Dynamics of Atmospheres and Oceans, 3 (1979) 501—521

LINEAR STATISTICAL MODELS

K. HASSELMANN
Max-Planck-Institut für Meteorologie, Hamburg (W. Germany)
(Received 3 May 1978; accepted 24 July 1978)

Hasselmann, K., 1979. Linear statistical models. Dyn. Atmos. Oceans, 3: 501—521.

1. INTRODUCTION

Climatic variability is defined, in current terminology, as the variability of the coupled atmosphere—ocean—cryosphere—land system on time scales longer than the theoretical limit of deterministic synoptic-scale weather forecasts. Thus climatic prediction, at least as regards the atmospheric component, is necessarily statistical in nature. It has often been surmised (e.g. Lorenz, 1959, 1977; Leith, 1975; Davis, 1976) that statistical forecasting over long time scales can be effectively accomplished by linear models, even when the deterministic equations of the system are strongly nonlinear. It can be argued that the loss of information on the detailed properties of the system implied by a statistical representation will limit the predictability of the reduced statistical system, and it is plausible that within these limitations linear models may then yield an adequate first order description.

A linear treatment is also appropriate when considering the response of the climatic system, or components of the system, to small external influences. These may represent either changes which are external to the entire climatic system (e.g. solar insolation, anthropogenic CO_2 emissions, changes in the dust content of the atmosphere due to volcanic activity, etc.) or variations of the internal transfer rates describing the coupling between individual components of the climatic system (heat transfer at the air—sea interface, air—sea—ice interactions, etc.). The linear transfer functions describing these responses largely characterise the dynamical structure of the climatic perturbations.

This may be illustrated, for example, by the Fokker—Planck model (cf. Hasselmann, 1976)

$$\partial p/\partial t + \sum_i (\partial/\partial y_i)(v_i p) = \sum_{i,j} (\partial/\partial y_i)\ D_{ij} \frac{\partial p}{\partial y_j} \tag{1.1}$$

for the evolution of the probability density $p(\mathbf{y})$ of climatic states in a climatic

phase space $\mathbf{y} = (y_1, y_2, \ldots)$ representing the instantaneous state of the "slow" components of the climatic system (ocean, cryosphere, land vegatation, etc.). Here the velocity v_i denotes the (deterministic) rate of change of $\mathbf{y}$ due to internal coupling within the system, and the diffusion coefficient D_{ij} arises from the stochastic forcing of the slow components of the climatic system by short time-scale atmospheric (weather) disturbances. The atmosphere is assumed to adjust to a statistically stationary equilibrium state (dependent on $\mathbf{y}$) on a time scale short compared with the characteristic time scales of the slow system $\mathbf{y}$. Thus the atmospheric variables are parameterised, for the time scales relevant for climatic variations, in terms of $\mathbf{y}$.

Equation (1.1) represents a closed evolution equation for the climatic system, provided the dependence of the coefficients v_i and D_{ij} on the climatic state $\mathbf{y}$ is known. In addition to knowledge of the internal dynamics of the slow parts of the climatic system, this requires information on the response of the atmosphere to changes of $\mathbf{y}$.

For small perturbations of the climatic states about an equilibrium mean value, the dependence of the coefficients v_i and D_{ij} on the perturbations of $\mathbf{y}$ can be linearised. In this case eqn. (1.1) can be solved analytically; the probability distribution is asymptotically stationary and normal, with moments which can be simply related to the linear response coefficients of the expansions of v_i and D_{ij} (Hasselmann, 1976). Thus the main problem in developing a quantitative description of climatic perturbations is to determine the linear response relations of the basic components of the climatic system.

In practice, it is difficult to derive linearised climatic equations directly from the full nonlinear equations of the atmosphere, ocean and other components of the climatic system. The standard approach is therefore to fit linear models to observed data, normally under constraints expressing particular physical preconceptions regarding the structure of the model. The main problems encountered with this technique, as has been pointed out by Lorenz (1959, 1977) and Davis (1976), lie not so much in the formal fitting procedure, as in the inherent statistical indeterminacy associated with finite data sets. A minimal-error model can be determined only within prescribable error bands. Typical questions which then arise are whether these error limits are sufficiently narrow to distinguish between competing models within a given class, or whether an alternative class of models may have yielded another, perhaps better defined optimal model.

This review will therefore be concerned primarily with the basic problems of statistical uncertainty and significance in model fitting. Although the emphasis will be on linear models, much of the analysis is directly applicable to arbitrary nonlinear models and will accordingly be presented, where appropriate, in a general form.

2. MODEL FITTING

(a) Deterministic models

As model we shall term generally any set of equations

$$r_\nu(\boldsymbol{\alpha}) = 0\ , \qquad \nu = 1, 2, \ldots n \tag{2.1}$$

interrelating the components of a data set $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots \alpha_m)$.

The model relations are often cast in the predictive form

$$r_\nu \equiv \beta_\nu - \mathrm{f}_\nu(\boldsymbol{\alpha}') = 0 \tag{2.2}$$

where $\beta_\nu \equiv \alpha_{\mu_\nu}$ represents a particular data value (the predictand) which is predicted from the remaining data values $\boldsymbol{\alpha}' = (\ldots \alpha_\mu \ldots)$, $\mu \neq \mu_\nu$ (the predictors) through the function f_ν. If the prediction corresponds to a causal physical model, additional side conditions must be satisfied. For example, if the data represent measurements at discrete times, the predictors must represent earlier data than the predictand.

The model-fitting or "inverse" problem arises when the model contains a number of free parameters $\mathbf{a} = (a_1, a_2 \ldots a_q)$, which can be chosen to yield an optimal fit of the model to the data. "Optimal" is generally defined in terms of an error function

$$\epsilon = \sum_{\nu,\mu=1}^{n} M_{\nu\mu} r_\nu r_\mu \tag{2.3}$$

where $M_{\nu\mu}$ is some positive-definite, symmetrical matrix. The optimal model is then given by the parameter vector $\mathbf{a}^0$ which minimizes ϵ:

$$\epsilon(\mathbf{a}^0) = \min \tag{2.4}$$

or

$$\tfrac{1}{2}(\partial\epsilon/\partial a_j) = \sum_{\nu,\mu=1}^{n} M_{\nu\mu} r_\nu (\partial r_\mu/\partial a_j) = 0\ , \qquad j = 1, 2, \ldots q \tag{2.5}$$

By a suitable linear recombination of the set of relations r_ν, the metric can be normalised to the unity matrix, $M_{\nu\mu} = \delta_{\nu\mu}$, and the optimal model becomes the usual least-square solution. However, we shall retain an arbitrary error metric, as we shall require the general form later when discussing a criterion for the choice of the error metric in connection with model validity tests.

For $q \leqslant n$, eqns. (2.5) may define a number of local minima or stationary points, but there will normally exist only a single absolute minimum (2.4).

