Analog (Nonlinear) Forecasts of the Southern Oscillation Index Time Series

WASYL DROSDOWSKY

Bureau of Meteorology Research Centre, Melbourne, Australia

(Manuscript received 9 March 1993, in final form 20 September 1993)

ABSTRACT

A nonlinear time series forecasting scheme developed by Sugihara and May has been applied to the Southern Oscillation index. Although forecast skill is comparable only to persistence or linear (autoregressive) methods, the scheme has the advantage of identifying close analogs to the current situation, if these exist. The operational implementation of the scheme in the *Seasonal Climate Outlook* issued by the National Climate Centre of the Australian Bureau of Meteorology is described and its performance during the 1991/92 El Niño–Southern Oscillation event is examined.

1. Introduction

The use of circulation analogs has a long history in climate and weather forecasting (Lorenz 1969; Gutzler and Shukla 1984). A common finding of these studies has been the difficulty in obtaining realistic analogs to a given situation, due to the large number of degrees of freedom in typical circulation patterns. Gutzler and Shukla (1984) found improved forecast skill by reducing the number of degrees of freedom by spatial and temporal averaging; however, this also increased the error doubling rate and decreased the value of the forecasts. Gutzler and Shukla speculated that closer analogs could be obtained by progressively decreasing the number of degrees of freedom by using fewer grid points. Ultimately, when reduced to a single point, such an analog would have little forecast value since it would not represent the large-scale flow. Van den Dool (1989), however, showed that reasonable analogs could be found for limited areas and that these could be overlapped to produce forecasts comparable in skill to single-level barotropic models.

In recent years study of dynamical systems has led to powerful new methods of forecasting nonlinear time series (Farmer and Sidorowich 1988; Sugihara and May 1990). These methods, which are essentially analog prediction methods, are based on the premise that for a deterministic dynamical system the future state of that system may be predictable from past states, even if all the relevant variables and the dynamical laws governing the system are not known and the behavior of the system is chaotic. The major difference between these techniques and those considered in meteorological applications is the use of the previous history of the circulation pattern in selection of the analog. Use of one prior state in selecting an analog was suggested by Barnett and Preisendorfer (1978). Consideration of past states, however, dramatically increases the number of degrees of freedom and would appear to make it even more difficult to find good analogs to complex circulation patterns, as found by Lorenz (1969). This procedure may be useful, however, for a single time series, such as the Southern Oscillation index. This index represents an extreme reduction in spatial degrees of freedom as described by Gutzler and Shukla (1984) while still maintaining a useful portion of the variance in the global monthly mean pressure pattern. In this paper the Sugihara and May (1990) methodology is explored and applied to the Troup (1967) Southern Oscillation index (SOI).

2. Methodology

The single time series x_t is "embedded" in an E-dimensional space defined by a sequence of lagged coordinates $(x_t, x_{t-\gamma}, x_{t-2\gamma}, \ldots, x_{t-(E-1)\gamma})$, where γ is the lag interval, usually taken as one time step. The $E + 1$ closest neighbors (analogs) to the current state [defined by the vector $(x_t, x_{t-\gamma}, x_{t-2\gamma}, \ldots, x_{t-(E-1)\gamma})$] are found and used to construct the smallest simplex containing the current state. Future states of the system are found by projecting each analog forward nT, where $n = 1, 2, \cdots$, time steps and taking a weighted average of the analogs. The weights employed by Sugihara and May are proportional to $\exp(-d_j^2)$, where d_j is the root-mean-square (rms) or Euclidean distance between the jth analog and the current state of the system. The Sugihara and May simplex method requires $E + 1$ nearest neighbors (analogs). In the present work, a simpler method (Euclidean or root-mean-square dis-

Corresponding author address: Mr. Wasyl Drosdowsky, Bureau of Meteorology Research Centre, P.O. Box 1289K, Melbourne, 3001, Australia.

