Multivariate Statistical Analysis of a Sea Surface Temperature Anomaly Experiment with the GISS General Circulation Model I

GERHARD HANNOSCHÖCK

Max-Planck Institut für Meteorologie, Hamburg, FRG

CLAUDE FRANKIGNOUL*

Laboratoire de Physique et Chimie Marines, Université Paris VI, 4 place Jussieu, Paris 5è, France

(Manuscript received 15 June 1984, in final form 22 February 1985)

ABSTRACT

The multivariate statistical analysis of sensitivity experiments with atmospheric GCMs is difficult because the sample size is always much smaller than the dimensionality of the GCM fields. Thus, Hasselmann has suggested using a hypothesis testing method, where the anticipated GCM response is represented by an *a priori* sequence of guessed patterns characterized by only a few parameters. Here we extend it to the more realistic case where the sample size is limited. When only a few GCM runs are available, it is shown that the statistical significance of the guessed patterns is best established in the full GCM space, using a Hotelling T^2-test and no optimization procedure. Only in the case of large sample size might it be advantageous to work in the subspace defined by the empirical orthogonal functions of the sample GCM noise field, and to consider rotated guess vectors leading to an optimal signal-to-noise ratio. However, the distribution of the test statistic is then only known asymptotically, and the method is sensitive to the correctness of the guesses and to sampling errors in the noise field.

The method is used to evaluate the sensitivity of the Goddard Institute for Space Studies (GISS) GCM Model I to a North Pacific sea surface temperature anomaly. After discussing standard univariate tests of significance, the multivariate procedure is applied, using a sequence of large-scale spherical harmonics as *a priori* guesses. The analysis is done both in the full GCM space and in the subspace of the sample noise field. It is found that the SST anomaly has a significant large-scale influence on the wintertime circulation of the model. A two mode linear wave model is then used to provide dynamical guesses for the GCM response, but only the barotropic response is consistent with the GCM data. This is due to uncertainties in the heating data, and to the oversimplicity of the linear model.

1. Introduction

General Circulation Models (GCMs) of the atmosphere are one of our most powerful tools for understanding the climatic impact of sea surface temperature or sea ice anomalies, volcanic eruptions, and CO_2 increase or other man-produced changes in the environment. However, the *interpretation* of climate sensitivity experiments with GCMs poses serious problems. Indeed, these models have, like the atmosphere, a large natural variability, and an impact on climate can only be decided by testing if the differences between "anomaly" and "control" runs are statistically significant. Since GCM experiments are costly, few runs are generally available and the signal-to-noise ratio is small. Furthermore, GCM variables have large correlation scales; hence data at different grid points are not statistically independent. This requires the use of multivariate tests of significance which are difficult to apply to the GCM case because of the limited sample size. Thus, although the analysis of sensitivity experiments has rested exclusively, until very recently, upon simpler univariate significance tests, it is becoming increasingly clear that such procedure can lead to erroneous results.

Although no statistical tests were used in the early sensitivity experiments with GCMs (e.g., Rowntree, 1972), univariate testing of the null hypothesis that there is no true climate change has been applied routinely since the midseventies. The most commonly used algorithm was introduced by Chervin and Schneider (1976); it is a t test, applicable when several independent runs are available (e.g., Kutzbach *et al.*, 1977; Rowntree, 1979; Chervin *et al.*, 1980). Other algorithms deal with the individual GCM time series and are more appropriate to the analysis of long simulations in the "perpetual" mode (Shukla, 1975; Katz, 1982, 1983). Typically, the null hypothesis is tested at some prescribed level of statistical significance (for example, 5%) at each grid point and for each

* Also affiliated with: Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

variable. When the null hypothesis is rejected, differences between anomaly and control runs are often interpreted *a posteriori* as evidence of a climate response without consideration of the *global* rejection rate. However, for a pattern to be significant as a whole (collective significance), the univariate null hypothesis generally has to be rejected in much more than 5% of the individual cases, because of the limited spatial network and data interdependence (Hasselmann, 1979). Livezey and Chen (1983) (see also von Storch, 1982) showed that the critical rejection rate of the null hypothesis could be inferred from the binomial distribution for a finite sample of *independent* variables. However, the strong interdependence of GCM data reduces the effective number of independent tests and thus increases the requirements for global significance. To evaluate statistical significance, the multivariate character of the GCM variables must be taken into account explicitly.

