Techniques of Linear Prediction, With Application to Oceanic and Atmospheric Fields in the Tropical Pacific

T. P. BARNETT

Scripps Institution of Oceanography, La Jolla, California 92093

K. HASSELMANN

Max-Planck-Institut für Meteorologie, Hamburg, Federal Republic of Germany

The problem of constructing optimal linear prediction models by multivariance regression methods is reviewed. It is well known that as the number of predictors in a model is increased, the skill of the prediction grows, but the statistical significance generally decreases. For predictions using a large number of candidate predictors, strategies are therefore needed to determine optimal prediction models which properly balance the competing requirements of skill and significance. The popular methods of coefficient screening or stepwise regression represent a posteriori predictor selection methods and therefore cannot be used to recover statistically significant models by truncation if the complete model, including all predictors, is statistically insignificant. Higher significance can be achieved only by a priori reduction of the predictor set. To determine the maximum number of predictors which may be meaningfully incorporated in a model, a model hierarchy can be used in which a series of best fit prediction models is constructed for a (prior defined) nested sequence of predictor sets, the sequence being terminated when the significance level either falls below a prescribed limit or reaches a maximum value. The method requires a reliable assessment of model significance. This is characterized by a quadratic statistic which is defined independently of the model skill or artificial skill, As an example, the method is applied to the prediction of sea surface temperature anomalies at Christmas Island (representative of sea surface temperatures in the central equatorial Pacific) and variations of the central and east Pacific Hadley circulation (characterized by the second empirical orthogonal function (EOF) of the meridional component of the trade wind anomaly field) using a general multiple-time-lag prediction matrix. The ordering of the predictors is based on an EOF sequence, defined formally as orthogonal variables in the composite space of all (normalized) predictors, irrespective of their different physical dimensions, time lag, and geographic position. The choice of a large set of 20 predictors at 12 time lags yields significant predictability only for forecast periods of 3 to 5 months. However, a prior reduction of the predictor set to 4 predictors at 10 time lags leads to 95% significant predictions with skill values of the order of 0.4 to 0.7 up to 6 or 8 months. For infinitely long time series the construction of optimal prediction models reduces essentially to the problem of linear system identification. However, the model hierarchies normally considered for the simulation of general linear systems differ in structure from the model hierarchies which appear to be most suitable for constructing pure prediction models. Thus the truncation imposed by statistical significance requirements can result in rather different models for the two cases. The relation between optimal prediction models and linear dynamical models is illustrated by the prediction of east-west sea level changes in the equatorial Pacific from wind field anomalies. It is shown that the optimal empirical prediction is statistically consistent in this case with both the first-order relaxation and damped oscillator models recently proposed by McWilliams and Gent (but with somewhat different model parameters than suggested by the authors). Thus the data do not allow a distinction between the two physical models; the simplest acceptable model is the first-order damped response. Finally, the problem of estimating forecast skill is discussed. It is usually stated that the forecast skill is smaller than the true skill, which in turn is smaller than the hindcast skill, by an amount which in both cases is approximately equal to the artificial skill. However, this result applies to the mean skills averaged over the ensemble of all possible hindcast data sets, given the true model. Under the more appropriate side condition of a given hindcast data set and an unknown true model, the estimation of the forecast skill represents a problem of statistical inference and is dependent on the assumed prior probability distribution of true models. The Bayesian hypothesis of a uniform prior distribution yields an average forecast skill equal to the hindcast skill, but other (equally acceptable) assumptions yield lower forecast skills more compatible with the usual hindcast-averaged expression.

1. INTRODUCTION

The prediction of short-time climate variations in the range from a month to a few years is a problem of great practical significance which has long challenged meteorologists and climatologists. Yet despite numerous studies and forecasting approaches, the ability to predict on these time scales has remained, on the whole, marginal. The question arises whether this low predictability is an intrinsic property of the climate system, or whether at least some climate variables could be predicted with higher skill than presently achieved through the development of improved modeling techniques. We suggest in this paper that in many cases improved predictions can indeed be obtained by the systematic application of a general modeling strategy designed to determine the optimal model, within a given model class, which maximizes the skill while satisfying the condition of statistical significance. Essential for the construction of an optimal model is the joint consideration of the competing requirements of skill and statistical significance (an aspect which has not always received sufficient attention in model construction).

Various types of models have been considered for short-time climatic prediction. High-resolution general circulation models of the atmosphere [cf. *Garp,* 1975], which are used routinely for 1- to 4-day weather forecasts, should in principle also be able to describe the development of climate on longer time scales, if suitably extended to include the dynamics of the more

Paper number 9R0319.
0034-6853/79/009R-0319$01.00

slowly varying components of the climatic system, such as the oceans and sea ice. However, these models become very expensive if integrated for longer periods and are more difficult to calibrate with respect to the longer time scale processes. An alternative approach has therefore been to develop simpler, lower-resolution dynamical models [cf. *Kurihara,* 1970] or energy balance models [cf. *Adem,* 1964, 1975; *Vernekar,* 1975; *Sellers,* 1976] which emphasize the dynamics of the slower parts of the system while parameterizing the more rapid weather scale processes which are of interest only statistically for climatic time scales (see reviews by *Schneider and Dickinson* [1974] and *Saltzman* [1978]). The stronger the simplifications (parameterizations) introduced into these models, the heavier the reliance which must be placed on data to calibrate (tune) the models. Successive simplification then leads naturally to the empirical prediction model, in which only the general structure of the model is specified and all essential parameters of the model are determined from the data. Empirical short-time climate prediction models have generally been formulated as linear models on the assumption that for the time scales of interest the anomaly fields are normally sufficiently small to be treated as perturbations [cf. *Lorenz,* 1956, 1977; *Davis,* 1976, 1977, 1978]. The construction of the model reduces in this case to a problem of linear regression. The limiting form of an empirical model, finally, is the analogue model [cf. *Elliott,* 1951; *Namias,* 1951; *Barnett and Preisendorfer,* 1977, 1978], in which not even the model structure is specified; the prediction of the future evolution of the system is taken simply as the repetition of a past development selected from a longer series of observations on the basis of the resemblance between the initial states of the analogue and the prediction.

We shall be concerned here only with models containing a number of adjustable coefficients, thereby excluding essentially high-resolution general circulation models and analogue methods at the two ends of the model hierarchy. To simplify the analysis, we shall further restrict the discussion to the case of linear prediction models. The basic approach, however, can also be generalized without difficulty to nonlinear models [cf. *Hasselmann,* 1979*a*].

The main difficulty in empirical model construction is that the data set from which the model must be determined is finite. For a hypothetical infinite data set, the best fit empirical model is normally uniquely defined. However, for a finite data sample the basic indeterminacy of the estimates of ensemble-averaged quantities induces an unavoidable indeterminacy of the estimated optimal model. This indeterminacy generally increases with the number of model parameters used and sets a natural limit to the complexity of a model which can be constructed from a finite data set [cf. *Lorenz,* 1956; *Davis,* 1976]. Since the predictive skill, on the other hand, increases with the number of model parameters, the central problem in constructing empirical prediction models is to arrive at a proper balance between the competing requirements of skill and significance. Although the basic methods for treating this problem are well known and can be found distributed in the statistical literature [e.g., *Kashyap and Rao,* 1976], they do not appear to have found general application in the field of climate prediction. In fact, many short-time climatic prediction models derived by a posteriori screening and stepwise regression methods from large initial sets of predictors would probably fail the appropriate multivariate significance test, if correctly applied to the complete predictor set (e.g., see the difficulties discussed by *Harnack and Landsberg* [1978] or the review by *Jones* [1977]).

In view of their general relevance for the climate modeling problem, the basic concepts needed for the construction of prediction models from data are reviewed and summarized in sections 2–3. The construction of optimal predictions should be distinguished here from the problem of linear system identification and simulation [cf. *Box and Jenkins,* 1976; *Kashyap and Rao,* 1976]. If the data set is infinite, so that the linear characteristics of the system can be completely determined from the data, the two problems become essentially identical. However, for finite data sets the structure of the model will depend on whether the model is optimized with respect to prediction (for a given lead time) or with respect to some measure of the fidelity of the model in reproducing the observed overall statistics of the system.

Section 4 describes the construction of optimal prediction models from a model hierarchy. The method differs basically from the frequently used methods of screening, ranking, and stepwise regression by requiring a priori definition of the predictor sequence, as opposed to the a posteriori selection of principal predictors used in the latter techniques.

Section 5 describes the application of the model hierarchy approach to the construction of linear predictions for anomalies in the tropical Pacific. Two examples are given: sea surface temperature (SST) anomaly at Christmas Island and changes (primarily displacements) of the Hadley circulation, for which 95% statistically significant predictions with skill values of the order of 0.5–0.7 are constructed for periods up to 8 and 6 months respectively, into the future. A more comprehensive analysis of predictability in the tropical Pacific, including statistically significant forecasts of El Niño properties, will be presented in a later paper (T. P. Barnett, manuscript in preparation, 1979).

Section 6 discusses the problem of interpreting empirical linear prediction models in terms of linear dynamical models, as represented by linear systems of ordinary differential equations. Since optimal prediction models and dynamic system simulations are normally constructed from different model hierarchies, a one-to-one correspondence between the finite-order truncations of the two types of model cannot generally be expected. Moreover, a basic asymmetry exists between the two modeling approaches: whereas a dynamical model attempts to reproduce the entire statistics of the data set in a single model, a prediction model is designed to filter out only those data properties which are most useful for prediction. These will generally depend on the prediction lead time chosen, so that a prediction model with variable lead time should be viewed as a series of models emphasizing different properties of the predictors and the system response for different lead times. Thus although it is possible to define an associated prediction model for any given dynamical model (simply by determining the optimal prediction model for the simulated data set) the inverse transformation of a prediction model into an equivalent (low order) dynamical model is not generally feasible. The interpretation of an empirical linear prediction model in terms of a (simple) dynamical model can therefore be formulated only as a consistency test: if a dynamical model exists for which the associated prediction model lies within the confidence region of the empirical prediction model, the dynamical model can be regarded as a statistically consistent physical interpretation of the empirical prediction model.

This is illustrated in section 7, in which the empirical prediction of east-west sea level variations in the equatorial Pacific in response to wind variations is interpreted dynamically in terms of the simple first-order relaxation and second-order damped-oscillator models recently proposed by *McWilliams and Gent*

[1978]. The linear prediction solution is found to be statistically consistent with both models, although the parameters of the models differ to some extent from the values proposed by McWilliams and Gent. The second-order damped-oscillator model does not yield a significantly improved fit over the simpler first-order damped system, so that the added complexity of this model (oscillating response characteristics) cannot be justified by the data.

Finally, section 8 addresses the problem of forecast skill. It is generally stated that the forecast skill, defined as the skill of the model when applied to a new independent data set, is lower on the average than the hindcast skill, defined as the skill of the model when tested against the data used to construct the model. The skill of the true model lies midway between the forecast skill and hindcast skill, differing from each by an amount approximately equal to the artificial skill. In a reanalysis of the problem it is noted that the usual form of ensemble averaging on which these relations are based does not in fact correspond to the given side conditions of the problem. It is normally assumed that the true model is given, and the averages of the hindcast and forecast skill are then formed over the ensemble of hindcast and independent data sets. In reality, the hindcast data set is given, whereas the true model is unknown. Under these side conditions, the estimation of the mean forecast skill appears as a problem of statistical inference and depends on the assumed prior distribution of possible true models [cf. *Savage*, 1962]. In particular, Bayes' hypothesis of a uniform prior distribution yields an (likelihood) averaged-forecast skill equal to the hindcast skill. However, equally acceptable alternative prior distributions are found to yield mean forecast skills which differ from the Bayesian estimates by amounts of the order of the artificial skill. It is therefore concluded that the deviation of the mean forecast skill from the hindcast skill cannot be meaningfully determined under the appropriate side conditions of a given estimated model and can only be estimated to be of the same order as the artificial skill.

2. Linear Prediction

We wish to predict the value of a discrete time series $y(t_j)$ $(t_{j+1} - t_j = \Delta t = \text{const})$ p time lags into the future from the past and present values of a set of n time series $x_i(t_j)$ $(i = 1, \cdots, l)$ with the aid of a general linear relation of the form

$$\hat{y}(t_j + p\Delta t) = \sum_{i=1}^{l} \sum_{k=0}^{m-1} D_{ik} x_i(t_j - k\Delta t) \tag{1}$$

Here $\hat{y}$ denotes the predicted or estimated variable as distinct from the true value y.

Equation (1), with constant coefficients D_{ik}, represents the general expression for a time-independent linear predictive system (the additional dependence of D_{ik} on p, which is regarded as fixed, is not shown explicitly). The number of past lags, which theoretically may be infinite, is regarded as finite for practical purposes.

The assumption of a time-dependent system may be questioned in applications to the ocean-atmosphere system, since the anomalies of interest are often superimposed on large seasonal signals. More appropriate to this case would be the generalization to a physical system with an annual periodicity T. This is obtained by replacing the constant coefficient D_{ik} in (1) by the Fourier series

$$D_{ik}(t_j) = \sum_{q=-\infty}^{\infty} D_{ikq} \exp(2\pi i q t_j/T) \tag{2}$$

Although the more general form (2) is probably necessary for successful predictions of ocean-atmosphere anomalies at mid-latitudes, where numerous studies indicate a seasonal dependence of the anomaly structures, in our later applications to tropical anomalies we shall consider only zero'th-order Fourier coefficient D_{ik0}, i.e., the form (1), since at low latitudes the ratio of the anomaly signals to the annual signal is relatively high. In this case the interaction between the seasonal cycle and the anomalies may perhaps be neglected to first order.

The form (1) includes the case in which the predictand y coincides with one of the predictor fields x_i. The equation may also be generalized to a set of simultaneous equations for a number of predictands, and if these are identical with x_i one obtains the problem of the autoprediction of a vector field x_i. With the exception of the discussion in section 6, however, we shall consider here only a single predictand, since the simultaneous prediction of a number of variables follows from the single-predictand problem by straightforward superposition (see, for example, *Jenkins and Watts* [1968]).

For simplicity of notation it is convenient to rewrite (1) in the form

$$\hat{y}(t_j) = \sum_{i=1}^{n} a_i z_i(t_j) \tag{3}$$

where the $l \times m = n$ time series

$$z_1(t_j) = x_1(t_j - p\Delta t), \cdots, z_m(t_j) = x_1(t_j - (p + m - 1)\Delta t)$$
$$\tag{4}$$
$$z_{m+1}(t_j) = x_2(t_j - p\Delta t), \cdots, z_n(t_j) = x_l(t_j - (p + m - 1)\Delta t)$$

are defined such that the predictand and the new predictor series z_i are all taken at the same time t_j.

The optimal linear prediction is defined in the usual least square sense as the set of coefficients $\mathbf{a} = (a_1, \cdots, a_n)$ which minimizes the mean square error

$$\langle \epsilon^2 \rangle = \langle (y - \hat{y})^2 \rangle \tag{5}$$

where $\langle \cdots \rangle$ denotes the average over a (hypothetical) statistical ensemble of time series. The solution is given by

$$a_i = \sum_{j=1}^{n} Z_{ij}^{-1} \langle z_j y \rangle \tag{6}$$

where $Z_{ij} = \langle z_i z_j \rangle$ is the predictor covariance matrix.