If the model contains more free parameters than model relations, $q > n$, the model is generally underdetermined. In this case a unique optimal model may be defined, however, by requiring that the relations (2.1) are satisfied

exactly, and that in addition some further positive-definite property η of the model is minimized. For example, if the model predicts a continuous function $\hat{\alpha}(t)$ in $0 \leqslant t \leqslant T$, and the data represent observed values $\alpha(t_j)$ at discrete measurement times t_j, $\alpha_j \equiv \alpha(t_j)$ (the model relations (2.1) being given simply by $r_j \equiv \alpha_j - \hat{\alpha}(t_j) = 0$), η may be defined as the mean-square deviation of $\hat{\alpha}$ from its mean value,

$$\eta = (1/T_0) \int_0^T \{\hat{\alpha}(t) - (1/T_0) \int_0^T \hat{\alpha}(t')\, dt'\}^2\, dt$$

or by some similar measure of the "noisiness" of the model (Backus and Gilbert, 1967; Gilbert, 1971).

More generally, both cases may be combined by minimising the sum $\epsilon + \eta$, thereby requiring both a good fit to the data and a "smooth" model. This method is applicable independently of the number of parameters of the model (Long and Hasselmann, 1979). If η takes the form of a quadratic expression

$$\eta = \sum_{\nu,\mu=n+1}^{n+n'} M_{\nu\mu} r'_\nu r'_\mu$$

in terms of n' model "output" parameters r'_ν (as in the example), the addition of η to the error function is formally equivalent to the extension of the model (and the associated metric) to include further constraints $r'_\nu = 0$ for $\nu = n + 1$, $n + 2$, ... $n + n'$, such that the total number of relations $n + n'$ is greater than the number of parameters.

For the following it is irrelevant whether the quadratic form (2.3) contains additional terms representing "smoothness" criteria, and we shall simply regard the net metric $M_{\nu\mu}$ as given, with $q \leqslant n$ = total number of model relations, with or without possible additional constraints.

(b) Statistical models

Up to this point the data $\boldsymbol{\alpha}$ have been treated as a single, unique set. In statistical modelling, however, the data of a particular experiment are regarded as only one realisation selected from a hypothetical infinite ensemble of possible realisations. The optimal model is accordingly defined with respect to the complete statistical ensemble, rather than a single realisation.

Two techniques for statistical model fitting can be considered, depending on whether ensemble averages (denoted in the following by cornered parentheses) are introduced before or after the definition of the model:
(1) Averaging equations (2.4) and (2.5) yields an optimal statistical model defined by $\langle \epsilon(\boldsymbol{\alpha}, \mathbf{a}) \rangle = \min$, or

$$\sum_{\nu,\mu=1}^{n} M_{\nu\mu} \langle r_\nu (\partial r_\mu / \partial a_j) \rangle = 0\,, \qquad j = 1, 2 \ldots q \tag{2.6}$$

(2) Alternatively, the data α_i can be ensemble averaged prior to model fitting. In this case the original minimal-error equations (2.4) and (2.5) apply

unchanged. The method normally requires the derivation of new data expressions from the original data, for example in the form of quadratic products, which are then "ensemble" averaged by time or space averaging to provide estimates of the input covariance functions or spectra for a deterministic-model fit.

The first technique is generally more useful in constructing maximum skill predictions, whereas the second technique has advantages in testing physical hypotheses. Adopting the terminology of linear time-series analysis we shall refer to the two types of models as filter models and spectral models, respectively. However, to the extent that our considerations apply also to nonlinear models, we shall interpret the term filter here generally to denote arbitrary nonlinear relations between the data of individual realisations, and the term spectrum to imply a spectrum of arbitrary order.

3. LINEAR MODELS

A model will be termed linear if the model relations are linear with respect to the data (but not necessarily with respect to the model parameters **a**). In the predictive form, the model relations are given by

$$r_\nu \equiv \beta_\nu - \sum_{\mu \neq \mu_\nu} A_{\nu\mu}\alpha_\mu - C_\nu = 0 \tag{3.1}$$

In the case of linear filter models, the set of coefficients $A_{\nu\mu}$ and C_ν are often chosen to be identical to the set of model parameters **a**. In this case the averaged least-squares equations (2.6) yield the usual linear regression solutions

$$A_{\nu\mu} = \sum_{\lambda=1}^{n} \langle \beta_\nu \alpha_\lambda \rangle \cdot N^\nu_{\lambda\mu} \qquad \nu, \mu = 1, \ldots, n \tag{3.2}$$

$$C_\nu = \langle \beta_\nu \rangle - \sum_{\mu=1}^{n} A_{\nu\mu} \langle \alpha_\mu \rangle \tag{3.3}$$

where $N^\nu_{\lambda\mu}$ is the inverse of the covariance matrix $\langle (\alpha_\lambda - \langle \alpha_\lambda \rangle)(\alpha_\mu - \langle \alpha_\mu \rangle) \rangle$ of the set of predictors for the predictand β_ν. We note that the solutions (3.2) and (3.3) are independent of the choice of error metric. Note also that the index ν occurs simply as a "tag"; the linear regression coefficients are determined independently for each predictand β_ν. Where notationally convenient we shall therefore suppress the index ν in (3.2), (3.3) and consider formally a single prediction equation.

A simple example of a linear filter model which is nonlinear with respect to its model parameter a is given by the set of relations:

$$r_1 = \beta_1 - a\alpha$$

$$r_2 = \beta_2 - a^2\alpha$$

$$\vdots$$

$$r_n = \beta_n - a^n\alpha \tag{3.4}$$

where $\alpha = x_i$, $\beta_1 = x_{i+1}$, $\beta_2 = x_{i+2}$, ..., $\beta_n = x_{i+n}$ and $x_1, x_2, \ldots$ is a discrete time series. The data set $\alpha, \beta_1, \ldots \beta_n$ represent a data sample of length $n + 1$ beginning at some arbitrary reference time i. The error expressions (3.4) result if a linear prediction is sought for all predictands x_{i+j}, $j = 1, 2, \ldots n$ in terms of the predictor x_i under the side condition that the predictions for all lags j should be mutually consistent with the solutions for a first-order Markov process

$$x_{i+1} = ax_i + z_i \tag{3.5}$$

where z_i represents uncorrelated white noise, $\langle z_i \rangle = 0$, $\langle z_i z_j \rangle = \text{const} \cdot \delta_{ij}$.

This case is actually an example of a model which is more conveniently treated by the spectral method than by filter equations. The (discrete) variance spectrum $F_x(\omega_j)$ of the process x_i defined by (3.5) can be shown to be

$$F_x(\omega_j) = \sum_{p=-\infty}^{\infty} A/(\lambda^2 + (\omega_j - 2p\omega_n)^2)\,, \qquad \omega_j = (2\pi/T)j\,, \qquad j = 1, \ldots n \tag{3.6}$$

where $A = F_z(\omega_j) = \text{const}$ represents the variance spectrum of the white-noise forcing, $\lambda = -\ln(a)/\Delta t$ is the relaxation (e-folding) time of the Markov process (3.5), Δt is the time increment between successive measurements and T^{-1} is the frequency resolution ($T = 2n\Delta t$ is then the length of the record pieces used to estimate the individual Fourier amplitudes by the Bartlett procedure, yielding the Nyquist frequency $\omega_n = \pi/\Delta t$). The equation can also be written in a simpler alternative form, but with a more complicated interpretation of the parameters, by performing the summation over the Nyquist folding frequencies (Reynolds, 1978). Equation (3.6) represents n linear-model relations for the data $F_x(\omega_j)$ which are linearly dependent on the noise parameter A and nonlinearly dependent on the relaxation parameter a. The principal advantage of eqn. (3.6) over the set (3.4) is that the covariance matrix of the sampling errors for spectral estimates is diagonal. This greatly simplifies the error analysis and the testing of model validity (cf. section 7).