Unfortunately, the standard multivariate techniques cannot be applied straightforwardly to the evaluation of GCM experiments, because the dimensionality must be much smaller than the sample size. Since the number of independent runs is small, typically of order 10, and the dimensionality of the GCM fields is very large, typically of order 10^4–10^5, there are severe limitations on the amount of information or details that can be effectively evaluated in sensitivity experiments, and the dimensionality must be reduced *a priori* in a considerable manner. Hasselmann (1979) has considered the GCM response problem as one of pattern recognition, and suggested making *a priori* hypotheses on the general structure of the expected GCM response, using some prior knowledge or simpler dynamical models. The guesses are patterns characterized by only a few parameters, and they must be ordered *a priori* in a sequence reflecting their anticipated contribution to the total response. The sequence is normally terminated when the measure of multivariate statistical significance falls below a prescribed level. Hasselmann also showed how to use the properties of the noise field (the GCM natural variability, characterized by its empirical orthogonal functions or EOFs) to optimize the signal-to-noise ratio. This leads to considering modified guesses which are "rotated" away from the directions of large noise.

The multivariate approach advocated by Hasselmann (1979) has both advantages and limitations. Indeed, if the assumption of normality holds, the multivariate test of significance can tell us whether the sample GCM response is consistent with a particular *a priori* assumption, but it does not tell us whether there is any significant response at all. In other words, if one guesses wrong, what does one learn? At the same time, this restrictive aspect of the method makes it a convenient tool for the testing of hypotheses, and therefore it could fill a gap between mechanistic studies, GCM experiments and, potentially, observations. Indeed, much effort has been made to understand, for example, the atmospheric response to SST anomalies, using both simple linear wave models (e.g., Egger, 1977; Webster, 1981) and general circulation models (e.g., Kutzbach *et al.*, 1977; Rowntree, 1979). Yet, attempts at combining the two approaches (Chervin *et al.*, 1980; Roads, 1980) have not been convincing due to lack of an appropriate methodology. Clearly, Hasselmann's method allows one to test directly if linear wave models and GCM simulations are compatible, and whether the former can be used to interpret the latter, which is a question of much interest.

Although Bell (1982) has applied the optimal weighting to the problem of detecting the warming due to increasing CO_2 from data, and Hayashi (1982) has calculated joined confidence intervals for GCM variables, Hasselmann's hypothesis-testing approach has not yet been applied to a GCM experiment. Here, we discuss and clarify the methodology, and we use it to evaluate the sensitivity of the GISS (Goddard Institute for Space Studies) general circulation Model I (Hansen *et al.*, 1983) to a sea surface temperature (hereafter SST) anomaly in the North Pacific. Since Model I is a preliminary version of the GISS GCM, our main emphasis is on the methodology rather than on the atmospheric response to the SST anomaly, and the applicability of our results to the climate problem must be viewed with caution.

Recently, alternative multivariate approaches based on nonparametric permutation techniques have been developed (see review in Livezey, 1983). Livezy and Chen (1983) suggested using random selections of control runs to establish the probability distributions of *a priori* hypotheses, and using these distributions to test the particular outcome of the anomaly experiments. More appropriately to the small-sample case, Preisendorfer and Barnett (1983) have used both the control and the anomaly runs to construct the reference probability distributions, and they have introduced new measures of location, spread and shape of multivariate fields on which significance decisions can be based. The significance is estimated by comparison with the anomaly and control runs separately. The advantage of the permutation techniques is that they do not require the assumption of normality. On the other hand, the hypothesis testing approach may yield more easily interpretable results, and it requires less computations. If suitably handled, both methods should give similar results. Storch and Kruse (1985) have applied both techniques to a GCM experiment, and reached similar conclusions in the two cases.