In practice, it is usually convenient in evaluating (6) to redefine the predictors as the coefficients of a set of normalized empirical orthogonal functions (EOF's) or principal components [cf. *Pearson*, 1901; *Hotelling*, 1933; *Lorenz*, 1956; *Davis*, 1976; *Kutzbach*, 1967; *Barnett and Preisendorfer*, 1977, 1978]. These are obtained by a rotation of the predictor space with subsequent scaling such that the transformed predictor set z_i' is statistically orthonormal, $\langle z_i' z_j' \rangle = \delta_{ij}$. The prediction coefficients in this new system are simply $a_i' = \langle y z_i' \rangle$.

The quality of the prediction is generally characterized by the skill

$$S = 1 - (\langle \epsilon^2 \rangle / \langle y^2 \rangle) \tag{7}$$

which represents the fraction of the variance of y predicted by the model. For zero predictability, $S = 0$, while for zero prediction error $\epsilon = 0$, i.e., perfect predictability, $S = 1$. Using (3), (5), and (6) in (7) gives

$$S = \sum_{i,j=1}^{n} \frac{a_i a_i \langle z_i z_i \rangle}{\langle y^2 \rangle} \tag{8}$$

or for the z_i' predictor set

$$S = \sum_{i=1}^{n} \frac{a_i'^2}{\langle y^2 \rangle} \tag{9}$$

3. MODEL SIGNIFICANCE

a. The ρ^2 Statistic

In practice, the hypothetical ensemble averages $\langle \cdots \rangle$ must be estimated from averages $[\cdots]$ (normally time averages) taken over a finite data sample. The sampling errors $\delta Z_{ij} = [z_i z_j] - \langle z_i z_j \rangle$ and $\delta K_i = [z_i y] - \langle z_i y \rangle$ incurred in the estimation of the covariance matrices will then induce errors $\delta a_i = \tilde{a}_i - a_i°$ of the estimated optimal coefficients $\tilde{a}_i$ relative to the true optimal model $a_i°$. Normally, the indeterminacy of the model associated with the sampling errors increases with the number of predictors used. Thus attempts to increase the model skill by increasing the predictor number will generally be offset in practice by the accompanying reduction in significance of the higher-order models [*Lorenz*, 1956; *Davis*, 1977].

The central problem in model construction is therefore finding a satisfactory balance between the conflicting requirements of significance and skill. For this purpose a reliable measure of model significance is required. A statistic which has often been used in this context is the 'artificial skill' [cf. *Davis*, 1977, 1978]. However, it will be shown below that this quantity does not provide a reliable measure of model significance, since it is not based directly on the probability density of the model variables. The appropriate statistic for significance tests follows from consideration of the joint probability distribution of the estimated model coefficients.

If the coefficients are estimated from averages over a fairly large number of quasi-independent samples, the errors will be small and, by the Central Limit Theorem, approximately Gaussian. Thus the probability distribution of estimated models $\tilde{\mathbf{a}}$, given the true model $\mathbf{a}°$, is of the form

$$p(\tilde{\mathbf{a}}/\mathbf{a}°)\, d\tilde{\mathbf{a}} = (2\pi)^{-n/2} |M|^{-1/2} \exp(-\rho^2/2)\, d\tilde{\mathbf{a}} \tag{10}$$

where

$$\rho^2 = \sum_{i,j=1}^{n} M_{ij}^{-1}(\tilde{a}_i - a_i°)(\tilde{a}_j - a_j°) \tag{11}$$

$$M_{ij} = \langle \delta a_i \delta a_j \rangle \tag{12}$$

and we have assumed $\langle \delta a_i \rangle = 0$ to first order.

By Taylor expansion of (6), the covariance matrix M_{ij} can be expressed in terms of the second moments of δZ_{ij} and δK_i, and these in turn can be estimated from the time-lagged covariance functions of the variables z_j and y using standard results of covariance sampling theory [cf. *Jenkins and Watts* [1968] and appendix, this paper). An exact evaluation of M_{ij} requires information on the fourth moments of z_j and y, but these are normally expressed in terms of the second moments by assuming that the processes z_j and y are approximately Gaussian.

Equation (10) represents the theoretical probability distribution of models $\tilde{\mathbf{a}}$, estimated from individual data realizations, given the true ensemble-averaged moments and the associated true model $\mathbf{a}°$. In practice, one faces the inverse situation in which one has only a single data realization and would like to determine the set of possible true models which are compatible, within given statistical limits, with the estimated moments and the associated estimated model $\tilde{\mathbf{a}}$. This is a general problem of statistical infeténce which we shall return to in more detail in section 8 in the context of the estimation of the forecast skill and its relation to the hindcast skill and true skill. For the purposes of this section, however, we need to recall only the well-known definitions of the confidence region and significance level of an estimated model.

For a given (unknown) true model $\mathbf{a}°$, the 'γ probability region' $R(\mathbf{a}°)$ of estimated models $\tilde{\mathbf{a}}$ around $\mathbf{a}°$ is defined as a region such that

$$\int_R P(\tilde{\mathbf{a}}/\tilde{\mathbf{a}}°)\, d\tilde{\mathbf{a}} = \gamma$$

where $\gamma < 1$ represents some prescribed probability level. Conversely, for a given estimated model $\tilde{\mathbf{a}}$ one can then introduce a 'γ confidence region' $\tilde{R}(\tilde{\mathbf{a}})$ around $\tilde{\mathbf{a}}$ as the set of all true models $\mathbf{a}°$ whose associated probability regions $R(\mathbf{a}°)$ contain the given estimated model $\tilde{\mathbf{a}}$. In practice, it can be assumed that for sufficiently small sampling errors $\delta\mathbf{a} = \tilde{\mathbf{a}} - \mathbf{a}°$, the probability distributions $p(\delta\mathbf{a})$ for the relative errors remain the same for different $\mathbf{a}°$ so that the probability distributions $p(\tilde{\mathbf{a}}/\mathbf{a}°)$ for different $\mathbf{a}°$ differ only in the position $\mathbf{a}°$ of the maximum.

The shapes of the regions R and $\tilde{R}$ have not yet been specified. To complete the definitions, we require now that the region R is limited by a surface of constant probability density p = const or, equivalently, ρ^2 = const. This yields an ellipsoid for R, and $\tilde{R}$ is then the same ellipsoid with its center displaced to $\tilde{\mathbf{a}}$ (see Fig. 1). The definition is optimal in the sense that it yields the smallest regions R, $\tilde{R}$ for a given confidence level γ, and ensures that all regions within R have higher probability density than regions outside. Additionally, the definition is invariant with respect to linear transformations of the predictors. (This follows from the fact that ρ is a scalar invariant obtained by contracting tensor products, or more simply because linear transformations affect probability densities only by a constant factor, the transformation Jacobian, so that surfaces of constant density remain invariant under such transformations.) The invariance property is clearly a necessary requirement for all linear models in which no distinction is made between different representations of the predictor set.

The mean radius ρ_γ of the ellipsoid can be found by considering the one-dimensional probability density associated with the n-dimensional distribution (10) with respect to the distance coordinate ρ. This is given by a χ^2 distribution with n degrees of freedom,

$$p(\rho^2)\, d\rho^2 = (\Gamma(n/2)2^{n/2})^{-1}(\rho^2)^{(n/2)-1} \exp(-\rho^2/2)\, d\rho^2 \tag{13}$$

and the γ confidence radius ρ_γ defined by

$$\int_0^{\rho_\gamma^2} p(\rho^2)\, d\rho^2 = \gamma$$

can be obtained from standard tables. For large n

$$\rho_\gamma^2 \simeq n + b_\gamma(n)^{1/2} \tag{14}$$

where b_γ is independent of n.

The confidence region $\tilde{R}$ is useful primarily for identifying acceptable true models in significance tests. In particular, if $\tilde{R}$ contains the trivial zero-prediction model $\mathbf{a}° = 0$, the estimated model is said to be statistically nonsignificant at the γ confidence level (in other terminology: at the $1 - \gamma$ confidence level). In the example shown in Figure 1*a*, the estimated model is therefore statistically significant at the γ confidence level, since $\tilde{R}$ does not include the point $\mathbf{a}° = 0$, whereas the estimated model in Figure 1*b* cannot be distinguished statistically from the zero-skill prediction model $\mathbf{a}° = 0$. Thus although the

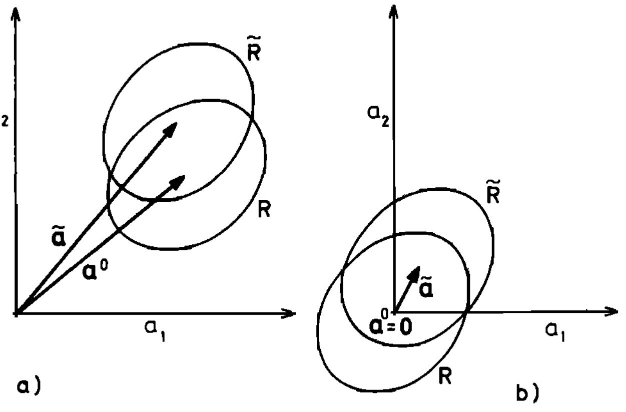

Fig. 1. The 95% probability region R and confidence region $\tilde{R}$ for true models $\mathbf{a}^\circ$ and estimated model $\tilde{\mathbf{a}}$, respectively: (*a*) corresponds to a statistically significant model, (*b*) to a statistically insignificant model, for which the zero-prediction true model $\mathbf{a}^\circ = 0$ lies within $\tilde{R}$.

second model can appear quite successful when evaluated in terms of skill, the skill values must be rejected as fictitious on the basis of the independent significance test.

In applying (13) and (14) in practice it should be remarked that for a more rigorous analysis one should consider the significance levels of the variable $\tilde{\rho}^2 = \tilde{M}_{ij}^{-1}\tilde{a}_i\tilde{a}_j$ rather than $\rho^2 = M_{ij}^{-1}\tilde{a}_i\tilde{a}_j$, where $\tilde{M}_{ij} = [\delta a_i \delta a_j]$ is the estimated rather than true covariance matrix of the model errors. The χ^2 distribution (13) is then replaced by Hotelling's distribution, the multidimensional generalization of Student's distribution [cf. *Kendall and Stuart*, 1966]. However, for reasonably large n, which is the main case of interest here, the relations (13) and (14) represent adequate approximations.

b. Hindcast Skill and Artificial Skill

Sampling errors will also affect the estimates of skill. Replacing the true moments and coefficients by the corresponding estimated values in the expression (8) for the true skill S_0 one obtains the estimated or hindcast skill

$$S_H = \sum_{i,j} \frac{\tilde{a}_i\tilde{a}_j[z_iz_j]}{[y^2]} \tag{15}$$

Since S_H is a positive definite expression, it will generally be positive even for a zero-prediction true model, $\mathbf{a}_i^\circ = 0$. The hindcast skill in this case is termed the artificial skill

$$S_A = S_H(\mathbf{a}^\circ = 0) = \sum_{i,j} \frac{\delta a_i \delta a_j[z_iz_j]}{[y^2]} \approx \sum_{i,j} \frac{\delta a_i \delta a_j\langle z_iz_j\rangle}{\langle y^2\rangle} \tag{16}$$

For small skill values one readily finds [*Lorenz*, 1956; *Davis*, 1976]

$$\langle S_H\rangle \approx S_0 + \langle S_A\rangle \tag{17}$$

If the model is statistically significant, the estimated hindcast skill S_H may be expected to exceed the estimated average artificial skill $\langle S_A\rangle$ by a 'significant' factor. However, while a comparison of S_H against $\langle S_A\rangle$ provides a general indication of model significance, the quantities S_H and $\langle S_A\rangle$ can be misleading if used as a quantitative measure of significance in place of ρ^2.

Statistical significance tests based on the ratio of the quadratic forms S_H and $\langle S_A\rangle$ are equivalent to defining probability regions R_S in the model parameter space bounded by surfaces of constant S_A rather than constant ρ^2. The difference between the two significance tests is indicated in Figure 2, which shows the corresponding 95% probability regions R_S and R for the null hypothesis of a zero-prediction model. Model A, although in a region of very small probability density well outside the 95% probability region, and therefore highly significant, would fail the S_A significance test, whereas the statistically insignificant model B within the 95% probability region would be accepted by this test.

The difference between the probability ellipsoids bounded by constant S_A or constant ρ^2 can be quite significant. In the cases considered in section 5, for example, differences in the ratios of the major axes of the quadratic forms S_A and ρ^2 by factors of 2 or 3, together with large changes in the axis orientations, were not uncommon. However, it should be noted that these differences are a characteristic property of time series with finite correlation lags, in which the ensemble averages are estimated by taking continuous time averages over a finite record length. If the ensemble averages are estimated from a finite number of statistically independent samples (e.g., if the processes y and z_i represent white noise) the artificial skill metric $\langle z_iz_j\rangle/\langle y^2\rangle$ and the metric $\langle \delta a_i \delta a_j\rangle^{-1}$ of the ρ^2 statistic can be shown to coincide. This may explain in part why statistics characterizing model significance and model skill are not always clearly distinguished (e.g., in the definition of 'stopping rules' in the stepwise construction of models [cf. *Mosteller and Tukey*, 1977]).

4. SIGNIFICANCE VERSUS SKILL: MODELING STRATEGIES

In the discussion in the preceding sections it was assumed that the structure of the model, i.e., the set of predictors, was given. We turn now to the problem of choosing the predictors. As has been pointed out, the skill of a model increases, whereas the statistical significance generally decreases, as the number of predictors is increased. Thus a strategy is needed to determine the largest number of predictors which can be incorporated in a model while still retaining a statistically significant solution. A number of such strategies have been suggested. Essential for a correct application of all methods is a careful distinction between the a priori and a posteriori selection of predictors. A posteriori predictor selection, or screening, is permissible only if the model has first passed a statistical significance test for the entire predictor set; both the skill and the statistical significance of the screened model are then always lower than for the original model. This fact has often been overlooked in constructing prediction models and has resulted in erroneously optimistic estimates of the statistical significance of screened models. A statistically consistent ordering strategy aimed at maintaining high statistical signifi-

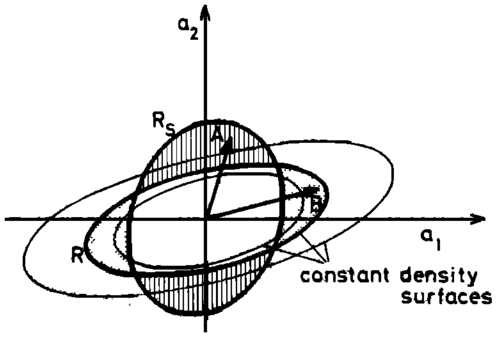

Fig. 2. Relation between 95% probability regions R and R_S limited by constant ρ^2 and S_A, respectively, for the zero-prediction model $\mathbf{a}^\circ = 0$. The model estimate A is statistically significant, but would fail a significance test based on the S_A statistic. Conversely, the model estimate B is statistically insignificant but would pass the S_A statistic test.

cance must therefore be based on a priori predictor selection. This requires, unavoidably, some a priori hypothesis regarding the relative importance of different predictors. In the present context the terms a posteriori and a priori simply denote, respectively, with and without use of correlation information between predictand and predictors; a priori selection using correlation information between predictors only (e.g., by ordering in an EOF sequence) is permissible.

a. Screening Strategies

The common starting point of all screening (winnowing, ranking, and stepwise regression) techniques is the (at least conceptual) existence of a comprehensive optimal prediction model containing a large number of predictors. From this a truncated model is then derived by retaining only the most 'important' components of the complete coefficient vector. The ordering of coefficients is generally based either on separate statistical significance tests for each coefficient (after transformation to a coefficient set with statistically orthogonal estimation errors) or on the individual contributions of the coefficients to the net skill (in the case of stepwise regression techniques). Screening techniques have been applied widely in the construction of geophysical models [cf. *Gilbert,* 1971], medium- and extended-range weather predictions [cf. *Jones,* 1977], and (implicitly) in the definition of the significant response regions in numerical atmospheric response experiments using general circulation models [cf. *Garp,* 1975; *Hasselmann,* 1979].