4. MODEL INDETERMINACY

Irrespective of whether the model is of the filter or spectral type, the minimal-error fitting technique yields a set of optimal model parameters

$$a_i^0 = \phi_i(\mathbf{m})\,, \qquad i = 1, 2, \ldots q$$

as functions ϕ_i of some set of ensemble-averaged data properties $\mathbf{m} = (m_1, m_2, \ldots m_s)$. Normally, the m_i represent moments, and we shall refer to them simply under this term.

The main difficulty in statistical model fitting is that ensemble-averaged moments cannot be determined exactly from measurements, but must be

estimated from finite data samples. The error $\delta m_j = \tilde{m}_j - m_j$ between the estimated moment $\tilde{m}_j$ and true moment m_j gives rise to a model error

$$\delta a_i = \tilde{a}_i^0 - a_i^0 \approx \sum_{j=1}^{s} (\partial \phi_i / \partial m_j) \delta m_j \tag{4.1}$$

Given the statistical properties of the data, the statistics of the model errors δa_i can then be determined. If the estimates $\tilde{m}_j$ are computed from a fairly large data set, which is normally the case, the joint probability distribution p of $\delta \mathbf{m}$, and therefore also of $\delta \mathbf{a}$, will be approximately Gaussian by the Central Limit Theorem:

$$p(\delta \mathbf{a}) = (2\pi)^{-q/2} T^{1/2} \exp(-\rho^2/2) \tag{4.2}$$

where

$$\rho^2 = \sum_{i,j=1}^{q} T_{ij} \delta a_i \delta a_j \tag{4.3}$$

and T_{ij} is the inverse of the covariance matrix

$$\langle \delta a_i \delta a_j \rangle = \sum_{k,l=1}^{s} (\partial \phi_i / \partial m_k)(\partial \phi_j / \partial m_l) \langle \delta m_k \delta m_l \rangle \tag{4.4}$$

The covariance matrix $\langle \delta m_i \delta m_j \rangle$ can be estimated from the data using standard methods (Jenkins and Watts, 1968).

Unfortunately, it is not possible to deduce the properties of the (hypothetical) statistical data ensemble exactly from a single realisation provided by a particular experiment. One can ask only whether a particular realisation is consistent with an assumed statistical ensemble at some prescribed confidence level. Given the statistical ensemble and the associated true optimal model $\mathbf{a}^0$, eqns. (4.2) and (4.3) can then be used to define a region R in the model phase space $\mathbf{a}$ such that 95%, say, of all optimal models $\tilde{\mathbf{a}}^0$ estimated from finite data sets lie within R. The shape of the region R is to some extent arbitrary, but it is customary to limit the region by a hypersurface of constant probability density. In the present case this corresponds to a hyperellipsoid $\rho^2 \leqslant$ const (Fig. 1). This choice of R is optimal in the sense that it yields the smallest confidence volume in $\mathbf{a}$ — space for given confidence limits, and regions of exceptionally low probability density are excluded. It also has the important property that it is invariant with respect to linear transformations of the variables. The q-dimensional probability distribution $p(\mathbf{a})$ induces for the variable ρ^2 the χ^2 probability distribution

$$p(\rho^2) d\rho^2 = [2^{q/2} \Gamma(q/2)]^{-1} (\rho^2)^{(q/2-1)} \exp(-\rho^2/2) d\rho^2 \tag{4.5}$$

with q degrees of freedom. The estimated optimal model $\tilde{\mathbf{a}}^0$ may then be regarded as consistent with the true optimal model $\mathbf{a}^0$ of an assumed statis-

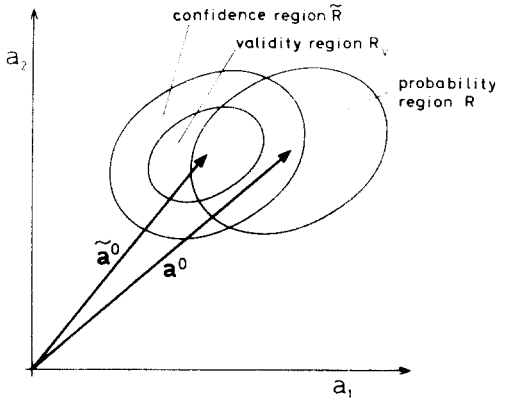

Fig. 1. Relation between the probability region R, confidence region $\tilde{R}$ and validity region R_v (cf. section 7) in the model parameter space $\mathbf{a} = (a_1, a_2, \ldots a_q)$. The true minimal-error model is represented by the vector $\mathbf{a}^0$, the model estimated from a finite data set by $\tilde{\mathbf{a}}^0$.

tical ensemble if the square distance $\rho^2(\tilde{\mathbf{a}}^0 - \mathbf{a}^0)$ is less than the appropriate confidence limit ρ_L^2 of the χ^2-distribution.

Conversely, for a given estimated optimal model $\tilde{\mathbf{a}}^0$ one can now define a confidence region $\tilde{R}$ of permissible true models $\mathbf{a}^0$ such that the distance $\rho^2(\tilde{\mathbf{a}}^0 - \mathbf{a}^0)$ lies within the appropriate confidence limit of the true model $\mathbf{a}^0$. It may be assumed to first order that the covariance matrix $\langle \delta m_i \delta m_j \rangle$, and therefore $\langle \delta a_i \delta a_j \rangle$, remains the same for all statistical ensembles considered, the probability distributions differing only in the positions of the mean value $\mathbf{a}^0$. In this case the confidence region $\tilde{R}$ is identical to the region R except for a shift of the center from $\mathbf{a}^0$ to $\tilde{\mathbf{a}}^0$ due to the interchange of fixed and variable parameters $\mathbf{a}^0$ and $\tilde{\mathbf{a}}^0$ (Fig. 1).

The estimated model $\tilde{\mathbf{a}}^0$ is the "maximum likelihood" model in the sense that the probability density $p(\tilde{\mathbf{a}}^0)$ with respect to $\tilde{\mathbf{a}}^0$, for fixed $\mathbf{a}^0$, is a maximum when the assumed "true" model $\mathbf{a}^0$ is chosen coincident with the estimated model $\tilde{\mathbf{a}}^0$. However, it should be remarked that the expression "maximum likelihood" must be understood here simply as a formal definition; we have not considered the relative likelihood, in the non-technical sense of the word implying probability, of different probability distributions characterised by different $\mathbf{a}^0$. The maximum of p refers to a probability density with respect to $\tilde{\mathbf{a}}^0$, not $\mathbf{a}^0$. In fact it is meaningless in the present context to consider the relative probability of different true models, since we have assumed only a single ensemble defining a single true model. Only the inverse question is well posed, namely whether the observed data set is statistically consistent, within prescribed confidence levels, with a given probability distribution. (The extended statistics needed to consider distributions of true models is discussed in standard text books, e.g. Martin (1971) or in the present framework in Barnett and Hasselmann (1979).)