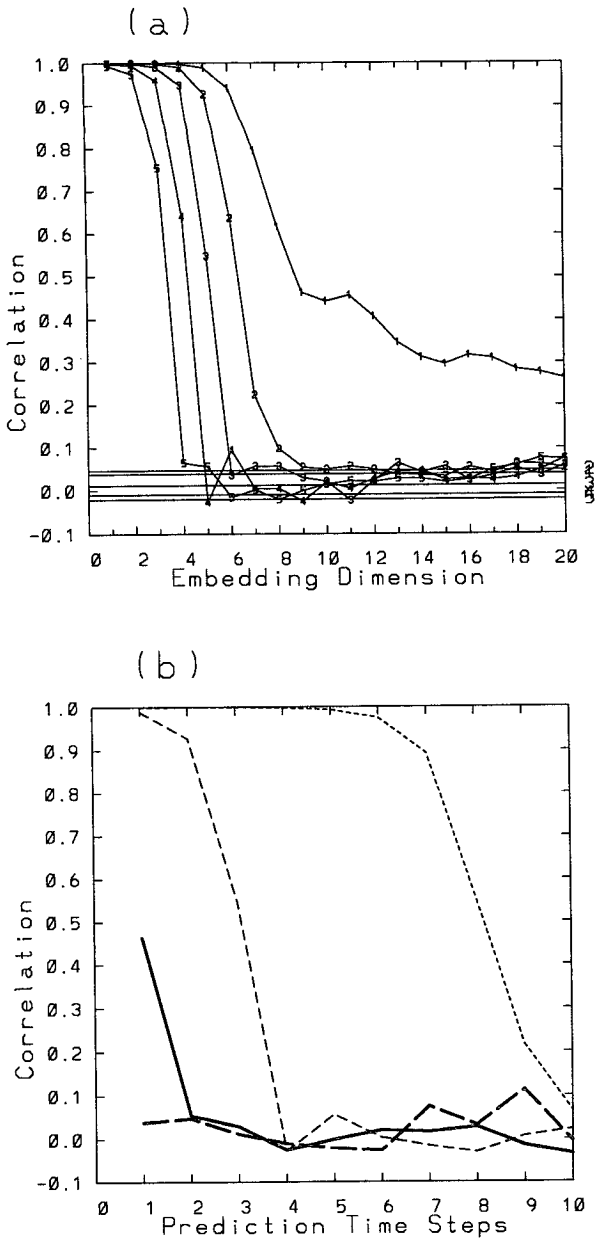

FIG. 1. (a) Correlation between observed and predicted values of the logistic map time series, as a function of embedding dimension for predictions of one to five time steps. Persistence forecasts, which are independent of the embedding dimension, are shown as thin horizontal lines for the same time steps. (b) Correlation between observed and predicted values of the logistic map time series, for persistence forecasts (heavy dashed curve) and for nonlinear forecasts with embedding dimensions of $E = 1$ (dotted), $E = 5$ (dashed), and $E = 9$ (heavy solid) as a function of prediction time.

tance) will be used to find the nearest neighbors. The requirement for $E + 1$ nearest neighbors will be retained, on the assumption that in most cases the nearest neighbors selected by this method will form a simplex around the current state. Most meteorological studies have used just one (the "best") analog, or an unweighted average of a few analogs, with no justification for the choice of number of analogs.

The optimal embedding dimension E is determined by a trial and error procedure, using the library of patterns formed by the first half of the time series to predict the evolution at each point of the last half of the time series. This then determines the window over which the analog is selected. The "correct" embedding dimension E is related to the dimension of the attractor of the system or the effective number of degrees of freedom D by $D \approx 2E + 1$ (Farmer and Sidorowich 1988).

The forecast system can be tested on time series with known properties. A chaotic time series can be simply generated using iterated maps such as the logistic map $x_{t+1} = rx_t(1 - x_t)$, with $r = 4$. [See May (1976) for a discussion of the properties of this simple mapping.] A 1000-point time series was generated using this mapping and the first 500 points were used to predict the evolution at each of the last 500 points. Figure 1a shows the correlation between observed and predicted values for one to five time steps, as a function of the embedding dimension. The predictability, measured by the correlation between observed and forecast values is seen to decrease from a maximum at $E = 1$ with increasing embedding dimension. Figure 1b shows the correlation between observed and forecast values for forecasts of one to ten time steps for persistence (heavy dashed curve) and embedding dimensions of $E = 1$ (light dotted curve), $E = 5$ (light dashed curve), and $E = 9$ (heavy solid curve). The sharp decrease in predictive skill of the nonlinear forecast as a function of prediction time and the poor performance of persistence are characteristic of chaotic or nonlinear systems.

