To appear in the *Journal of Geophysical Research*, 2001.

# Use of data assimilation via linear low-order models for the initialization of El Niño – Southern Oscillation predictions

Rafael Cañizares[1], Alexey Kaplan, Mark A. Cane, and Dake Chen

Lamont–Doherty Earth Observatory of Columbia University, Palisades, New York, USA

Stephen E. Zebiak

International Research Institute for Climate Prediction (IRI), Lamont–Doherty Earth Observatory of Columbia University, Palisades, New York, USA

**Abstract.** The utility of a Kalman filter (KF) for initialization of an intermediate nonlinear coupled model for El Niño – Southern Oscillation prediction is studied via an approximation of the nonlinear coupled model by a system of seasonally dependent linear models. The low-dimensional nature of such an approximation allows one to determine a sequence of "perfect" initial states that start a trajectory segment best fitting the observed data. Defining these perfect initial conditions as "true" states of the model, we compute a priori parameters of the KF and test its ability to produce an estimate of the "truth" superior to the less theoretically sound estimates. We find that in this application such a KF does not produce an estimate outperforming a pure observational projection as an initial condition for the coupled model forecast. The violation of standard KF assumptions on temporal whiteness of observational errors and system noise is identified as the reason for this failure.

## 1. Introduction

The first coupled ocean-atmosphere numerical model applied to El Niño – Southern Oscillation (ENSO) prediction was the model by *Zebiak and Cane* [1987] (ZC model). This model is an anomaly model with regard to climatology specified from 1970-1985 observations. Predictions of ENSO on an experimental basis began in 1985 [*Cane et al.*, 1986], providing forecasts for up to 1 year or more. The ocean component was initialized by forcing with analyzed surface wind data (Florida State University (FSU) winds) [*Goldenberg and O'Brien*, 1981], while the initial conditions of the atmospheric component were obtained as a response to the ocean model fields. *Blumenthal* [1991] used a reduced set of variables to build a linear autoregressive (Markov) model from the output of a free run of the ZC model. The linear model was used to study the fastest growing disturbances in the tropical Pacific that limit the model predictability. Another important finding was that these disturbances grow differently depending on the starting season, with maximum growth for February starts. On the basis of this work, *Xue et al.* [1994] created a linear Markov model from a suite of 3-year forecast runs of the ZC model from monthly starts within the interval (1972-1991). When this linear model was applied to ENSO prediction, it demonstrated equal or better skill than the ZC model, especially for short-term forecasts. It showed variation of predictability with starting season similar to that shown by *Blumenthal* [1991].

In recent years the forecast skill of the ZC model has been improved by applying more sophisticated initialization methods. *Chen et al.* [1995,1997] assimilated FSU wind data using a nudging technique. In a later work, Chen et al. [1998] added to the previous method the assimilation of subsurface data. These data have been obtained from the assimilation of tide gauge data into the ocean component of the ZC model via a reduced-space Kalman filter [*Cane et al.*, 1996]. The assimilation of analyzed ocean fields made little difference in the prediction skill for the entire period starting from 1970. However, the prediction skills for the

[1] Presently at Moffat & Nichol Engineers, New York, New York, USA

period 1992-1997, where the original initialization [*Chen et al.*, 1995] with FSU winds was producing incorrect forecasts, were greatly improved. *Chen et al.* [2000] applied an internal statistical correction to the ZC model in order to reduce its systematic biases. The corrections were calculated using available data from the period 1972-1985, and then validated on the period 1986-1999. The ZC model with this internal correction (LDEO4) has an improved forecast skill and more realistic internal variability, particularly for the wind stress anomalies.

Other authors have applied simplified models for ENSO prediction as well. *Graham et al.* [1987a,1987b] applied a canonical correlation analysis for prediction of sea surface temperature from sea level pressure and winds in the tropical Pacific. Following this research path, other statistical models [*Barnston and Ropelewski*, 1992; *Latif et al.*, 1994; *Penland and Magorian*, 1993; *Penland and Sardeshmukh*, 1995; *Jiang et al.*, 1995; *Johnson et al.*, 2000] have shown useful prediction skill. Because the statistical models are constructed from the available data, and the records are not very long, the model's construction may be affected by artificial skill. This is a serious problem that all these studies have tried to overcome.

Following previous work by *Blumenthal* [1991] and *Xue et al.* [1994,1997a,1997b], we build a linear Markov autoregressive model from a multivariate set of data obtained from a free 100-year run of the latest version of the Lamont coupled ocean-atmosphere prediction model, LDEO4 [*Chen et al.*, 2000]. Because it is based on a free coupled model run, the linear model is expected to have no artificial skill when used for ENSO prediction. We will use this model to explore data assimilation techniques with the goal of finding a set of initial conditions that provides the best ENSO predictions.

The low-dimensional nature of our model allows us to determine via an inverse calculation a set of initial conditions which gives the best possible prediction for 6 months. We then treat this set as if it represented the true state of the system and compute directly all statistical parameters required by the Kalman filter. We then attempt to use this perfectly tuned procedure for estimation of the initial states for the model forecasts. We discover that the results do not outperform a direct initialization from the observations. We identify the violation of standard assumptions of Kalman filter as the reason for that and infer the necessity to use the state-dependent error models. We expect this problem to be quite general in applications of standard data assimilation schemes for initialization of imperfect models.

The paper is organized as follows. Section 2 describes the approach used for the linear model construction. A number of sensitivity tests are presented in section 3, where the main features and parameters of the linear model that lead to the best forecast skills are determined. Section 4 studies the forecast performance of the linear model for an initialization derived from observations and contrasts it with the performance of the "best" initial conditions that are found with the knowledge of the target forecast. Section 5 evaluates how close one can get to these best initial conditions using sequential assimilation techniques like the Kalman filter. Final discussion and conclusions are presented in section 6.

## 2. Linear Markov Model

*Chen et al.* [2000] have shown that adding an internal statistical bias correction to the Zebiak-Cane coupled ocean-atmosphere model can help hold it in states closer to observations. A detailed description of the Zebiak-Cane model and its variables has been given by *Zebiak and Cane* [1987]. The coupled model $\mathcal{L}$ can be defined via a state vector representation $\mathcal{T}$ with a transition from the month $t$ to the month $t+1$:

$$\mathcal{T}^{t+1} = \mathcal{L}(\mathcal{T}^t). \tag{1}$$

The coupled model state space is similar to that used by *Xue et al.* [1997a], which contains variables from both the ocean and atmospheric components of the model. In our case the state vector contains 13 different variables and has a dimension of $N \approx 2.4 \times 10^4$:

$$X = (a_k, h, u, h_{\mathrm{bndy}}, u_{\mathrm{bndy}}, \mathrm{ssta}_o, \mathrm{ssta}_m, \tau_x, \tau_y, U_a, V_a, d, q)\,. \tag{2}$$

The oceanic variables are represented by the amplitude of the equatorial Kelvin wave $a_k$, the Rossby component of the upper layer depth $h$, the Rossby mode zonal velocity $u$, and the boundary components $h_{\mathrm{bndy}}$ and $u_{\mathrm{bndy}}$ of the last two variables. The meridional component is not included in the state space because it is a diagnostic variable in this model [see *Cane and Patton*, 1984]. The sea surface temperature anomalies (SSTA) are represented by two variables: $\mathrm{ssta}_o$ is the dynamical model SSTA, and $\mathrm{ssta}_m$ is the bias-corrected SSTA that is used internally in LDEO4 to force the atmospheric model. The atmospheric component is represented by the two components of the surface wind anomaly $U_a$ and $V_a$, the wind convergence anomaly $d$, and the atmospheric heating anomaly $q$. Although the two components of the wind stress anomaly $\tau_x$ and $\tau_y$ are diagnostic variables, they are included in the state vector in order to make the connection of the model state with the observed wind stresses more straightforward.

