Skill assessment for ENSO using ensemble prediction

By A. M. MOORE1* and R. KLEEMAN2

1Nova Southeastern University Oceanographic Center, USA
2Bureau of Meteorology Research Centre, Australia

(Received 12 June 1996; revised 25 February 1997)

SUMMARY

A crucial component of any prediction system is the ability to estimate the predictive skill of a forecast so that the degree of confidence that can be placed in an individual forecast can be assessed. In this paper we have used ensemble prediction techniques to develop a means of estimating a priori the predictive skill of forecasts of El Niño Southern Oscillation (ENSO) using an intermediate coupled ocean–atmosphere model. Each member of an ensemble forecast is perturbed using either noise forcing fields or perturbations that are known to increase the low-frequency variability of the coupled model. These are respectively the so-called stochastic optimals and optimal perturbations of the coupled system; they are added to the model to mimic the presence of initial-condition errors and high-frequency stochastic noise and their effect on the predictability of the coupled system in the tropics.

By performing ensemble predictions in hindcast mode, we have identified a usable relation between the skill of a model hindcast and the spread of the ensemble measured relative to some control hindcast. The practical nature of ensemble prediction is demonstrated by computing the relationship between model skill and the spread of an ensemble from hindcasts of ENSO for the period 1972–86, and then comparing the actual hindcast skills for the period 1987–93 with those suggested by the skill-spread relation from the preceding period. The relationship that we find between the model skill and spread of an ensemble appears to be robust in the sense that it is relatively insensitive to variations in the ensemble prediction procedure and to changes in some model parameters. However, crucial to the success of the ensemble predictions is the use of perturbations in the ensembles that are known to increase the low-frequency variability of the coupled model. These perturbations can efficiently probe the probability density function of possible coupled-model states. If randomly chosen perturbations are used in the ensembles, however, no practical relation between model skill and the spread of an ensemble emerges.

The relationship identified here, between model skill and the spread of an ensemble prediction, offers a practical means of estimating the confidence that we can place in future forecasts of ENSO using the same coupled model.

KEYWORDS: El Niño Ensemble prediction Optimal perturbations Predictability

1. INTRODUCTION

Forecasts of the El Niño Southern Oscillation (ENSO) have been made routinely at a number of institutes around the world for the past decade. These forecasts are made in a variety of ways using coupled models, statistical techniques, and predictor methods. In each case, a forecast for the sea surface temperature (SST) anomalies in the tropical Pacific Ocean and for the Southern Oscillation Index (SOI) are issued for the next 12 to 24 months, and these are published routinely for use by the scientific community (e.g. ‘Experimental Long-Lead Forecast Bulletin’). Although forecasts of SST and SOI are useful, their value is limited because no indication is given about the likely skill of a forecast or its reliability. If the community is to act upon the information gained from a forecast, then some a priori estimate is needed of the likelihood that the forecast is correct.

The problem of forecasting the skill of a forecast has received much attention in relation to numerical weather prediction (NWP). The first papers to appear dealing with this issue were those of Epstein (1969) and Gleeson (1970) who considered, in detail, the problem of computing the evolution of the probability density function of the state of the atmosphere. Epstein’s idea, which still forms the basis of our thinking today, was that the skill of a weather forecast depends upon the rate at which forecast errors grow in the system. Epstein was concerned mainly with the growth of errors and uncertainties in the initial conditions. As we shall see later, forecast skill also depends upon the influence of

* Corresponding author: Program in Atmospheric and Oceanic Sciences, University of Colorado, Campus Box 311, Boulder, CO 80309-0311, USA.
model errors and errors associated with stochastic (i.e. unpredictable) events in the system, but the framework proposed by Epstein still applies essentially in this case.

Epstein argued that the initial conditions of a NWP model, as determined from an analysis of observations and previous forecasts, are not unique, and there are an infinite number of possible initial conditions consistent with the observations. If one of these describes the true state of the system, then all other possible initial conditions can be viewed as the true state plus some initial error. The growth rate of the initial errors depends on the system dynamics, and determines the skill of a forecast. If we view each possible initial condition as a single point in a multi-dimensional state-space, then the infinite set of possible initial conditions form a cloud of points. The density of points within the cloud will not be uniform, and we can view it as a cloud of probability or more precisely as the probability density function (PDF) which increases towards the centre of the cloud. If the model is a good representation of the real system, then the true state of the system is likely to be near the point of maximum probability. If a forecast is made from each point in the initial cloud, then the collective evolution of the forecasts describes how the PDF evolves in time. An initially spherical PDF will evolve into an ellipsoid, indicating that some directions in state-space are more probable than others because of the growth of errors. As the volume occupied by the PDF increases, there is less certainty of where the true state of the system is likely to reside.

Epstein derived equations that describe the evolution of the PDF of a dynamical system. However, the value of the low-order moments of the PDF, such as the mean, standard deviation and skewness, depend increasingly upon its higher-order moments. To make the problem tractable, it is necessary to make some assumptions about the higher-order moments and develop a closure relation for the system*. The solution of the PDF equations of the atmosphere would be very demanding computationally.

Since the seminal works of Epstein and Gleeson, other more tractable techniques have been proposed for estimating the shape of the PDF of a dynamical system. The most popular, which has received much attention in meteorology, is ensemble prediction (e.g. Leith 1974; Murphy 1988; Hoffman and Kalnay 1983; Molteni et al. 1996). The idea behind ensemble prediction is to estimate the shape of the atmospheric PDF by repeating a forecast many times, each time perturbing the initial conditions of a forecast model. This amounts to sampling a subset of all possible initial conditions described by the initial PDF. The first and second moments of the evolving PDF can be estimated from the resulting ensemble. The first moment (the ‘ensemble mean’) represents a best estimate of the true state of the system, while the second moment (the ‘variance’) is a measure of the spread of the PDF. If the spread of the members of the ensemble relative to one another increases slowly in time, the atmospheric state we are trying to predict is probably relatively insensitive to errors and uncertainties in the initial conditions, so that the skill of the original, unperturbed forecast is probably high. If, however, the ensemble members diverge rapidly, then the atmospheric state we are trying to forecast is susceptible to the growth of errors in the initial conditions, and the skill of the original, unperturbed forecast is likely to be poor. This is the guiding philosophy behind ensemble prediction.

Obviously, the larger the size of the ensemble the better an estimate of the evolving PDF is likely to be. If a forecasting model has \( N \) grid points describing \( M \) variables, then the system is said to have \( N_p = M \times N \) degrees of freedom. In theory, we must choose \( N_p \) independent perturbations for the initial conditions of the model in order to obtain a best estimate of the evolving PDF of the system. For NWP models, \( N_p \approx 10^6 \), although

* These ideas are related to those developed much earlier in statistical mechanics by Liouville and Boltzmann (e.g. see Sommerfeld (1964)) and in turbulence theory (see Batchelor (1953) and Kraichnan (1959)). More recently they have been applied to other branches of meteorology (e.g. Frederiksen and Bell (1988)).
the maximum affordable ensemble size is \( N_e \approx 10 - 50 \) is much less than \( N_f \). Even if we consider only the physically-realistic and dynamically-balanced atmospheric states, \( N_b \), this is still likely to be very much larger than the maximum affordable ensemble size. Since \( N_e < N_b \), we must choose perturbations for the initial conditions of ensemble forecasts wisely. Ideally, these perturbations should project onto the \( N_b \) balanced degrees of freedom of the model that account for most of the variance \( \sigma^2 \) in the system, so that the resulting ensemble will give us some hope of reliably estimating the shape of the PDF of the system.

The approach currently adopted for operational ensemble prediction at the European Centre for Medium-Range Weather Forecasts (ECMWF) and the US National Centers for Environmental Prediction (NCEP) is to perturb the initial conditions of a NWP model with perturbations that are most likely to project on to the dominant balanced degrees of freedom of the atmosphere. The methods used at ECMWF and NCEP for computing these perturbations are somewhat different. At ECMWF (Buizza et al. 1993; Mureau et al. 1993; Molteni et al. 1996) the singular vectors or so-called 'optimal perturbations' of the system are used and are the fastest-growing disturbances that can exist in the system during the early stages of linear error-growth. At NCEP, so called 'breeding vectors' are used which correspond to approximations of the Lyapunov vectors of the time evolving forecast (Toth and Kalnay 1993, 1995; Tracton and Kalnay 1993). The optimal perturbations of the atmosphere can sustain very rapid growth for only a few days, while the Lyapunov vectors may sustain much slower growth for longer periods. Despite these and other differences in the approaches to ensemble prediction employed at ECMWF and NCEP, both groups have reported significant improvements in forecasting ability as a result of these practices. It is found that the ensemble mean forecast can extend the range of useful skill, and that the spread of the ensemble of forecasts appears to be a good indicator of the skill of the resulting ensemble mean.