Sugihara and May (1990) noted that the results do not depend strongly on γ, provided it is not too large. The decrease in predictive skill with increasing prediction times (Fig. 1a) is due to the divergence of the analog trajectories in the E-dimensional phase space and is a characteristic of a chaotic system. In contrast, Sugihara and May show that a strongly periodic signal with added noise, whose time series may resemble that of a chaotic signal, shows no marked decrease in predictive skill with increased prediction time.

3. The Southern Oscillation index time series

The nonlinear forecasting technique is applied to a time series of 696 monthly values of the Troup SOI (Troup 1967), covering the period 1933–1990. This index is the Tahiti minus Darwin pressure difference normalized to a standard deviation of 10. Extensions of this index back to 1876 have been produced by Ropeleweski and Jones (1987) and Allan et al. (1991). The period prior to 1933, unfortunately, contains segments of missing data and is not used in the development of the forecast method but is used in the operational implementation.

As with the developmental chaotic series described in the previous section, the first half of the time series (348 months) is used as a library of patterns to obtain predictions for each point in the second half of the series. Figure 2a shows the correlation between observed and predicted values for one to five time steps as a function of the embedding dimension. The time delay γ was set at 1 month, the resolution of the SOI time series. Experiments were carried out with 2- and 3-month delays, yielding generally similar but less skillful results. For 1-month predictions the largest correlation is found for an embedding dimension of 2, with secondary peaks at $E = 5$ and in the range 8–11. For longer prediction times, this last peak becomes dominant, with overall best results for $E = 9$. Also shown in Fig. 2a as thin horizontal lines are persistence forecasts for one to five time steps. Unlike the deterministically chaotic time series produced by the logistic map, the SOI time series is strongly autocorrelated, so that only for a 5-month time step with $E = 9$ is the performance of the analog scheme actually superior to persistence. This is highlighted in Fig. 2b, which shows the correlation between observed and forecast values for forecasts of one to ten time steps for persistence (heavy dashed curve), $E = 1$ and 5 (light broken curves) and $E = 9$ (heavy solid curve). The decrease in predictive skill as a function of prediction time shows the decrease characteristic of chaos, although this in itself does not prove that the SOI series is chaotic, only that it may be. In this case a similar decline is shown by persistence, consistent with a purely stochastic process. Tsonis and Elsner (1992) have examined the scaling of the correlation for increased prediction time and concluded that the SOI is in fact chaotic. The optimal embedding dimension of order 9–12 suggests a system with four to six degrees of freedom, consistent with the results of Hense (1987) for east Pacific sea surface temperature anomalies and central Pacific rainfall anomalies.

The skill of the analog forecasts as measured by rms errors exceeds that of persistence for a wide range of embedding dimensions for forecasts of two to five time steps (Figs. 2c, 2d). A possible reason for the poor performance of analog forecasts relative to persistence for one time step is the short length of the time series and lack of good analogs. This lack of good analogs is evident in the low correlation (not shown) and large rms errors at $t = 0$. This error generally increases as the embedding dimension increases, since the rms or Euclidean distance is being minimized over a greater number of lagged coordinates or prior time steps.

Van den Dool (1989) has suggested that analog forecast errors can be reduced if the initial fit at time $t = 0$ is improved. For persistence this is always zero, so a combination of persistence and some other forecast such as an analog may perform better than either forecast alone. In the present scheme it is relatively trivial to include persistence by adjusting the weighted analog so that the $t = 0$ value agrees exactly with the current observed base value. Applying this procedure results in a marked increase in the skill of the analog as measured by the correlation between observed and forecast