The basic shortcoming of these techniques is that they are based on the a posteriori selected of predictors. To avoid biasing the statistics, however, the significance test of a model must be based on the complete predictor set, prior to truncation. In the full predictor space the complete coefficient vector is statistically significant at the γ significance level if the end point of the vector lies outside the ellipsoidal γ probability region, centered at the origin, associated with the null hypothesis of zero predictability. Coefficient screening then defines a new coefficient vector in which the smallest components of the original vector, as determined with respect to some particular coordinate system, are set equal to zero. While the new vector may still be statistically consistent with the data, provided it lies within the confidence region $\tilde{R}$ of the original coefficient vector, both the skill and statistical significance of the new model, defined with respect to the original predictor space, will necessarily be reduced. In terms of the subspace of predictors retained in the truncated model the apparent significance of the model is increased, but this is a statistically impermissible measure of significance, since it is based on a posteriori, i.e., biased data selection.

To illustrate the inherent shortcomings of a posteriori screening from another viewpoint, consider the case of n orthogonal predictors z_i, of which all except the first represent noise which is uncorrelated with y. The estimated prediction coefficients are accordingly of the form $\tilde{a}_1 = a_1{}^0 + \delta a_1$, $\tilde{a}_1 = \delta a_i$ for $i \geq 2$. For simplicity, we assume that the error metric M_{ij} of the quadratic form ρ^2 is unity.

For a statistically significant model, the generalized distance ρ of the estimate $\tilde{\mathbf{a}}$ from the origin must satisfy the inequality (14)

$$\rho^2 = (a_1{}^0 + \delta a_1)^2 + \sum_{j=2}^{n} (\delta a_j)^2 > \rho_\gamma{}^2 \approx n + b_\gamma(n)^{1/2}$$

For large n the first term in the expression for ρ^2 is small in comparison with the other terms in the equation, so that ρ^2 is dominated by the noise, and the model will normally fail the significance test. According to the usual screening philosophy, one could now argue that the significant first predictor could nevertheless have been readily distinguished from the irrelevant noise components simply by testing the statistical significance of each prediction coefficient individually. However, this technique is not invariant with respect to linear transformations of the predictor coordinate system. An arbitrary rotation, for example, will generally immerse the first predictor in linear combinations containing a large number of noise components; in the new coordinate system all prediction coefficients would then appear statistically insignificant. Conversely, given an unpredictable system, it is always possible to rotate the predictor space such that, for a particular realization, the new axis for the first predictor component lies in the direction of the estimated (spurious) coefficient vector $\tilde{\mathbf{a}}$. In the new coordinate system only the first coefficient is then nonzero, and, if tested by itself, will normally appear highly significant. We conclude that individual coefficient screening is meaningless in a system in which no significance is attached a priori to a particular coordinate system.

b. Nested-Model Hierarchies

From the above example it is clear that in testing the statistical significance of individual predictors or predictor sets the tested components must be specified a priori. Thus the first predictor could have been identified in this example as a statistically significant predictor in the presence of a large number of noise components if, and only if, the hypothesis had been made a priori that the first predictor, as opposed to some other linear combination of predictors, was statistically significant when tested for itself (i.e., in a prediction model containing only this component as predictor).

The requirement of a priori predictor ordering leads naturally to a nested-model hierarchy. To determine the model with the highest skill which is still statistically significant at some prescribed confidence level one can consider a nested sequence of models of increasing order, in which each model of the sequence is constructed from the previous member of the sequence by the inclusion of one (or more) additional predictor(s). If the ordering of the predictors is chosen in accordance with the expected contribution of the predictors to the net skill, the statistical significance of the resultant sequence of optimal models (as determined by a χ^2 test on ρ^2) may be expected to decrease more or less monotonically with the order of the model.

Beyond some order, the significance will then fall below the preselected confidence level, and the cutoff point then defines the highest-order model and thereby, since the skill increases monotonically with the model order, the highest-skill model, which is still statistically acceptable at the prescribed significance level. The success of the technique clearly depends strongly on the degree of a priori insight into the relative significance of candidate predictors. 'Incorrect' ordering can mask all potential predictors by noise, and the entire sequence of models can become statistically insignificant. Unfortunately, it lies in the nature of objective statistical tests that once an 'incorrect' ordering choice is made there exists no a posteriori technique for recovering the 'true' predictors from the noise.

Alternative cutoff criteria other than a fixed critical significance level can be chosen to terminate the model sequence. For example, the cutoff point can be taken as the model which yields the highest significance value. This is a rather stringent condition which eliminates not only noise but also predictors which, although less significant than the lower-order components, may still yield a significant contribution to the prediction. A less stringent cutoff, although still more conservative than a fixed significance level, may be defined in terms of the rate of decrease of significance with model order. In this case the model sequence is terminated when the critical significance level is still exceeded, but the rate of decrease of significance becomes comparable with the rate of decrease which would result from the addition of predictors consisting of uncorrelated noise.

Another class of cutoff rules focuses on the incremental increase in skill obtained by sequentially adding predictors or on the confidence limits of individual predictor coefficients [cf. *Jenkins and Watts*, 1968; *Box and Jenkins*, 1976; *Kashyap and Rao*, 1976; *Mosteller and Tukey*, 1977]. However, these techniques are not based on the joint probability distribution of the coefficient estimation errors, as characterized by the ρ^2 statistic, and therefore suffer from the basic shortcomings discussed under section 3*b*.

Finally, the number of predictors in a model can be limited by further criteria, in addition to the significance cutoff condition. In working with EOF predictor sequences we have found it useful to limit the set of predictors to the EOF components whose eigenvalues (at the 95% significance level) exceed the values which would be expected by chance for a set of statistically independent predictors [cf. *Preisendorfer and Barnett*, 1977, 1978]. In practice, the combination of this a priori predictor filtering technique with a fixed-significance level cutoff was found to be roughly equivalent to applying a maximum significance cutoff, or a cutoff based on the rate of decrease of significance, to the complete EOF predictor set.

c. Reordering of Predictors

Besides the entirely a posteriori screening methods based on the unrestricted selection of suitable linear combinations of predictors and the converse technique of completely ordering the predictors a priori in a nested model hierarchy an intermediate approach is sometimes adopted in which the a posteriori selection of predictors is permitted, but the set of linear transformations employed in selecting the predictors is restricted to permutations. This reordering technique appears appropriate if the predictors represent genuinely distinct, statistically independent physical quantities. In this case it may be expected that some of the predictors are directly coupled with the predictand, whereas others represent independent variables of no predictive value, and one would naturally like to separate the predictors into these two classes. However, the statistical significance of the resultant reordered predictor sequence and, in particular, the cutoff point separating the statistically significant predictors from the statistically insignificant components must again be judged with respect to the joint probability distribution of the complete set of predictor coefficients.

Consider, for example, the case that all predictors and all coefficient estimation errors are statistically independent. Let the predictors be normalized such that the coefficient estimation errors are unity, $\langle \delta a_i^2 \rangle = 1$, and let α be the value of the largest estimated coefficient. The probability p_α that the largest coefficient is greater or equal to α by chance (i.e., assuming that the null hypothesis $a_i^0 = 0$ is valid) is equal to one minus the probability that all estimated coefficients are less than α, or

$$p_\alpha = 1 - w_\alpha^n$$

where

$$w_\alpha = \int_{-\alpha}^{\alpha} \frac{e^{-x^2/2}}{(2\pi)^{1/2}}\, dx$$

and n is the number of predictors.

The largest coefficient can therefore be regarded as statistically significant at the γ significance level if $\alpha \geq \beta$ where

$$p_\beta = (1 - \gamma)$$

or

$$w_\beta = \gamma^{1/n}$$

In contrast, a test of the statistical significance of a single predictor which had been specified a priori would yield the less stringent condition $\alpha \geq \beta'$, where $w_{\beta'} = \gamma$. For example, at the 95% confidence level, the a priori significance test yields a critical level $\beta' \approx 1.96$, whereas for the largest coefficient of a set of n predictors one obtains for $n = 20$, $\beta = 3.0$, and for $n = 100$, $\beta = 3.5$. Thus a priori ordering is to be preferred, if physical selection criteria can be found.

In the general case one needs to consider the statistical significance of the set of p largest coefficients $\alpha_1 \geq \alpha_2 \geq \alpha_3 \cdots \alpha_p$. With increasing p, the probability calculations rapidly become very complex. However, an appropriate technique for estimating the significance cutoff point of a reordered predictor sequence can be developed using Monte Carlo simulations of the statistics of the null hypothesis, following the approach used by *Preisendorfer and Barnett* [1977] in their similar investigation of the distribution of eigenvalues of EOF's estimated from a finite data sample of uncorrelated variables.

In the following examples the reordering technique was not applied since all predictors were intercorrelated, and there appeared to be no justification for ordering predictors a posteriori on the basis of one particular predictor representation, as opposed to some alternative linearly transformed representation. In this case the only recourse for eliminating statistically insignificant predictors from a large predictor set is to consider an a priori ordered predictor sequence in a predefined model hierarchy.

5. Prediction of SST and Wind Field Anomalies in the Tropical Pacific

As application, we consider now the construction of maximum-skill predictions for SST and wind field anomalies in the tropical Pacific by using a nested model hierarchy.

Long-term interactions in the tropical Pacific have been the subject of numerous investigations, both because of their relevance to the El Niño phenomenon and, more generally, because air-sea interactions in the tropics play an important role in controlling the principal energy fluxes driving the global atmospheric circulation. Past studies of these interactions have been concerned largely with the general structure of the anomaly fields, either in terms of specific event analyses or in terms of overall anomaly statistics, and with simple feedback models which were then proposed to explain the principal features observed. However, there appears to have been no systematic attempt to construct maximum-skill prediction models within a general statistical framework independent of the detailed structure of particular physical interaction models.

We shall present here only a few typical examples of optimal model construction, four of which are discussed in this section and a further case in section 7. A general application of the

TABLE 1. Characteristics of Predictors

Predictor Number	Variable	Station	Location	Time Span
1	sea surface temperature	Tumaco, Colombia	2°N, 79°W*	1951–1970
2		Talara, Peru	4°S, 81°W	1942–1970
3		Galapagos, Ecuador	1°S, 90°W	1951–1970
4		Christmas Island	2°N, 157°W	1954–1970
5–7	sea level	eigenmodes 1, 2, and 3 of equatorial sea level†		1950–1970
8–17	wind field	areally averaged u and v components for the regions shown in Figure 3		1950–1972
18	other	equatorial wind stress τ_χ*		1950–1972
19		southern oscillation index [*Quinn,* 1974]; the sea level pressure difference between Easter Island and Darwin, Australia		1950–1972
20		North Equatorial Countercurrent index [*Wyrtki,* 1974]; the sea level difference between Christmas and Kwajalein islands		1950–1970

Predictor set C: all 20 predictors at time lags of 1, 2, 3, 4, 5, 6, 9, 12, 15, 18, 24, and 30 months. Predictor set R: SST at Talara, first equatorial sea level EOF, τ_χ, and δu_2 (Figure 3) at time lags of 1, 2, 3, 4, 5, 6, 9, 12, 15, and 18 months.

*See Figure 3.

†See *Barnet* [1977*b*] for definition and Figure 3 for station locations.

technique to a comprehensive set of anomaly fields will be given in a later paper (T. P. Barnett, manuscript in preparation 1979), in which it is shown that significnt predictions with maximum lead times ranging from 6 to 12 months, and sometimes beyond, can be constructed for a variety of anomaly fields, including variables which are directly associated with El Niño.

Our first four prediction examples involve two predictands T and V and two sets of predictors C and R. The predictands represent the anomaly T of monthly sea surface temperature at Christmas Island and a wind anomaly variable V associated with meridional displacements of the Hadley circulation. The variable T is representative of mid-ocean, near-equatorial water temperatures over a large region of the central and eastern Pacific. The variable V formally represents the amplitude of the second EOF of the meridional component of the Trade Wind Field; its close relation to the position of the Intertropical Convergence Zone (ITCZ) and the associated changes of the Hadley circulation in the central and eastern Pacific is shown in *Barnett* [1977*a*].

The first predictor set C consists of a comprehensive set of 20 time series of oceanic and atmospheric variables obtained from various stations (see Table 1 and Figure 3) which were considered as potentially useful predictors, without regard to existing physical feedback hypotheses. Where good spatial coverage was available for a given data set (e.g., wind field), the number of predictors was reduced by spatial averaging over coherent high-variance regions determined by an EOF analysis, or by retaining only the first few components of an EOF representation of the field. Each of the resultant 20 series was taken at $m = 12$ different time lags (1 to 30 months), yielding a total of $20 \times 12 = 240$ (highly correlated) predictors z_i as defined by (3). The second set of predictors R represents a strongly reduced subset of the comprehensive set C and contained only four time series considered to be most important to the prediction of T and V on the basis of the physical feedback theories developed by *Bjerknes* [1966] and *Wyrtki* [1975], and tested in *Barnett* [1977*b*]. Ten time lags were taken for each series, yielding a total of $4 \times 10 = 40$ (again correlated) predictors z_i. (The objection that the physical models were proposed after visual inspection of data and that our choice of the set R cannot therefore be regarded as a strictly a priori selection must be accepted. This typifies a general dilemma in constructing unbiased prediction models from limited data (cf. discussion following paper by *Hasselmann* [1979] at the Helsinki conference). In the present case, however, the physical hypotheses originally proposed by Bjerknes were based on data sets other than those used here, namely, a few years of sea level pressures, rather than 20-year wind field records. Nevertheless, the possibility of some indirect biasing means that our statistical significance limits for the experiments TR and VR should be regarded conservatively as upper limits.)

All series z_i were normalized to unit variance and were then transformed by rotation to a new set of orthogonal predictors. This set, ordered in decreasing variance, then defined the nested-model sequence. Although ordering with respect to variance is standard practice when no other criteria are available, it should be noted that the structure of the EOF's in the present case depends on the (arbitrary) common normalization of anomalies of different dimension, location, and lag time. The physical interpretation of the EOF's in such a composite variable space-time space is accordingly more difficult than in the more familiar EOF analysis with respect to a single field at zero relative time lag. It may be expected, however, that such a composite variable EOF analysis will identify the principal space-time variance patterns of the predictor set and that these will prove most useful for constructing predictions. For the reduced predictor set R this ordering did indeed produce a successful nested model sequence yielding a reasonably well-defined maximum-skill statistically significant model at a critical cutoff order, as defined by the criteria discussed in section 4*b*.