After describing the usual univariate tests of significance in Section 2, we introduce the multivariate problem in Section 3 and show how the testing of hypotheses can reduce the dimensionality to a tractable level. In Hasselmann (1979), the problem

was simplified by considering the asymptotic case where the sample size is large, and the covariance matrix and the EOFs can be assumed to be known. Then, the measure of significance is distributed like a χ^2 variable, and an optimal signal-to-noise ratio can be easily obtained, since the rotation of the guess vectors is well-determined. Here we consider the more realistic case where the sample size may be small, and only an estimate of the error covariance matrix can be found. The number of guesses that can be tested must then be well under the number of independent experiments. When only a few GCM runs are available (small sample size), it is preferable to apply the multivariate method directly in the full GCM space, and the measure of significance follows a Hotelling T^2 distribution. Only in the case of large sample size might it be advantageous to work in the EOF subspace and to consider rotated guess vectors leading to optimal significance. However, the exact distribution of the test statistic has not been established, and more work is needed for determining an optimal methodology.

In Section 4, the univariate and multivariate tests of significance are applied to the sea surface temperature-anomaly experiment with Model I. We use "nondynamical" guesses (an *a priori* ordered sequence of spherical harmonics) to show that there are significant large-scale changes in the atmospheric circulation. In Section 5, an attempt is made at using a simple linear wave model to provide a dynamical guess for the anticipated GCM response.

2. Univariate tests of significance

a. The two-sample t-test

Chervin and Schneider (1976) first used the two-sample t-test to evaluate the significance of prescribed change GCM experiments. The test has been applied routinely thereafter. Suppose N independent control and M independent anomaly runs are available. Let $\bar{X}^c(n)$ denote the sample mean of a GCM variable at some grid point in control run n and $\bar{X}^a(m)$ the corresponding sample mean in anomaly run m. Then, assuming normality, it is easy to test whether the two populations have the same true mean (i.e., whether there is no true change due to the prescribed change), since the quantity

$$t = \frac{\bar{X}^a - \bar{X}^c - (\mu^a - \mu^c)}{s(1/N + 1/M)^{1/2}} \tag{1}$$

is a t random variable with $\nu = N + M - 2$ degrees of freedom. Here μ^a and μ^c are the true mean of the anomaly and control runs, respectively; X^a and X^c are the corresponding sample means, given by

$$\left.\begin{aligned} \bar{X}^a &= \frac{1}{M}\sum_{m=1}^{M} \bar{X}^a(m) \\ \bar{X}^c &= \frac{1}{N}\sum_{n=1}^{N} \bar{X}^c(n) \end{aligned}\right\}, \tag{2}$$

and s is an unbiased pooled estimate of the variance, given by

$$s^2 = \frac{\sum_{n=1}^{N} (\bar{X}^c(n) - \bar{X}^c)^2 + \sum_{m=1}^{M} (\bar{X}^a(m) - \bar{X}^a)^2}{N + M - 2}. \tag{3}$$

In the absence of an *a priori* hypothesis on the GCM response, a two-sided t-test must be used, and the null hypothesis $\mu^a = \mu^c$ is rejected at the α level of significance if

$$|t| > t_\nu\left(1 - \frac{\alpha}{2}\right), \tag{4}$$

where $t_\nu(1 - \alpha/2)$ represents the upper percentage point of the t distribution. Note that these definitions require that the variances in the control and anomaly runs are not too different, which is a reasonable assumption if the prescribed change is small.

The evaluation of the collective significance of the individual significance tests has been largely ignored by GCM users who prefer to discuss the features that are significant according to the test (4), independently of the overall rate of rejection of the null hypothesis. By definition, one expects that the null hypothesis will be falsely rejected in α% of the cases on the average, if an α% level of significance is used. However, *global significance* requires that a larger percentage of individual tests yield rejection of the null hypothesis at the α% level in the case of finite grid. Livezey and Chen (1983) have shown that the critical rejection rate could be inferred from the binomial distribution for a finite sample of independent variables and that rather stringent conditions are found when the number of independent tests is not large. For instance, the threshold for rejection of the null hypothesis at the 5% level is about 14% of the individual 5% significance tests if there are $n = 30$ independent tests, 10% for $n = 80$, 7% for $n = 500$. Clearly, taking into account the interdependence between the GCM variables would reduce the number of independent tests and increase the effective threshold. However this increase cannot be estimated quantitatively, and the significance problem must be considered instead as a multivariate one.

b. A t-test for autocorrelated variables

If only a few long runs in the perpetual mode are available for the control and anomaly experiments, one may use more fully the available information in

dealing with the red noise character of the GCM variables and taking into account their finite correlation time when estimating the standard errors of time averages (Leith, 1973).