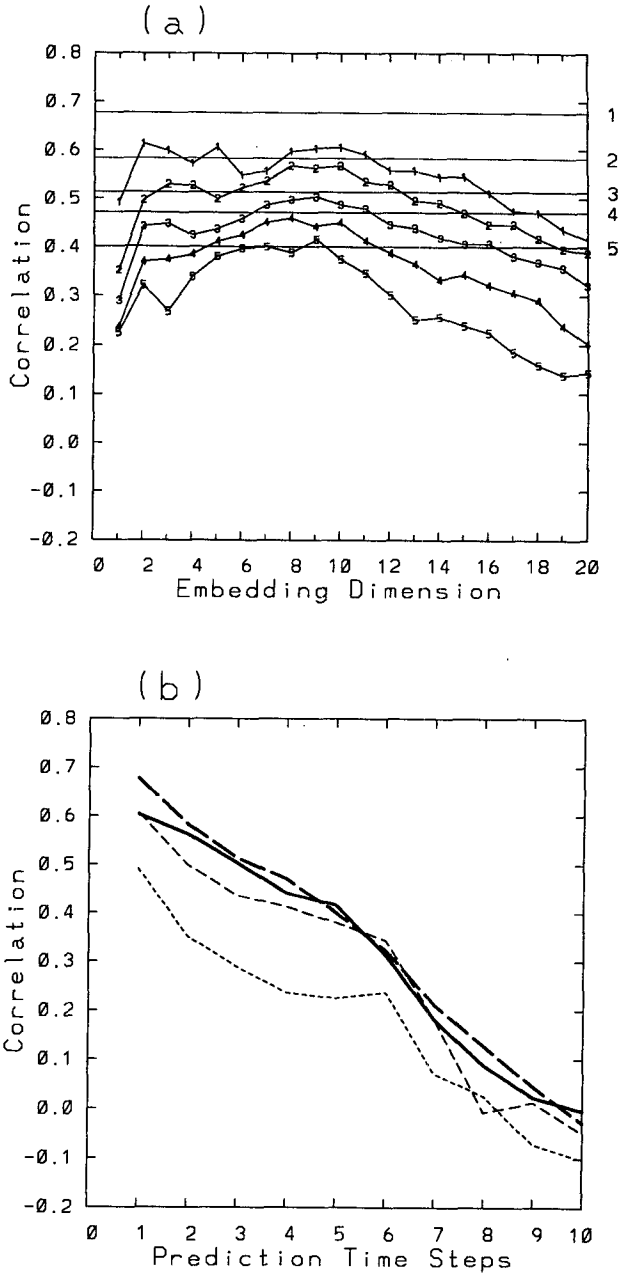

FIG. 2. (a) Correlation between observed and predicted values of the SOI time series, as a function of embedding dimension for predictions of one to five time steps. Persistence forecasts, which are independent of the embedding dimension, are shown as thin horizontal lines for the same time steps. (b) Correlation between observed and predicted values of the SOI time series, for persistence forecasts (heavy dashed curve) and for nonlinear forecasts with embedding dimensions of $E = 1$ (dotted), $E = 5$ (dashed), and $E = 9$ (heavy solid) as a function of prediction time.