In fitting models to data it is normally desirable to retain a large number

of model parameters, to ensure that the model class encompasses a close description of the real system, while at the same time minimising the statistical uncertainty of the optimal model fit. Unfortunately, the two requirements are generally incompatible. As is familiar from power spectral analysis, high resolution, requiring a large number of model parameters, generally implies low statistical significance. Moreover, the introduction into a model of parameters which do not significantly improve the model fit can degrade the statistical significance of the more important model parameters and is thus actually harmful, rather than simply not helpful.

To determine the degree of detail which can be statistically supported by a given data set, it is useful to consider a nested sequence of model classes. This is illustrated in the following sections.

5. MODEL NESTING

The degradation of the statistical significance of a model by the inclusion of noisy parameters is demonstrated by the simplest case in which only one parameter, a_1, say, is statistically significant. Assume that all parameters have been ortho-normalised,

$$\langle \delta a_i \delta a_j \rangle = \delta_{ij}$$

Let $(\tilde{a}_1^0)^2 = 10$, say, and $(\tilde{a}_j^0)^2 = 1$ for $j \geqslant 2$. We test the hypothesis that the true optimal model is given by $\mathbf{a}^0 = 0$ (for a linear regression model (3.1)–(3.3), this implies zero predictability).

If the model class is defined to contain only the single free parameter a_1, the estimated optimal value $(\tilde{a}_1^0)^2 = 10$ is found to be significantly different from zero beyond the 99% confidence level. However, as the number of parameters q introduced into the model is increased, the statistical significance of the test variable $\rho^2 = \Sigma_{i=1}^{q} (a_i^2)$ is successively degraded by the addition of noise (Fig. 2). For $q > 9$, the entire model (including the parameter a_1^0) can no longer be distinguished from the zero-predictability model at the 95% confidence level.

The apparent paradox that a model which is statistically significant in its simplest form must be rejected in its entirety when embedded in a larger model class — even though the added parameters are clearly suspect as noise — can be resolved by distinguishing between a priori and a posteriori nesting. If the nesting sequence is specified prior to the analysis of the data, it is permissible to terminate the sequence of models at some value q (in the present case between 1 and 8) for which the resultant optimal model is still statistically significant. However, it is not permissible to terminate a model class sequence which has been defined a posteriori — for example, by reordering the parameters in a decreasing sequence with respect to their individual significance levels. This technique is often applied in various schemes of coefficient screening, whereby coefficients which fail to satisfy individual statistical significance criteria are rejected.

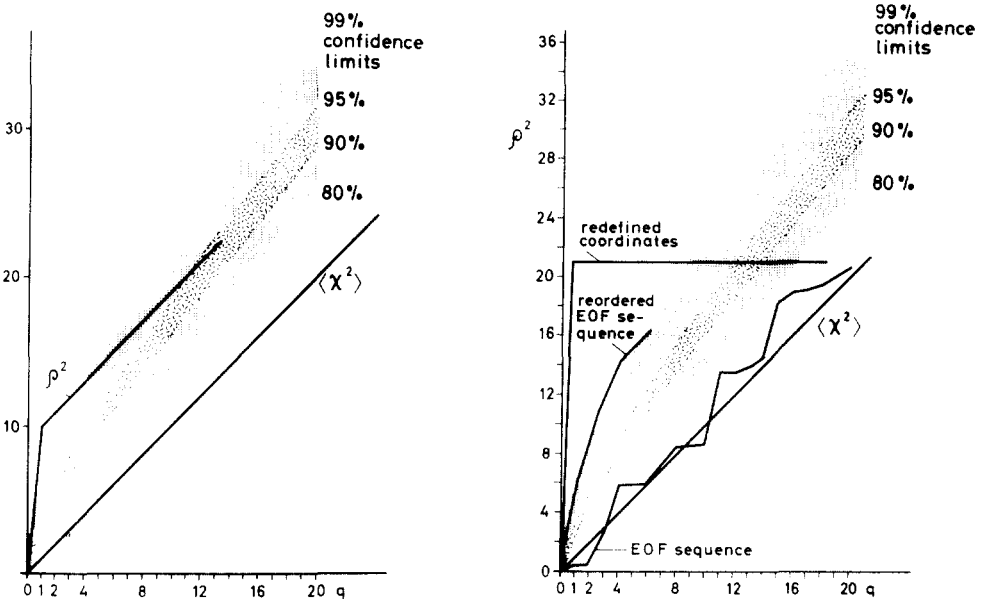

Fig. 2. Degradation of the statistical significance of a model containing a single significant predictor a_1 by inclusion of noisy predictors $a_2, a_3, \ldots$. For $q > 9$ the entire model must be rejected at the 95% confidence level.

Fig. 3. Construction of sequences of apparently significant models from statistically insignificant predictions (EOF sequence) by a posteriori reordering of the predictors, or rotation of the predictor space. The EOFs were formed from 400 predictors from the Equatorial Pacific (20 time lags, 20 time series); the predictand is SSTA at Christmas Island 8 months in the future (from Barnet and Hasselmann, 1979).

The dangers of coefficient screening or reordering are illustrated by a second example (Fig. 3), showing an attempt to predict the sea-surface temperature anomaly (SSTA) at Christmas Island eight months in advance using past and present data from twenty stations in the equatorial Pacific (Barnett and Hasselmann, 1979). The predictor fields consisted of six SSTA stations, various series representing anomalies of trade winds, sea-level and sea-surface pressure, and the Southern Oscillation Index. For each series, data values were taken at 20 time lags extending back two years into the past, yielding a total of $20 \times 20 = 400$ predictors. The covariance matrix of this 400-component predictor vector was orthogonalised by a suitable rotation, and the amplitudes of the resultant empirical orthogonal functions were then taken as the new predictor variables. The nested sequence of model classes obtained by taking the first q EOF amplitudes as predictors yielded the curve ρ^2 versus q shown. The models are statistically indistinguishable from the zero-predictability model $\mathbf{a}^0 = 0$ for all q.

However, if only the most "important" components are retained by

reordering the coefficients with respect to their statistical significance, one obtains a sequence of models which are apparently significant at the 95% confidence level for values of q up to and beyond $q = 8$ (the reordering procedure actually leads to representations in terms of modified functions which are linear combinations of EOFs, rather than the EOFs themselves, since these do not in general yield an exactly diagonal covariance matrix $\langle \delta a_i \delta a_i \rangle$ for the coefficient perturbations, as required for the assessment and ordering of the statistical significance of individual predictors).