Using an intermediate coupled ocean–atmosphere model of ENSO, Moore and Kleeman (1996, 1997a,b) have demonstrated that the optimal perturbations of the coupled system act as efficient precursors for ENSO, and so perturbations of this form will increase the variance of the system. Moore and Kleeman (1997b) have also demonstrated that the optimal perturbations bear a striking resemblance in many ways to westerly and easterly wind bursts over the tropical western Pacific Ocean, disturbances also thought to increase the variance of the real coupled ocean–atmosphere system in the tropics. For this reason, we have developed an ensemble forecasting system for ENSO in which the optimal perturbations of a coupled model are used to perturb the model during each forecast of the ensemble.

There are essentially three forms of error that limit the skill of a numerical forecast model and which contribute to the growth of forecast errors. These are (a) initial-condition errors resulting from errors in forcing fields, boundary conditions, and observation and analysis errors, (b) model errors resulting from errors in parametrizations of physical processes, and the truncation errors of numerical methods used to solve the discretized equations of the system, and (c) errors arising from processes that are unpredictable or stochastic in nature but which are nonetheless an important component of the system. We shall attempt to address the effect of each form of error on the skill of an operational forecast model of ENSO.

The forecast model used here is an intermediate coupled ocean–atmosphere model and is described briefly in section 2. The philosophy behind our approach to ensemble prediction is described in section 3. In sections 4 and 5, we consider the influence of initial condition errors on model skill, and examine alternative methods for perturbing the initial conditions of the model in ensemble forecasts. Throughout these sections, a general relationship emerges between the spread of the ensemble and the skill of a forecast, and
this relationship is robust to many changes in the way the ensemble is performed. Using this relationship, we shall demonstrate how a priori estimates of forecast skill can be made. We conclude, in section 6, with a summary of our results.

2. THE COUPLED MODEL

The model employed in the present study is an intermediate coupled model that is currently used operationally at the Australian Bureau of Meteorology for making routine forecasts of ENSO (Kleeman 1994). The coupled model computes the anomalous atmospheric and oceanic circulations relative to the mean seasonal cycle of the system which is specified from an observed climatology. Only a brief description of the model is given here and interested readers are referred to the detailed model description given by Kleeman (1993).

The atmospheric component of the model is a simple global, steady-state atmosphere similar to that of Gill (1980), and describes the dynamics of the first baroclinic mode. A parametrization for surface latent heating and latent heating due to tropical deep penetrative convection is included. Convective heating is nonlinear and is generally active only when SST exceeds 28 °C (Hirst 1986). The observed seasonal cycle is imposed in the model, and the model computes the anomalies about the seasonally varying fields. In the model, latent-heating anomalies develop in response to SST anomalies which are controlled by ocean dynamics.

The ocean component of the coupled model describes the upper-ocean circulation in the tropical Pacific due to the first baroclinic mode. SST anomalies develop through the action of vertical displacements of the oceanic thermocline, but, in accordance with observation, their effect on SST is limited when the thermocline is very deep or very shallow. Although changes in vertical and horizontal advection are also known to influence SST, Kleeman (1993) has demonstrated that including only the influence of thermocline displacements on SST produces the most skillful coupled-model forecasts. Including the effects of changes in the rate of vertical and horizontal advection on SST can degrade the skill of the model, so these effects are not included in the coupled model used here.

The coupled model possesses a high level of skill in hindcasting ENSO events that is comparable to that of other models currently used by the international community. Figure 1, from Kleeman et al. (1995), shows the anomaly correlation of the model NINO3 index (i.e. the average SST anomaly between 5°N and 5°S, 90°W and 150°W) with the observed NINO3 index as a function of hindcast lag time. When subsurface thermal observations from the tropical Pacific are assimilated into the model, anomaly correlations in excess of 0.6 are achieved for forecast periods out to 16 months.

3. OPTIMAL PERTURBATIONS AND STOCHASTIC OPTIMALS

In order to understand the philosophy behind the ensemble prediction schemes described in section 4, we shall briefly review some current ideas regarding error growth in dynamical systems. We denote by $\Psi$ the state vector that describes low-frequency ENSO variability in the coupled ocean–atmosphere system. In particular, $\Psi$ will denote the true state vector of the coupled system, and we assume it evolves according to

$$\frac{\partial \Psi}{\partial t} = L(\Psi),$$

where $L$ is a nonlinear operator. Ignoring model error for the moment, we shall assume that a forecast of ENSO, denoted $\Psi_f$, also evolves according to (1). We shall return to
the subject of model error later. In general, however, the model forecast will be in error because of uncertainties in the model initial conditions, boundary conditions etc., so that \( \Psi_t = \Psi + \psi \) where \( \psi \) denotes the forecast errors in the low-frequency component of the system. Assuming that \( \psi \) is initially small compared to \( \Psi \), then a Taylor expansion of (1) shows that, to first order, \( \psi \) evolves according to

\[
\frac{\partial \psi}{\partial t} = (\partial L/\partial \Psi)\psi, \tag{2}
\]

where \( (\partial L/\partial \Psi) \) is in general non-autonomous and depends upon the evolution of \( \Psi \). Equation (2), commonly referred to as the ‘tangent linear equation’ (TLE), describes the evolution of errors \( \psi \) during their early stages of linear development when \( \psi^2 \ll \psi \).

Since we are concerned with numerical models that solve discrete forms of (1), we shall consider the discrete form of (2), namely

\[
\psi_n = \psi_{n-1} + \Delta t A_{n-1} \psi_{n-1}, \tag{3}
\]

where \( \Delta t \) is the time-step, the subscript \( n \) refers to the time level \( t = n \Delta t \), and the matrix \( A_n = (\partial L/\partial \Psi)_{t=n,\Delta t} \). In (3) we have used a simple time-stepping scheme to simplify the algebraic expressions that follow, however more complex time-stepping procedures can be used without loss of generality.

If a forecast \( \Psi_t \) is subject to errors \( \psi_0 \) in the initial conditions, then from (3) it is easy to show that

\[
\psi_n = \left( \prod_{k=1}^{n} (1 + \Delta t A_{n-k}) \right) \psi_0 = R_{0,n,\Delta t} \psi_0. \tag{4}
\]

The linear operator \( R_{0,n,\Delta t} \) in (4) is the so-called propagator of the TLE (3) and advances the initial errors \( \psi_0 \) forward in time \( n \) time-steps. Typically, when assessing the skill of a forecast \( \Psi_t \), we measure the growth of errors \( \psi \) in terms of some norm \( V_X \) given by

\[
V_X = \psi^T X \psi, \tag{5}
\]

where the superscript \( T \) denotes a vector transpose*, and the matrix \( X \) defines the norm of interest. If \( X = I \), the identity matrix, then (5) denotes the L-2 norm.

* Formerly, conjugate matrix: the matrix whose rows are the columns of the given matrix.
It can be shown (Farrell 1982, 1990) that the errors which grow most rapidly during the interval $\tau$ are the eigenvectors of the operator $R_{0,t}^1 XR_{0,t}$. These eigenvectors, denoted $\hat{\Psi}_\lambda$, are by definition the singular vectors of the propagator $R_{0,t}$ (Noble 1969). A more common name for $\hat{\Psi}_\lambda$, introduced by Farrell (1982), is 'optimal perturbations', i.e. $\hat{\Psi}_\lambda$ are optimal in the sense that they optimize the growth of $V_X$ over the time interval $\tau$.

In addition to uncertainties in the model initial conditions, forecast errors can also grow under the influence of stochastic processes in the system. These processes are processes that are essentially unpredictable over the timescales of interest, so for the seasonal and interannual timescales of interest to ENSO forecasters, the stochastic processes are represented by events such as 30–60 day waves, day-to-day variations in weather, diurnal variations in convective activity, and so forth. These processes are not described by the simple coupled model considered here in which a steady-state atmosphere is used; however, their influence on the low-frequency variability is important in limiting the predictability of the real coupled system so they cannot be ignored. As shown by Kleeman and Moore (1997), stochastic noise represents an unavoidable source of error in the coupled system and its effects must be included in any estimate of the predictability of the coupled system.