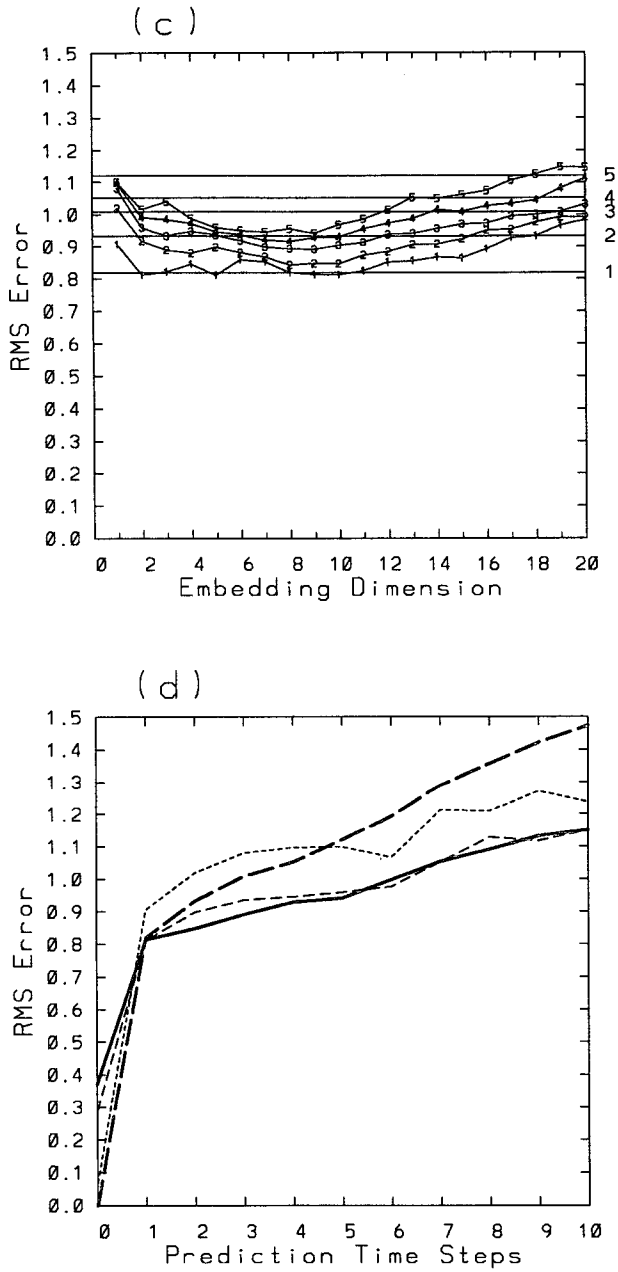

FIG. 2. (c) Root-mean-square errors between observed and predicted values of the SOI time series, as a function of embedding dimension and for persistence forecasts (thin horizontal lines), for predictions of one to five time steps. (d) Root-mean-square errors between observed and predicted values of the SOI time series, for persistence forecasts (heavy dashed curve) and for nonlinear forecasts with embedding dimensions of $E = 1$ (dotted), $E = 5$ (dashed), and $E = 9$ (heavy solid) as a function of prediction time.

values. The one time step forecasts are now equal to persistence for embedding dimensions E of 8 or 9, and the two or more time step forecasts exceed persistence by an increasing amount (Figs. 3a,b). Root-mean-square errors are improved for forecasts up to three time steps ahead and degraded slightly for forecasts of five time steps or longer (Figs. 3c,d).

4. Operational forecasts

a. Implementation

The nonlinear prediction scheme was implemented in mid-1991 in the monthly *Seasonal Climate Outlook* issued by the National Climate Centre of the Australian Bureau of Meteorology. It replaced an earlier scheme of analog selection based on the correlation between the previous 12 months SOI values and all other 12-month periods, and the mean value over that 12-month period.

To select analogs, we set $\gamma = 1$ month and $E = 9$ (which implies selecting the $E + 1 = 10$ best analogs) and use the entire available SOI time series from 1876 to the present time except that points with missing values in the prior 9 months (the analog selection period) or the following 3 months (the forecast period) are not used. Since the Southern Oscillation is believed to be tied to the seasonal cycle, the analogs are restricted to those points within 1 month of the same calendar month as the current state. Only the best of the three possible states in a given year is used, so that each analog represents a different trajectory or event. The initial operational implementation used a stricter restriction to the same calendar month, but this was found to place limits that were too severe on the number of possible analogs, since the monthly values may contain an aliased signal of 2-month period due to the tropical intraseasonal (or Madden–Julian) oscillation. This signal can also be reduced by low-pass filtering the data with a simple weighted running mean. This procedure results in a slight increase in skill of both the analog and persistence forecasts; however, this is offset by the decreased lead time.

b. Verification

A set of hindcasts was performed on the 348 months in the second half of the development time series, from January 1962 to December 1990, using the operational implementation with the analog selection restricted to within 1 month of the same calendar month as the current month. This restriction in the number of possible analogs was partially offset by allowing the hindcasts to make use of all the years of record, extending back to 1876, and was updated to the current year. The skill of these hindcasts was comparable to that shown in Fig. 3. A more interesting aspect of these hindcasts was the seasonal variation in skill, which followed that achieved by persistence or linear regression, with least skill (as measured by the correlation between observed and forecast values of 0.3–0.4) in March–May and greatest skill (correlations of 0.8) in August–October.