The variance of the sample mean $\bar{x}$ of a statistically stationary process $x(1), \ldots, x(L)$ is given by

$$\sigma_{\bar{x}}^2 = \frac{1}{L} \sum_{l=-(L-1)}^{L-1} \left(1 - \frac{|l|}{L}\right) \sigma_x^2(l), \tag{5}$$

where $\sigma_{\bar{x}}^2(l)$ is the lagged covariance of $x(t)$. If the averaging time is much larger than the correlation time of the natural variability of the GCM variables, it can be assumed that $\bar{X}^a$ and $\bar{X}^c$ are approximately normally distributed (using the central limit theorem). Then, if estimators $s_{\bar{X}^c}^2$ and $s_{\bar{X}^a}^2$ can be found for the variance of the mean of the control and anomaly runs, respectively, the null hypothesis may again be tested since

$$t = \frac{\bar{X}^a - \bar{X}^c - (\mu^a - \mu^c)}{(s_{\bar{X}^a}^2 + s_{\bar{X}^c}^2)^{1/2}} \tag{6}$$

is a t-variable with an approximate equivalent number of degrees of freedom given by

$$\nu = \frac{s_{x^a}^2(0)}{s_{\bar{X}^a}^2} + \frac{s_{x^c}^2(0)}{s_{\bar{X}^c}^2} - 2, \tag{7}$$

where $s_{x^a}^2(0)$ and $s_{x^c}^2(0)$ are variance estimators for the anomaly and control runs, respectively. The main problem with this method is that no *unbiased* estimator of the lagged covariance $\sigma_{\bar{x}}^2(l)$ exists if the true mean is not known, and that the use of traditional estimators in (5) results in an unacceptable bias (Anderson, 1971). We should note that Laurmann and Gates (1977) have suggested considering differences between anomaly and control runs with the underlying hypothesis that the true mean $\mu = \mu^a - \mu^c$ is zero. In this case, there is an unbiased estimator of $\sigma_x^2(l)$ and thus of (5). Laurmann and Gates suggested that the latter could be used in a t-test as before. However, the definition of a t-variable requires the independence of the random variables in the numerator and denominator, which would not be the case. It is easy to show that with such an improper t-test, the null hypothesis would be most easily accepted when the true mean difference is largest.

The variance of the sample mean x can also be written as

$$\sigma_{\bar{x}}^2 = \frac{1}{L} \int_{-\pi}^{\pi} \frac{\sin^2 \frac{\omega L}{2}}{2\pi L \sin^2 \frac{\omega}{2}} f_x(\omega) d\omega, \tag{8}$$

where $f_x(\omega)$ is the spectral density at frequency ω, and the time interval between successive values of $x(t)$ is taken to be unity. Equation (8) can be estimated using spectral analysis since one has asymptotically for large L,

$$\sigma_{\bar{x}}^2 \sim \frac{1}{L} f_x(0), \tag{9}$$

where $f_x(0)$ is the spectral density near zero frequency. Frequency spectra of atmospheric variables are approximately white at low frequencies, hence $f_x(0)$ can be estimated if the time series are long enough. If $F_x(0)$ is a pooled estimator of $f_x(0)$ with ν degrees of freedom, the null hypothesis for the GCM response can then be verified with a t-test, since

$$t = \frac{\bar{X}^a - \bar{X}^c - (\mu^a - \mu^c)}{[(1/L_a + 1/L_c)F_x(0)]^{1/2}} \tag{10}$$

is a t-variable with ν degrees of freedom (Jones, 1976). Here L_c and L_a are the total number of data points in the control and anomaly runs, respectively. Since it is important to use an unbiased estimator of the spectral density at zero frequency, we favor averaging raw spectral density estimates obtained for each run by Fourier Transform in the frequency interval $0 < \omega < \omega_{\max}$. The zero frequency estimates are not used since they are biased (the true mean is not known) and since the t-test requires that the signal and the variance estimates are uncorrelated random variables. Thus a smoothed spectral estimator of $f_x(0)$ with $\nu = 2n$ degrees of freedom can be obtained from

$$F_x(0) = \frac{1}{n} \sum_{i=1}^{n} a(\omega_i) a^*(\omega_i), \tag{11}$$

where $a(\omega_i)$ is the complex Fourier amplitude at frequency ω_i in one control or anomaly run, suitably scaled, and where n is the total number of estimates in the interval $0 < \omega < \omega_{\max}$. In practice, however, GCM time series may not be long enough, and the frequency averaging needed to have enough degrees of freedom may include frequencies for which the spectrum is not entirely flat. This will generally cause an underestimation of $f_x(0)$, hence an overestimation of the percentage of rejection of the null hypothesis (Jones, 1976).