Following Farrell and Ioannou (1993a) and Kleeman and Moore (1997), we can include the effect of high-frequency, stochastic processes on the predictability of the low-frequency ENSO timescales (given by $\hat{\Psi}$) by adding a noise forcing term $F(t)$ to (2) so that (3) becomes

$$\Psi_n = \Psi_{n-1} + \Delta t A_{n-1} \Psi_{n-1} + \Delta t F_n,$$

where $F_n \equiv F(n\Delta t)$ is described by the statistics

$$\langle F_\lambda \rangle = 0; \quad \langle F_\lambda, F_\mu^T \rangle = C_{\lambda,\mu},$$

where $\langle \ldots \rangle$ denotes the ensemble average, and $C_{\lambda,\mu}$ is the covariance matrix of the ubiquitous stochastic noise inherent in the system. Following Kleeman and Moore (1996), the variance $\|\Psi_n\|^2_X$ of $\Psi$ at time $n$ associated with many different realizations of the noise forcing $F$ is given by

$$\|\Psi_n\|^2_X = \langle (\Psi_n - \overline{\Psi})^T X (\Psi_n - \overline{\Psi}) \rangle = (\Delta t)^2 \text{trace}\left[\sum_{\lambda=0}^{n-1} \sum_{\mu=0}^{n-1} R_{\lambda,n}^T X R_{\mu,n} C_{\lambda,\mu}\right],$$

where $\overline{\Psi} = \langle \Psi \rangle$. If the noise covariance $C_{\lambda,\mu}$ is separable in space and time, where $D_{\lambda,\mu}$ and $C$ respectively describe the temporal decorrelation and spatial covariance of $F$, then

$$C_{\lambda,\mu} = D_{\lambda,\mu} C$$

and (8) can be written as

$$\|\Psi_n\|^2_X = (\Delta t)^2 \text{trace}\left[\sum_{\lambda=0}^{n-1} \sum_{\mu=0}^{n-1} D_{\lambda,\mu} R_{\lambda,n}^T X R_{\mu,n} C\right] = \text{trace}\{ZC\},$$

where $Z = (\Delta t)^2 \sum_{\lambda=0}^{n-1} \sum_{\mu=0}^{n-1} D_{\lambda,\mu} R_{\lambda,n}^T X R_{\mu,n}$. If $Z$ and $C$ have the eigenvalue/eigenvector sets $\{q, Q\}$ and $\{p, P\}$ respectively, then (10) can be rewritten as

$$\|\Psi_n\|^2_X = \sum_i \sum_j q_i p_j Q_i^T P_j,$$

which has a simple geometric interpretation. The eigenvectors $P_j$ of $C$ are the 'empirical orthogonal functions' (EOFs) of the stochastic noise forcing of the system. The eigenvectors
Q' of Z are the ‘stochastic optimals’ (Kleeman and Moore 1997) or ‘forcing orthogonal functions’ (FOFs) (Farrell and Ioannou 1993a). If the noise EOFs P^j project onto the stochastic optimals Q' of the system, then the noise can increase the system variance on ENSO timescales. The stochastic optimals depend on the dynamics of the low-frequency ENSO variability as shown in (10). The summations in (10) are a discrete representation of time integrals, so, when the noise is white in time (i.e. D_{k,\mu} = \delta_{k,\mu}), Z represents the average over time of the optimal operators R^T_{k,b} XR_{\mu,n} whose eigenvectors are the ‘optimal perturbations’ of the system. The ability of the system to amplify the noise F depends, therefore, upon the linearized dynamics of the system and the time history of the basic-state Ψ, both of which define the propagator R. As discussed by Farrell and Ioannou (1993a,b,c) and Ioannou (1995), it is the nature of (\partial L/\partial \Psi) in (2) that governs the amplification of stochastic noise in dynamical systems. For all realistic systems, the linear operator (\partial L/\partial \Psi) is non-normal. What (11) shows is that the uncertainties associated with high-frequency stochastic events in the coupled ocean–atmosphere can be amplified by the dynamics that describe the low-frequency variability, which in turn leads to forecast-error growth and a reduction in the skill of a forecast of the low-frequency ENSO signal.

Our aim is to use ensemble prediction to estimate the first and second moments of the PDF of the coupled ocean–atmosphere system and so attempt to estimate a priori the probable skill of a model forecast. This we shall do by choosing perturbations for the model initial conditions that increase the variance of the coupled system so that the ensemble that we obtain (whose size is usually constrained by computational demands) will probe a large portion of the state-space occupied by the PDF, including its extremes, thus allowing us to estimate its shape. Clearly, we can use the ideas relating to optimal perturbations and stochastic optimals presented above to guide us in our choice of ensemble perturbations since, by definition, the optimal perturbations and stochastic optimals increase the low-frequency variability of the coupled system.

4. THE ENSEMBLE PREDICTION SCHEME

Guided by the ideas of section 3, we have used stochastic optimals and optimal perturbations to perturb the coupled model during ensemble prediction experiments. In the present section we shall investigate a number of different ensemble prediction strategies, and examine the relationship between the spread of the ensemble and the skill of a forecast. We shall consider the 22-year period January 1972 to December 1993 with two-year forecasts initialized at the start of each calendar month.

The normal operational procedure for making an ENSO forecast using the model of section 2 is as follows

- Step (A) Spin up the ocean model alone with observed wind-stress anomalies derived from the Florida State University (FSU) wind data-set of Legler and O'Brien (1984) for the two years before the start date of the forecast.
- Step (B) Following step A, the atmosphere and ocean models are coupled and run forward in time for two years in forecast mode.

For the period 1972–93, the forecasts that we describe here are obviously hindcasts since we know what really happened to the coupled system during that period. In the following, a hindcast made using steps (A) and (B) will be referred to as a ‘control hindcast’.

(a) Ensemble hindcasts using stochastic optimal forcing

In our first approach to ensemble prediction, the stochastic optimals of the coupled model are used to introduce noise into the system to mimic the effects of unpredictable
events on ENSO. By choosing the stochastic optimals as the noise forcing functions, we are assured of increasing the low-frequency variance of the system since the noise is amplified by the system dynamics. The purpose of adding such a noise forcing is two-fold: firstly, the stochastic noise forcing induces variability in the system which is likely to span a large region of the state-space occupied by the coupled model PDF and so probe its extremes; secondly, the noise forcing introduces unavoidable errors into the system which must be included in any assessment of its predictability.

The wind-stress and heat-flux patterns of the first two stochastic optimals of the coupled model are shown in Fig. 2. These are the dominant eigenvectors of $Z$ in (10), where the basic state consists of the observed seasonal cycle only. In computing $Z$, a decorrelation time of 3.5 days was assumed for the noise forcing. This is a typical decorrelation time of the stochastic variability present in the ECMWF daily wind analyses that are assumed to be representative of the observed state of the atmosphere (see Kleeman and Moore (1997)). However, the shape of the stochastic optimals is relatively insensitive to the value of the decorrelation time. The stochastic optimals of Fig. 2 were computed using the observed seasonal cycle as the background state for the six-month period August–January. The operator $X$ in (10) was chosen to define the ($\text{NINO3 Index}^2$)-norm. The structure of the stochastic optimals is insensitive to the time of year, although the rate of growth of system variance arising from the stochastic optimals varies seasonally, being largest during boreal spring and summer. The dynamics responsible for these variations is discussed in detail by Moore and Kleeman (1996). For periods of six months or more, the structure of the
stochastic optimals remains reasonably unchanged. Consequently, in the following, we have used the six-month stochastic optimals as forcing functions for the coupled model.

Ensemble hindcasts were performed as follows. Firstly, a two-year control hindcast was performed as described in steps (A) and (B) above. Secondly, each member of the ensemble was generated by re-running the control hindcast with a stochastic forcing term added to the wind-stress and heat-flux forcing. The stochastic forcing term, composed of a linear combination of the first two stochastic optimals shown in Fig. 2, was applied to the system only during the ocean spin-up step (A). Kleeman and Moore (1996) showed that the first two stochastic optimals account for 96% of the variability that would result in the coupled model if noise forcing were present with characteristics similar to the noise present in ECMWF daily wind analyses. The first stochastic optimal alone would account for 90% of this variability. The first two stochastic optimals of time intervals longer than six months display similar characteristics, although the percentage of explained variability tends to decrease slowly with increasing time. The time series of the stochastic optimal amplitudes were in the form of noise with a decorrelation time of 3.5 days, in accordance with the characteristics of the ECMWF analyses. The relative contribution of the first and second stochastic optimals to the noise forcing was chosen to be 1:0.07 which is the same as their contribution to the perturbation growth expected from synoptic noise with characteristics similar to those of ECMWF daily wind analyses (Kleeman and Moore 1996). The variance of the heat-flux component of the noise forcing averaged over the entire tropical Pacific Ocean was equivalent to a basin-wide standard deviation of 27 W m\(^{-2}\), and the wind component of the noise had a basin-wide standard deviation of 0.35 m s\(^{-1}\). This level of noise is well within the uncertainty in SST and wind stress. A total of fourteen different hindcasts were performed in this way using different time series of noise forcing composed of the first two stochastic optimals in each case. With the control hindcast, this gives a total ensemble size of fifteen. This is significantly smaller than the total number of degrees of freedom in the model which is 1860. The results presented below are insensitive to the size of the ensemble (see subsection 5(c)).

It remains for us to choose a measure of skill \(S_k\) of a coupled-model hindcast, and a measure of the spread \(S_p\) of an ensemble of forecasts. There are a number of possible choices for both \(S_k\) and \(S_p\). We shall consider the following as measures of skill.