A still higher apparent significance can be constructed by rotating the model parameter space such that the new axis a_1' lies in the direction of the vector $\tilde{\mathbf{a}}^0$ representing the minimal error solution. For $q = 20$, this would yield in the present example $(a_1')^2 = \rho^2 = 21$, and a truncation of the series after $q = 1$ would yield an extremely high (but entirely ficticious) significance level.

It is clear from these examples that if the model parameter space has no a priori preferred coordinates, the significance level of a model must be judged in terms of the probability density in the complete q-dimensional parameter space. A posteriori projection on to data-dependent subspaces results in biased statistics.

Screening is equivalent to replacing the original vector $\tilde{\mathbf{a}}^0$ representing the maximum likelihood solution in the q-dimensional model space by an alternative vector $\tilde{\mathbf{a}}_s^0$, in which the smallest components of the vector $\tilde{\mathbf{a}}^0$ (in a space whose axes have been arbitrarily chosen) are set equal to zero. Although this may yield a model which is also consistent with the data (namely if $\mathbf{a}_s^0$ still lies within the likelihood region), there is no a priori reason to regard the screened model as superior to the original solution. The projection on to parameter subspaces is a legitimate, unbiased procedure only if the subspaces have been decided on by data independent, a priori criteria.

In contrast to Fig. 3, in which a possible predictability was lost in the noise of a large number of irrelevant predictors, Fig. 4 shows an example of a statistically significant prediction for the same predictand, SSTA at Christmas Island using a smaller number of predictors. EOFs were again used to define a nested-model class sequence, but in this case only those SSTA stations and atmospheric variables were retained which were anticipated to be effective predictors by a priori physical arguments (Barnett and Hasselmann, 1979). It must be recognised, however, that practically all physical mechanisms proposed to explain long-term, ocean—atmosphere interactions have actually been influenced to some extent by at least cursory inspection of the data. Thus the assumption of a genuine a priori data selection must be questioned also in this example.

This points to a basic dilemma in the objective statistical testing of prediction models for climate studies. In most cases only a rather limited data sample is available and will become available in the near future. In addition to the danger of biasing by a priori data inspection, if a sufficient number of data subsets are considered successively as predictors, ultimately

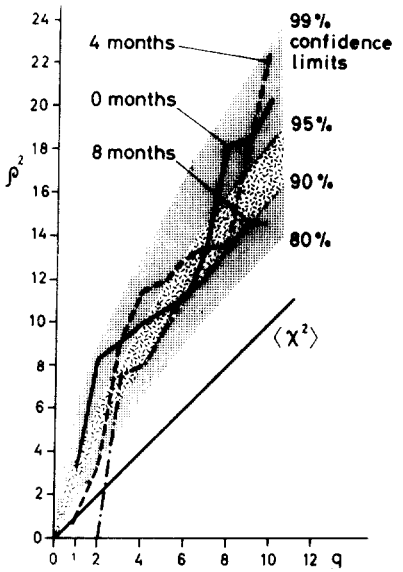

Fig. 4. Construction of statistically significant predictions of SSTA at Christmas Island 0, 4 and 8 months in the future by a priori selection of predictors. The EOFs were constructed in this case from only 4 time series, but again using 20 time lags (from Barnett and Hasselmann, 1979).

one will happen by chance on the most "significant" components of the data. These are the same as one would have found by a posteori screening after an analysis of the complete data set. If the results of the complete regression analysis must be rejected as statistically insignificant, the significance of predictions using predictor subsets, even when chosen a priori, must be interpreted with caution. The number of data subsets which have been tried and rejected before retaining a "statistically significant" subset of predictors must be taken into account. In practice, this can be rather difficult to quantify, in particular since several investigators may be contributing to the model building. A frequent suggestion is to use one part of the data to fit a model and then the remainder for an independent test. This provides some protection against undue high claims of statistical significance resulting from a priori data inspection (provided the second data set is not available for inspection). However, it is powerless against trial-and-error selection, since the probability of success with both data sets simultaneously is essentially the same as for the combined data set.

It appears that, in practice, the validity of prediction models can be only partially supported by purely "objective" statistical tests and must depend to an important extent on the inherent physical credibility of the model.

6. SKILL

The above discussion has been concerned only with the region of statistically acceptable minimal-error models in the model parameter space **a**, without any evaluation of the quality of the resulting models. Various measures of model performance can be considered, depending on the type and purpose of the model.

Filter models are normally designed for prediction. A measure of predictability is given by the skill parameter

$$S = 1 - \langle \epsilon \rangle / (\sum_{\nu,\mu} N_{\nu\mu} \langle \beta_\nu \beta_\mu \rangle) \tag{6.1}$$

which varies between zero, for a zero prediction function f_ν in (2.2), and unity, for vanishing residual error. For true minimal-error models, $S \geqslant 0$, but negative S can occur in estimated minimal-error models (e.g. in the zero prediction case).

For linear regressive models, the residual r_ν and the predictors α_ν are statistically orthogonal, and eqn. (6.1) can then be written in the alternative form

$$S = \langle (\sum_{\mu} A_\mu \alpha_\mu)^2 \rangle / \langle \beta^2 \rangle \tag{6.2}$$

in which S is expressed as the ratio of the predictable variance to the total variance of the predictand. (We have retained here only a single predictand — see the remark following equation (3.3) — and have assumed zero means for β and α_ν.)

In practice, S must be estimated from finite data samples. The hindcast skill S_H is then defined as the estimate which results if the same data sample is used to estimate both the coefficients and the skill,

$$S_H = [\sum \tilde{A}_\mu \alpha_\mu]^2 / [\beta^2] \tag{6.3}$$

where

$$\tilde{A}_\mu = \sum_{\lambda} [\beta \alpha_\lambda] \tilde{N}_{\lambda\mu} , \tag{6.4}$$

$$\tilde{N}_{\lambda\mu} = [\alpha_\lambda \alpha_\mu]^{-1}$$

and the square parentheses denote time or space averages over the finite data sample. Equations (6.3) and (6.4) are identical to eqns. (6.2) and (3.2) except that the ensemble means $\langle \ldots \rangle$ have been replaced by data averages $[\ldots]$.

Since (6.3) is a positive-definite quantity, any errors in estimating the coefficients $A_{\nu\mu}$ will yield a finite hindcast skill, even when the true skill is zero.