This test for nonindependent variables has not been applied satisfactorily to GCM experiments [Shukla (1975) and Moura and Shukla (1981) estimated expression (5) but did not calculate confidence intervals], and it will be applied to the GISS data in Section 4c. It should be noted that Katz (1982, 1983) has devised an alternative statistical procedure based on parametric time series modeling which also takes into account finite correlation times.

3. Multivariate analysis and hypothesis testing

a. Signal significance

Like the atmospheric fields, the GCM fields are highly correlated in space, and a correct assessment

of the statistical significance of GCM prescribed change experiments should be made with multivariate methods. Multivariate tests of significance are basically straightforward generalizations of the univariate tests, but their application to GCM data is nonetheless much more difficult. They are discussed here in the case where several independent runs are available, since most GCM operate with a seasonal cycle, rather than with perpetual month forcing as in the application below.

Let all variables and grid points in the GCM be denoted by the n-dimensional vector $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$. Using the notations of Section 2a in vector form, we have N independent control runs of sample means $\bar{\mathbf{X}}^c(r)$, $r = 1, N$ and M independent anomaly runs of sample means $\bar{\mathbf{X}}^a(m)$, $m = 1, M$. If $\boldsymbol{\mu}^c$ and $\boldsymbol{\mu}^a$ denote the true means, the null hypothesis $\boldsymbol{\mu}^a = \boldsymbol{\mu}^c$ that the anomaly and control runs have the same true mean can in principle be tested by considering the statistic (e.g., Morrison, 1976)

$$T^2 = \left(\frac{1}{N} + \frac{1}{M}\right)^{-1} (\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c)'\mathbf{S}^{-1}(\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c), \quad (12)$$

where $\bar{\mathbf{X}}^a$ and $\bar{\mathbf{X}}^c$ are the estimates of $\boldsymbol{\mu}^a$ and $\boldsymbol{\mu}^c$ as in (2), the prime denotes the transpose vector, and normality has been assumed. The matrix $\mathbf{S}$ is an unbiased estimate of full rank n of the true error covariance matrix $\mathbf{C}$ and is given by

$$\mathbf{S} = \frac{1}{N+M-2}\left\{\sum_{m=1}^{M} [\bar{\mathbf{X}}^a(m) - \bar{\mathbf{X}}^a][\bar{\mathbf{X}}^a(m) - \bar{\mathbf{X}}^a]' + \sum_{r=1}^{N} [\bar{\mathbf{X}}^c(r) - \bar{\mathbf{X}}^c][\bar{\mathbf{X}}^c(r) - \bar{\mathbf{X}}^c]'\right\}. \quad (13)$$

The two-sample Hotelling T^2 statistic (12) is the direct analogue of the univariate t^2, and it is invariant with respect to linear transformations of coordinates in the n-dimensional space. The null hypothesis is rejected at the α-level if

$$T^2 > \frac{(N+M-2)n}{N+M-n-1} F_{\alpha;n,N+M-n-1}, \quad (14)$$

where $F_{\alpha;n,N+M-n+1}$ denotes the 100α upper percentage point of Fisher's F distribution with n and $N + M - n - 1$ degrees of freedom. (When the covariance matrix $\mathbf{S}$ is estimated from low-frequency Fourier coefficients as in Section 2b, the formulation for the test statistic is the direct multidimensional analogue of (10), and the test (14) holds with the appropriate number of degrees of freedom.)

If the covariance matrix $\mathbf{C}$ were known or could be treated as such, as considered by Hasselmann (1979), the test statistic would keep the same form as in (12),

$$\rho^2 = \left(\frac{1}{N} + \frac{1}{M}\right)^{-1} (\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c)'\mathbf{C}^{-1}(\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c), \quad (15)$$

and would be distributed as an χ^2 variable with n degrees of freedom if the null hypothesis holds. The null hypothesis would therefore be rejected at the α-level when

$$\rho^2 > \chi^2_{\alpha;n}. \quad (16)$$

In the GCM case, however, the true error covariance matrix is not known and (12) must be used.