(i) The correlation of the control hindcast NINO3-index, denoted by \(\mathcal{N}_c\), with the observed NINO3-index, \(\mathcal{N}_o\), over a 24-month hindcast period, namely

\[
S_k = \frac{\sum_{i=1}^{24}(\mathcal{N}_{ci} - \overline{\mathcal{N}_c})(\mathcal{N}_{oi} - \overline{\mathcal{N}_o})}{\left(\sum_{i=1}^{24}(\overline{\mathcal{N}_{ci}} - \overline{\mathcal{N}_c})^2 \sum_{r=1}^{24}(\mathcal{N}_{or} - \overline{\mathcal{N}_o})^2\right)^{\frac{1}{2}}},
\]

(12)

where an overbar denotes the mean value of \(\mathcal{N}\) over a 24-month hindcast period, and time is denoted by the subscripts \(i\), \(j\) and \(r\).

(ii) The correlation of the ensemble mean hindcast NINO3-index, denoted \(\mathcal{N}_e\), with the observed NINO3-index, \(\mathcal{N}_o\), over a 24-month hindcast period

\[
S_k = \frac{\sum_{i=1}^{24}(\mathcal{N}_{ei} - \overline{\mathcal{N}_e})(\mathcal{N}_{oi} - \overline{\mathcal{N}_o})}{\left(\sum_{j=1}^{24}(\overline{\mathcal{N}_{ej}} - \overline{\mathcal{N}_e})^2 \sum_{r=1}^{24}(\mathcal{N}_{or} - \overline{\mathcal{N}_o})^2\right)^{\frac{1}{2}}},
\]

(13)

where \(\mathcal{N}_{ei} = 1/15 \sum_{m=1}^{15} \mathcal{N}_{im}\), and \(\mathcal{N}_{im}\) is the NINO3-index of the \(m\)th ensemble member at month \(i\).
(iii) The root-mean-square (r.m.s.) difference between the NINO3-index of the control hindcast and the observed NINO3-index over a 24-month hindcast period

\[ S_k = \left\{ \frac{1}{24} \sum_{i=1}^{24} (\mathcal{N}_{ci} - \mathcal{N}_{oi})^2 \right\}^{\frac{1}{2}}. \]  

(iv) As (iii) except using the ensemble mean \( \mathcal{N}_{ci} \) in place of the control hindcast \( \mathcal{N}_{oi} \) in (14).

The following will be considered as measures of the spread of the ensemble.

(a) The mean absolute deviation (MAD) of the correlation of the NINO3-index of the ensemble members with the control hindcast NINO3-index over a 24-month hindcast period

\[ S_p = \frac{1}{14} \sum_{m=1}^{14} \left| C_{em} - \overline{C_e} \right|, \]  

where \( C_{em} \) denotes the correlation of the NINO3-index of the \( m \)th ensemble member with the control hindcast, and \( \overline{C_e} = \sum_{k=1}^{14} C_{er} / 14 \) is the mean value of the correlations \( C_{er} \). This measure of spread can also be used when \( S_k \) is defined in terms of the r.m.s. difference as in (iii) and (iv) above, in which case the correlations in (15) are replaced by r.m.s. errors.

(b) The standard deviation (SD) of the correlation of the NINO3-index of the ensemble members with the control hindcast NINO3-index over a 24-month hindcast period

\[ S_p = \left\{ \frac{1}{13} \sum_{m=1}^{14} (C_{em} - \overline{C_e})^2 \right\}^{\frac{1}{2}}. \]  

This measure of spread can also be used when \( S_k \) is defined in terms of the r.m.s. difference as in (iii) and (iv).

(c) The average correlation of the NINO3-index of each member of the ensemble with the NINO3-index of the control hindcast over a 24-month hindcast period

\[ S_p = \frac{1}{14} \sum_{m=1}^{14} C_{em} = \overline{C_e}, \]

where \( C_{em} \) is defined in (a). To understand this measure of \( S_p \), consider the case where all ensemble members are highly correlated with the control hindcast. In this case, \( \overline{C_e} \approx +1 \) which would correspond to a low spread in the ensemble members about the control. Alternatively, if each ensemble member differs considerably from the control and from the remaining members of the ensemble, then \( \overline{C_e} \approx 0 \), which would correspond to a high spread in the ensemble members. If all of the ensemble members are anti-correlated with the control hindcast, then \( \overline{C_e} \approx -1 \) and the spread of the ensemble could also be considered small. We shall discuss these scenarios in more detail later.

Figure 3(a,b) shows time series of the control hindcast skill defined by (12) and (14) respectively as a function of the start date of the hindcast. Also shown are the skills of the best and worst members of the ensemble. Clearly, the skill of the control hindcast varies
Figure 3. (a) A time series of the skill of each 2-year control hindcast as a function of hindcast start-date (heavy solid line). Also shown are time series of the skill of the most skilful and least skilful members of each ensemble as a function of ensemble start-date (light solid lines). The region bounded by the most and least skilful ensemble members is stippled. Skill is defined by the correlation equation (12). (b) As (a) but when skill is defined by the r.m.s. error (Eq. (14)). The dashed line indicates the average r.m.s. error. (c) As (a) but for the case where the model is assumed to be perfect.

considerably during the 22-year period considered. However, the ensemble is well behaved in the sense that in nearly all cases the skill of the control hindcast lies between that of the best and worst members of the ensemble. Experience in NWP suggests that a hindcast with $S_c \geq 0.6$ possesses useful skill. In Fig. 3(a) there are times when all members of the ensemble have skill that is well below 0.6, and at the same times Fig. 3(b) shows a tendency for the r.m.s. errors of most ensemble members to be larger than the average. In
Figure 4. Scatter plots of $S_k$ versus $S_p$ for ensemble hindcasts perturbed by noise forcing in the form of stochastic optimals during the spin-up step (A). The line labelled H marks the approximate position of the hypotenuse of the triangular distribution of points and represents the "line of minimum skill". Different measures of $S_k$ and $S_p$ are used in each case: (a) $S_k$ from Eq. (12), $S_p = \text{MAD}$; (b) $S_k$ from Eq. (12), $S_p = \text{SD}$; (c) $S_k$ from Eq. (12), $S_p = \overline{C_r}$; (d) $S_k$ from Eq. (13), $S_p = \text{MAD}$; (e) $S_k$ from Eq. (13), $S_p = \text{SD}$; (f) $S_k$ from Eq. (13), $S_p = \overline{C_r}$; (g) $S_k$ from Eq. (14), $S_p = \text{MAD}$; (h) $S_k$ from Eq. (14), $S_p = \text{SD}$; (i) $S_k$ from Eq. (14) for ensemble mean, $S_p = \text{MAD}$; (j) $S_k$ as in (i) and $S_p = \text{SD}$. 
such instances, perturbing the initial conditions never causes a member of the ensemble to show skill; consequently, we conclude that errors in the model cause these poor hindcasts, not errors in the initial conditions. However, we cannot rule out the possibility that wind errors present during step (A) may be so large at these times that the perturbations added to the ensemble members cannot offset the effect of the large errors that the winds induce in the initial conditions.

Using each measure of skill defined in (i)–(iv) and each measure of spread defined in (a)–(c), we shall examine the relationship between the spread of an ensemble and the skill of the control hindcast or ensemble mean. Figure 4 shows scatter plots of $S_k$ against $S_p$ for the different measures of skill and spread defined above. (In Fig. 4, panels (a)–(f) use correlation-based measures of $S_k$ and panels (g)–(j) use r.m.s.-based measures of $S_k$.) When using (i), (iii) or (iv) as measures of skill, we assume that each member of the ensemble is an equally likely hindcast, so each member is treated in turn as the control hindcast and $S_k$ and $S_p$ are computed accordingly. If the skill of the ensemble-mean hindcast is plotted against $S_p$, there are many fewer points on the scatter plots. To eliminate the effects of model error on the scatter plots, we have rejected from Fig. 4(a–f) all ensembles that are most likely poor due to model error as discussed above. If all members of the ensemble
have a skill $S_k < 0.6$ then they are rejected. As a result of this, only approximately 20% of the hindcasts were rejected from Fig. 4(a–f). No such condition was applied when r.m.s.-based measures of skill were used as in Fig. 4(g–j).

Panels (a)–(f) in Fig. 4 show that, when correlation-based measures of $S_k$ are used, the points in the scatter plots are approximately distributed in the form of a right-angled triangle, with the hypotenuse of the triangle sloping upward from the high-skill–low-spread region to the low-skill–high-spread region. The approximate position of the hypotenuse is indicated by the line labelled H. The position of H was determined by identifying the smallest value of $S_k$ that results in the scatter plot for each value of $S_p < 0.2$ in Fig. 4(a, b, d, e), and $S_p > 0.4$ in Fig. 4(c, f), and drawing the best least-squares-fit straight line. The triangular distribution of points indicates that there is a relation between $S_k$ and $S_p$, and we shall examine the reasons for the distribution shortly. R.m.s.-based measures of $S_k$ and $S_p$ do not yield a triangular distribution in Fig. 4(g–j). Instead, there is a tendency for points to be clustered over a fairly modest range of $S_p$ regardless of the values of $S_k$. We note that if panels of Fig. 4 are plotted for each season, instead of for all months of the year as shown, the resulting point distributions are very similar to those of Fig. 4, and the slope of H changes very little throughout the year.