The expectation value of S_H can be evaluated by expanding (6.3) with respect to the small perturbations

$$\delta A_\nu = \tilde{A}_\nu - A_\nu \ ,$$

$$\delta m_{\nu\mu} = [\alpha_\nu \alpha_\mu] - \langle \alpha_\nu \alpha_\mu \rangle$$

Since the perturbations are approximately Gaussian, the expectation values of the linear perturbations vanish to first order, and $\langle S_H \rangle$ is determined by quadratic terms. For small values of the coefficients A_ν, i.e. small S, the dominant quadratic terms are those containing the perturbation product $\delta A_\nu \delta A_\mu$, as all other terms retain the small factors A_ν. The expression (6.3) then reduces simply to

$$\langle S_H \rangle = S + \langle S_A \rangle \tag{6.5}$$

where the mean artificial skill

$$\langle S_A \rangle = \sum_{\nu,\mu} \langle \delta A_\nu \delta A_\mu \rangle \langle \alpha_\nu \alpha_\mu \rangle / \langle \beta^2 \rangle \tag{6.6}$$

The estimation of the skill from a second data sample b independent of the data sample a used to estimate the coefficients yields the forecast skill S_F. In this case the residual and predictors are not exactly orthogonal with respect to the averages $[...]_b$, so that the original form (6.1) must be used rather than (6.2). Considering still a single predictand, we have

$$S_F = 1 - [\epsilon]_b / [\beta^2]_b \tag{6.7}$$

where

$$[\epsilon]_b = [(\beta - \sum_\nu (\tilde{A}_\nu)_a \alpha_\nu)^2]_b \tag{6.8}$$

Expanding (6.7) and (6.8) again in a perturbation series, one obtains in analogy with the derivation of (6.5) and (6.6) for small S

$$\langle S_F \rangle = S - \langle S_A \rangle \tag{6.9}$$

Thus the hindcast skill $\langle S_H \rangle$ is increased and the forecast skill $\langle S_F \rangle$ decreased by the same amount, the artificial skill $\langle S_A \rangle$ (Lorenz, 1959, 1977; Davis, 1976). The relations (6.5), (6.6) and (6.9) between $\langle S_F \rangle$, $\langle S_H \rangle$ and $\langle S_A \rangle$ can be readily understood; the deviations from the true prediction coefficients which are introduced to yield an artificially improved hindcast fit to a particular data sample must necessarily yield errors of magnitude comparable to the improvement in the hindcast fit when the model is applied to an independent data sample to which the coefficients were not tuned.

The mean artificial skill $\langle S_A \rangle$ can be computed from the known covariances of the moment estimates using (4.4). The hindcast skill S_H for a particular data realisation can then be compared against $\langle S_A \rangle$ to test if the minimal-

error solution can be distinguished statistically from the zero-predictability model, $A_\nu = 0$. The probability distribution of

$$S_H = S_A = \sum_{\nu,\mu} \delta A_\nu \delta A_\mu \langle \alpha_\nu \alpha_\mu \rangle / \langle \beta^2 \rangle \tag{6.10}$$

for the zero-predictability case is approximately a χ^2-distribution. However, the equivalent number of degrees of freedom is less than q, since — in contrast to the expression (4.3) for ρ^2 — the quadratic form (6.10) is not normalised with respect to the covariance matrix $\langle \delta A_\nu \delta A_\mu \rangle$. (The equivalent number of degrees of freedom of an approximately χ^2-distribution may be defined as the number of degrees of freedom of a χ^2-distribution which has the same ratio of standard deviation to mean.) Thus the inequality $S_H \leqslant$ constant does not define a minimal phase-volume region bounded by a constant probability density surface in the model parameter space A_ν. For this reason the quadratic form ρ^2 is preferable to S_H as a test variable for establishing statistical significance.

7. MODEL VALIDITY

For filter models a zero residual error cannot normally be expected, even for high skill values. The error is zero in the mean only if it is zero for each individual realisation. Thus if the optimal model yields a non-zero residual, this cannot be attributed to sampling errors and must be accepted (apart from measurement errors) as real. In many cases the residuals r_ν actually represent meaningful physical processes, such as the noise input necessary to maintain a dissipative system in a statistically stationary state.

In the case of spectral models, however, where the model relations are expressed in terms of averaged quantities, it is generally conceivable that the model could, in principle, satisfy perfect data exactly. The observed residuals can then be attributed entirely to sampling inaccuracies in the estimation of the moments from finite data samples. The most useful measure of model performance in this case is not the skill, but rather the validity of the model, as inferred from the statistics of the residual error. Typically, spectral models are used to test physical hypotheses rather than to predict, and for this reason also the skill is a less relevant parameter than the model validity. (However, an interesting proposal for using spectral model properties to construct prediction filter models has recently been proposed by Leith (1975).)

Let us assume that for the optimal model $\tilde{\mathbf{a}}^0$, estimated from a finite data sample, there exists a set of moments $\mathbf{m}$ (different from the estimated moments $\tilde{\mathbf{m}}$) for which the model would be exactly valid, $\epsilon(\mathbf{m}, \tilde{\mathbf{a}}^0) = 0$. Under the hypothesis that $\mathbf{m}$ represents the true moments the probability distribution of the error $\tilde{\epsilon}$ for an estimated optimal-fit model, as determined from a finite data set, can be calculated. This then yields a confidence limit ϵ_L for the rejection ($\tilde{\epsilon} > \epsilon_L$) or acceptance ($\tilde{\epsilon} < \epsilon_L$) of the valid-model hypothesis.

For small deviations $\delta m_i = \tilde{m}_i - m_i$ of the estimated moments $\tilde{m}_i$ from their true values m_i, the estimated error is

$$\tilde{\epsilon} = \sum_{\nu,\mu=1}^{n} M_{\nu\mu}\delta r_\nu \delta r_\mu \tag{7.1}$$

where

$$\delta r_\nu = \delta r'_\nu + \delta r''_\nu \tag{7.2}$$

consists of the variation

$$\delta r'_\nu = \sum_{j=1}^{s} (\partial r_\nu/\partial m_j)\delta m_j \tag{7.3}$$

induced by the deviations in the moments and the variation

$$\delta r''_\nu = \sum_{k=1}^{q} (\partial r_\nu/\partial a_k)\delta a_k \tag{7.4}$$

which arises from the errors δa_k in the model parameters incurred by fitting the model to the estimated moments rather than the true ones.

The variations δa_k can be expressed as linear functions of the r'_j by making use of the minimal condition $\partial\tilde{\epsilon}/\partial\delta a_j = 0$. In matrix notation, one obtains

$$\delta\mathbf{a} = -\mathbf{P}^{-1}\mathbf{R}^{+}\mathbf{M}\delta\mathbf{r}' \tag{7.5}$$

where $\mathbf{P} = \mathbf{R}^{+}\mathbf{M}\mathbf{R}$, $R_{\nu k} = \partial r_\nu/\partial a_k$, and $\mathbf{R}^{+}$ denotes the transpose of $\mathbf{R}$.

Substituting (7.5) (7.4) and (7.3) in (7.2), $\delta\mathbf{r}$ is seen to be a linear function of $\delta\mathbf{m}$. For large sample sizes, the errors δm_j and therefore δr_ν are approximately jointly Gaussian. Thus (7.1) defines a variable which has approximately a χ^2-distribution. However, the equivalent number of degrees of freedom is in general less than the number of variables q, since the matrix $M_{\nu\mu}$ in the quadratic form (7.1) is not defined as the inverse of the covariance matrix $\langle \delta r_\nu \delta r_\mu \rangle$. In fact, this is not possible, as it can readily be seen that the error $\tilde{\epsilon}$ vanishes in the subspace spanned by the q column vectors of the matrix $R_{\nu k}$, so that the rank of the covariance matrix $\langle \delta r_\nu \delta r_\mu \rangle$ is maximally $n - q$. If the covariance matrix $\langle \delta r'_\nu \delta r'_\mu \rangle$ is non-singular, the maximal number of degrees of freedom is $f = n - q$, and is attained if $M_{\nu\mu}$ is chosen as the inverse of this matrix (Linnik, 1961; Olbers et al., 1976).