In practice, the test (12) cannot be applied directly to GCM experiments because the number of independent runs is small and therefore the estimate of the error covariance matrix has only a small number of degrees of freedom, typically of order 10. Since the dimensionality of the GCM fields is much larger, the covariance matrix $\mathbf{S}$ is of very reduced rank and information about the "noise" is only available in a subspace of much lower dimension.

To apply the multivariate tests, the dimensionality must be strongly reduced. Let n_c be the rank of the matrix $\mathbf{S}$, equal to the number of degrees of freedom ν of the covariance estimates. If the GCM fields are projected onto a set of n_c new basis vectors, the test (12) can then be applied in the truncated space, since the covariance matrix has become of full rank n_c. However, the significance conditions encountered in multidimensional tests are very stringent when the number of degrees of freedom does not largely exceed the dimensionality of the fields. Hence, a signal of reasonable magnitude would only be accepted as statistically significant if the dimensionality has been reduced further. This severely limits the amount of detail of the GCM experiment that can be evaluated in practice. Furthermore, it obliges one to make *a priori* assumptions on the general structure of the GCM response to the prescribed change, because of the subjective choice of a highly truncated representation. It is therefore equivalent to use the *hypothesis testing* strategy suggested by Hasselmann (1979). If *a priori* hypotheses regarding the expected GCM response can be formulated, the signal can in this case be characterized by a limited number of parameters (much smaller than the number of degrees of freedom) and the significance tests can be applied, even in the full GCM space.

b. *A priori guesses*

Let us assume that because of prior knowledge of the expected structure of the atmospheric response to the prescribed change, we can formulate a first guess $\mathbf{g}_1 = (g_{1,1}, g_{1,2}, \ldots, g_{1,n})$ of the expected response. In many cases, we will also be able to estimate the likely structure of the deviations of the first guess from the true response, hence an improved guess of the latter may be represented by a linear combination of the two vectors $\mathbf{g}_1$ and $\mathbf{g}_2$. In general, there will be a sequence of guesses $\mathbf{g}_\alpha$, $\alpha = 1, 2, \ldots, p (p < \nu)$, which have been ordered *a priori*. The guesses could

be the predictions of some simplified dynamical model (for instance $\mathbf{g}_1$ could be the linear response of a barotropic model, $\mathbf{g}_2$ the first baroclinic correction, etc.) or simply a sequence of nondynamical guesses based on prior knowledge (for instance a sequence of large-scale spherical harmonics).

The hypothesis that the true GCM response can be represented by a linear combination of the p guesses $\mathbf{g}_\alpha$ is written in the n-dimensional space as

$$\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c = \sum_{\alpha=1}^{p} \gamma_\alpha \mathbf{g}_\alpha + \mathbf{r}, \tag{17}$$

where the γ_α are scalar parameters and $\mathbf{r}$ is a residual noise. The true value of γ_α is unknown but an estimate $\tilde{\gamma}_\alpha$ can be determined by minimizing the residual error $|\mathbf{r}|^2$. This yields

$$\tilde{\gamma}_\alpha = \sum_{\beta=1}^{p} G_{\alpha\beta}^{-1} (\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c)' \mathbf{g}_\beta, \tag{18}$$

where $\mathbf{G}^{-1}$ is the inverse of the $p \times p$ matrix with elements $G_{\alpha\beta} = \mathbf{g}_\alpha' \mathbf{g}_\beta$. Since $(\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c)$ has a multivariate normal distribution, the parameters $\tilde{\gamma}_\alpha$ are also normally distributed, being linear combinations from normal variables. The null hypothesis that there is no atmospheric response is that the true parameters γ_α are all equal to zero, and it can be tested by verifying that the p-dimensional vector of estimated parameters $\tilde{\gamma} = (\tilde{\gamma}_1, \tilde{\gamma}_2, \ldots, \tilde{\gamma}_p)$ lies inside a fixed region—the noise ellipsoid—around the zero vector. Thus, we consider the test statistic

$$T^2 = \left(\frac{1}{N} + \frac{1}{M}\right)^{-1} \tilde{\gamma}' \Gamma^{-1} \tilde{\gamma}, \tag{19}$$

where the $(p \times p)$ error covariance matrix Γ is computed from the $(n \times n)$ error covariance matrix $\mathbf{S}$ using

$$\Gamma_{\alpha\beta} = \mathbf{d}_\alpha' \mathbf{S} \mathbf{d}_\beta. \tag{20}$$