The results for the four predictions TC, TR, VC, and VR are summarized in Figures 4–12. We discuss the results here primarily from the standpoint of applications of linear prediction theory; a more detailed consideration of general predictability in the tropics, key predictors, etc. will be presented by T. P. Barnett (manuscript in preparation, 1979).

Figures 4 and 5 show the confidence parameter ρ^2 ((11) with $\mathbf{a}^\circ = 0$) as a function of model order for the prediction of T 8 months into the future and V 6 months into the future for the C and R predictor sets. Also shown are the expected values of ρ^2, $\langle\rho^2\rangle = n$, and the 90% and 95% confidence levels. The figures clearly demonstrate that the inclusion of too many predictors in the initial formulation of a prediction model will generally result in a statistically meaningless prediction, unless the anticipated effective predictors are specified a priori. The a posteriori extraction of significant predictors from the comprehensive predictor set C, which failed to yield statistically significant predictions for either T or V at any model order (for 8- and 6-month lead times), is not permissible. However, statistically meaningful predictions were obtained in these cases by starting from a new reduced predictor set R.

Figures 6 and 7 show the hindcast and estimated artificial skill for the 8- and 6-month predictions of T and V as a function of model order. For the comprehensive data set C the artificial skills S_A are comparable with the hindcast skills S_H, again suggesting but not proving (cf. section 3*b*) that the predictions with this set are statistically insignificant.

It may be noted that Figures 4 and 5 indicate that the first few predictors in both models *TR* and *VR* are not, by themselves, significant. The significance is built up by the following predictors before decreasing again for higher orders. Thus the ordering of predictors with respect to variance was not 'optimal.' This again demonstrates the critical nature of the unavoidably arbitrary a priori ordering of predictors. It is naturally tempting to 'improve' Figures 4 and 5 by reordering the predictors with respect to their contribution to the net significance. However, the statistical significance and cutoff point of the reordered sequence would then need to be investigated by independent tests, as discussed in section 4*c*. After the significance of a model of a particular order n has been established for the prespecified predictor sequence, it is of course permissible to formally reorder the predictors for that model with respect to their significance or skill contribution (in accordance with *Davis*' [1978] use of 'principal predictors'). However, a truncation of this reordered sequence would decrease both the skill and significance of the model. Within an a priori chosen predictor space, the optimal model is represented by the complete n-dimensional coefficient vector, and all truncation techniques by projection on to a posteriori defined linear subspaces necessarily represent distortions of the optimal model.

The relation between significant predictability, lead time,

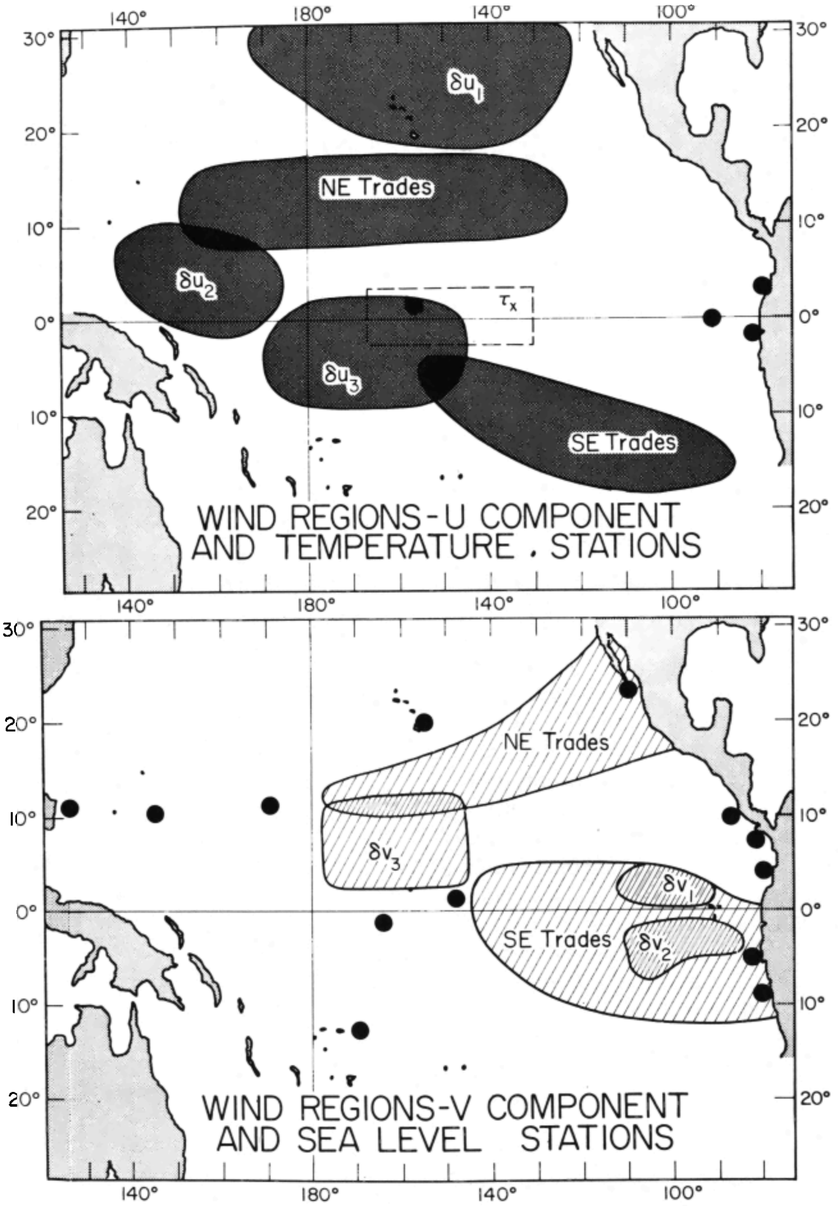
140°
180°
140°
100°
30°
20°
10°
0°
10°
20°
δu1
NE Trades
δu2
τx
δu3
SE Trades
WIND REGIONS-U COMPONENT AND TEMPERATURE . STATIONS
NE Trades
δv3
δv1
SE Trades
δv2
WIND REGIONS-V COMPONENT AND SEA LEVEL STATIONS

Fig. 3. Distribution of stations and averaging areas used in prediction models. See Table 1 and *Barnett* [1977*a*, *b*] for more detail.

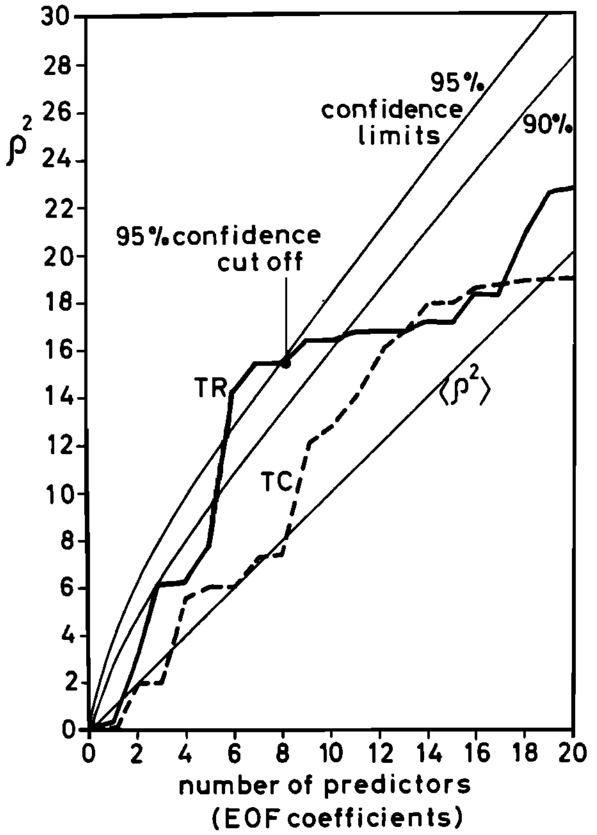

Fig. 4. Confidence parameter ρ^2 for prediction of Christmas Island SST anomaly T 8 months in advance. Case TC: comprehensive predictor set C, 20 time series with 12 time lags per series. Case TR: reduced predictor set R, 4 time series with 10 time lags.

and number of predictors for the four prediction models is shown in Figures 8 and 9. (For simplicity, we have used a fixed 95% significant level cutoff. The results are not significantly affected by the use of alternative cutoff criteria.) The limit of significant predictability is seen to be greatly enhanced by the introduction of the reduced predictor set. The hindcast and artificial skill curves associated with Figures 8*b* and 9*b* are shown in Figure 10. Examples of the hindcast time series versus the actual observations are shown in Figures 11 and 12. In all cases the agreement between hindcast and observation is quite good. However, they also demonstrate the danger of estimating the quality of a prediction from visual inspection of hindcast time series or from the computed skill, without regard to the model significance. Thus Figure 12 shows the two 6-month predictions of V using the first 20 EOF coefficients of C and the first four EOF coefficients of R; the former hindcast accounts for 67% of the record variance, but the ρ^2 value is below even the average value expected by chance, and the model must therefore be rejected as insignificant already at the 50% confidence level. In contrast, the 6-month prediction of V using only four prediction coefficients from the R set is significant well above the 95% level, although it accounts for less of the variance (37%) than the VC model. Thus a model's ability to 'account for variance' is not directly related to model reliability (significance).

In summary, the comprehensive models C, in conjunction with the EOF nesting strategy, failed to reveal significant predictability for either T or V beyond a few months, even though these models accounted for a large proportion of the variance in T and C. Thus the predictability of the system, if present, was being masked by the inclusion of irrelevant predictors. However, a reduction of the predictor set based on a physical feedback model suggested by *Bjerknes* [1966] and others led to a considerable enhancement of statistical significance, prediction lead time, and (significant) skill. These results underline the need, already emphasized by *Lorenz* [1956] and *Davis* [1977], to limit the set of predictors at the outset of model construction, and demonstrate that a successful application of a sequential model construction strategy is dependent on the suitable a priori choice and ordering of the predictors.

6. Relation Between Empirical Prediction Models and Linear Dynamical Models

The interpretation of an empirical prediction model in terms of the general dynamics of the system is not a straightforward problem. If the system is nonlinear, the exact dynamical solution will often be unknown. Indeed, the motivation for the construction of an empirical prediction model in the first place is usually the inability to cope with the full set of dynamical equations. However, it may be useful to ask whether an empirical linear prediction model may be interpreted in terms of a simple linear dynamical system (i.e., a set of linear differential equations), which may then be regarded as an approximate description of the true dynamics of the system. But even this is not easily addressed without additional restrictions. Thus in the examples considered in the previous section, the unit-lead prediction for T, Christmas Island SST, may be interpreted formally as the Green function solution of a differential equa-

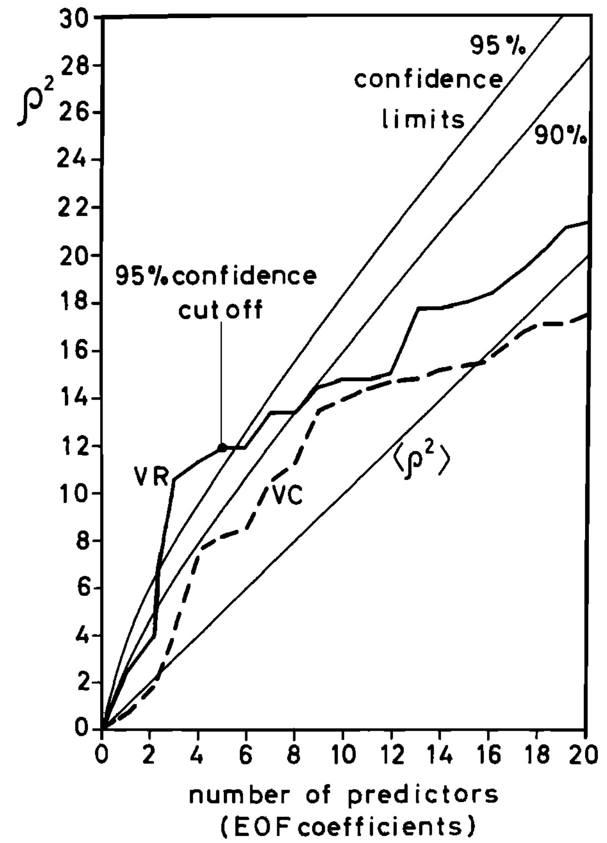

Fig. 5. Confidence parameters ρ^2 for predictions of second eigenfunction V of meridional trade wind component 6 months in advance. Case VC: comprehensive predictor set C, 20 time series with 12 time lags per series. Case TR: reduced predictor set, 4 time series with 10 time lags.

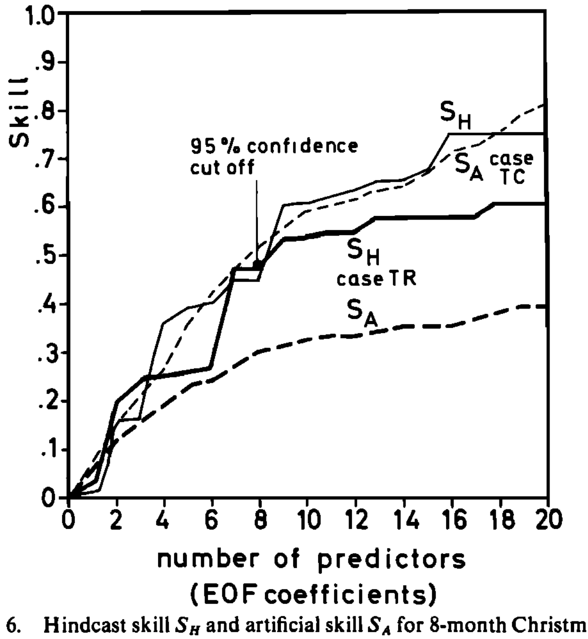

Fig. 6. Hindcast skill S_H and artificial skill S_A for 8-month Christmas Island SST predictions (see Figure 4 caption).

tion for T driven by a linear combination of the predictor fields. However, already for the simpler prediction model TR, using only four predictors and 10 time lags, the equivalent dynamical system would generally contain derivatives of high order and be far from simple. The dynamical model derived from the unit-lead prediction would also normally fail to reproduce the empirical predictions for lead times greater than unity. In general, a simple physical interpretation of an empirical linear prediction model is possible only if the linear dynamical model is first specified and the structure of the prediction model is then chosen to match the structure of the proposed dynamical model. However, one then still faces the difficulty that the finite order prediction model generally contains a set of free coefficients different from the dynamical model, so that an exact one-to-one mapping between the two types of model is not possible.

To illustrate the relation between empirical linear prediction models and linear dynamical models, consider a dynamical system described by a set of linear ordinary differential equations with respect to time. It is assumed that spatial derivatives, if present in the continuous description of the system, no longer appear explicitly but have been expressed, after spatial discretization, in terms of linear combinations of a finite number of separate components. Discretizing also the time derivatives, and introducing transformations analogous to (4), the evolution of the system may then be expressed in the form of the general first-order process

$$\zeta_i(t_{k+1}) = \sum_{j=1,\cdots,n} C_{ij}\zeta_j(t_k) + \sum_{l=1,\cdots,m} B_{il}f_l(t_k) \qquad (18)$$

where $\zeta = (\zeta_i)$ is the state vector of the system, $\mathbf{f} = (f_l)$ is an external forcing field (which need not be of the same dimension as ζ), and C_{ij} and B_{il} are matrices characterizing the internal dynamical structure of the system and the form of the coupling to the external fields.