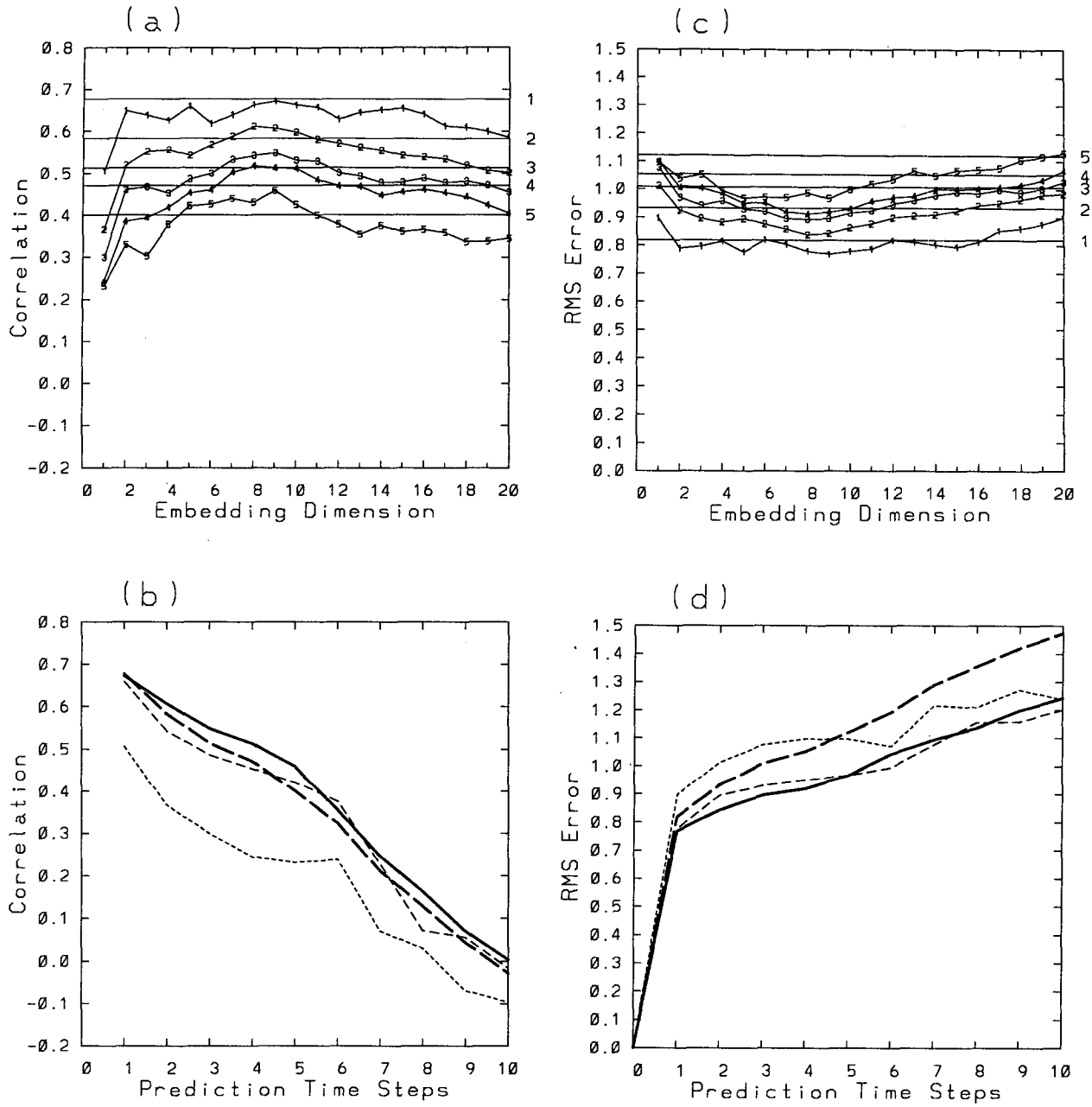

FIG. 3. As for Fig. 2 but for combined persistence–analog forecasts.