Here $\mathbf{d}_\alpha'$ are the row vectors from the n-dimensional dual space projecting $(\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c)$ on $\hat{\gamma}_\alpha$, since (18) can be written

$$\tilde{\gamma}_\alpha = \mathbf{d}_\alpha' (\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c), \tag{21}$$

with

$$\mathbf{d}_\alpha = \sum_{\beta=1}^{p} G_{\alpha\beta}^{-1} \mathbf{g}_\beta. \tag{22}$$

The null hypothesis is rejected at the α-level if

$$T^2 > \frac{(N + M - 2)p}{N + M - p - 1} F_{\alpha;p,N+M-p-1}. \tag{23}$$

A model selection criterion (of order p to stop the hierarchy of guesses if $p > 1$) can be obtained by choosing the highest model satisfying (23) for increasing values of p. There are different selection criteria that may put more or less emphasis on skill and significance (e.g., considering the incremental changes in significance associated with the addition of each new guess vector) but it is important that the selection criteria be defined *a priori* (Barnett et al., 1981). The procedure can be generalized to the statistical testing of a nonzero value for γ_α.

c. The optimal testing procedure

The above procedure should lead to statistically significant results if the *a priori* guesses are appropriate. However, the least squares estimates (18) of the parameter γ_α provide the optimal representation of the signal in terms of the guesses with respect to the total variance of the expected response, but not with respect to the statistical significance, since the properties of the GCM noise field were not taken into account to achieve a larger signal-to-noise ratio. Hasselmann (1979) has therefore suggested using a pattern recognition method which eliminates as much of the noise as possible while still retaining a high proportion of the expected signal. To apply the method, it is convenient to work in the subspace where information on the noise is available, i.e. in the truncated space which has as basis vectors the principal components, or EOFs, $\mathbf{e}_k$ of the error covariance matrix estimate $\mathbf{S}$. Since the rank n_c of the matrix $\mathbf{S}$ is much smaller than its dimension, $\mathbf{S}$ has only n_c nonzero eigenvalues $\lambda_1^2, \lambda_2^2, \ldots, \lambda_{n_c}^2$, with $\lambda_1^2 \geqslant \lambda_2^2 \geqslant, \ldots, \lambda_{n_c}^2 > 0$. The n_c EOFs are orthonormal and satisfy

$$\mathbf{S}\mathbf{e}_k = \lambda_k^2 \mathbf{e}_k \tag{24}$$

for $k = 1, 2, \ldots, n_c$. As summarized, e.g., by Morrison (1976), the EOFs provide the most efficient method of compressing data and may be regarded as uncorrelated modes of variability of the field, in order of decreasing variance.

In the n_c-dimensional space of the new basic vectors $\mathbf{e}_k$, the guesses $\mathbf{g}_\alpha$ are described by the components $h_{\alpha,k}$, $k = 1, 2, \ldots, n_c$, which are chosen to minimize the error $|\mathbf{E}_\alpha|^2$ in the expansion

$$\mathbf{g}_\alpha = \sum_{k=1}^{n_c} h_{\alpha,k} \mathbf{e}_k + \mathbf{E}_\alpha. \tag{25}$$

Using the orthonormality of the EOFs, one finds

$$h_{\alpha,k} = \mathbf{e}_k' \mathbf{g}_\alpha. \tag{26}$$

Similarly, the mean GCM signal is projected onto the EOFs by

$$\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c = \sum_{k=1}^{n_c} s_k \mathbf{e}_k + \epsilon, \tag{27}$$

which yields the least squares solution

$$s_k = \mathbf{e}_k' (\bar{\mathbf{X}}^a - \bar{\mathbf{X}}^c). \tag{28}$$

Before discussing the method further, it should be noted that it is only meaningful to work in the EOF subspace if the available number of degrees of freedom (the number of EOFs) is large. Otherwise, the projections (25) and (27) will normally cause a substantial loss of variance for the signal and the guesses in directions where the GCM noise is small, which jeopardizes the power of the optimization procedure. The problem did not arise in the asymptotic case treated by Hasselmann (1979), since the EOFs then formed a complete orthonormal basis, but it could be very limiting in practice.

