regularized by minimization of the generalized cross-validation function. *Mon. Wea. Rev.,* **128,** 1456–1473.
Zebiak, S. E., M. A. Cane, and D. Chen, 1999: Forecast of Tropical Pacific SST using a simple ocean–atmosphere dynamical model. *Exp. Long-Lead Forecast Bull.,* **8,** 1–5.