Repeated iteration of (18) yields the integral form of the evolution equation, in matrix notation

$$\zeta(t_{k+h}) = C^h\zeta(t_k) + G\mathbf{f} \qquad (19)$$

where the Green function G, given by the sum

$$G\mathbf{f} = \sum_{q=1}^{h} C^{q-1}B\mathbf{f}(t_{k+h-q}) \qquad (20)$$

represents the effect of the external forcing from time t_k to t_{k+h}, and the first term on the right-hand side describes the dependence on the initial state at time t_k.

We shall adopt the viewpoint that equations (18) or (19) are not rigorously satisfied by the data, but represent approximations to the real system which, just as in a prediction model, must be fitted to the data through appropriate choice of the model parameters. The relation between the coefficients of the optimal dynamical model and the optimal prediction model will then depend on (1) how many of the variables (ζ_i, f_l) occurring in the dynamical model are actually available as measured time series, (2) the way in which (18) or (19) is fitted to the data, and (3) the way in which the variables (ζ_i, f_l) of the dynamical model are related to the predictors and predictands of the prediction model.

We consider first the optimal situation in which measurements for all time series ζ_i and f_l exist. In this case the best fit dynamical model may be defined, for example, as the set of model coefficients which minimizes the error function

$$\epsilon = \langle|\zeta(t_{k+1}) - C\zeta(t_k) - B\mathbf{f}(t_k)|^2\rangle \qquad (21)$$

appropriate to the differential form (18) or, alternatively, the corresponding error expression

$$\epsilon = \langle|\zeta(t_{k+h}) - C^h\zeta(t_k) - G\mathbf{f}|^2\rangle \qquad (22)$$

for the integral form (19).

The minimal solution of (21) for arbitrary matrices C and B is clearly identical to the solution for the prediction model (1), in which the predictors $\mathbf{x}$ are defined as the combined set $\{\zeta(t_k), \mathbf{f}(t_k)\}$ and the predictands $\mathbf{y}$ as the set $\zeta(t_{k+1})$. In most cases, however, the components of the matrices C and B cannot be regarded as independently adjustable parameters, but will be determined by the combination of time derivatives and internal coupling terms assumed for the dynamical system. These will normally be governed by a rather small number of free physical parameters, and the error expression (21) must there-

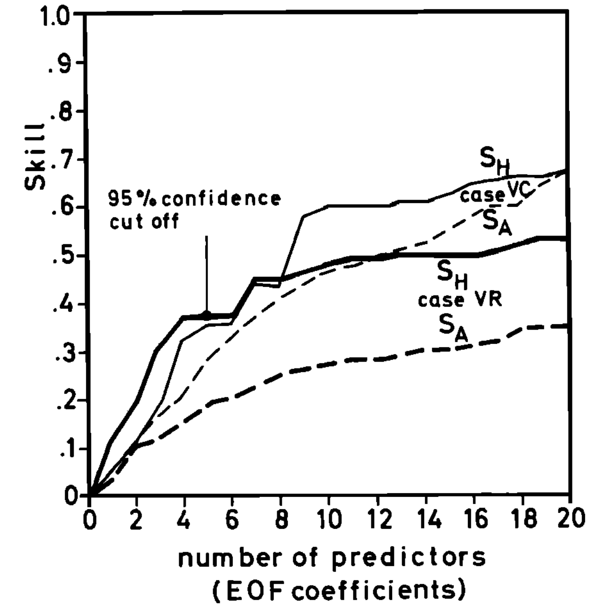

Fig. 7. Hindcast skill S_H and artificial skill S_A for 6-month trade wind anomaly prediction (see Figure 5 caption).

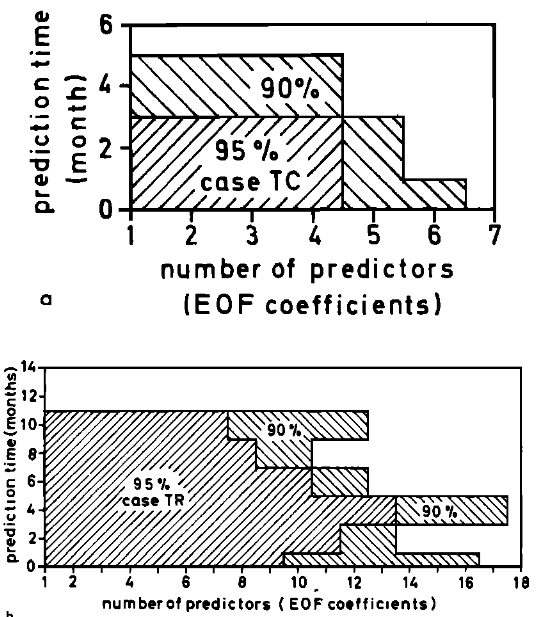

Fig. 8. Regions of statistically significant predictions for Christmas Island SST anomaly T as function of prediction time and model order. (a) Case TC, 20 time series with 12 time lags. (b) Case TR, 4 time series with 10 time lags.

fore be minimized with respect to these parameters, rather than the complete set of independent matrix components.

Similarly, the minimal solution of the error function (22) for the integral form of the dynamical system is formally identical to the solution of an optimal prediction problem in which the predictors are identified with the set of variables $\zeta(t_k)$, $\mathbf{f}(t_k)$, $\mathbf{f} = t(_{k+1}),\cdots$, $\mathbf{f}(t_{k+h-1})$ and the predictands with the vector $\zeta(t_{k+h})$, provided the individual components of the matrix $C' \equiv C^h$, and the components of the matrices occurring in the series G can again be regarded as separately adjustable coefficients. In fact, this is not the case, since C' and the matrices of the series G depend (nonlinearly) on only two matrices, C and B, the components of which in turn will depend on the smaller number of original physical parameters.

Thus despite the formal equivalence of the first-order dynamic system with a unit-lead prediction model a basic difference exists in the way in which the parameters are introduced into the model. In constructing a dynamical model the goal is generally to minimize error expressions of the form (21) or (22) using the simplest acceptable physical model (e.g., of lowest possible order in the spatial and time derivatives). Methods of constructing such models, using a hierarchy of models of increasing complexity, are discussed in detail by *Box and Jenkins* [1976], *Kashyap ar ' Rao* [1976], and *Mosteller and Tukey* [1977]. In the empiricai ;rediction models discussed in this paper, however, the model hierarchy is determined by the a priori selection of predictors, based, for example, on the variance properties of the predictor set, independent of the physical interpretation of the resultant model. For an infinite data set it may be presumed that both model hierarchies will ultimately converge to the same model (18) or (19). However, a truncation of the model hierarchies, as required by statistical sampling considerations, yields different models for the two cases, which in general are not directly related. (It is, of course, possible to select the predictors of the prediction model such that they represent lower-order spatial or temporal finite difference expressions. In the case of the differential form (18) the prediction model hierarchy can then be chosen to coincide with a given dynamical model hierarchy. However, this is not possible for the integral form (19), which depends nonlinearly on the coefficient matrices C and B.)

The disparity between finite-order dynamical models and empirical prediction models becomes more evident if prediction lead times greater than unity are considered. Since a dynamical model completely specifies the input-response structure of the system, it also defines an optimal prediction model. Thus once a dynamical model has been fitted to the data, for example by minimizing the unit-lead prediction error (21), the associated optimal prediction models for all other lead times are also specified. The coefficients of empirical prediction models, however, are determined independently for each lead time. Since different predictor combinations may be expected to be more useful for different prediction lead times, a strategy in which the model is optimized independently for each lead time clearly has advantages when working with approximate, truncated representations. However, it also implies that the physical interpretation of empirical prediction models is made more difficult by requiring the identification of a series of different prediction models, depending on the lead time as a parameter, with a single dynamical model.

For these reasons an exact one-to-one correspondence between finite-order prediction models and dynamical system simulations cannot generally be established. However, the physical interpretation of empirical prediction models can nevertheless be attempted in the form of a statistical consistency test: a dynamical model may be regarded as a statistically consistent interpretation of an empirical prediction model if the prediction coefficients associated with the dynamical model lie within the confidence ellipsoid of the coefficients derived for the optimal empirical prediction.

This interpretation can be applied also to more general cases in which the predictors and predictands of the prediction model represent mixed or incomplete combinations of the forcing and response functions of the dynamical model. A common example is the case in which only some of the postulated inputs of the dynamical model are available as measured inputs for a prediction model, but the statistical properties of the missing inputs are postulated as part of the dynamical model specification (e.g., white noise forcing). In these cases it may be more effective to construct dynamical models by fitting observed and predicted variance and covariance spectra, rather than by attempting to reproduce the actual time series by minimizing prediction-error expressions of the form (21) and (22) [cf. *Hasselmann,* 1979*a*]. Regardless of the method of fitting, however, a given dynamical model always defines an associated optimal prediction model with respect to a specified predictor set. Thus the physical interpretation of empirical prediction models may again be carried out by testing the statistical consistency of the empirical prediction coefficients with the prediction coefficients inferred from a dynamical model.

7. EMPIRICAL PREDICTION AND DYNAMICAL MODELS OF SEA LEVEL VARIATIONS IN THE EQUATORIAL PACIFIC

The prediction examples considered in section 5 represent typical cases in which useful predictions could be constructed without reference to a (mathematically formulated) dynamical model. As pointed out, a simple physical interpretation of the empirical prediction solutions in these cases would have been rather difficult, since the predictors and predictands were not chosen to match the input and response variables of a particular dynamical model. To illustrate the interrelationship of prediction and dynamical models discussed in the previous section we consider now a further example in which the struc-

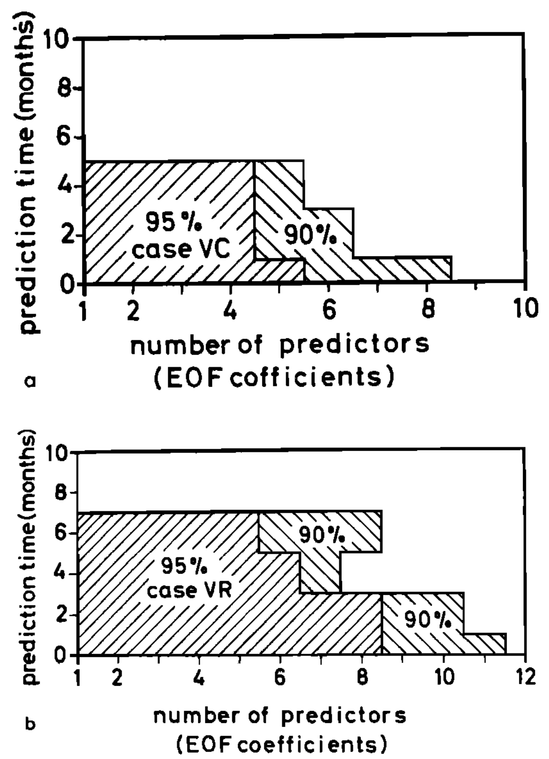

Fig. 9. Regions of statistically significant predictions of the trade wind anomaly component V as function of prediction time and model order. (a) Case VC, 20 time series with 12 time lags. (b) Case VR, 4 time series with 10 time lags.

ture of the prediction model was specifically chosen to correspond to a particular dynamical model.

A suitable case for such a comparison is the variation of east-west sea level in the equatorial Pacific in response to changes in the wind field, for which various simple dynamical models have been proposed. *Bjerknes* [1966], and later *Wyrtki* [1975] and others, have suggested that the El Niño phenomenon may be explained as an oscillation in the east-west equatorial sea level induced by long-term fluctuations in the strength of the trade winds. The basin-wide oscillation set up by this mechanism produces an influx of warm surface waters and a dynamic adjustment at the eastern edge of the Pacific Basin that leads to El Niño conditions (cf. also *McCreary* [1976] and *Hurlburt et al.* [1976]). *McWilliams and Gent* [1978] have attempted to summarize the principal interactions involved in the form of a simple set of coupled dynamical equations for the key variables of the system. With respect to east-west sea level difference h, the authors propose a response to variations of the zonal wind stress (represented by some suitably average wind anomaly variable u) in accordance with (1) a first-order relaxation equation

$$dh/dt + \lambda h = b \cdot u \qquad (23)$$

with constant feedback factor λ and wind-coupling coefficient b or, alternatively, (2) a damped oscillator equation

$$\frac{d^2h}{dt^2} + 2r\frac{dh}{dt} + \omega_0^2 h = b \cdot u \qquad (24)$$

with constant damping coefficient r and frequency ω_0.

The integral (Green function) representations of the systems (23) and (24) may be written

$$h(t) = b\int_{t_0}^{t} u(t')e^{-\lambda(t-t')}dt' + h(t_0)e^{-\lambda(t-t_0)} \qquad (25)$$

and

$$h(t) = b\int_{t_0}^{t} u(t')\frac{\sin\omega(t-t')e^{-r(t-t')}}{\omega}dt' + h(t_0) \cdot \cos\omega(t-t_0)e^{-r(t-t_0)} + \left(r \cdot h(t_0) + \frac{dh}{dt}(t_0)\right) \cdot \frac{\sin\omega(t-t_0)e^{-r(t-t_0)}}{\omega} \qquad (26)$$

respectively, with $\omega = (\omega_0^2 - r^2)^{1/2}$ or, in discretized form,

$$h_{l+1} = \sum_{k=0}^{\infty} (\Lambda\Delta t)^k b U_{l-k} \qquad (27)$$

and

$$h_{l+1} = \sum_{k=0}^{\infty} \frac{R^k\{\sin\omega_1(k+1)\Delta t\}}{\sin\omega_1\Delta t} U_{l-k} \qquad (28)$$

where the subscript l is a discrete time counter, Δt is the time step, and the coefficients Λ, R, and ω_1 are related to the coefficients λ, r, and ω_0 in a manner depending in detail on the form of finite differences used [cf. *Jenkins and Watts*, 1968]. It has been assumed in (28) that we are dealing with an underdamped oscillator, ω_1 real (cf. Figure 13).

On the basis of these dynamical models, two prediction models for h were constructed. The first model (I) was chosen to correspond as closely as possible to the dynamical models and used only u as predictor. In the second model (II), six additional wind predictors were introduced. (The total set of wind predictors consisted of the 10 wind components shown in Table 1 and Figure 5 without δu_1, δv_1, and δv_2.) The purpose of the second model was to test if a simple dynamical model with only one input was consistent with a more complex prediction model containing enough degrees of freedom to adapt to a number of different physical processes. In both cases the prediction was formulated as a single-lead predictand, multiple-lag predictor system in accordance with the discrete forms (27) and (28) of the dynamical models, by using first-order centered differences. The number of predictor lags was assumed to be sufficiently large ($m \cdot \Delta t = 18$ months) that the initial values of h in (27) and (28) could be neglected and the prediction constructed from the forcing function u only. (This was confirmed by tests including the initial values and by the damping coefficients found for the best fit dynamical models. In fact, it can be shown generally that even when the initial values contribute to the prediction their exclusion has no influence on the prediction coefficients for the forcing function, provided the forcing is fairly white. For the same reason, deletion of the $(k - 1)$ most recent terms of the forcing function in (27) and (28), as required for a k-lead rather than a unit-lead prediction, should have no influence on the remaining prediction coefficients of the forcing function at earlier times. This was verified by constructing k-lead predictions for various values of $k > 1$.) In both cases the optimal prediction models represented maximum-skill models at the 95% confidence level, as determined by a nested-model hierarchy using an EOF representation of the predictor field, after removal of the white noise EOF components according to the technique of *Preisendorfer and Barnett* [1977]. Prefiltering of the EOF set yielded about the same results as using the complete EOF set and applying a maximal significance cutoff condition or as using a cutoff based on the rate of significance decrease. However, a fixed-significance level cutoff without prior EOF filtering was found to include too many noisy predictors of higher order.