The calculation of the linear seasonal Markov model follows *Blumenthal* [1991] and *Xue et al.* [1994]. Both studies use a monthly independent basis (MIB), that is, a single multivariate empirical orthogonal function (MEOF) basis calculated from the set of data encompassing all seasons. Output

fields from the 100-year run have been stored in the matrix $A$ which has dimensions $N \times t$, $N$ being the dimension of the state vector and $t$=1200 the number of monthly time steps in the data. Each variable has been normalized with respect to its total variance before the MEOF decomposition was performed. In the spirit of the work by *Xue et al.* [1997a], the normalized atmospheric model variables ($U_a$, $V_a$, $d$, and $q$) and the ocean boundary variables ($h_{\rm bndy}$ and $u_{\rm bndy}$) are down-weighted by a factor of 10, while the rest of the model variables are left unchanged. The basis is calculated from the matrix $A$:

$$A = EX^T. \tag{3}$$

Columns of $E$ are the MEOFs of the data, columns of $X$ are principal components (PCs), and superscript $T$ denotes matrix transposition. Columns of $E$ are orthogonal, and columns of $X$ are orthonormal in this notation. For further development we truncate expansion (3) to the first $n$ terms. *Blumenthal* [1991] calculated seasonal transitions and found that the seasonal model fits the ZC model better than a non-seasonal one. As in the work by *Xue et al.* [1994], we separate the PCs into monthly blocks, in order to calculate monthly transition matrices $L^i$ for each of the 12 calendar months:

$$x_{i+1} = L^i x_i + \xi_i. \tag{4}$$

Here $x_i$ is a subset of rows in $X$ which correspond to the month $i$, and $\xi_i$ is the error in the model fit. *Xue et al.* [1994] calculated the linear Markov model from

$$L^i = \langle x_{i+1} x_i^T \rangle \langle x_i x_i^T \rangle^{-1}, \tag{5}$$

where angle brackets denote time averaging.

*Blumenthal* [1991] used a small diagonal taper in the calculation of the linear Markov model to avoid the influence of numerical singularity in the autocovariance matrix (the denominator of equation (5)). He interpreted the taper $\Delta$ as the covariance matrix of the uncertainty in the EOFs. When tapering is applied, equation (5) becomes

$$L^i = \langle x_{i+1} x_i^T \rangle \langle x_i x_i^T + \Delta \rangle^{-1}. \tag{6}$$

*Blumenthal* [1991] showed that the taper reduces the growth of the error in the initial conditions by filtering low-energy modes. The use of the taper results in a more diagonally dominant transition matrix but increases the rate of decay in the linear model. *Xue et al.* [1994] constructed the linear model without the use of the taper. They found that when the linear model is initialized with the projections of the observed SSTAs on the MEOFs, the use of projections on the modes with numbers higher than 1 results in degraded predictions. (The first MEOF is a characteristic ENSO pattern which can be initialized from the observed data in an unambiguous way; the patterns with higher numbers receive erratic initialization because of a large portion of observed SSTA variability which is not being represented by the model.) We obtained a similar result when no taper was applied in the calculation of the transition matrices. In that case the low-energy modes show very fast error growth.

## 3. Sensitivity Tests and Parameters for the Markov Model

We first define the projection of observational data onto the reduced space spanned by the basis $E$. These low-dimensional representations of the observed system states will be used as the baseline initial conditions and as the verification data set for the forecast experiments with different versions of the Markov model. In order to choose a model with the best forecast skill, we subject different model versions to the following tests: Using observational projections as initial conditions, we compare predictions of NINO3 (SSTA area average for ($5^o$S–$5^o$N, $90^o$-$150^o$W)) of up to a year ahead with the observed values. Comparison results are presented in terms of correlation coefficients and RMS differences separately for initial conditions from 1972-1985, which was the LDEO4 training period, and from 1986-1999, the LDEO4 validation period [*Chen et al.*, 2000].

We introduce a vector $\mathcal{T}^o$ which contains the following data: (1) zonal and meridional components of the wind stress from Florida State University (FSU) [*Goldenberg and O'Brien*, 1981]; (2) SSTA from the Climate Prediction Center of the National Centers for Environmental Prediction (NCEP); and (3) analyzed ocean model fields obtained from the assimilation of tide gauges into the Cane-Patton model [*Cane and Patton*, 1984] using a reduced space Kalman filter [*Cane et al.*, 1996]. These fields are interpolated to the same grid as their model counterparts. The vector of observations can be connected to the full state vector using a sampling matrix $H$ (a submatrix of the identity matrix which includes only rows corresponding to the variables which are observed):

$$\mathcal{T}^{\rm o} = H\mathcal{T}. \tag{7}$$

The estimate of the projection of $\mathcal{T}^o$ onto the reduced space will follow the projection method described by *Kaplan et al.* [1997], with an observational error covariance matrix equal to the identity as in the work by *Smith et al.* [1996]. The projection then can be written as

$$x_{\rm obs} = (E^T H^T H E)^{-1} E^T H^T \mathcal{T}^o. \tag{8}$$

We have analyzed the utility of defining the basis $E$ as monthly dependent or independent. For the monthly-dependent case (monthly-dependent basis, MDB) a different basis

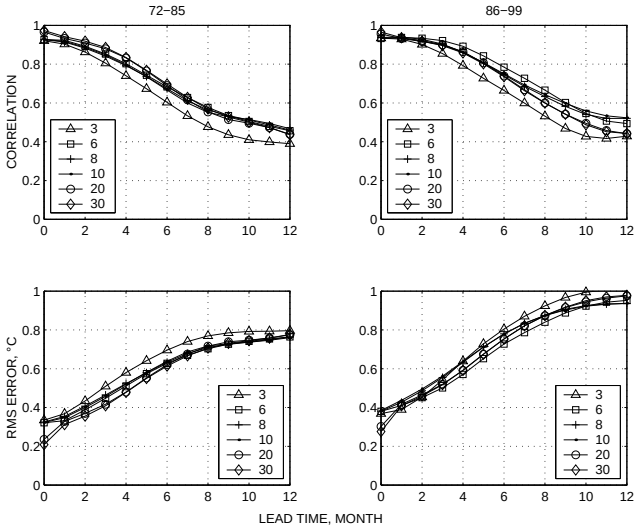


**Figure 1.** Correlations and RMS error for forecasted NINO3 index for the 1972-1985 and 1986-1999 periods. Different symbols represent forecast skill for linear Markov models of different dimensions.