Correlation-based measures of $S_k$ award a high skill to hindcasts that predict the correct trends in Niño3-index, regardless of their amplitude. R.m.s.-based measures of skill, on the other hand, are less forgiving and award high skills to hindcasts only if the amplitude and phase of the Niño3-index is correct. The scatter plots of Fig. 4 resulting from r.m.s.-based measures of skill (Fig. 4(g–j)) perhaps indicate the presence of a systematic error in the coupled-model Niño3-index. This is typical of intermediate coupled models like that used here. Given the simplistic nature of such a model, we are usually content with good predictions of the phase and trends in the Niño3-index (which, after all, are a good indication of an approaching ENSO episode), rather than expecting to capture precisely all the details of both amplitude and phase. Therefore, in the remainder of the present paper, we consider only correlation-based measures of $S_k$ and $S_p$, and do not concern ourselves further with model systematic error.

For correlation-based measures of $S_k$ and $S_p$, Fig. 4(a–f) shows that a similar triangular distribution of points results regardless of whether MAD, SD or $\bar{C}_e$ is used as a measure of $S_p$. Since we are using correlation-based measures of $S_k$, we know that any ensemble of correlations cannot be normally distributed because there are upper and lower bounds (+1 and −1 respectively) on any correlation measure. In this case, MAD is more appropriate than SD as a measure of $S_p$, and in the remainder of this paper we shall use MAD as a measure of ensemble spread. The implications of $\bar{C}_e$ will be discussed shortly. In addition, we note that analyses presented shortly suggest that the PDF of the coupled model may be bimodal. In this case, the arithmetic ensemble-mean may be misleading, particularly if each mode of the PDF is equally populated. Therefore, in the remainder of the paper we shall measure skill in terms of the skill of the control hindcast. This has the added advantage (over using the ensemble mean) that there are many more points on the $S_k$ versus $S_p$ scatter plots (e.g. contrast panels (a) and (d) in Fig. 4) which will define more clearly the relationship between $S_k$ and $S_p$. The results and conclusions to be presented are relatively insensitive to the choices of $S_k$ and $S_p$ made here.

To understand the triangular distribution of points in panels (a)–(f) of Fig. 4, it is instructive to consider the case when the model is assumed to be perfect and free from model error. In this case, the unperturbed control hindcast from steps (A) and (B) (see the second paragraph of this section 4) is taken to represent the true state of the coupled system; the skill and spread of each ensemble are then computed as before, treating each member
Figure 5. (a) A scatter plot of $S_k$ versus $S_p$ (defined by Eqs. (12) and (15)) for ensemble hindcasts perturbed by noise forcing in the form of stochastic optimals during the spin-up step (A) assuming that the model is perfect; (b) a scatter plot of $C_e$ versus $S_p$ (defined by Eq. (15)) assuming that the model is perfect. The extent of the 'boomerang' distribution of points is indicated by the dashed lines; (c) as (a) for the case where the perfect model is degraded by introducing a 30% error in the equatorial ocean waves speeds of the perfect model.

of the ensemble as an equally likely hindcast. Figure 3(c) shows time series of the 'control' hindcast skill (defined by (12)) for the 'perfect model', and the skill of the best and worst ensemble members, where the control hindcast was chosen to be one of the perturbed ensemble-members. In this case, the skill of the best ensemble-member generally exceeds 0.9, consistent with the assumption of a perfect model, and there are no cases where the skill of all members of an ensemble falls below 0.6. The resulting skill-spread scatter plot (defined by (12) and (15)) of the 'perfect model' is shown in Fig. 5(a). The triangular distribution of points in this case is similar to that shown in Fig. 4(a), where the same measures of $S_k$ and $S_p$ are used. For a given $S_p$, there are many points below H, so hindcast skills greater than the minimum skill on H are possible, and indeed most likely. Figure 5(a) shows that, in general, the minimum skill H decreases as $S_p$ increases.

The triangular distribution of points in Fig. 5(a) (and in Fig. 4(a–f)) can be understood by considering the different states in which the control hindcast and the ensemble members
can reside. Broadly speaking, the coupled model resides for most of the time in one of two possible states, namely (i) a warm, El-Niño-like state described by $N_c > 0$, and (ii) a cold, La-Niña-like state described by $N_c < 0$. In between, conditions are ‘normal’ where SST is close to its climatological values and $N_c \approx 0$. These states are illustrated schematically in Fig. 6(a–c).

The two-state analogy of the coupled system has also been advanced by Palmer (1993) who likened the ENSO states to Lorenz attractors. In either case, we can visualize the PDF of the coupled model as described by a double-peaked function as indicated schematically in Fig. 6(d). A similar view of the coupled system was described by Webster (1995) who considered the preferred states of the coupled system as analogous to potential wells. In the case considered here, there will be two wells, corresponding to the two preferred states of the system illustrated in Fig. 6(e) The depth of each well will be related to
the strength of a given ENSO episode, and will depend on many environmental factors.
Consider, in Fig. 6(e), a control hindcast that resides at point A deep within one of the
wells. The ensemble members are formed by perturbing the control. If the well is deep,
then perturbations with a realistic amplitude are unable to dislodge the system from the
well, so all the ensemble members describe the same kind of event. Because of this, we
view the interior of the wells as relatively stable states of the system, the degree of stability
depending on the depth of the well. A good indicator of whether or not the ensemble
members reside in the same state or potential-well as the control is the mean correlation
$C_e$ of the ensemble members with the control, which can also be considered as an alternative
measure of ensemble spread (see (17)). If all the ensemble members describe the same kind
of event as the control, then $C_e \approx +1$ and $S_p$ is likely to be small. This case is illustrated
schematically in Fig. 6(f). If, in Fig. 6(e), the well in which A resides is shallow (or if the
perturbations are very large) then it is possible for the majority of the ensemble members
to hop into the well which describes the other state of the system, so $C_e \approx -1$ and $S_p$ is
small. This situation is illustrated in Fig. 6(h). Alternatively, if the control resides between
the two wells at point B in Fig. 6(e), the perturbed ensembles are likely to fall into either
well so $C_e \approx 0$. Because of this, we view point B as a rather unstable state of the system.
This situation is illustrated in Fig. 6(g).

Figure 5(b) shows a scatter plot of $C_e$ versus $S_p$ (defined by MAD in (15)) for the
ensemble hindcasts. Using the imagination, one can see that the points belong to a sort
of ‘fat boomerang’ distribution. The points in the regions $C_e \approx +1$, $C_e \approx 0$ and $C_e \approx -1$
correspond to the situations shown schematically in panels (f), (g) and (h) respectively in
Fig. 6. In Fig. 5(b), there are many more cases when $C_e > 0$ than when $C_e < 0$, which
indicates that the ensemble members have a tendency to reside in the same state (or
potential well) as the control hindcast. Furthermore, since the model exhibits a moderately
high level of predictive skill (cf. Fig. 1), this suggests that the control hindcast resides
for most of the time in the same state as the real coupled system. Therefore, Fig. 5(b)
also suggests that the ensemble members have a tendency to reside in the same state as
the real coupled system. Based on this argument, we can, broadly speaking, consider two
classes of ensemble hindcasts based on the sign of $C_e$; it is of interest to consider how
they influence the distribution of points in Fig. 5(a). Figure 7(a) shows the scatter plot
assuming a perfect model when only ensembles with $C_e \geq 0$ are considered. In this case,
the ensemble members have a tendency to reside in the same state as the control hindcast
(and the same state as the real coupled system), the right-angled triangle distribution of
points is very pronounced and the line of minimum skill $H$ is well defined. As in Figs.
4(a) and 5(a), there is a tendency for $S_k$ to decrease with increasing $S_p$. Figure 7(b) shows
the scatter for ensembles with $C_e < 0$ which indicates a tendency for ensemble members
to reside in the opposite state to the control (and the opposite state to the real system).
Although there are far fewer points than in the previous case, the tendency is for them
to form a triangular distribution, but in this case the hypotenuse slopes upward from the
low-$S_k$–low-$S_p$ region to the high-$S_k$–high-$S_p$ region. This suggests a tendency for $S_k$ to
increase with increasing $S_p$, which for this class of ensembles makes sense because any
increase in $S_p$ will most probably be accompanied by an increase in $C_e$ and $S_k$ as more
and more members of the ensemble move into the same state as the control. In this case,
the hypotenuse in Fig. 7(b) behaves like a line of maximum skill.