Following McWilliams and Gent, the zonal wind stress field

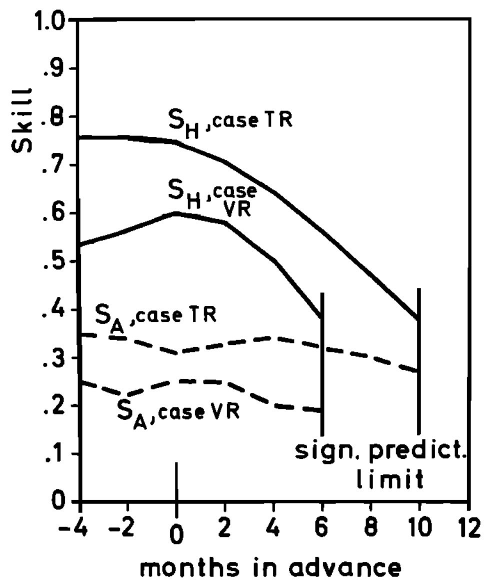

Fig. 10. Maximum hindcast skill (at 95% significance cutoff) of Christmas Island temperature and meridional trade wind anomaly predictions for the reduced predictor set R as function of forecast period. Also shown are the artificial skills.

driving the east-west pressure difference in the dynamical model was assumed proportional to the zonal component of the wind anomaly. Specifically, the forcing function u was defined as the average of the zonal wind anomaly over the region between 0°-10°N and 150°-180°E denoted by δu_2 in Figure 3. This choice is in accordance with the recent findings of *McCreary* [1976] and *Barnett* [1977*b*], who showed that major equatorial oceanic features respond to wind changes in this region rather than to changes in the core strength of the trade winds as suggested by *Bjerknes* [1966] and *Wyrtki* [1975]. The additional wind field predictors used in the second prediction model describe the general features of the Pacific Trade Wind System. The variations h in the east-west slope of sea level were represented by the amplitude of the first eigenmode of the sea level anomaly field [*Barnett*, 1977*b*].

To test the consistency of the dynamical models with the prediction models, the square deviation

$$\rho_d{}^2 = \sum_{i,j} (\tilde{a}_i - a_i{}^d)(\tilde{a}_j - a_j{}^d)\langle\delta a_i \delta a_j\rangle^{-1} \tag{29}$$

between the prediction coefficient vectors $\tilde{a}_i$ of the optimal prediction model and the equivalent prediction coefficient vector $a_i{}^d$ of the best fit dynamical model was compared against the mean square radius $\rho_{95}{}^2$ of the 95% confidence region computed for the optimal prediction model. The dynamical model was regarded as statistically consistent or inconsistent with the prediction model at the 95% confidence level according to $\rho_d{}^2 < \rho_{95}{}^2$ or $\rho_d{}^2 > \rho_{95}{}^2$.

The square deviation $\rho_d{}^2$ is defined in (29) with respect to the coefficient space of the prediction model. For a complete prediction model, without truncation, this may be identified with the set of coefficients occurring in the discretized form of the integral representation of the dynamical model. With this choice of predictors, the best fit dynamical model can then be translated immediately into an equivalent prediction model, since the Green function representing the best fit dynamical model immediately defines an equivalent set of prediction coefficients with respect to the complete set of predictors $u(t_{k-1}), u(t_{k-2}), \cdots, u(t_{k-m})$. However, in the truncated prediction model only a linear subspace of the predictors, represented by a relatively small number of EOF's, was acutally used. Thus to carry out the consistency test, the dynamical model must first be projected on to the predictor subspace used in the prediction model. In practice, this was done by creating a synthetic time series h' from the best fit dynamical model, using the observed forcing u as input in the integral form (27) or (28) and then determining the optimal prediction for this series h' by using the same predictors as were used for the optimal empirical prediction of the real series h. The best fit dynamical model itself was obtained by minimizing the mean square error of the discrete differential form rather than the integral form.

Each of the first- and second-order dynamical models were found to be consistent with each of the empirical prediction models I and II. A comparison of the empirical prediction coefficients and the theoretical forms (27) and (28) (after projection on to the truncated EOF space) is shown in Figure 14; other results of the experiments are listed in Table 2. Within the 95% error bounds on the coefficients, all four models are indistinguishable from each other. The difference between the first- and second-order best fit dynamical models is particularly small, and it may be concluded that the data does not justify the use of a second-order finite difference process to describe the local (differential) evolution of h. However, both empirical prediction models suggest a damped-oscillator form of the Green function. Although the overshoot is not quite significant at the 95% level, a second-order dynamical model in accordance with the integral form (28) was therefore also fitted to the empirical prediction coefficients. (This was done by inspection: a formal least squares fit was not carried out because for the integral form of the dynamical model the minimization problem is no longer linear in the model parameters.)

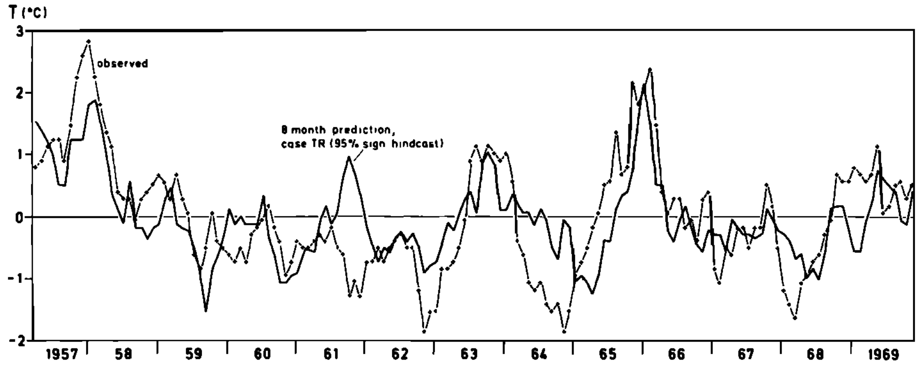

Fig. 11. Predicted and observed anomalies of trade wind component V 6 months in advance, cases VC (statistically insignificant at 50% confidence level) and VR (statistically significant at the 95% confidence level).

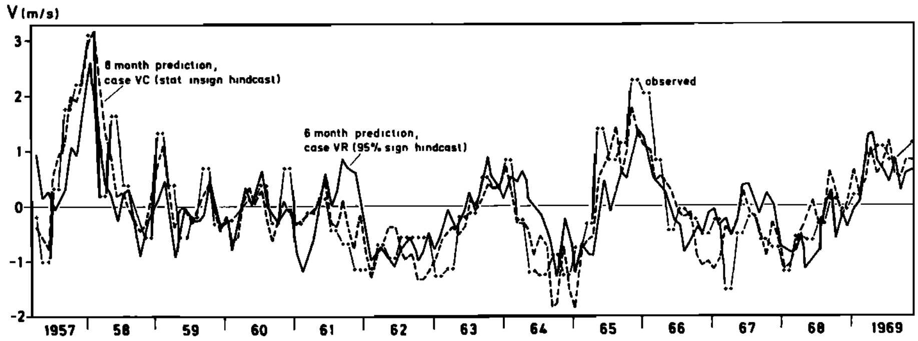

Fig. 12. Predicted and observed SST anomalies at Christmas Island 8 months in advance, case TR (four time series with 10 lags).

This yielded a prediction model with higher skill (20%) than the skill values (≈ 10%) of either of the dynamical models obtained by best fitting the differential forms to the data. The parameters of the Green function-fitted dynamical model were found to be $\omega_0 = 1.4 \cdot 10^{-7}$ s^{-1} and $r = 2.5 \cdot 10^{-8}$ s^{-1}, as compared with the values $\omega_0 = 2.6 \cdot 10^{-7}$ s^{-1} and $r = 2.8 \cdot 10^{-8}$ s^{-1} estimated by McWilliams and Gent (the former value represents a revised 'composite' frequency based on later work [*McWilliams and Gent,* 1979]). Considering the stated uncertainty of order 40% in this 'composite' frequency and the rather large error bands in Figure 14, we conclude that the two estimates of ω_0 are not inconsistent.

In summary, we were able to forecast variations in east-west slope of Pacific sea level 1 month in advance from past wind field anomalies, with skills of 38% (one wind field predictor) and 49% (seven wind field predictors) at a significance level of 95%. Alternative prediction models equivalent to a (differentially best fitted) first-order or second-order damped dynamical system were also consistent with the optimal prediction model within the error bounds of the coefficients. However, these models accounted for only about 10% of the observed sea level variance. A higher skill of approximately 20% could be achieved by fitting the integral response function of a second-order dynamical model to the empirical response function. Nevertheless, the conclusion remains that a simple first- or second-order dynamical model driven by a single wind field variable is inadequate to explain most of the variance in east-west sea level slope. A more general model allowing an arbitrary integral response to a larger number of wind field variables is able to account for a considerably higher fraction of the variance. This would imply that the wind field is an important driving term, but the sea level response is more complex than assumed by McWilliams and Gent. However, the alternative hypothesis that the sea level response to the wind is correctly modeled by McWilliams and Gent, but accounts for only 10–20% of the sea level variance, cannot be excluded at the 95% confidence level with the available data.

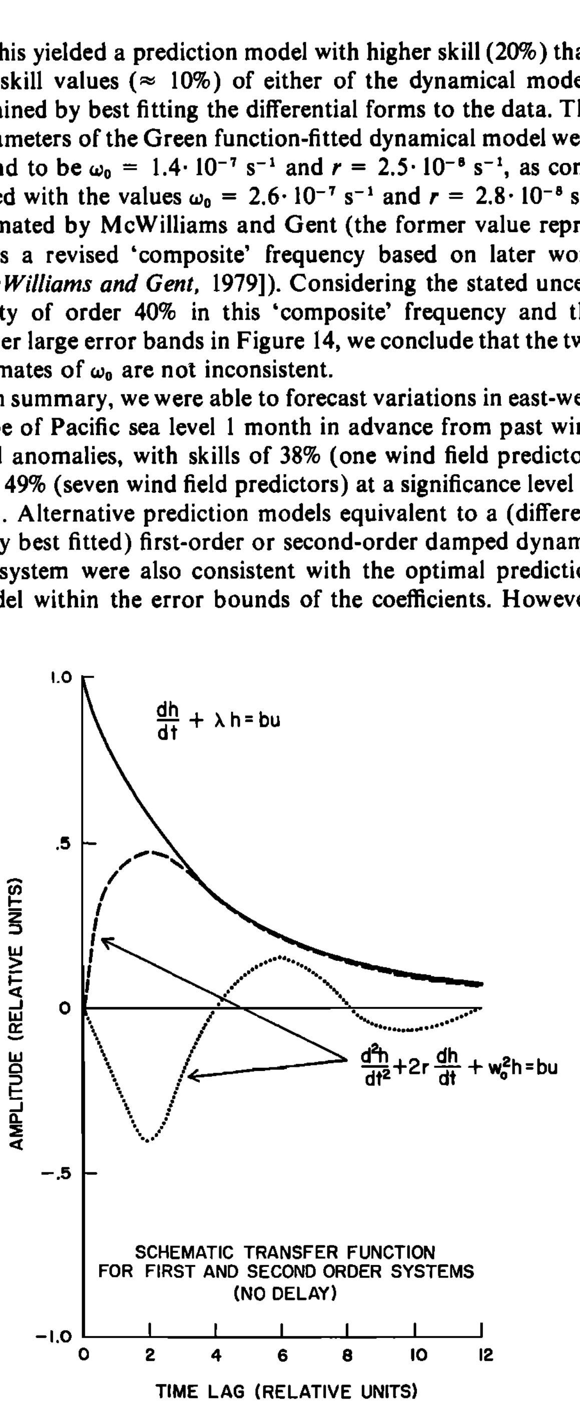

Fig. 13. Theoretical transfer functions for first- and second-order dynamical systems.

8. Forecast Skill

Up to this point we have discussed the performance of models only in terms of an abstract infinite data ensemble, or the given finite data set from which the model was constructed. A measure of model skill which is relevant for applications, however, should be defined with respect to the performance of the model when applied to a second data set, independent of the set used to construct the model. Intuitively, one may expect the average forecast skill S_F for an independent prediction to be lower than the hindcast skill S_H for the original data set, since the estimated optimal model $\hat{\mathbf{a}}$ was chosen to yield maximal skill specifically for the first data set. Thus the deviation of the estimated model from the true model, which produced an artificial enhancement of the hindcast skill S_H relative to the true skill S_0 (17), may be expected to have an opposite adverse effect on the skill when applied to a second independent data set. Indeed, it can be readily shown [cf. *Lorenz,* 1956; *Davis,* 1976] that for a given true model $\mathbf{a}^0$, the ensemble average of the forecast skill is given to first order for small true skill S_0 by

$$\langle S_F \rangle \approx S_0 - \langle S_A \rangle \tag{30}$$

Substituting (17), this yields

$$\langle S_F \rangle = \langle S_H \rangle - 2\langle S_A \rangle \tag{31}$$

Although the derivation and interpretation of equations (17), (30), and (31) is straightforward, there is nevertheless some question whether they really provide relevant estimates of the average forecast skill. It is assumed in deriving the relations that the true model is known, and the averages are then formed over all possible realizations of the first (hindcast) and second (independent) data sets. In fact, the true model is not

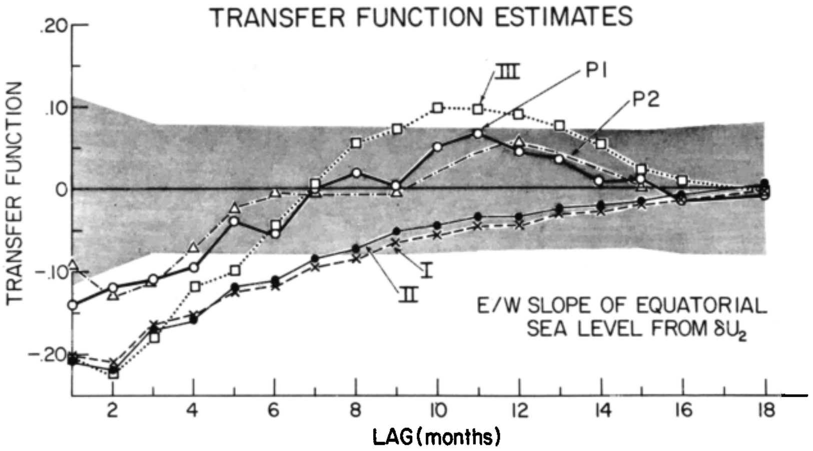

Fig. 14. Empirical transfer functions for prediction of east-west sea level change in response to equatorial wind forcing. The shaded region shows the approximate 90% confidence limits on the estimates of the transfer function. The symbols denote (triangles) optimal prediction p_1 with δU_2 only, (open circles) optimal prediction p_2 with all winds, (crosses) best-fit first-order (I) dynamic model, (solid circles) best-fit second-order (II) dynamic model, and (squares) empirically fitted second-order dynamic model (III).

known, whereas the hindcast data set is given. It could therefore be argued that the average forecast skill should be defined as an average under the side condition of a given hindcast data set and unknown true model, rather than the side condition of a given true model and unknown hindcast data set. In this form the problem of estimating the mean forecast skill, given the hindcast skill, represents a problem of statistical inference: what can be inferred about the true model (and therefore the true forecast skill), given only a single finite data realization? (Note that 'true' refers here, as before, to the optimal empirical prediction for an infinitely long data set, regardless of whether or not the empirical prediction provides a good physical description of the real evolution of the system.)