**Table 1.** Total Normalized Variance of Each Model Variable and its Percentage Accounted for by six Multivariate Empirical Orthogonal Functions

| Variable | Total Normalized Variance | % of Variance |
|---|---|---|
| $a_k$ | 1.00 | 97 |
| $h$ | 1.00 | 87 |
| $u$ | 1.00 | 76 |
| $h_{\mathrm{bndy}}$ | 1.00 | 87 |
| $u_{\mathrm{bndy}}$ | 1.00 | 73 |
| $\mathrm{ssta}_o$ | 1.00 | 93 |
| $\mathrm{ssta}_m$ | 1.00 | 93 |
| $\tau_x$ | 1.00 | 81 |
| $\tau_y$ | 1.00 | 84 |
| $U_a$ | 0.01 | 88 |
| $V_a$ | 0.01 | 94 |
| $d$ | 0.01 | 94 |
| $q$ | 0.01 | 92 |

is calculated for every calendar month. In this case the autocovariance in (6) is diagonal, and so is its inverse. The transition matrix is more diagonally dominant, but if the range between its largest and smallest eigenvalues is too large it has the same problems as the MIB when a taper is not used. Since the MDB-based models did not show better skill than those with the MIB, for convenience and simplicity we chose the MIB-based models for further use.

We experiment with different values of the diagonal elements of the tapering matrix $\Delta$ in (6). Sensitivity tests were carried out, and the best forecasting performance was obtained for values below 1% of the maximum eigenvalue of the autocovariance matrix. This value was smaller than the leading 11 eigenvalues for all seasonal autocovariances.

For statistical models constructed from observations, the forecast skills increase with the number of leading MEOFs included in the model state, if the model is initialized using the same data that were used for its construction. Because we construct our statistical model from a free coupled model run and evaluate its forecasting skill from observations, this behavior is not expected. Moreover, the taper acts like a truncation of the singular values of the transition matrices, suppressing growth of the modes that do not show a strong presence in the model run. Figure 1 shows the prediction skill for linear Markov models of different dimensions, when the projection of the observations (8) is used for the initialization. The model dimension at which the forecast skill becomes invariant depends on the value of the taper. If $\Delta = 0$, the forecast skill gets worse without showing any convergence as the dimension increases. For the value of the taper we use, the forecast skill is improved up to the model dimension of 5 and 6; then it worsens slightly and becomes almost unchanging for dimensions higher than 10 or 12.

On the basis of this analysis we choose for further use the linear Markov model of dimension 6 with an MIB and a taper below 1% of the largest eigenvalue of the sample covariance. We refer to this model as the LDEO4-MK6.

Table 1 shows how much of the variance for each independent variable is accounted for by six MEOFs. The percentage of the variance that is accounted for is less than 80% only for the variables associated with the Rossby mode zonal velocity.

## 4. Applications of the Linear Markov Model

### 4.1. Forecast Performance

To study the forecast performance of the LDEO4-MK6 model we use two different sets of initial conditions: (1) the LDEO4 standard initial conditions as in the work by *Chen et al.* [2000], which uses the full insertion of FSU wind stresses into the model together with the nudging of the ocean fields analyzed by the method of *Cane et al.* [1996]; and (2) projection of observations $x_{\mathrm{obs}}$ defined by (8). Figure 2 compares the performance of the full LDEO4 model with that of the LDEO4-MK6 initialized with the projection of the standard LDEO4 initial conditions and with the projection

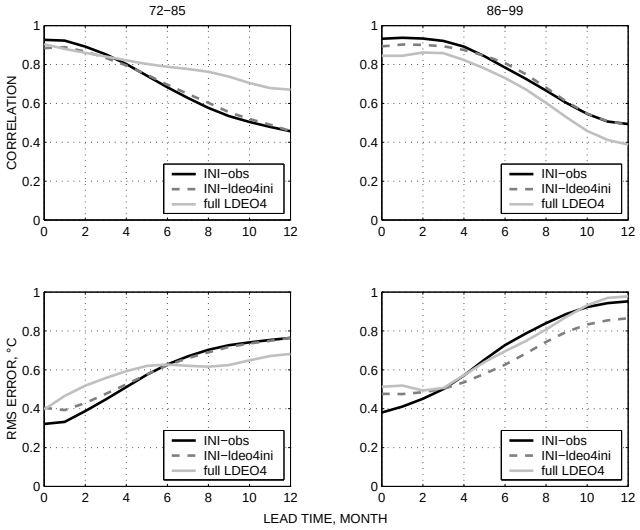


**Figure 2.** Same as Figure 1, but for two LDEO4-MK6 forecasts (initialized by observations (INI-obs) and initialized by the LDEO4 initial conditions(INI-ldeo4ini)) and the standard LDEO4 forecast (full LDEO4).

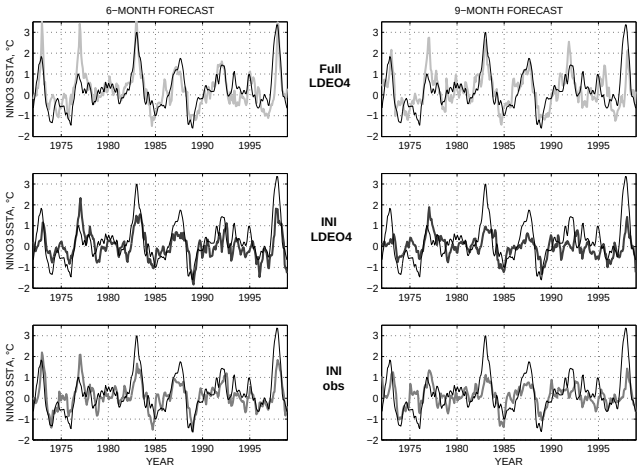


**Figure 3.** Time series of observed (thin lines) and predicted (thick lines) NINO3 for lead times of 6 (left) and 9 (right) months. Shown are predictions for the full LDEO4 and for the LDEO4-MK6 initialized with the standard LDEO4 initial conditions (INI LDEO4) or with observations (INI obs).

of observations. As could be expected, the full model correlations are higher than those for the linear model for the full model training period (1972-1985), though the RMS error is smaller only after 6 months. For the validation period (1986-1999) the linear model forecasts are better than those of the full model. For both periods the differences are not very large. Note that the full LDEO4 model predicts the amplitude of the 1997-1998 event better than the LDEO4-MK6 model, although it overestimates small events (Figure 3).

To assess the linear model's decay, we carried out the following test. For the period 1972-1999, we compute the mean and standard deviation of the six MEOF amplitudes representing the projection of the observations and compare them with the same parameters for each forecast lead time (our analysis sample has a nonzero mean as it comes from a period different from the climatological one). To provide a standard of comparison, we analyzed the time evolution by the LDEO4-MK6 of an ensemble of a normally distributed random field with the same initial mean and standard deviation as our sample of observational projections. The first MEOF amplitude decays substantially, losing about half of its standard deviation in 12 months, while the other five amplitudes experience only small decay. Only the first two mean amplitudes show a significant change, and it is small compared to the values of their respective standard deviations. The decay of the first mode affects significantly the prediction of the NINO3 index, because it represents a dominant contribution to this index. On the basis of the decay found in this experiment, it is possible to correct the first element of the state vector by multiplying it by a constant factor after each propagation (the corrected model is not the best fit to the free coupled model run as the LDEO4-MK6 is). The use of this correction improves the prediction of large events (especially 1997-1998) without generating any unreal events. The amplitudes of the corrected linear model predictions are still smaller than those from the full LDEO4 model, but the unreal large events predicted by the full LDEO4 model in 1977-1978 and 1991-1992 are not present in either of the linear model predictions (Figure 4). Overall, neither the correlation coefficient nor the RMS error changes significantly with the correction, and we proceed to use the uncorrected version of the LDEO4-MK6 for the rest of our study.