The greater population of ensembles that fall into the $C_e > 0$ class suggests that
Fig. 7(a) is the preferred distribution of skill-spread points for this model system, and in
some sense represents the best that we can ever hope to achieve in the limit of a perfect
model. The class of ensembles described by $C_e < 0$ tend to increase the scatter of the
skill-spread points in Figs. 5(a) and 4(a–f), so obscuring the relation between $S_k$ and $S_p$. 
Figure 7. Scatter plots of $S_k$ versus $S_p$ (defined by Eqs. (12) and (15)) assuming a perfect model for the cases (a) $C_e \geq 0$, and (b) $C_e < 0$. The hypotenuse H of the approximate triangular distribution of points is shown in each case.

This is consistent with the idea that when $C_e < 0$ then, by and large, the control hindcast lies outside the PDF defined by the ensemble members (Fig. 6(h)) which is an undesirable configuration of the ensemble. On the other hand, when $C_e > 0$, the control hindcast will lie within the PDF defined by the ensemble members (Fig. 6(f,g)) which can be regarded as a necessary ensemble configuration if we are to estimate the relationship between the second moment $(S_p)$ of the PDF and the control hindcast skill reliably.

The presence of model error also significantly increases the scatter of $(S_p, S_k)$ points in Fig. 4. To illustrate this, the 'perfect model' of Fig. 5(a) was degraded by introducing a 30% error in the equatorial ocean-wave speeds of the coupled model. The ensemble hindcasts were repeated, and the resulting skill-spread scatter plot is shown in Fig. 5(c). Compared to the 'perfect model' case in Fig. 5(a), the scatter of points in Fig. 5(c) has increased considerably. Clearly, the distribution of skill-spread points in Fig. 4(a–f) is influenced by the structure of the model PDF, and by model error.

Figure 4(a–f) shows that, for a given ensemble spread, the distribution of points along the $S_k$-axis is not uniform, indicating that some range of $S_k$ values is more likely than any other. We therefore converted the scatter plots of Fig. 4(a–f) into probability diagrams in the following way. The $S_k$-axis was divided into $i = 20$ equally spaced intervals $y_i$ of width 0.1, and the $S_p$-axis into $j = 10$ equally spaced intervals $x_j$ of width 0.1. Thus, the scatter plot was divided into 200 equal sized bins with $(S_p, S_k)$ pairs $(x_i, y_j)$. The number of points $n_{i,j}$ falling in each bin was determined, and the conditional probability $P_{i,j}$ of a hindcast with ensemble spread $x_j$ having skill $y_i$ (i.e. $P_r(y_i|x_j)$) was computed according to

$$P_{i,j} = n_{i,j} / \sum_{i=1,20} n_{i,j}. \quad (18)$$

Figure 8(a) shows the conditional probability diagram associated with Fig. 4(a), which takes the form of a tongue sloping upwards from the high-$S_k$–low-$S_p$ region towards the low-$S_k$–high-$S_p$ region. The conditional probabilities associated with Fig. 4(b–f) have a very similar structure (not shown). As described earlier, if $S_k < 0.6$ for all ensemble members, the ensembles are not included in the determination of $P$. If these cases are
Figure 8. (a) Contours of the conditional probability $P$ that the control hindcast will have a skill $S_k$ when the spread of an ensemble hindcast is $S_p$. Values of $P$ were computed for all ensemble hindcasts from the period between January 1972 and December 1993. (b) As (a) but contours of $P$ were computed using only ensemble hindcasts from the period January 1972 to December 1986 inclusive. The actual values of $S_p$ and $S_k$ for the ensemble hindcasts during the period January 1987 to December 1993 are also plotted as points. The contour interval for $P$ in (a) and (b) is 0.05. Darker stippling indicates increasing $P$, and the minimum value of $P$ plotted is 0.05. Lines H and M are defined in the text.

retained, Fig. 8(a) changes little. The line labelled M in Fig. 8(a) is a best fit to all points at which $P_{i,j} = 0.04$; it can be considered as an alternative line of minimum skill in probability space. From Fig. 8(a), we can obtain two valuable pieces of information about the control hindcast: the minimum $S_k$ we expect for the control hindcast (from line H or line M), and, secondly, the most likely range of $S_k$ for the control hindcast. Figure 8(a) shows that, in general, as $S_p$ increases the minimum $S_k$ expected for the control hindcast decreases. In addition, as $S_p$ increases the range of most likely $S_k$ (as measured by $P$) also increases. However, as $S_p$ increases the probability $P$ of the most likely range of $S_k$ decreases. In general, if $S_p \leq 0.2$ then the most likely $S_k$ is relatively high, and the possible range of most likely $S_k$ is small and well defined.

Obvious questions concern the usefulness of a conditional probability diagram like that in Fig. 8(a), and the way it might be used operationally to estimate a priori the skill of a forecast. To answer these questions, we computed the conditional probability diagram from the ensemble hindcasts for the period 1972–86, and compared the actual control hindcast skill for the 1987–93 ensemble hindcasts with the most likely $S_k$ suggested by the probability diagram from the previous fifteen years. Figure 8(b) shows $P$ for 1972–86, and superimposed are the actual ($S_p, S_k$) points for 1987–93. In general, when $S_p \leq 0.2$, the majority of the ensemble hindcasts have a skill that falls within the tongue of maximum probability, so estimates of control forecast skill based upon the range of most likely $S_k$ would be good. For $S_p > 0.2$, the skill of the hindcasts varies considerably, and is generally poor (i.e. $S_k < 0.6$), in agreement with the broad region occupied by the lower probability contours. If $P_{\text{max}}(S_p)$ is the maximum value of $P$ for a given $S_p$, an objective measure of the agreement between the actual skill attained by the 1987–93 hindcasts and their most likely skill is the fraction of actual ($S_p, S_k$) points, denoted by $f_p$, that fall within $P_{\text{max}}(S_p) \pm 0.05$ during this period. For $0 \leq S_p \leq 0.2$, $f_p = 74\%$, while for $0.2 < S_p \leq 0.4$, $f_p = 62\%$. The range $P_{\text{max}} \pm 0.05$ defines a narrow band of $S_k$ for $0 \leq S_p \leq 0.2$ so $f_p = 74\%$ represents
good agreement between the actual and expected \( S_\delta \) for \( S_p \leq 0.2 \). On the other hand, for \( 0.2 < S_p \leq 0.4 \), the range \( P_{\text{max}} \pm 0.05 \) spans a large range of \( S_\delta \) since the contours of \( P \) are more diffuse (cf. Fig. 8(b)) which is reflected in a fairly large value of \( f_P \). The fractions \( f_P \) change respectively to 60% and 80% if the ensembles believed to be adversely affected by model error are included Fig. 8(b).

(b) Ensemble prediction using optimal perturbations

The stochastic optimals used to perturb each member of the ensemble hindcasts in subsection 4(a) are computationally expensive, and, for complex coupled GCMs, the computational cost would be prohibitive. It is, therefore, desirable to identify a cheaper and more efficient means of ensemble prediction. The stochastic optimals of Fig. 2 increase the variance of the coupled model by inducing in the system perturbations with structures that are similar to the optimal perturbations of the system, as discussed in section 3. These perturbations are favourably configured for rapid transient growth and so perform the job of elevating the system variance above the level expected in the absence of stochastic noise forcing. Figure 9 shows the wind stress and SST of a typical optimal perturbation that maximises the growth of \( N^2 \). This is the fastest growing member of the optimal perturbation spectrum that maximises the growth of \( N^2 \) during a 3-month period* using only the observed seasonal cycle as the basic state in (4). We consider only the seasonal cycle because, as shown by Moore and Kleeman (1996), the structure of the optimal perturbations is insensitive to the state in which the coupled model resides. The SST structure of the optimal perturbation in the west and central Pacific is similar to the surface heat flux of the dominant stochastic optimal shown in Fig. 2(a).

The similarity between the stochastic optimals and the optimal perturbations suggests an alternative strategy for ensemble forecasting. The vehicles for the amplification of stochastic noise in the coupled system are perturbations with structures similar to the optimal perturbations, so we can use optimal perturbations instead of stochastic optimals to perturb the model during ensemble hindcasts. The optimal perturbations of the coupled model can be computed very cheaply, so this approach to ensemble forecasting is computationally less expensive overall than that described in subsection 4(a). In practice, we

* A 3-month interval was chosen because this is the timescale over which forecast errors for ENSO are typically found to grow (Palmer and Anderson 1994). However the results presented here are relatively insensitive to the optimal growth period chosen.
Figure 10. (a) A scatter plot of $S_k$ versus $S_p$ (defined by Eqs. (12) and (15)) for the ensemble hindcasts perturbed by optimal perturbations during the spin-up step (A). (b) contours of the probability $P$ that the control hindcast will have a skill $S_k$ when the spread of an ensemble hindcast is $S_p$. Values of $P$ were computed for all ensemble hindcasts from the period January 1972 to December 1986. The actual values of $S_p$ and $S_k$ for the ensemble hindcasts during the period January 1987 to December 1993 are also plotted as points. The contour interval for $P$ is 0.05. Darker stippling indicates increasing $P$, and the minimum value of $P$ plotted is 0.05.

perturbed the model with optimal perturbations only during the spin-up step (A) of the ocean model. This was done once every month during the spin-up by adding to the ocean model SST and FSU wind anomalies the fastest growing optimal perturbations of the seasonal cycle whose growth begins during the month in question. Perturbations were added with a random amplitude such that the r.m.s. errors in wind speed, SST and thermocline depth were 0.8 m s$^{-1}$, 1.2 deg C and 1.5 m respectively. At first sight, these perturbations may appear to be larger than the corresponding level of stochastic noise applied to the ensemble hindcasts in subsection 4(a). However, here the coupled model is perturbed with optimal perturbations only once each month instead of being forced continuously in time as in subsection 4(a). In practice, we used the optimal perturbations that maximize the growth of perturbation energy instead of $N^2$. This was done because, in its operational configuration, the optimal perturbations of the $N^2$-norm generally grow more slowly than the perturbation energy optimals, and, since we are adding perturbations only once each month rather than continuously, we want to ensure that we shall induce sufficient variance in the system for it to give the best estimate of the PDF. In addition, the optimal perturbations of the perturbation energy norm elevate the levels of variance in both the ocean and atmosphere rather than in just the NINO3-index.