In this case $\tilde{\epsilon}$ represents an optimal variable for testing the hypothesis of a valid model, just as ρ^2 yielded an optimal test variable for the zero-predictability hypothesis. The inequality $\tilde{\epsilon} < \epsilon_L$ defines an $(n - q)$ — dimensional hyperellipsoid in the $(n - q)$ — dimensional error space orthogonal to the q vectors $R_{\nu k}$. For a given confidence value, the ellipsoid has a minimal volume if $M_{\nu\mu} = \langle \delta r'_\nu \delta r'_\mu \rangle^{-1}$, and the surface of the ellipsoid then represents a surface of constant probabiiity density. These conditions provide a criterion for

choosing the error metric, which in the previous considerations remained unspecified. A different choice of $M_{\nu\mu}$ yields a larger confidence ellipsoid which normally contains rather extended regions of very low probability density which should be excluded from the validity region.

In analogy with the definition of the confidence region for prediction models, we can now ask further: in which region of the model parameter space **a** can a model be regarded as a valid model at a given confidence level? To make the problem meaningful, it must be assumed that a set of moments **m** corresponding to a perfect-fit model uniquely determines the associated model parameters $\mathbf{a}^0$, and vice versa.

An important example in which this is not the case is in "consistency testing" (Fofonoff, 1969; Müller and Siedler, 1976). For particular classes of flow fields, such as internal waves, quasi-geostrophic currents or simply incompressible flow, the auto- and cross-spectra for different components of motion must satisfy certain restraints. These are specified by the general structure of the flow field, independent of the spectral distribution of energy. Similar restraints exist if the flow exhibits certain symmetries. In these cases the error expressions r_ν involve only the moments **m** and are independent of **a**. Thus the model-fitting problem does not arise. Although the following discussion is then irrelevant, the statistical tests for the validity of the zero-error hypothesis remain applicable.

We consider now the hypothesis that the perfect-fit model $\tilde{\mathbf{a}}^0$ associated with the true moments **m** does not coincide with the estimated model $\tilde{\mathbf{a}}^0$ of our finite data sample, but deviates from this by a small quantity $\Delta\mathbf{a} = \mathbf{a}^0 - \tilde{\mathbf{a}}^0$. For small $\Delta\mathbf{a}$, the confidence ellipsoid in the error phase space, which is determined by the derivatives $\partial r_\nu/\partial m_j$ and $\partial r_\nu/\partial a_k$ at the parameter values $\mathbf{m}, \mathbf{a}^0$ of the perfect-fit model, will remain approximately constant, independent of the shift $\Delta\mathbf{a}$. However, the calculation of $\tilde{\epsilon}$ for a given data sample is affected, as the errors of the individual model relations r_ν must be defined now with respect to the new perfect-fit model $\mathbf{a}^0$ rather than the optimal-fit model $\tilde{\mathbf{a}}^0$, as previously. Thus in expression (2.3) the individual errors $r_\nu = r_\nu(\tilde{\mathbf{m}}, \tilde{\mathbf{a}}^0)$ must be replaced by $r_\nu(\tilde{\mathbf{m}}, \mathbf{a}^0) = r_\nu(\tilde{\mathbf{m}}, \tilde{\mathbf{a}}^0) + \Sigma_{j=1}^{q} (\partial r_\nu/\partial a_j) \cdot \Delta a_j$. Noting that $\tilde{\mathbf{a}}^0$ is defined as the parameter set which minimizes the net error $\tilde{\epsilon}$, so that the linear terms in the expansion of $\tilde{\epsilon}$ with respect to $\Delta\mathbf{a}$ vanish, one obtains then for the estimate of the net error relative to the model $\mathbf{a}^0$,

$$\tilde{\epsilon}(\tilde{\mathbf{m}}, \mathbf{a}^0) = \tilde{\epsilon}(\tilde{\mathbf{m}}, \tilde{\mathbf{a}}^0) + \Delta\epsilon \tag{7.6}$$

with

$$\Delta\epsilon = \sum_{\nu,\mu,j,k} M_{\nu\mu} R_{\nu j} R_{\mu k} \Delta a_j \Delta a_k \tag{7.7}$$

The model $\mathbf{a}^0$ is then accepted as valid at a given confidence level if $\tilde{\epsilon}(\tilde{\mathbf{m}}, \tilde{\mathbf{a}}^0) < \tilde{\epsilon}_L$, which yields the elliptic relation

$$\sum_{j,k} N_{jk} \Delta a_j \Delta a_k < \tilde{\epsilon}_L - \tilde{\epsilon}(\tilde{\mathbf{m}}, \tilde{\mathbf{a}}^0) \tag{7.8}$$

with the positive definite matrix

$$N_{jk} = \sum_{\nu,\mu}' M_{\nu\mu} R_{\nu j} R_{\mu k} \tag{7.9}$$

If the maximum likelihood model $\tilde{\mathbf{a}}^0$ is accepted as valid, $\tilde{\epsilon}(\tilde{\mathbf{m}}, \tilde{\mathbf{a}}^0) < \epsilon_L$, eqn. (7.8) defines a hyperellipsoid region R_v of models which are equally acceptable at the given confidence level. If $M_{\nu\mu}$ is chosen optimally as the inverse of the covariance matrix $\langle \delta r'_\nu \delta r'_\mu \rangle$, the matrices N_{jk} in (7.8) and T_{jk} in the expression (4.3) for ρ^2 can be shown to be identical.

Figure 5, from Reynolds (1978), gives an example of a validity test for a first-order Markov model (3.5) of sea surface temperature anomalies in the North Pacific. The model corresponds physically to a constant-depth mixed layer driven by local white noise fluctuations of the heat transfer across the air—sea interface (Frankignoul and Hasselmann, 1977). A minimal-error model was determined for each 5°-square of the region shown by a (unit metric) least-squares fit of the logarithms of the predicted spectra, given by (3.6), to the observed spectra. A unit-matrix error metric is optimal in this case, since the covariance matrix $\langle \delta r'_\nu \delta r'_\mu \rangle$ is also proportional to the unit matrix. The model is seen to be valid in the central ocean, but fails along the boundaries and near the equator, where horizontal advection and upwelling may be expected to become important.