### 4.2. Inverse Solution for the "Best" Initial Conditions

As Figures 2-4 demonstrate, a linear Markov model initialized with either standard LDEO4 or observed initial conditions can produce a reasonably good forecast up to lead times of 6 months (correlation coefficient 0.8, and $0.7^oC$ RMS). After 6 months the prediction skill degrades quickly (correlation coefficient of 0.6 and a RMS error of about 0.85 after 9 months). Is it possible to find a set of initial conditions which yields a good forecast for longer lead times? To answer this question, we define a set of initial conditions of the model trajectory best fitted to the observations for a given length of time in a least squares sense.

The set of initial conditions we are trying to estimate will

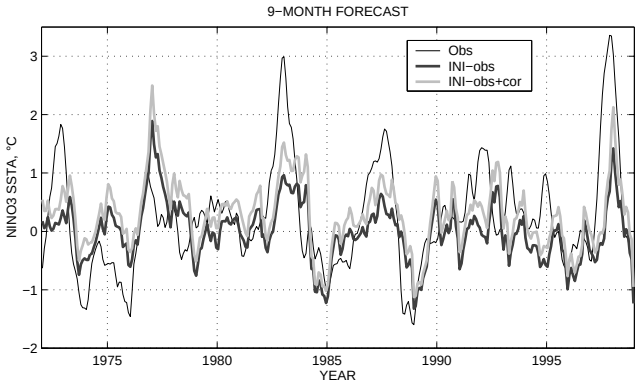


**Figure 4.** Time series of observed and predicted NINO3 for a 9-month lead. Shown are predictions for the LDEO4-MK6 model initialized with the projection of observations (INI-obs) and the same with an internal correction for the decay of the first component (INI-obs+cor).

be obtained by minimizing the functional

$$\mathcal{J}[x] = \sum_{i=0}^{m} (x^i_{\rm obs} - L^{*,i}x)^T (x^i_{\rm obs} - L^{*,i}x) \qquad (9)$$

with respect to $x$, where $L^{*,i}$ is the transition matrix from month 1 to month $i+1$ calculated recursively:

$$L^{*,0} = I\,; \qquad L^{*,i} = L^i L^{*,i-1}, \quad i = 1, 2, \ldots. \qquad (10)$$

Here $m$ is the number of months in the segment for which the trajectory is fitted to the observations. The initial conditions $x_{\rm in}$ obtained as a minimizer of $\mathcal{J}$ (we will call them the "inverse solution") show better prediction skills than any of the other two sets presented above, and in the sense of the cost function (9) its skill is as good as it gets. At lead times up to $m$ the correlation of the forecasts with the observations is very close to that of the initial conditions. The RMS error is slightly larger at month $m$ than at month 0 due to the decay of the linear model and other model imperfections (Figure 5). Note that this method is not a forecasting procedure since in order to issue a forecast starting at month $t$ observations up to month $t+m$ are needed. However, it gives us an ideal initial state to reach for.

Time series of six initial conditions from the 6-month trajectory best fit ($m = 6$) and from the observations are shown in Figure 6. The differences among the components of $x_{\rm in}$ and $x_{\rm obs}$ are small, and the correlation between these two states is very high except for the third and fourth components. The study of the 6-month propagator $L^{*,6}$ shows that the 6-month prediction of the first component (the one that mainly represents the NINO3 index) is influenced very

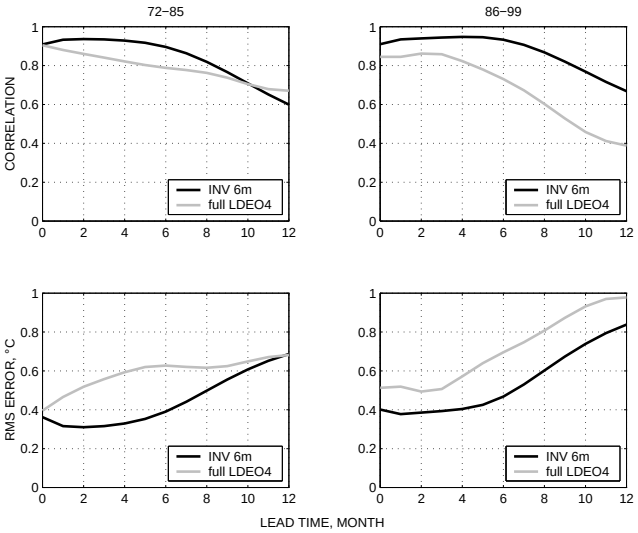


**Figure 5.** Same as Figure 1, but for the LDEO4-MK6 model initialized with the inverse solution for a 6-month optimization interval (INV 6m) and the standard LDEO4 model forecast (full LDEO4).

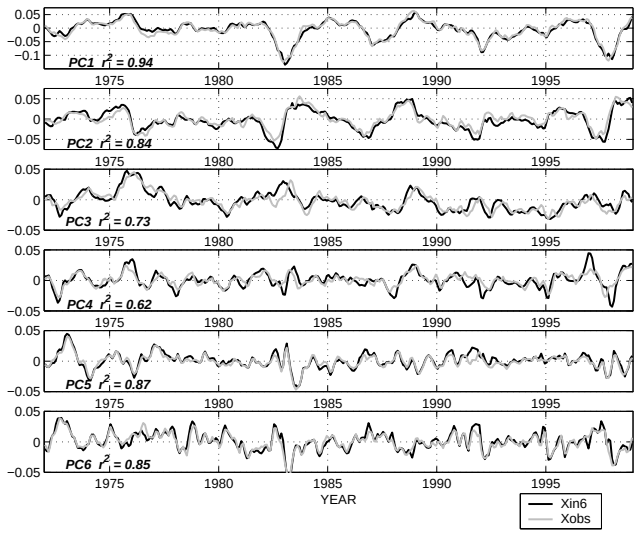


**Figure 6.** Time series of all six components of the inverse solution (Xin6) and the projection of observations (Xobs). Squared correlation coefficients are shown in the lower left corner of the panels.

strongly by the third or fourth component in the initial conditions depending on whether the starting month is in the summer-fall or in the winter-spring. The first and second components contribute significantly for all starting months.

## 5. Can the Optimal Solution Be Obtained From the Sequential Assimilation?

The inverse solution for the initial conditions that we calculated in the previous section makes use of the observations up to lead time $m$, and therefore cannot be considered a forecast for time $m$. Our goal here is to use a sequential assimilation technique in an attempt to find an initial state for the actual forecast which would be close to the inverse one. That is, the initial conditions will be estimated from observations up to and including the initialization time, but not from the future observations. We will use the inverse solution $x^t_{\rm in}$ for starting month $t$ as if it were a true state of the system within the period of available data.

The observed vector $x_{\rm obs}$ is related to the true state by the measurement equation:

$$x^t_{\rm obs} = x^t_{\rm in} + \varepsilon^t_{\rm obs}, \tag{11}$$

where $x^t_{\rm obs}$ is the projection of the observations to the reduced space determined by (8), $\varepsilon^t_{\rm obs}$ represents the observational error, or the difference between the "true" (6-month inverse solution $x_{\rm in}$) and observed states, and $t$ here is the time index. We write the propagation equation as

$$x^t_{\rm in} = L^i x^{t-1}_{\rm in} + \varepsilon^t_{\rm mod}, \tag{12}$$

where superscript $i$ indicates the season corresponding to the time $t-1$ and $\varepsilon_{\rm mod}$ is the model error in a single-month transition, often called "the system noise". Note that in this equation we use the same LDEO4-MK6 model operator $L^i$ as before, even though the inverse "true" states are in fact governed by more complicated (not AR(1)) dynamics.