The theory of statistical inference has been the subject of considerable discussion. The origin of the debate is well known and basically simple: classical axiomatic probability theory starts from a given abstract infinite data ensemble with known statistical properties; in practice, however, only a finite data set is available, and this therefore has to be first imbedded in an abstract infinite ensemble before one can proceed with statistical theory. It can be argued that the choice of imbedding is basically arbitrary, provided it is consistent with the given finite data set within reasonable statistical confidence bounds. If, however, one now addresses statistical questions which depend, already to first order, on differences between possible alternative choices of the infinite statistical ensemble, the 'proper' choice of imbedding becomes a problem. The estimation of the mean forecast skill or, more precisely, the difference between the mean forecast skill and hindcast skill is a question of this nature.

According to one school of thought, such questions cannot be rigorously answered and are therefore meaningless. However, various attempts have been made to deal with the basic arbitrariness in the choice of statistical imbedding by assigning some form of (necessarily subjective) a priori weighting to the different imbedding possibilities [cf. *Savage,* 1962]. Since the true model $\mathbf{a}^\circ$ is not known, it is natural, for example, to consider an ensemble of true models characterized by a probability density function $p_0(\mathbf{a}^\circ)d\mathbf{a}^\circ$. In the absence of other information, it is then often assumed that in a local region around the estimated model $\hat{\mathbf{a}}$ the distribution of true models is uniform (Bayes' hypothesis)

$$p_0(\mathbf{a}^\circ)d\mathbf{a}^\circ = \text{const } d\mathbf{a}^\circ \qquad (32)$$

For each true model $\mathbf{a}^\circ$ of this ensemble there exists, as before, a probability distribution of estimated models $p(\hat{\mathbf{a}}/\mathbf{a}^\circ)d\hat{\mathbf{a}}$

TABLE 2. Comparison of Dynamical and Prediction Modeling Experiments

Model Type	Hindcast Skill, %	Variance Reproduced After EOF Truncation*	Artificial Skill, %	ρ_d^2†	ρ_{95}^2	Consistent
Optimal prediction (I)	38		13		14.1 (n = 7)	
Optimal prediction (II)	49		34		16.9 (n = 9)	
First-order dynamic, differentially fitted	10	94 (I) 88 (II)		9.5 (I) 9.7 (II)		Yes (I) Yes (II)
Second-order dynamic, differentially fitted	11	94 (I) 89 (II)		9.9 (I) 8.8 (II)		Yes (I) Yes (II)
Second-order dynamic, response-curve fitted	20	98(I) 93(II		10.1 (I) 9.5 (II)		Yes (I) Yes (II)

*Applies only to dynamic models. A truncation of the formal prediction of the synthetic time series h' by using only the EOF's used in predictions I or II reduces the skill of 100% achieved with the full EOF set to the values shown.

†For a statistically consistent dynamical model, ρ_d^2 must be less than the ρ_{95}^2 values of the corresponding prediction model I or II.

which is given by the Gaussian form (10) (the covariance matrix $\langle \delta a_i \delta a_j \rangle$ is assumed to be independent of $\mathbf{a}^\circ$). As $p(\tilde{\mathbf{a}}/\mathbf{a}^\circ)$ depends only on the difference $\delta\mathbf{a} = \tilde{\mathbf{a}} - \mathbf{a}^\circ$, we may write $p(\tilde{\mathbf{a}}/\mathbf{a}^\circ) = p_g(\tilde{\mathbf{a}} - \mathbf{a}^\circ)$, and the joint probability distribution of estimated and true models therefore takes the form

$$p_j(\mathbf{a}^\circ, \tilde{\mathbf{a}}) d\tilde{\mathbf{a}} d\mathbf{a}^\circ = p_g(\tilde{\mathbf{a}} - \mathbf{a}^\circ) p_0(\mathbf{a}^\circ) d\tilde{\mathbf{a}} d\mathbf{a}^\circ \qquad (33)$$

To determine the mean forecast skill, one now averages first the forecast skill over the ensemble of independent second data sets, given the true model $\mathbf{a}^\circ$ and the first (hindcast) data set (i.e., $\tilde{\mathbf{a}}$). This step is the same as in Lorenz' derivation of the mean forecast skill and yields for small skill values

$$\langle S_F \rangle_2 = S_0 - S_A \qquad (34)$$

Equation (34) is identical to (30) except that averages over the hindcast data set have not been taken; the average $\langle \cdots \rangle_2$ on the left-hand side is taken over the second independent data sets, and the artificial skill S_A in the right-hand side represents an unaveraged quantity determined from the given estimated model and given true model. Up to this point both approaches yield the same (expected) result, namely that an error $\delta\mathbf{a}$ between the true and estimated model results in a reduction of order S_A in the forecast skill relative to the true skill.

In the next step, however, the Bayesian analysis differs from the usual derivation. Instead of averaging (34) over the ensemble of hindcast data sets for a given true model $\mathbf{a}^\circ$, the average is taken over the set of all possible true models $\mathbf{a}^\circ$, given the hindcast data set (i.e., given $\tilde{\mathbf{a}}$).

According to (33 and 32), the relevant conditional probability (or likelihood) distribution which enters in this averaging operation is given by

$$p(\mathbf{a}^\circ/\tilde{\mathbf{a}}) d\mathbf{a}^\circ = p_g(\tilde{\mathbf{a}} - \mathbf{a}^\circ) d\mathbf{a}^\circ \qquad (35)$$

Thus $p(\mathbf{a}^\circ/\tilde{\mathbf{a}})$ is Gaussian and is identical to the conditional probability distribution $p(\tilde{\mathbf{a}}/\mathbf{a}^\circ)$ with the variable $\mathbf{a}^\circ$ and $\tilde{\mathbf{a}}$ interchanged. Denoting the likelihood average of a quantity Q by

$$\{Q\} = \int Q p(\mathbf{a}^\circ/\tilde{\mathbf{a}}) d\mathbf{a}^\circ \qquad (36)$$

one then obtains for the likelihood-averaged forecast skill

$$S_F{}^L \equiv \{\langle S_F \rangle_2\} = \{S_0\} - \{S_A\} \qquad (37)$$

where $\{S_A\} = \langle S_A \rangle$ (through the assumption that $\langle \delta a_i \delta a_j \rangle$ is independent of $\mathbf{a}^\circ$) and the likelihood-averaged true skill is given by

$$\{S_0\} = S_H + \langle S_A \rangle \qquad (38)$$

(In (37) and (38) and later in (40) and (43) the subscript 1 or 2 is dropped from averages $\langle \cdots \rangle$ which are independent of the side condition of a given true model or a given hindcast data set.)

Equation (38) follows from (17) by noting that the expression for the likelihood-averaged true skill for a given estimated model $\tilde{\mathbf{a}}$ is formally identical to the expression for the mean hindcast skill averaged over $\tilde{\mathbf{a}}$ for fixed $\mathbf{a}^\circ$, except that the variables $\tilde{\mathbf{a}}$ and $\mathbf{a}^\circ$ are interchanged. The equality of the conditional distributions $p(\tilde{\mathbf{a}}/\tilde{\mathbf{a}}^\circ)$ and $p(\mathbf{a}^\circ/\tilde{\mathbf{a}})$ explains also the result (which at first sight appears rather surprising) that the likelihood-averaged true skill is greater than the hindcast skill. The surface of constant probability (likelihood) $p(\mathbf{a}^\circ/\tilde{\mathbf{a}})$ correspond to the confidence ellipsoids $\tilde{R}$ in Figure 1 and are centered on the given estimated model $\tilde{\mathbf{a}}$. Since the skill is a positive definite quadratic function of the model coefficients, the likelihood-averaged skill, obtained by averaging over all true models $\mathbf{a}^\circ$ for a fixed estimated model $\tilde{\mathbf{a}}$, must always exceed the (hindcast) skill of the estimated model, just as the mean hindcast skill obtained by averaging over all hindcast data sets with respect to a probability distribution $p(\tilde{\mathbf{a}}/\mathbf{a}^\circ)$ centered on the given true model $\mathbf{a}^\circ$ is always greater than the skill of the true model. The interchange symmetry of the likelihood-averaged and hindcast-averaged relations between S_0 and S_H is illustrated geometrically by Figure 1.

Substituting (38) in (37) one obtains

$$S_F{}^L = S_H \qquad (39)$$

This should be compared with the usual ensemble-averaged relation (31)

$$S_F{}^E \equiv \langle \langle S_F \rangle_2 \rangle_1 = \langle S_H \rangle_1 - 2\langle S_A \rangle \qquad (31')$$

The difference between (39) and (31') is due not only to the different forms of averaging but, more fundamentally, to the different ways in which the given finite data set is assumed to be imbedded in an infinite statistical ensemble. The disadvantage of the usual ensemble-averaged expression is that it yields a relation only between the average hindcast skill and average forecast skill, and therefore makes no use of the fact that the hindcast skill is actually known. The likelihood approach, on the other hand, yields a mean forecast skill under the side condition of a known hindcast skill, but requires an imbedding of each finite data series not simply in a hypothetical infinite time series, but in an additional ensemble of possible true models. The Bayes hypothesis of a uniform distribution of true models, while perhaps the simplest, is basically arbitrary. Although it is invoked here only for a limited region around the estimated model $\tilde{\mathbf{a}}$, it cannot be justified in the present case simply by continuity considerations.

This can be seen by generalizing (38) and (39) to the case of an arbitrary distribution $p_0(\mathbf{a}^\circ)$. It is instructive to consider this generalization, as it brings out more clearly the relation between the two forms of estimating the forecast skill and helps to resolve the apparent discrepancy between the expressions (31') and (39). In the following we therefore make no assumptions regarding $p_0(\mathbf{a}^\circ)$, other than that it is a smooth function which varies slowly in relation to $p_g(\delta\mathbf{a})$.

Restricting the discussion, as before, to small skill values so that the principal variations in skill arise from the variations in the model coefficients, we have

$$\begin{aligned} S_H - S_0 &= \sum_{i,j} (\tilde{a}_i \tilde{a}_j - a_i{}^\circ a_j{}^\circ) \frac{\langle z_i z_j \rangle}{\langle y^2 \rangle} \\ &= \sum_{i,j} (-\delta a_i \delta a_j + 2\delta a_i \tilde{a}_j) \frac{\langle z_i z_j \rangle}{\langle y^2 \rangle} \qquad (40) \end{aligned}$$

Forming the likelihood average of (40) according to (36) and expanding the factor $p_0(\mathbf{a}^\circ)$ occurring in the joint probability distribution (33) in the form $p_0(\mathbf{a}^\circ) = p_0(\tilde{a}) - \sum_j [\partial p_0(\tilde{a})/\partial a_j] \delta a_j$, we obtain as generalization of (38)

$$S_H - \{S_0\} \approx \sum_{i,j} \left[\left(-\langle \delta a_i \delta a_j \rangle - \sum_k 2 \frac{\langle \delta a_i \delta a_k \rangle \tilde{a}_j}{p_0(\tilde{\mathbf{a}})} \frac{\partial p_0(\tilde{a})}{\partial a_k} \right) \frac{\langle z_i z_j \rangle}{\langle y^2 \rangle} \right] \qquad (41)$$

Substitution of (41) into (37) then yields as the corresponding generalized expression for the likelihood-averaged forecast skill

$$S_F{}^L = S_H + 2 \sum_{i,j,k} \frac{\langle \delta a_i \delta a_k \rangle \tilde{a}_j}{p_0(\tilde{\mathbf{a}})} \frac{\partial p_0(\tilde{\mathbf{a}})}{\partial a_k} \frac{\langle z_i z_j \rangle}{\langle y^2 \rangle} \qquad (42)$$

In contrast to the Bayesian form (39), the more general relation (42) satisfies the requirement

$$\langle S_F{}^L \rangle_1 = \{S_F{}^E\} = \iint S_F p_j(\mathbf{a}^\circ, \tilde{\mathbf{a}}) d\mathbf{a}^\circ d\tilde{\mathbf{a}} = \langle S_H \rangle_1 - 2\langle S_A \rangle \tag{43}$$

which states that the likelihood-averaged and normal ensemble-averaged forecast skills are related by the condition that the mean of the likelihood-averaged forecast skill $S_F{}^L$, integrated over all first data sets, is equal to the average, with respect to the set of all true models, of the ensemble-averaged forecast skill $S_F{}^E$.

Equations (42) and (43) show that while the strict Bayesian hypothesis of constant p_0 yields a likelihood-averaged forecast skill which is equal to the hindcast skill, the generalization of the Bayesian approach to an arbitrary distribution of true models yields a likelihood-averaged forecast skill which depends on $\tilde{\mathbf{a}}$ and, averaged over all estimated models $\tilde{\mathbf{a}}$, is smaller than the hindcast skill by the quantity $2\langle S_A \rangle$, in accordance with the expression for the usual ensemble-averaged forcast skill. The apparent contradiction between the strict Bayesian relation and the more general result is resolved by noting that a constant probability density $p_0(\mathbf{a}^0)$ cannot exist for all $\mathbf{a}^0$ and that on averaging (42) over $\tilde{\mathbf{a}}$ the second term containing the derivative $\partial p_0(\tilde{a})/\partial a_j$, which is neglected in the strict Bayesian analysis, will always yield an additional contribution $-2\langle S_A \rangle$, regardless of the form of p_0.

In summary, (42) and (43) show that the difference between the measured hindcast skill and the mean forecast skill depends entirely on the structure of the unknown prior probability density function p_0 of true models. For a distribution p_0 which varies smoothly near $\tilde{\mathbf{a}} = 0$, the linear dependence of the second term in (42) on $\tilde{\mathbf{a}}$ suggests that for small $\tilde{\mathbf{a}}$ $S_F{}^L$ will be approximately equal to S_H. The integral constraint (43) then requires that $S_F{}^L$ must be smaller than S_H for larger values of $\tilde{\mathbf{a}}$. However, if there exists no possibility of estimating the structure of $p_0(\mathbf{a}^0)$ a priori it must be concluded generally that it is not possible to infer the mean forecast skill $S_F{}^L$, given the hindcast skill S_H of a particular experiment, more accurately than $S_F{}^L \approx S_H - O(S_A)$.