Operational forecasts on independent data have been run from January 1991. The performance of the scheme for 1-, 2-, and 3-month predictions for the January 1991 to November 1992 period is shown in Fig. 4. The scheme produced very poor forecasts during the first 6 months of 1991, and this affects the skill, as shown by the low correlation of less than 0.1 between the observed and forecast values for 1-, 2-, or 3-month forecasts, for the full 23-month period. These poor forecasts may be due to the unusual development of the 1991/92 El Niño Southern Oscillation event during the first half of 1991. During this period the scheme selected several very close analogs (Fig. 5a); however, these tended to diverge during the forecast period. In contrast, the analogs were not as close a fit during the remainder of the period—for example, at the peak of the event in February 1992 (Fig. 5b)—but usually followed a similar evolution over the forecast period and correctly indicated, in this case, the rising trend in the index. Consequently, the scheme shows reasonable skill over the last 17 months with correlations of 0.66, 0.48, and 0.73 for the 1-, 2-, and 3-month forecasts. Some

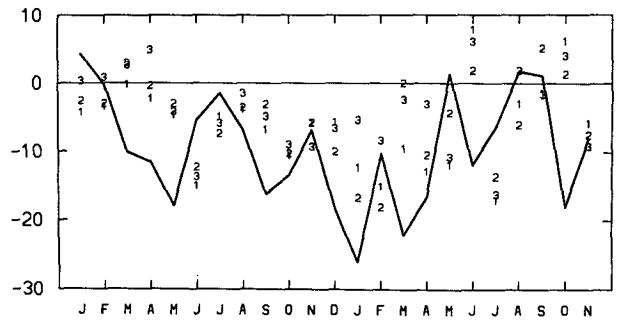

FIG. 4. Observed SOI time series for January 1991 to November 1992 (solid curve) and forecasts for 1-, 2-, and 3-month lead times.

of this skill may also be due to the existence of one very close analog. Most of the mature phase of the 1991/92 event, (from September 1991 through to September 1992) closely paralleled the strong 1911/12 event lagged by 1 month (i.e., June 1992 corresponds to July 1912), as shown by Fig. 6.

5. Discussion and conclusions

The nonlinear forecasting technique provides forecasts comparable to those of persistence or linear regression schemes. The method can produce much better results, relative to persistence, if the SOI time series is high-pass filtered by taking first differences. This procedure reduces the skill of 1-month forecasts (measured as a correlation between observed and forecast values) for persistence to near zero and that of the analog forecasts to approximately 0.3. The major advantage of the nonlinear prediction scheme, compared to persistence or linear regression, lies in the ability to select realistic analogs to the current situation, if these exist. This advantage is lost if the low-frequency components of the time series are removed by first differencing.

The ability to select realistic analogs is useful since the Southern Oscillation exerts a dominant influence on the behavior of other variables, such as rainfall and the large-scale circulation, in the Australia/southwest Pacific region. The future behavior of these variables may then be inferred from examination of the SOI analogs. Attempts to directly find analogs to these large-scale circulation patterns have failed due to the large number of degrees of freedom.

The spread of the analogs during the forecast period can provide a measure of the confidence level of the forecast. In February 1991 (Fig. 5a) there is little agreement between the analogs during the forecast period, suggesting that low confidence should be placed on the forecast or any individual analog. In contrast, in February 1992 (Fig. 5b) the analogs show a consistent pattern and suggest a higher level of confidence in the forecast. A similar assessment can be applied to the rainfall or temperature patterns during the selected analog years. In this case the consistency of the rainfall or temperature patterns during the analog selection period itself is a measure of the strength of the relationship between that variable and the SOI.