Since for the time interval 1972-1999 we know $x^t_{\rm obs}$ and $x^t_{\rm in}$, we can use equations (11) and (12) to compute actual realizations of error sequences $\varepsilon^t_{\rm obs}$ and $\varepsilon^t_{\rm mod}$. We then use these sequences to compute error covariances (averaging here is done separately for each calendar month, due to the model seasonality):

$$\begin{aligned} Q^i &= \langle \varepsilon^i_{\rm mod} \varepsilon^{i\,T}_{\rm mod} \rangle, \\ R^i &= \langle \varepsilon^i_{\rm obs}\ \varepsilon^{i\,T}_{\rm obs} \rangle, \quad i = 1, 2, \ldots, 12. \end{aligned} \tag{13}$$

At this stage we have all the necessary elements for building a sequential assimilation scheme based on the Kalman filter in order to estimate the true state of the system. At every time step, the system error covariance and the Kalman gain matrix can be calculated using:

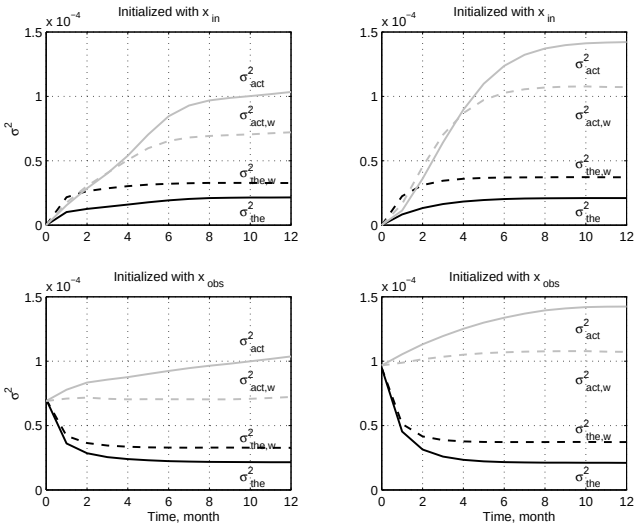


**Figure 7.** Evolution of the theoretical $\sigma^2_{\rm the}$ (dark) and actual $\sigma^2_{\rm act}$ (light) Kalman filter error variance for the amplitudes of the (left) first and (right) second modes. The nonweighted cases are shown with solid lines, and the weighted cases are shown with dashed lines. Top plots are for initialization with $x_{\rm in}$, and bottom plots are for $x_{\rm obs}$.

$$P^t_f = L^i P^{t-1}_a L^{i\,T} + Q^i, \tag{14}$$

$$K^t = P^t_f (P^t_f + R^i)^{-1}, \tag{15}$$

$$P^t_a = (I - K^t) P^t_f, \tag{16}$$

where subscripts $f$ and $a$ denote Kalman filter forecast and analysis respectively, and $K$ is the Kalman gain matrix (see *Cane et al.* [1996] and *Kaplan et al.* [1997] for the formalism; in the present case the observational map $H = I$). Index $i$ here denotes the season corresponding to the month $t-1$. The Kalman filter corrections can be written as

$$x^t_a = x^t_f + K^t (x^t_{\rm obs} - x^t_f). \tag{17}$$

We perform 12-month-long Kalman filter assimilation runs starting from every month in 1972-1999, and we experiment with two initialization schemes for the assimilation runs: (1) we start from the "truth" $x_{\rm in}$ and assume initial error $P = 0$, or (2) we start from the observational projection $x_{\rm obs}$ and assume the initial error $P = R$. Since the values of $x_{\rm in}$ are known at all months, we can compare the actual error variance in a Kalman filter analysis with its theoretical estimate ($\mathrm{diag}[P_a]$). Figure 7 presents the 12-month evolution of these values for the amplitudes of the first two components (the comparison is presented as a function of time elapsed since the beginning of a Kalman filter run, in

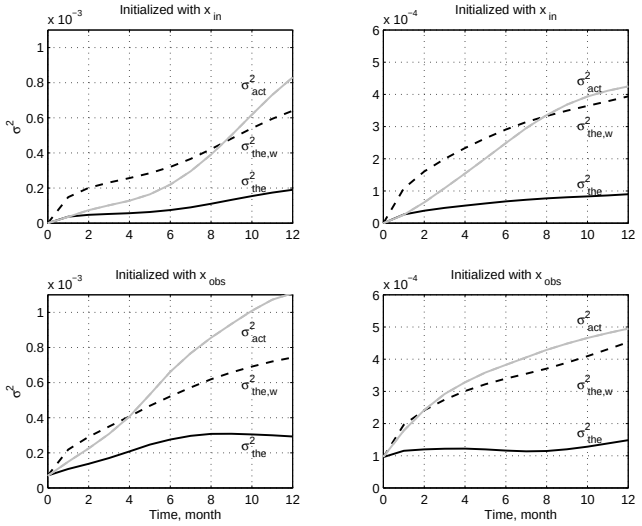


**Figure 8.** Evolution of the theoretical $\sigma^2_{\rm the}$ (dark) and actual $\sigma^2_{\rm act}$ (light) error variance of model forecasts with no data assimilated. Shown are the amplitudes of the (left) first and (right) second modes. The nonweighted cases are shown with solid lines, and the weighted cases are shown with dashed lines. Top plots are for initialization with $x_{\rm in}$, and lower plots are for $x_{\rm obs}$.

averages over the entire sample of runs). The actual error $\sigma^2_{\rm act}$ grows much faster and toward a larger value than the estimated error $\sigma^2_{\rm the}$; that is, the filter significantly underestimated the system error. In the case when the filter is initialized with the observed data, the theoretical error decreases with time, while the actual error increases. However, the end-of-period values for actual and theoretical error do not depend on the initialization. The same behavior is observed for amplitudes of other components.

This discrepancy between theoretical and actual error growth can be partially traced to the nonwhiteness of $\varepsilon_{\rm mod}$ in time. Figure 8 shows error evolution with no data assimilation (a forecast). The actual error $\sigma^2_{\rm act}$ grows much faster than the theoretically estimated one $\sigma^2_{\rm the}$, for both initialization cases. Equation (14) correctly describes the error evolution with $Q$ defined by (14) only if $\varepsilon_{\rm mod}$ is uncorrelated for different times. In fact, this is not the case, as is demonstrated by Figure 9, in which the temporal autocorrelations of $\varepsilon^i_{\rm mod}$ and $\varepsilon^i_{\rm obs}$ ($i = 1, 2, 3$) are shown. Note the precise match between the actual and theoretical error for 1-month lag in the top two panels of Figure 8: It is guaranteed by (14). In order to obtain a better fit to the error in the nonassimilation run, we attempt to increase the system noise matrix $Q$. We multiply $Q$ from both sides by the diagonal weighting matrix $w = \mathrm{diag}[4, 4, 2.5, 2, 1.5, 2]^{1/2}$ chosen to force the

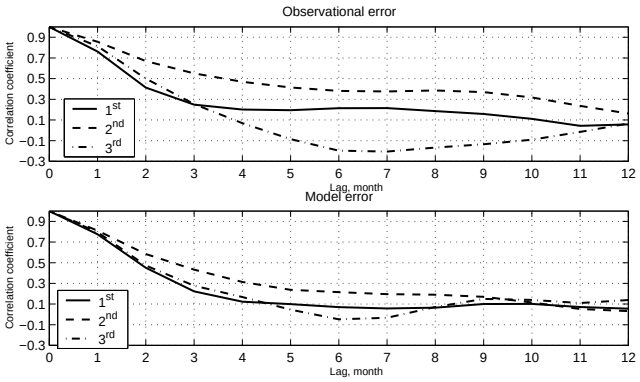


**Figure 9.** Temporal autocorrelation of the first three components of the (top) observational $\varepsilon_{\rm obs}$ and (bottom) model $\varepsilon_{\rm mod}$ errors.