The physical hypotheses of the model were more readily tested in this example in terms of the spectra than the corresponding filter relations (3.4). This is generally the case when the model includes assumptions both about

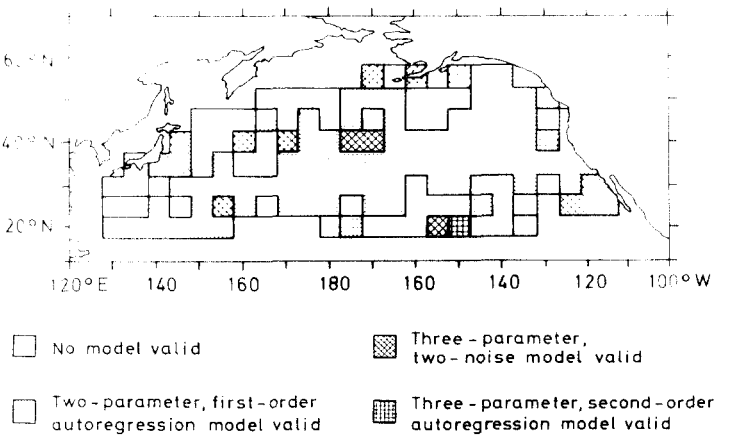

Fig. 5. Regions of validity of the first-order Markov (autoregression) model (3.5) for SSTA in the North Pacific. Also shown are the regions in which the first-order Markov model is invalid, but extended models containing an additional free parameter are valid (from Reynolds, 1978).

the form of the forcing (in this case white noise) and the response (first-order linear relaxation). Filter models are more convenient for constructing optimal prediction models without side conditions regarding the structure of the forcing or the response.

8. CONCLUSIONS

In fitting models to data, one is normally faced with the problem of model indeterminacy due to data uncertainty. In the case of statistical models the uncertainty is associated primarily with finite estimation errors. These can be relatively large, particularly in climate applications, but have the advantage that they can be estimated theoretically. Although not considered here explicitly, instrumental errors can, of course, be regarded simply as a contribution to the total error and treated in the same framework as the sampling errors (provided they are Gaussian).

Two types of statistical models were discussed: filter models, in which the model was defined for individual data realisations and the net model error was obtained by taking ensemble averages over the errors for the individual realisations, and spectral models, in which the model was formulated for ensemble-averaged data variables. In both cases the probability distribution in the model parameter space of the minimal-error models estimated from finite data realisations is approximately Gaussian. Confidence limits of models, discrimination between competing models, etc. can then be discussed in terms of the quadratic form ρ^2 occurring in the exponent of the Gaussian, which has a χ^2 probability distribution with q degrees of freedom where q is the number of model parameters.

Other quadratic forms were found to be important in connection with the question of model performance. The performance of filter models is normally measured in terms of the predictive skill. The hindcast skill S_H, which is estimated from the same finite data sample used to estimate the optimal model, exceeds the true skill S, which in turn is greater than the forecast skill S_F obtained when an estimated optimal model is applied to an independent data sample. For linear regression models and small skill, $\langle S_H \rangle \approx S + \langle S_A \rangle$, $\langle S_F \rangle \approx S - \langle S_A \rangle$, where the artificial skill S_A is given by a positive definite quadratic form in the model coefficient perturbations.

The artificial skill S_A can be used as an alternative variable to ρ^2 for testing the zero-prediction hypothesis. However, the test variable S_A has less resolution and is less reliable than the optimal quadratic form ρ^2 based on the covariance matrix of the coefficient perturbations.

For spectral models the skill is normally less relevant than the validity of the model. The hypothesis of zero model error for the true (ensemble-averaged) moments can be tested using the model error $\tilde{\epsilon}$ for finite data samples as test variable. The parameter $\tilde{\epsilon}$ is given by a quadratic form in the individual model residuals r_ν, which have a joint-normal distribution. If the error metric $M_{\nu\mu}$ is chosen as the inverse of the error covariance matrix $\langle r'_\nu r'_\mu \rangle$ computed

for fixed model parameters, the probability distribution of $\tilde{\epsilon}$ is a χ^2-distribution with a maximal number of degrees of freedom $n - q$ (n = number of model relations). Other choices of the error metric also yield an approximately χ^2-distribution, but with a smaller equivalent number of degrees of freedom and less discrimination.

In applying these concepts to climate data it must be emphasized that objective statistical tests are possible only if there has been no a priori screening of data with respect to the model properties which are to be tested. Because of the limited number and length of climatic time series, this requirement is difficult to fulfill in practice. Most physical models which have been proposed to explain climate variability have been guided to some extent by a priori inspection of the data. The exclusion of data simply on the basis of the observation that there appears to be no obvious correlation between the rejected data and the predictand already represents a biasing of the data. These difficulties become more pronounced when searching for subtle interactions, such as tele-connections, between a large number of fields.

In conclusion, we may have to accept the fact that the conditions for purely objective statistical tests of model hypotheses are often not satisfied in practice and that the credibility of a model will have to rest to a large part also on the intrinsic credibility of the physics of the model. Nevertheless, a careful analysis of the statistical significance and determinacy of a model under clearly stated data-selection conditions remains a necessary, if not always sufficient, requirement for assessing model performance.

ACKNOWLEDGMENTS

The author is grateful for a number of helpful discussions with Tim Barnett, Dirk Olbers, Peter Lemke and Peter Müller.

REFERENCES

Backus, G.E. and Gilbert, J.F., 1967. Numerical applications of a formalism for geophysical inverse problems, Geophys. J.R. Astron. Soc., 13: 247—276.

Barnett, T.P. and Hasselmann, K., 1979. Techniques of linear prediction, with application to oceanic and atmospheric fields in the tropical Pacific. Rev. Geophys. Space Phys., in press.

Davis, R.W., 1976. Predictability of sea surface temperature and sea level anomalies over the North Pacific. J. Phys. Oceanogr. 6: 249—266.

Fofonoff, N.P., 1969. Spectral characteristics of internal waves in the ocean. Deep Sea Res., Suppl., 16: 59—71.

Frankignoul, C. and Hasselmann, K., 1977. Stochastic climate models, Part 2. Application to sea surface temperature anomalies and thermocline variability. Tellus, 29: 289—305.

Gilbert, J.F., 1971. Ranking and winnowing gross earth data for inversion and resolution, Geophys. J.R. Astron. Soc., 23: 125—128.

Hasselmann, K., 1976. Stochastic climate models, Part 1. Theory. Tellus, 28: 473—485.

Jenkins, G.M. and Watts, P.G., 1968. Spectral Analysis and its Applications. Holden-Day, San Francisco, 521 pp.

Leith, C.E., 1975. Climate response and fluctuation dissipation, J. Atmos. Sci., 32: 2022—2026.

Linnik, Yu.V., 1961. Method of least squares and principles of the theory of observations. Pergamon, Oxford.

Long, R.B. and Hasselmann, K., 1979. A variational technique for extracting directional spectra from multi-component wave data. J. Phys. Oceanogr., in press.

Lorenz, E.N., 1959. Empirical orthogonal functions and statistical weather prediction, Scient. Rep. 1. Statist. Forecasting Project, MIT.

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

Martin. B.R.. 1971. Statistics for Physicists. Academic Press, London, New York, 209 pp.

Müller, P. and Siedler, G., 1976. Consistency relations for internal waves. Deep Sea Res., 23: 613—628.

Olbers, D.J., Müller, P. and Willebrand, J., 1976. Inverse technique analysis of a large data set. Phys. Earth Planet. Inter., 12: 248—252.

Reynolds, R.W., 1978. Sea surface temperature anomalies in the North Pacific Ocean, Tellus, 30: 97—103.