Figure 10(a) shows the skill-spread scatter plot (based on (12) and (15)) for ensemble hindcasts that use the optimal perturbations to perturb each member. The distribution of points in Fig. 10(a) is very similar to that in Fig. 4(a) in which the stochastic optimals were used to perturb the model. Figure 10(b) shows contours of the conditional probability $P$ given by (18), computed from Fig. 10(a) for the ensemble hindcasts of 1972–86, and superimposed are the skill-spread points for 1987–93. $P$ is similar to that derived in subsection 4(a), and for $S_p \leq 0.2$ many of the hindcasts fall within the region of highest $P$ ($f_P = 53\%$ for $0 \leq S_p \leq 0.2$, $f_P = 70\%$ for $0.2 < S_p \leq 0.4$). In general the results of Fig. 10(b) and Fig. 8(b) are consistent with each other.
Figure 11. Contours of the probability $P$ that the control hindcast will have a skill $S_p$ when the spread of an ensemble hindcast is $S_k$ when the initial conditions were perturbed by optimal perturbations immediately after the spin-up step (A). Values of $P$ were computed for all ensemble hindcasts from the period January 1972 to December 1986. The actual values of $S_k$ and $S_p$ for the ensemble hindcasts during the period January 1987 to December 1993 are also plotted as points. The contour interval for $P$ is 0.05. Darker stippling indicates increasing $P$, and the minimum value of $P$ plotted is 0.05.

(c) Ensemble prediction with data assimilation

Operationally, data assimilation is used in conjunction with observed forcing to produce the best initial condition for a coupled model forecast (cf. Fig. 1). The methods described in subsections 4(a) and 4(b) for perturbing initial conditions during ensemble hindcasts do not easily lend themselves to the case when data assimilation is used since we do not want to perturb the model and force it away from the observations while trying to assimilate them. In addition, coupled GCMs are computationally expensive to run so spinning-up such a model many times to generate each member of an ensemble forecast may not be affordable. Therefore, in this section we explore an alternative approach to ensemble prediction that is amenable to data assimilation, and which could be used in coupled GCMs. In this case, we spin the ocean-model up only once with the observed wind anomalies, and then construct members of the ensemble by perturbing the initial conditions of the spin-up using a linear combination of growing optimal perturbations. In practice, we used combinations of the fastest growing optimal perturbations at different stages of their evolution. This was done in an attempt to recreate the situation described in subsection 4(a) where the stochastic optimals are used to perturb the model and where perturbations develop through the excitation of rapidly growing disturbances that resemble optimal perturbations. An advantage of this approach is that data assimilation can be used to constrain the initial conditions of the coupled model during the spin-up phase of the ocean model before the perturbations are added.

Figure 11 shows contours of the conditional probability $P$ of (18), computed from the skill-spread scatter (based on (12) and (15)) of ensemble hindcasts from 1972–86 perturbed in the above manner. The r.m.s. errors due to the perturbations were about 1 deg C in SST and about 5 m for the thermocline depth. These are typical of the errors that one would expect to find in the model initial conditions. Superimposed in Fig. 11 are the skill-spread points for 1987–93. The tongue of $P$ in Fig. 11 is somewhat more elongated than in Figs. 8(b) and 10(b), suggesting that we may need to experiment further with the optimal perturbation amplitudes used in this case, but in general the relationship between
the ensemble spread and model skill appears to be robust. For $0 \leq S_p \leq 0.2$, $f_p = 59\%$, and for $0.2 < S_p \leq 0.4$, $f_p = 70\%$.

5. Sensitivity Experiments

The results of section 4 show that, if noise forcing functions or perturbations that are known to increase the variance of the coupled system are used to perturb a coupled model during ensemble predictions, then a clear relation emerges between the skill of a hindcast and the spread of the ensemble members. In addition the skill-spread relation is relatively insensitive to the way that the ensembles are perturbed when using either optimal perturbations or stochastic optimals. In this section we shall briefly describe the effect of variations in the ensemble hindcast procedure and variations in model parameters on the skill-spread relation.

(a) Ensemble hindcasts in the presence of stochastic noise

In section 4, we were concerned primarily with the influence on model skill of errors in the initial conditions of the hindcast. However, as argued in section 1, in nature, high-frequency stochastic noise forcing represents a source of unavoidable error and will influence the predictability of the coupled system. With this in mind, we repeated the ensemble hindcast experiments of subsection 4(a), except that in this case the stochastic optimal noise forcing was applied during both the model spin-up step (A) and the hindcast step (B). The resulting relation between $S_k$ and $S_p$ (not shown) is very similar to that shown in Fig. 4(a–f) and Fig. 8. This demonstrates that high-frequency stochastic noise forcing in the coupled system and its effect on the predictability of the low-frequency ENSO variability can be viewed in a similar way to the influence of initial condition errors on the system. Ideally, we should like to consider the effects of both initial-condition errors and stochastic-noise errors on ENSO predictability. This would allow the operational use of a variation of the procedure described in subsection 4(c) where optimal perturbations are used to perturb the initial conditions after data assimilation, and stochastic optimals are used to perturb the resulting coupled-model forecasts.

(b) Random perturbations

To demonstrate that optimal perturbations and stochastic optimals are an efficient means of probing the PDF, we ran ensemble hindcast experiments using randomly chosen perturbations to perturb each member of the ensemble. A total of fifteen perturbations were chosen at random from the ocean model spin-up step (A) and represent physically meaningful and dynamically balanced perturbations. These perturbations were rescaled to have unit energy and then orthogonalized using a Gram–Schmidt method. This procedure generates perturbations that are consistent with the optimal perturbations in that they have initial unit energy and are orthogonal to one another and so perturb independent directions of state-space. Different combinations of the random perturbations were added to the model initial conditions to yield an r.m.s. error of approximately 0.5 deg C in SST and approximately 5 m in thermocline depth. Figure 12 shows that randomly perturbed ensemble hindcasts yield no clear relation between the ensemble skill and ensemble spread, suggesting that unlike the optimal perturbations, a small number of random perturbations do not yield a good estimate of the second moment of the PDF. For most values of $S_k$ in Fig. 12, the ensemble hindcasts have a low $S_p$, indicating that most of the ensemble members strongly resemble the control hindcast. This suggests that, to increase $S_p$, random perturbations with unrealistically large amplitudes are required.
Figure 12. A scatter plot of $S_k$ versus $S_p$ (defined by Eqs. (12) and (15)) for the ensemble hindcasts perturbed by random perturbations immediately after the spin-up step (A).

(c) **Ensemble size and hindcast period**

In all cases examined in section 4, an ensemble size of 15 (control plus 14 perturbed hindcasts) was used. Further experiments with an ensemble size of 25 yield essentially identical results. When correlation-based measures of skill are used, the main effect of increasing the ensemble size is to 'sharpen' the definition of the triangular distribution of skill-spread scatter points. This occurs because the ensemble spans more of the region occupied by the PDF in state-space.

The skill of a hindcast is assessed according to (12)–(14). As the period over which skill is assessed decreases, the conditional probability $P$ of (18) begins to resemble more and more that of a perfect model. This occurs because as the hindcast time decreases the influence of model errors diminishes and the coupled model behaves as though it were close to perfect.

(d) **Stochastic forcing/perturbation amplitude**

When the model is forced with stochastic optimals, an increase in the forcing amplitude leads to a shift of the triangular skill-spread scatter up the $S_p$-axis when correlation-based measures of skill are used. In other words, $S_p$ has a tendency to increase for each value of $S_k$ as the forcing amplitude is increased. At high skill (i.e. $S_k \approx +1$), $S_p$ can become non-zero, meaning that even strong warm and cold events can be significantly perturbed by the forcing. When optimal perturbations are used to perturb the initial conditions and their amplitude is increased, there is relatively little effect on the skill-spread relation. Only when the amplitude is increased beyond the limit of realistically sized errors does the distribution of skill-spread points change significantly. At this stage, the hypotenuse of the triangular distribution becomes less well defined as more ensembles move from the class $C_e > 0$ to the class $C_e < 0$ discussed in subsection 4(a).