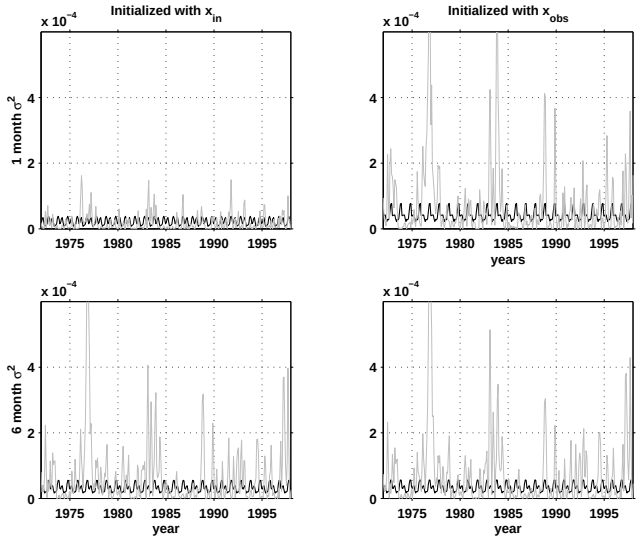


**Figure 10.** Time series of the theoretical $\sigma^2_{\rm the}$ (dark) and actual $\sigma^2_{\rm act}$ (light) error variance for the amplitude of the first mode, after (top) 1 and (bottom) 6 months of assimilation. Left plots correspond to the case initialized with $x_{\rm in}$, and right plots correspond to $x_{\rm obs}$.

theoretical values to approximate the actual ones.

This increase in Q results in a better, though still far from perfect agreement between the theoretical and actual errors (Figure 8). Use of increased Q in assimilation results in a reduction of the actual error and an increase in its theoretical estimate, although the gap between the two is still quite large (Figure 7). Figure 10 shows the variation in time of the actual and estimated error variances of the amplitude of the first component after 1 and 6 months of assimilation for the two initialization cases with increased matrix $Q$. While the

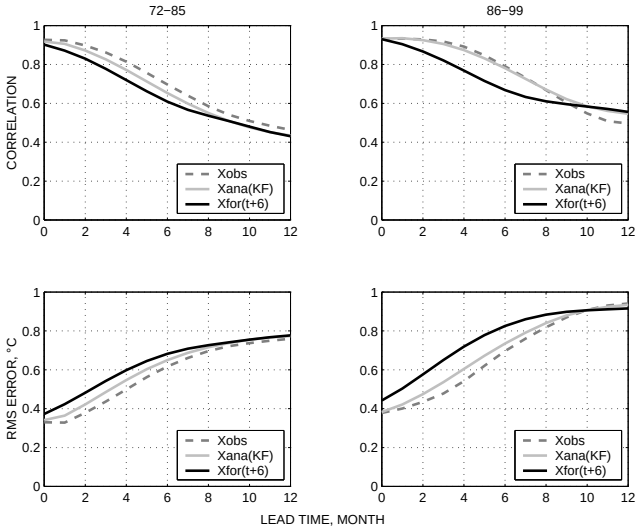


**Figure 11.** Same as Figure 1, but for the LDEO4-MK6 model forecast initialized with three different sets of initial conditions: projection of observations (Xobs), 6-month forecast from the inverse solution (Xfor(t+6)), and the Kalman filter analysis (Xkf).

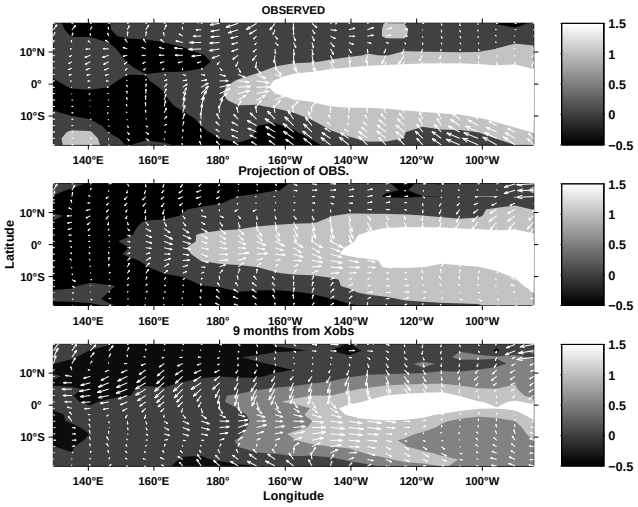


**Figure 12.** Sea surface temperature and wind stress anomaly fields in December 1997 from observations, the part of the observations that is projected onto the 6-dimensional LDEO4-MK6 model space, and the 9-month LDEO4-MK6 forecast initialized with the projection of observations.

theoretically estimated error exhibits a stable annual cycle, the actual error has periods of very large and very small values. Although errors at 1 month are very different for the two initializations, they are quite similar after 6 months, indicating that in both cases the Kalman filter provides almost the same solution at 6 months (it "forgets"an initial condition). Figure 11 compares the forecast skill of a model initialized by the end state of a 6-month KF run, $x_a^{t+6}$, with those which utilize the information less completely: the projection of observations at time $t+6$, $x_{\rm obs}^{t+6}$ (corresponding to a Kalman gain $K = I$), and the 6-month model forecast from $x_{\rm in}^t$ (corresponding to a Kalman gain $K = 0$ if the filter is initialized by $x_{\rm in}^t$). The prediction from the Kalman filter state lies in between the other two. Its skill is marginally worse than the initialization from observations. It clearly does not improve on observations alone.

Figure 12 shows for December 1997 (the peak of the El Niño of 1997-1998) the SSTA and wind stress anomalies from observations, the part of the observations that is projected onto the the reduced space, and the 9-month forecast initialized by observations. The linear model initialized by observations can predict the peak of the event reasonably well at least 9 months in advance, although the amplitude of the event is significantly underestimated (cf. Figure 3).

## 6. Discussion and Conclusions

A linear Markov model has been constructed from the output of a 100-year run of the Zebiak-Cane coupled ocean-atmosphere model with an internal bias correction (LDEO4) [*Chen et al.*, 2000]. The model variability can be successfully represented by a small number of MEOFs. It has been found that the reduced state of dimension 6 and the linear model created on the associated amplitudes present the best prediction skills when initialized with observations (or the standard LDEO4 initial conditions).

The reduced space approach allows us to find an initial state which best fits a model trajectory segment to the observations. By construction, such "inverse" solutions show excellent prediction skills. Of course, since one needs to know the observations from the future to calculate the inverse solution, this method cannot be used for true predictions. It does provide us with a set of "perfect" (in the sense of predictions with this model) initial conditions useful for predictability and initialization studies.