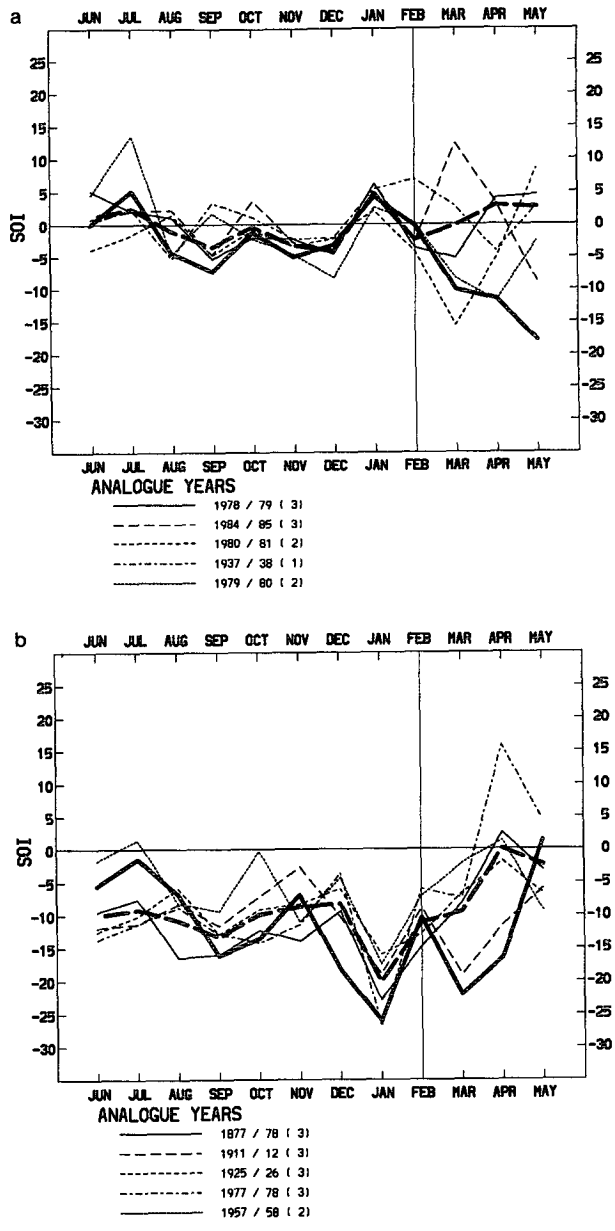

FIG. 5. Operationally selected analogs and forecasts for (a) February 1991 and (b) February 1992. Points corresponding to January, February, or March initial conditions are used. For clarity, only the best five analogs are shown (light dashed or dotted lines), labeled with the year and month corresponding to the current month, together with the current values (solid curve) and the composite weighted analog (solid dashed curve).

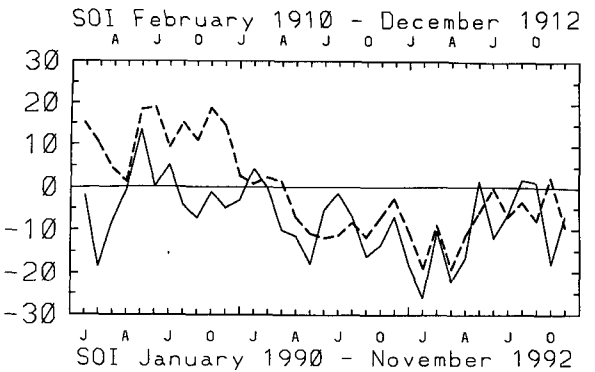

FIG. 6. Southern Oscillation index time series for January 1990 to November 1992 (light solid curve) and February 1910 to December 1912 (heavy dashed curve).

REFERENCES

Allan, R. J., N. Nicholls, P. D. Jones, and I. J. Butterworth, 1991: A further extension of the Tahiti-Darwin SOI, early ENSO events and Darwin pressure. *J. Climate,* **4,** 743–749.

Barnett, T. P., and R. W. Preisendorfer, 1978: Multifield analog prediction of short-term climate fluctuations using a climate state vector. *J. Atmos. Sci.,* **35,** 1171–1787.

Farmer, J. D., and J. J. Sidorowich, 1988: Exploiting chaos to predict the future and reduce noise. Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, LA-UR-88-901, 54 pp.

Gutzler, D. S., and J. Shukla, 1984: Analogs in the wintertime 500 mb height field. *J. Atmos. Sci.,* **41,** 177–189.

Hense, A., 1987: On the possible existence of a strange attractor for the Southern Oscillation. *Beitr. Phys. Atmos.,* **60,** 34–47.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.,* **26,** 636–646.

May, R. M., 1976: Simple mathematical models with very complicated dynamics. *Nature,* **261,** 459–467.

Ropeleweski, C. F., and P. D. Jones, 1987: An extension of the Tahiti—Darwin Southern Oscillation index. *Mon. Wea. Rev.,* **115,** 2161–2165.

Sugihara, G., and R. M. May, 1990: Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature,* **344,** 734–741.

Troup, A. J., 1967: Opposition of anomalies in upper tropospheric winds at Singapore and Canton. *Aust. Meteor. Mag.,* **15,** 32–37.

Tsonis, A. A., and J. B. Elsner, 1992: Nonlinear forecasting as a way of distinguishing chaos from random fractal sequences. *Nature,* **358,** 217–220.

Van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.,* **117,** 2230–2247.