(e) **Optimal perturbation basic-state**

In the experiments of subsections 4(b) and 4(c), the optimal perturbations used to perturb the ensemble members were computed using only the observed seasonal cycle as the basic state of the coupled system. Therefore, in computing the structure of the optimal
perturbations, the influence of the coupled-model anomalies on optimal perturbation development was ignored. However, the initial structures of the optimal perturbations were found by Moore and Kleeman (1996) to be relatively insensitive to both the phase of the seasonal cycle and the details of the coupled-model anomalies. Therefore, an optimal perturbation computed from the observed seasonal cycle alone evolves in a very similar way to an optimal perturbation computed with non-zero coupled-model anomalies. The rate at which the optimal perturbations can grow, however, depends on the anomalous circulations in the coupled model, although these circulations will exert an influence on all optimal perturbations added to the initial conditions of an ensemble member, regardless of how they are computed. In general, optimal perturbation growth is suppressed by thermocline nonlinearities during the mature phase of an El Niño episode, and suppressed by lower SSTs during La Niña episodes. Increasing SST can enhance growth during the onset of an El Niño (see Moore and Kleeman (1996) for more details). The experiments of subsection 4(b) were repeated using the optimal perturbations of the actual control hindcast to perturb the initial conditions of the ensembles in each case. The resulting distribution of skill-spread scatter points was little affected as compared to the case where the optimal perturbations of the seasonal cycle were used.

(f) Stability of the coupled system

In its normal mode of operation, the dominant coupled oscillation of the model which accounts for ENSO variability is damped in time. One of the parameters in the model which strongly influences the stability of the ENSO coupled oscillation is the ocean–atmosphere coupling strength $\alpha$. Typically, as $\alpha$ increases, the coupled model passes through a bifurcation point where the oscillation that describes the model ENSO becomes unstable*. Kleeman and Power (1994) showed that the skill of the coupled model is relatively insensitive to variations in $\alpha$ even when the amplitude of the ENSO oscillation does not decay in time. To investigate the influence of non-decaying oscillations on the ensemble prediction, we repeated some of the experiments described in section 4, using versions of the coupled model which had a non-decaying ENSO oscillation. In general, when the ENSO oscillation does not decay, the triangular distribution of points becomes more diffuse and the line of minimum skill $H$ less well defined. This occurs because the distribution of the ensembles between the two classes $C_e > 0$ and $C_e < 0$ changes. As the growth rate of the ENSO oscillation increases, more ensembles move into the class $C_e < 0$. This can be understood in terms of a reduction in the height of the peaks in the PDF of Fig. 6(d), or equivalently a reduction in the depth of the potential wells in Fig. 6(e).

6. Summary and conclusions

We have used ensemble forecasting techniques in an attempt to estimate the first and second moments of the PDF of the coupled ocean–atmosphere system in the tropics. Stochastic noise forcing and perturbations were used to perturb each member of the ensembles. The noise forcing and perturbations used were optimal in the sense that they increased the low-frequency ENSO variability of the coupled model and effectively probed the extremes of the PDF by projecting onto the dominant degrees of freedom of the system. The structures of the noise forcing fields and perturbations have particular dynamical significance because they resemble, in a number of ways, the latent–heat flux associated

* The ENSO oscillation is unstable in the traditional linear sense of normal-mode theory when a time-independent basic-state is considered. When a time-dependent basic-state is considered, the ENSO oscillation does not decay in time with increasing $\alpha$, and in this case it must be viewed as an approximation of the most unstable Lyapunov vector (Farrell and Ioannou 1996).
with intraseasonal variability such as the Madden and Julian oscillation and westerly and easterly wind bursts which occur frequently over the western tropical Pacific Ocean (Kleeman and Moore 1997). Using these forcing fields and perturbations, we have identified a relationship between the skill of a model forecast and the spread of an ensemble of forecasts. This relationship was shown to be practically useful for estimating a priori the skill of a forecast of ENSO, and appears to be very robust, in the sense that it is little affected by variations in the ensemble prediction procedure or by changes in some of the model parameters. The only aspect of the ensemble prediction scheme that is critical to its success is the use of stochastic optimals and optimal perturbations to perturb the model. If randomly chosen perturbations (which are sub-optimal) are used to perturb forecast initial conditions, the resulting ensembles are unlikely to span the PDF of the system effectively, and no practically useful relation emerges between the model skill and the ensemble spread.

We have found that scatter plots of the ensemble spread versus the model skill can be interpreted as indicating that, broadly speaking, the coupled model has two preferred states, a warm El Niño state and a cold La Niña state. There are obvious parallels between the two-state PDF of such a system and the many so called ‘regime centroids’ of the atmosphere identified by NWP. In more complex coupled GCMs and in nature, there may be other preferred coupled states or ‘cluster centroids’ of the system, in addition to those considered in the simple model used here. It is likely that these preferred states will be related to the regime centroids of the atmosphere since they are influenced by air–sea interaction. It is known that ENSO strongly influences some of the dominant centroids in the atmosphere, particularly the Pacific–North-American pattern. However, it is also likely that transitions between the atmospheric regime-centroids may also influence the ocean in subtle ways. Experience in NWP has shown that the predictability of the atmosphere is lowest at times when the system undergoes such transitions. The same is likely to be true in the coupled ocean–atmosphere system when the system moves from one preferred state to another. The degree to which the predictability of the system will be affected will depend on how nonlinear these transitions turn out to be. These important issues will have to be investigated using comprehensive coupled GCMs that model all timescales of variability.

Using an empirically derived relationship between the skill of model hindcasts of ENSO and the spread of an ensemble of hindcasts, we can estimate a priori the most likely skill of future forecasts made with the same model. The performance of this skill-spread relation will be tested in an operational setting and the results reported at a future date. Whether the skill-spread relation displayed by the intermediate coupled model used here is general and likely to be found in other models remains to be seen. It would be interesting if the experiments reported here were performed in other coupled models to see if similar skill-spread relations emerge in other operational forecasting systems for ENSO.

ACKNOWLEDGEMENTS

We are indebted to Professor Dan Sorensen at Rice University for making available to us the ARPACK software library used to compute the optimal perturbations described in this paper. Many thanks also to Tim Palmer for his comments and encouragement and for helping to publicize this work. Thanks also to Brian Farrell and two anonymous reviewers for helpful comments on the manuscript, and to Kevin Kohler for his assistance with the graphics. Andrew Moore gratefully acknowledges the support of the Nova Southeastern University Oceanographic Center.
REFERENCES

Batchelor, G. K.

Buizza, R., Tribbia, J., Molteni, F., and Palmer, T. N.

Epstein, E. S.

Farrell, B. F.

Farrell, B. F. and Ioannou, P. J.

Frederiksen, J. C. and Bell, R. C.

Gleeson, T. A.

Gill, A. E.

Hirst, A. C.

Hoffman, R. N. and Kalnay, E.

Ioannou, P. J.

Kleeman, R.

Kleeman, R. and Moore, A. M.

Kleeman, R. and Power, S. B.

Kleeman, R., Moore, A. M. and Smith, N. R.

Kraichnan, R. H.

Legler, D. M. and O’Brien, J. J.

Leith, C. E.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T.

Moore, A. M. and Kleeman, R.

Mureau, R., Molteni, F. and Palmer, T. N.

Murphy, J. M.

Noble, B.

Palmer, T. N.

1953 The theory of homogenous turbulence, Cambridge University Press

1993 Computation of optimal unstable structures for a numerical weather prediction model. Tellus, 45A, 388–407

1969 Stochastic dynamic predictions. Tellus, 21, 739–759

1982 The initial growth of disturbances in baroclinic flow. J. Atmos. Sci., 39, 1663–1686

1990 Small error dynamics and the predictability of flows. J. Atmos. Sci., 47, 2409–2416


1988 ‘Simulations and closure theory for interacting barotropic waves’ in Computational Techniques and Applications CTAC 87, Nooye and Fletcher Eds., North Holland


1983 Lagged average forecasting, an alternative to Monte Carlo forecasting. Tellus, 35A, 100–118

1995 Nonnormality increases variance. J. Atmos. Sci. 52, 1155-1158

1993 On the dependence of hindcast skill on ocean thermodynamics in a coupled ocean–atmosphere model. J. Climate, 6, 2012–2033


1997 A theory for the limitations of ENSO predictability due to stochastic atmospheric transients. J. Atmos. Sci., 54, 753–767

1994 Limits to predictability in a coupled ocean–atmosphere model due to atmospheric noise. Tellus, 46A, 529–540


1959 The structure of isotropic turbulence at very high Reynolds numbers. J. Fluid Mech., 5, 497–543