A sequential data assimilation method (Kalman filter) has been constructed in an attempt to calculate a model state which approximates the inverse solution using only past data. Although all the errors are known a priori (for the 1972-1999 period) the solution provided by the Kalman filter is notably inferior to the optimal one. In fact, it does not even outperform the one initialized purely from the latest observational data, that is, the one which uses no dynamical

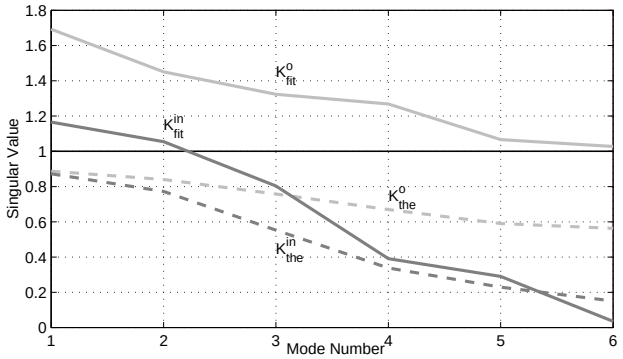


**Figure 13.** Singular values of the theoretical Kalman gain $K_{\rm the}$ (dashed), and the best fit Kalman gain $K_{\rm fit}$ (solid). Superscripts "in" and "o" denote the cases initialized with $x_{\rm in}$ and $x_{\rm obs}$, respectively.

information from the model.

We identified the nonwhiteness of the model error as a problem. This problem can be only partially corrected by the simple inflating of the system noise covariance $Q$ which we use in this work.

Here we attempt to find out if we are at all able to set up a sequential data assimilation system which linearly combines model forecasts with observations for results superior to both sources. Since we want $x_a^t$ to be the same as $x_{\rm in}^t$, we obtain from (17) that an "ideal" Kalman gain should satisfy

$$-\varepsilon_f^t = K^t(\varepsilon_{\rm obs}^t - \varepsilon_f^t), \tag{18}$$

where $\varepsilon_f^t = x_f^t - x_{\rm in}^t$ is the forecast error. Because the actual time series of the errors are available, we can determine $K$ (assuming it time-independent) by a linear best fit procedure:

$$K_{\rm fit} = \langle \varepsilon_f^t(\varepsilon_f^t - \varepsilon_{\rm obs}^t)^T\rangle\langle(\varepsilon_f^t - \varepsilon_{\rm obs}^t)(\varepsilon_f^t - \varepsilon_{\rm obs}^t)^T\rangle^{-1} \tag{19}$$

(angle brackets again denote averaging over time $t$). If forecast and observational errors are uncorrelated ($\langle\varepsilon_f, \varepsilon_{\rm obs}\rangle = 0$), then expression (19) becomes a familiar formula for Kalman gain:

$$K_{\rm the} = \langle\varepsilon_f\varepsilon_f^T\rangle(\langle\varepsilon_f\varepsilon_f^T\rangle + \langle\varepsilon_{\rm obs}\varepsilon_{\rm obs}^T\rangle)^{-1} = P(P+R)^{-1}. \tag{20}$$

Note that the theoretical Kalman gain has all its singular values between 0 and 1.

We analyze Kalman filter corrections for two types of 1-month predictions: from the inverse solution ("truth") and from observations. Figure 13 shows all six singular values of matrices $K_{\rm the}$ and $K_{\rm fit}$ obtained from these two initializations. While singular values of $K_{\rm fit}$ do not exceed 1 by much for the forecast from the inverse solution, the singular values of $K_{\rm fit}$ are considerably larger than 1 when the forecast is done from observations. Since $K_{\rm fit}$ has singular values larger than 1 it cannot be produced theoretically with any settings of the Kalman filter for our model of the dynamical system: The closest possible approximation would be $K = I$, which corresponds to replacing the forecast with observations at every time step. We tried to use the best fit Kalman gain to correct the initial conditions for predictions. The forecast skill obtained is not a significant improvement over that from the observations (not shown).

The reason for the large difference between $K_{\rm fit}$ and $K_{\rm the}$ is a correlation between $\varepsilon_{\rm obs}$ and $\varepsilon_f$. It is caused by the correlation between $\varepsilon_{\rm obs}$ and $\varepsilon_{\rm mod}$ (sample correlations are positive for all components: 0.68, 0.72, 0.66, 0.68, 0.38, and 0.56). This correlation (as well as a serial correlation in these errors, see Figure 9) occurs because the system is getting in and out of regimes, and the autoregressive model can not capture this behavior correctly (cf. Figure 10). While the residuals of the original fit of the LDEO4-MK6 to the free 100-year run of the coupled model show some autocorrelation, they do not reach the values and temporal extent of the correlations in Figure 9. However, when our definition of "truth" uses the 6-month predictive performance of this imperfect model as a term of reference, the imperfections of the model get translated into a common contribution toward errors of both types ($\varepsilon_{\rm obs}$ and $\varepsilon_{\rm mod}$) and cause their cross-correlation and autocorrelation.

The same problem affects the $\chi^2$ test for innovations in our Kalman filter runs. Figure 14 presents the statistics

$$J(t) = \langle(x_{\rm obs}^t - x_f^t)^T(P_f^t + R)^{-1}(x_{\rm obs}^t - x_f^t)\rangle \tag{21}$$

as a function of time since the beginning of the Kalman filter run, averaged over the entire set of 12-month-long assimilation runs. Under standard Kalman filter assumptions, $J$ is an average of individual $\chi^2$ variables with six degrees of freedom each, so its expected value is equal to 6. Sample elements that are averaged in the right-hand side of equation (21) display the 1-month lagged autocorrelation of 0.7. An effective sample size used for computing quantiles of the theoretical distribution for $J$ was reduced accordingly. The values of $J$ closest to its expected mean are produced by the runs where $Q$ was not scaled up, particularly for their months 7-12. However, even these values fall between the lower 10% and 1% quantiles of the theoretical distribution. Empirical values of $J$ are significantly smaller than their theoretical expectation, because the correlation between observational and model error makes innovations $x_{\rm obs} - x_f$ too small compared to their theoretical covariance $P_f + R$. This discrepancy increases when $Q$ is scaled up in order to compensate for the serial correlation in the model error, because

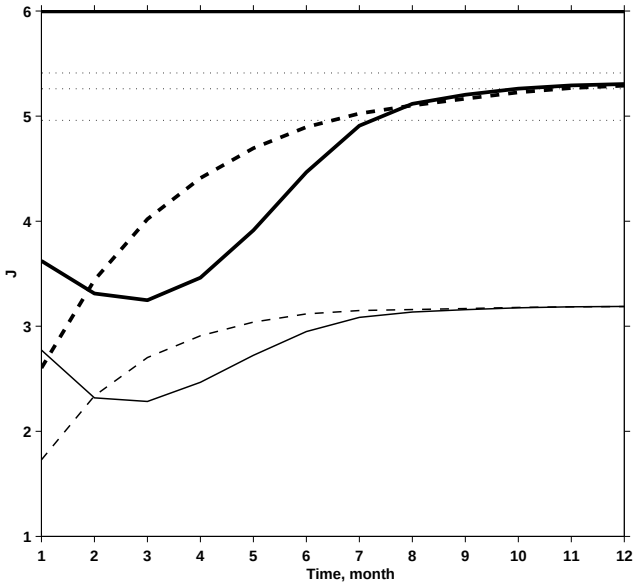


**Figure 14.** Averaged $\chi^2$ statistics for the Kalman filter innovation sequences, as a function of time since the beginning of the Kalman filter run. Shown are averages for the entire sample of runs for the cases of the nonweighted (thick lines) and weighted (thin lines) system noise covariance matrix $Q$. Initializations with $x_{\rm in}$ and $x_{\rm obs}$ are shown as dashed and solid lines, respectively. The top boundary of the plot is the expected theoretical mean for $J$ (6.00). Dotted horizontal lines show lower 10%, 5%, and 1% quantiles of the theoretical distribution (5.41, 5.26, and 4.96, respectively).

increase in $Q$ results in larger covariances $P_f$ and smaller innovations.