9. CONCLUSIONS

The main results of our presentation may be summarized as follows:

1. With increasing numbers of predictors, the skill of a linear prediction model increases, whereas the significance generally decreases. The central problem in constructing linear prediction models from data therefore is establishing a proper balance between skill and significance. This requires, in particular, a reliable measure of model significance independent of the measure of model skill. To determine the maximum-skill model which can be constructed from a given data set at a given significance level, a sequential model construction strategy can be used, in which a hierarchy of models is generated by successively adding new predictors according to a predefined sequence, the hierarchy being terminated when the statistical significance of the model falls below a prescribed limit. Alternatively, the sequence can be cut off at the model with the highest significance or at the point where the rate of significance falloff exceeds some critical value, thereby excluding predictors which do not contribute incrementally to the model significance. When an EOF predictor sequence is used, the same effect can generally be achieved by a priori filtering of the principal components to remove uncorrelated white noise, by using the technique of *Preisendorfer and Barnett* [1977]. The success of the technique depends critically on a reliable measure of significance and the judicious a priori selection of the predictor sequence such that the more important predictors are introduced early in the sequence. For complex systems some a priori physical insight into the structure of the system is generally necessary to define an effective model sequence yielding statistically significant predictions.

2. Alternative techniques in which the 'statistically most significant' predictors are selected a posteriori by linear combination from a larger set of candidate predictors, although widely used, yield no improvement in either skill or significance over the complete model constructed from the full predictor set. Thus a posteriori screening or stepwise regression cannot be used to recover significant models from a comprehensive model which contains too many predictors for statistical significance. Mixed techniques based on a posteori reordering of an a priori defined set of predictors, without linear recombination, may be useful for predictions involving physically distinct statistically independent predictors. However, the statistical significance of the reordered sequence must again be judged with respect to the complete predictor set.

3. Applications of the nesting strategy to predictions of SST, sea level, and wind field anomalies in the tropical Pacific yielded 95% significant predictions with skills in the range 0.4–0.7 for prediction lead times of 6 to 12 months and beyond. The examples discussed in this paper are typical of a number of similar predictions for this region, including fields representative of El Niño, which will be presented in more detail in a later paper (T. P. Barnett, manuscript in preparation, 1979).

4. To apply the technique to anomaly fields in mid-latitudes, a straightforward generalization to include the modulation of the system by the annual cycle will probably be needed.

5. The transformation of a given finite-order empirical linear prediction model into an equivalent finite-order linear dynamical model is in most cases not possible, since the dynamical model (if restrained to correspond to a lower-order differential system) normally exhibits a different dependence on the model parameters than the prediction model. However, a given dynamical model always defines an associated prediction model. The physical interpretation of empirical prediction models can therefore be formulated as a consistency test, a prediction model being regarded as statistically consistent with a dynamical model if the associated prediction model derived from the dynamical model lies within the confidence region of the empirical model.

6. The empirical prediction of east-west sea level variations from wind field anomalies in the tropical Pacific was found to be statistically consistent with either of the two simple dynamical models proposed by *McWilliams and Gent* [1978] on the basis of the theories of *Bjerknes* [1966], *Wyrtki* [1975], and others. However, the empirical model parameters differ from those proposed by McWilliams and Gent. A distinction between the two models could not be made with the current data base. The data base also does not exclude the alternative hypothesis that the empirical prediction model yields a more realistic description of the complex dynamical response of the ocean than the simple first- or second-order models of McWilliams and Gent, as evidenced by the considerably higher skill of the empirical model.

7. For statistically significant predictions the Bayesian likelihood-averaged forecast skill, under the side condition of a given hindcast model and unknown true model, is approximately equal to the hindcast skill. Previous discussions of the forecast skill have been largely concerned with the mean forecast skill averaged over the ensemble of hindcast models as-

suming a given true model. This is less than the similarly averaged hindcast skill by an amount approximately equal to twice the artificial skill. Under the side conditions of the problem, a general statistical inference approach, in accordance with a Bayesian analysis, appears more appropriate. However, the inferred mean forecast skill is found to depend strongly on the assumed prior distribution of true models, nonuniform prior distributions yielding inferred forecast skills which are generally lower than the Bayesian result and more in keeping with the hindcast-averaged estimate of the forecast skill. We conclude that a reliable estimate of the difference between the mean forecast skill and the hindcast skill, under the side condition of a given hindcast skill, cannot be meaningfully made without additional information on the prior probability distribution of true models.

APPENDIX

Estimation of $\langle \delta a_i \delta a_j \rangle$

The error of the estimated model coefficient $\tilde{a}_i$ relative to the true coefficient a_i^0 (6) is given by

$$\delta a_i = \tilde{a}_i - a_i^0 = \sum_j ([yz_j][z_i z_j]^{-1} - \langle yz_j \rangle \langle z_i z_j \rangle^{-1}) \qquad \text{(A1)}$$

We evaluate (A1) in an orthonormal coordinate system for which $\langle z_i z_j \rangle = \delta_{ij}$. In this case,

$$[z_i z_j]^{-1} = \delta_{ij} - \delta Z_{ij} + \cdots \qquad \text{(A2)}$$

where $\delta Z_{ij} = [z_i z_j] - \langle z_i z_j \rangle$.

Since the error covariance matrix $\langle \delta a_i \delta a_j \rangle$ is normally needed for model significance tests based on the null hypothesis of a zero prediction model, we assume that $\langle yz_j \rangle = 0$. From equation (A1) we obtain then

$$\langle \delta a_i \delta a_j \rangle = \langle [yz_i][yz_j] \rangle + \sum_{k,l} \langle [yz_k][yz_l] \delta Z_{ik} \delta Z_{jl} \rangle$$
$$- \sum \{ \langle [yz_k][yz_j] \delta Z_{ik} \rangle + \langle [yz_k][yz_i] \delta Z_{jk} \rangle \} + \cdots \qquad \text{(A3)}$$

The neglected higher-order terms indicated by $+ \cdots$ in (A3) arise from the neglected higher-order terms in (A2). Note, however, that the second term on the right-hand side of (A3) containing a fourth-order product of estimation errors has been retained, since the matrix product contains a large number of individual terms which can yield a net contribution on summation comparable to the first second-order term on the right-hand side of (A3).

The expectation values of the products of estimation errors occurring in (A3) can be evaluated under the usual approximation that the processes y and z_i are Gaussian. For large averaging times in comparison with the integral correlation times of the variables y and z_i, the third term on the right-hand side of (A3) can be neglected, and the second term decomposes into two separately averaged factors yielding

$$\langle \delta a_i \delta a_j \rangle = \langle [yz_i][yz_j] \rangle + \sum_{k,l} \langle [yz_k][yz_l] \rangle \langle \delta Z_{ik} \delta Z_{jl} \rangle \qquad \text{(A4)}$$

where [cf. *Jenkins and Watts,* 1968]

$$\langle [yz_i][yz_j] \rangle \approx \frac{\Delta t}{N} \sum_{p=-M}^{M} R_{yy}(p) R_{ij}(p) \qquad \text{(A5)}$$

$$\langle \delta Z_{ik} \delta Z_{jl} \rangle \approx \frac{\Delta t}{N} \sum_{p=-M}^{M} (R_{ij}(p) R_{kl}(p) + R_{il}(p) R_{kj}(p)) \qquad \text{(A6)}$$

and R_{yy} and R_{ij} represent the autocovariance functions of y and the cross-covariance function of z_i and z_j, respectively. In evaluating (A5) and (A6), the true covariance functions must be replaced by their estimates. N represents the number of data points and the summation limit M some limiting lag beyond which the covariance functions can be regarded as zero within the confidence limits of the estimate. (Instead of a cutoff at M, a fading function factor which tapers smoothly to zero may be introduced into (A5) and (A6), in accordance with the *Blackman and Tukey* [1958] spectral estimation methods.)

In practice, the true orthonormal coordinate system is not known. However, in accordance with the definition of the confidence region $\tilde{R}$ the covariance matrix $\langle \delta a_i \delta a_j \rangle$ should be determined for a true model given by the maximum-likelihood solution. Thus in evaluating (A4), (A5), and (A6) it may be assumed that $\langle z_i z_j \rangle$ is given by the estimated covariance matrix $[z_i z_j]$ of the data realization used to construct the model, and the orthonormal predictor coordinate system z_i is accordingly taken to be the orthonormal system with respect to this estimated covariance matrix.

Acknowledgments. This work was supported by the National Science Foundation's Office of Climate Dynamics, grant ATM77-20602, through the Norpax and International Decade of Ocean Exploration Program, grant OCE77-22890, and by the Max-Planck-Gesellschaft. The work was initiated while the senior author was visiting the Max-Planck-Institute of Meteorology. Sincere thanks are due to R. Preisendorfer for many useful and stimulating discussions.

REFERENCES

Adem, J., On the physical basis for the numerical prediction of monthly and seasonal temperatures in the troposphere, *Mon. Weather Rev., 91,* 375–386, 1964.

Adem, J., Numerical-thermodynamic prediction of mean monthly ocean temperatures, *Tellus, 27,* 541–551, 1975.

Barnett, T. P., The principal time and space scales of the Pacific Trade Wind fields, *J. Atmos. Sci., 34*(2), 221–236, 1977*a*.

Barnett, T. P., An attempt to verify some theories of El Niño, *J. Phys. Oceanogr., 7*(5), 633–647, 1977*b*.

Barnett, T. P., and R. W. Preisendorfer, Short-term prediction of global climate using multi-dimensional analog methods, in *Preprint Volume, Fifth Conference on Probability and Statistics,* American Meteorological Society, Boston, Mass., 1977.

Barnett, T. P., and R. W. Preisendorfer, Multifield analog prediction of short term climate fluctuations using a climate state vector, *J. Atmos. Sci., 35*(10), 1771–1787, 1978.

Bjerknes, J., Survey of El Niño, 1957–1958 and its relation to tropical Pacific meteorology, *Bull. Amer. Trop. Tuna Comm., 12,* 3–62, 1966.

Blackman, R. B., and J. W. Tukey, *The Measurement of Power Spectra,* 190 pp., Dover, New York, 1958.

Box, G. E. P., and G. M. Jenkins, *Time Series Analysis, Forecasting and Control,* Holden-Day, San Francisco, Calif., 1976.

Davis, R. E., Predictability of sea-surface temperature and sea-level pressure anomalies over the North Pacific Ocean, *J. Phys. Oceanogr., 6*(3), 249–266, 1976.

Davis, R. E., Techniques for statistical analysis and prediction of geophysical fluid systems, *Geophys. Astrophys. Fluid Dyn., 8,* 245–277, 1977.

Davis, R. E., Predictability of sea level pressure anomalies over the North Pacific Ocean, *J. Phys. Oceanogr., 8,* 233–246, 1978.

Elliott, R. D., Extended-range forecasting by weather types, in *Compendium of Meteorology,* pp. 834–840, American Meteorological Society, Boston, Mass., 1951.

Garp, Modelling for the first Garp global experiment, *Publ. 16,* International Council of Scientific Unions, World Meteorol. Organ., Geneva, 1975.

Gilbert, F., Ranking and winnowing gross earth data for inversion and resolution, *Geophys. J. Roy. Astron. Soc., 23,* 125–128, 1971.

Harnack, R. P., and H. E. Landsberg, Winter season temperature outlooks by objective methods, *J. Geophys. Res., 83*(7), 3601–3616, 1978.

Hasselmann, K., Linear statistical models, JOC/SCOR Study Conference on General Circulation Models of the Ocean and Their Relation to Climate, *Dyn. Oceans Atmos.,* in press, 1979*a*.

Hasselmann, K., On the signal-to-noise problem in the analysis of the atmospheric response to external forcing, in *Meteorology of Tropical Oceans*, 251-259, Royal Meteorological Society, London, 1979*b*.

Hotelling, H., Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.*, *24*, 417-441, 1933.

Hurlburt, H. E., J. C. Kindle, and J. J. O'Brien, A numerical simulation of the onset of El Niño, *J. Phys. Oceanogr.*, *6*, 621-631, 1976.

Jenkins, G. M., and D. G. Watts, *Spectral Analysis and Its Application*, 525 pp., Holden-Day, San Francisco, Calif., 1968.

Jones, R. H., Multivariate statistical problems in meteorology, in *Multivariate Analysis IV*, edited by P. R. Krishnaik, pp. 473-481, North-Holland, Amsterdam, 1977.

Kashyap, R. L., and A. R. Rao, *Dynamic Stochastic Models From Empirical Data*, 334 pp., Academic, New York, 1976.

Kendall, M. G., and A. Stuart, *The Advanced Theory of Statistics*, vol. 3, *Design and Analysis, and Time Series*, Charles Griffin and Co. Ltd., London, 1966.

Kurihara, Y., A statistical-dynamical model of the general circulation of the atmosphere, *J. Atmos. Sci.*, *27*, 847-870, 1970.

Kutzbach, J., Empirical eigenvectors of sea level pressure, surface temperature and precipitation complexes over North America, *J. Appl. Meteorol.*, *6*, 791-802, 1967.

Lorenz, E. N., Empirical orthogonal functions and statistical weather prediction, *Rep. 1*, Statist. Forecasting Proj., Mass. Inst. of Technol., Cambridge, 1956.

Lorenz, E. N., An experiment in nonlinear statistical weather forecasting, *Mon. Weather Rev.*, *105*, 590-602, 1977.

McCreary, J., Eastern tropical ocean response to changing wind systems application to El Niño, *J. Phys. Oceanogr.*, *6*, 634-645, 1976.

McWilliams, J. C., and P. R. Gent, A coupled air and sea model for the tropical Pacific, *J. Atmos. Sci.*, *35*(6), 962-989, 1978.

McWilliams, J. C., and P. R. Gent, Corrigendum to 'A coupled air and sea model for the tropical Pacific,' *J. Atmos. Sci.*, *36*(1), 181, 1979.

Mosteller, F., and J. W. Tukey, *Data Analysis and Regression*, 527 pp., Addison-Wesley, Reading, Mass., 1977.

Namias, J., General aspects of extended-range forecasting, in *Compendium of Meteorology*, pp. 802-813, American Meteorological Society, Boston, Mass., 1951.

Pearson, K., On lines and planes of closest fit to systems of points in space, *Phil. Mag.*, *2*, 559-572, 1901.

Preisendorfer, R. W., and T. P. Barnett, Significance test for empirical orthogonal functions, in *Proceedings of the Fifth Conference on Probability and Statistics in Atmospheric Sciences*, pp. 169-172, American Meteorological Society, Boston, Mass., 1977.

Quinn, W. H., Monitoring and predicting El Niño invasions, *J. Appl. Meteorol.*, *13*, 825-830, 1974.

Saltzman, B., A survey of statistical dynamical models of the terrestrial climate, *Adv. Geophys.*, *20*, 183-304, 1978.

Savage, L. J., *The Foundations of Statistical Inference*, Methuen, London, 1962.

Schneider, S. H., and G. E. Dickinson, Climate modeling, *Rev. Geophys. Space Phys.*, *12*, 447-493, 1974.

Sellers, W. D., A two-dimensional global climate model, *Mon. Weather Rev.*, *104*, 233-248, 1976.

Vernekar, A. D., A calculation of normal temperature at the earth's surface, *J. Atmos. Sci.*, *32*, 2067-2081, 1975.

Wyrtki, K., Equatorial currents in the Pacific, 1950 to 1970 and their relation to Trade Winds, *J. Phys. Oceanogr.*, *4*, 372-380, 1974.

Wyrtki, K., El Niño, the dynamic response of the equatorial Pacific Ocean to atmospheric forcing, *J. Phys. Oceanogr.*, *5*, 572-584, 1975.

(Received October 17, 1978;
accepted February 20, 1979.)