Our attempts to account for serial correlations in errors via state space extension [e.g., *Gelb*, 1974; *Cañizares*, 1999] have not resulted in appreciable improvements for the prediction [*Cañizares et al.*, 2000]. Space extension amounts to the use of higher-order autoregressive models for the errors, while in the current application the errors seem to require state-dependent modeling.

We demonstrated that the actual model and observational errors deviate from textbook error assumptions in the Kalman filter (white noise process, uncorrelated errors, etc.) to such an extent that the Kalman filter does not outperform pure observational projection. When the Kalman filter assumption of white noise processes for the system noise and the observational error are violated, errors must be defined with more faithful models in order to enable sequential data assimilation algorithms to find initial states for better prediction.

**Acknowledgments.** This work was supported by the NOAA through grants NA86GP0515 and NA06GP0414 and by the NASA through the IDS grant NAG5-4058 and its renewal. We are very grateful to Benno Blumenthal for his comments and suggestions and to two anonymous reviewers for their remarks and constructive criticism. Discussions with Richard Kleeman were very useful for this work. This is LDEO contribution 6240.

## References


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

Blumenthal, M. B., Predictability of a coupled ocean-atmosphere model, *J. Clim., 4*, 766-784, 1991.

Cane, M. A., and R. J. Patton, A numerical model for low-frequency equatorial dynamics, *J. Phys. Oceanogr., 14*, 1853-1863, 1984.

Cane, M. A., S. E. Zebiak, and S. C. Dolan, Experimental forecast of El Niño, *Nature*, *321*, 827-832, 1986.

Cane, M. A., A. Kaplan, R. N. Miller, B. Tang, E. C. Hackert, and A. J. Busalacchi, Mapping tropical Pacific sea level: Data assimilation via a reduced state space Kalman filter, *J. Geophys. Res., 101*, 22,599-22,617, 1996.

Cañizares, R., On the application of data assimilation in regional coastal models, 133 pp., A.A. Balkema, Brookfield, Vt., 1999.

Cañizares, R., A. Kaplan, and M.A. Cane, Initialization of a model for El Niño prediction via data assimilation: Failure of the "textbook" approach. *Eos Trans. AGU, 81*(48), Fall Meet. Suppl., abstract NG71A-29, 2000.

Chen, D., S.E. Zebiak, A. J. Busalacchi, and M. A. Cane, An improved procedure for El Niño forecasting: Implications for predictability, *Science, 269*, 1699-1702, 1995.

Chen, D., S.E. Zebiak, M. A. Cane, and A. J. Busalacchi, Initialization and predictability of a coupled ENSO forecast model, *Mon. Weather Rev., 125*, 773-788, 1997.

Chen, D., M. A. Cane, S.E. Zebiak, and A. Kaplan, The impact of sea level data assimilation on the Lamont model prediction of the 1997/98 El Niño, *Geophys. Res. Lett., 25*, 2837-2840, 1998.

Chen, D., M. A. Cane, S.E. Zebiak, R. Cañizares, and A. Kaplan, Bias correction of an ocean-atmosphere coupled model, *Geophys. Res. Lett., , 27*, 2585–2588, 2000.

Gelb, A. (Ed.), *Applied Optimal Estimation.*, 374 pp., MIT Press, Cambridge, Mass., 1974.

Goldenberg, S. B., and J. J. O'Brien, Time and space variability of tropical Pacific wind stress, *Mon. Weather Rev., 109*, 1190-1205, 1981.

Graham, N. E., J. Michaelsen, and T. P. Barnett, An investigation of the El Niño-Southern Oscillation cycle with statistical models, 1, Predictor field characteristics, *J. Geophys. Res., 92*, 14,251-14,270, 1987a.

Graham, N. E., J. Michaelsen, and T. P. Barnett, An investigation of the El Niño-Southern Oscillation cycle with statistical models, 2, Model results, *J. Geophys. Res., 92*, 14,271-14,289, 1987b.

Jiang, N., M. Ghil, and D. Neelin, Forecasts of equatorial Pacific SST anomalies by using an autoregressive process and singular spectrum analysis, *Exp. Long-Lead Forecast Bull., 4*, pp. 24-27, 35-36, Natl. Cent. for Environ. Predict., Natl. Oceanic and Atmos. Admin., U.S. Dep. of Commer., Washington, D.C., 1995.

Johnson, S. D., D. S. Battisti, and E. S. Sarachik, Empirically derived Markov models and prediction of tropical Pacific sea surface temperature anomalies, *J. Clim., 13*, 3-17, 2000.

Kaplan, A., Y. Kushnir, M. A. Cane, and M. B. Blumenthal, Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures, *J. Geophys. Res., 102*, 27,835-27,860, 1997.

Latif, M., M. A. Cane, M. Flugel, N. E. Graham, H. von Storch, J.-S. Xu, and S.E. Zebiak, A review of ENSO prediction studies, *Clim. Dyn., 9*, 167-179, 1994.

Penland, C., and T. Magorian, Prediction of NINO3 sea surface temperatures using linear inverse modeling, *J. Clim., 6*, 1067-1076, 1993.

Penland, C., and P. D. Sardeshmukh, The optimal growth of sea surface temperature anomalies, *J. Clim., 8*, 1999-2024, 1995.

Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes, Reconstruction of historical sea surface temperatures using empirical orthogonal functions, *J. Clim., 9*, 1403-1420, 1996.

Xue, Y., M. A. Cane, S.E. Zebiak, and M. B. Blumenthal, On the prediction of ENSO: A study with a low-order Markov model, *Tellus, Ser. A, 46* 512-528, 1994.

Xue, Y., M. A. Cane, and S.E. Zebiak, Predictability of a coupled model of ENSO using singular vector analysis, I, Optimal growth in seasonal background and ENSO cycles, *Mon. Weather Rev., 125*, 2053-2056, 1997a.

Xue, Y., M. A. Cane, S.E. Zebiak, and T.N. Palmer, Predictability of a coupled model of ENSO using singular vector analysis, II, Optimal growth and forecast skill, *Mon. Weather Rev., 125*, 2057-2073, 1997b.

Zebiak, S. E., and M. A. Cane, A model El Niño Southern Oscillation, *Mon. Weather Rev., 115*, 2262-2278, 1987.

M. A. Cane, D. Chen, and A. Kaplan, Lamont–Doherty Earth Observatory of Columbia University, P.O. Box 1000, 61 Route 9W, Palisades, NY 10964, USA. (mcane@rosie.ldgo.columbia.edu; dchen@ldeo.columbia.edu; alexeyk@ldeo.columbia.edu)

R. Cañizares, Moffatt & Nichol Engineers, 104 West 40th Street, 14th Fl., New York, NY 10018, USA. (rcanizares@moffattnichol.com)

S.E. Zebiak, International Research Institute for Climate Prediction, P.O. Box 1000, 61 Route 9W, Palisades, NY 10964, USA. (steve@iri.columbia.edu)

Received August 31, 2000; revised June 4, 2001; accepted July 16, 2001.

This preprint was prepared with AGU's LaTeX macros v5.01, with the extension package 'AGU$^{++}$' by P. W. Daly, version 1.6b from 1999/08/19.