In the n_c-dimensional space, the linear model becomes

$$\mathbf{s} = \sum_{\alpha=1}^{p} \varphi_\alpha \mathbf{h}_\alpha + \mathbf{n}. \tag{29}$$

Minimizing the residual error $|\mathbf{n}|^2$ as before yields the standard estimates

$$\tilde{\varphi}_\alpha = \sum_{\beta=1}^{p} H^{-1}_{\alpha\beta} \mathbf{s}' \mathbf{h}_\beta, \tag{30}$$

where $\mathbf{H}^{-1}$ is the inverse of the $p \times p$ matrix with elements $H_{\alpha\beta} = \mathbf{h}_\alpha' \mathbf{h}_\beta$. The test statistic is now

$$\tau^2 = \left(\frac{1}{N} + \frac{1}{M}\right)^{-1} \tilde{\boldsymbol{\varphi}}' \boldsymbol{\Phi}^{-1} \tilde{\boldsymbol{\varphi}}, \tag{31}$$

where the $(p \times p)$ error covariance matrix Φ can be computed from the diagonal error covariance matrix Λ ($\Lambda_{ij} = \lambda_i^2 \delta_{ij}$) of the mean response as in (20).

Because of the projections onto the EOFs, the test statistic (31) does not follow a Hotelling distribution. Indeed, the projection directions are random variables which were estimated *a posteriori* as the directions that diagonalize the sample error covariance matrix **S**. The exact distribution of τ^2 has not been established. Asymptotically, τ^2 should tend to be distributed as a χ^2-variable with p degrees of freedom, if p is large enough. We have verified by comparing control runs in the experiments discussed in Section 4 below, that τ^2 was indeed behaving like χ^2 when there was no true change. Lacking a rigorous probability model, we shall simply assume here that the null hypothesis can be rejected at the α-level if

$$\tau^2 > \chi^2_{\alpha,p}. \tag{32}$$

The EOF subspace is the space of the sample noise field; directions of little noise are not included in the truncated space, whereas the directions of large noise define its basis. Thus, the signal-to-noise ratio will normally be degraded in this space, and it is only advantageous to work in the EOF subspace if some optimization of the statistical significance is undertaken.

Larger statistical significance may be achieved by considering modified or "rotated" guess vectors. Hasselmann (1979) has shown that the statistical significance will be maximum for any response **s** lying in the space spanned by the guess vectors $\mathbf{h}_\alpha$ when the signal is represented as a line superposition of the set of rotated guess vectors $\mathbf{h}^*_\alpha$ defined by the nonorthogonal linear transformation,

$$h^*_{\alpha,k} = h_{\alpha,k} \lambda_k^{-2}. \tag{33}$$

The individual components of $\mathbf{h}^*_\alpha$ are thus reduced relative to $\mathbf{h}_\alpha$ by a factor inversely proportional to the noise variance. Hence, $\mathbf{h}^*_\alpha$ is skewed away from the high noise components toward the low noise components. The signal is then projected into the $\mathbf{h}^*_\alpha$,

$$\mathbf{s} = \sum_{\alpha=1}^{p} \varphi^*_\alpha \mathbf{h}^*_\alpha + \mathbf{n}^*, \tag{34}$$

where φ^*_α is estimated as in (30). The null hypothesis is tested as before, with φ^* replacing φ in (31).

The optimization procedure can be interpreted in an alternative more traditional way which does not involve a modification of the guess vectors. Let us consider the projection (29) of the mean response **s** into the original guess vectors $\mathbf{h}_\alpha$. To evaluate the parameter φ_α, we use a weighted least squares method and introduce a structure matrix **N** where the quadratic form which is to be minimized is given by $\mathbf{n}'\mathbf{N}\mathbf{n}$. The structure matrix **N** is a positive definite symmetric matrix to be defined below. Then, the new estimate of φ_α is given by

$$\tilde{\varphi}_\alpha{}^N = \sum_{\beta=1}^{p} K^{-1}_{\alpha\beta} \mathbf{s}' \mathbf{N} \mathbf{h}_\beta, \tag{35}$$

where

$$K_{\alpha\beta} = \mathbf{h}_\alpha' \mathbf{N} \mathbf{h}_\beta.$$

Consistent with the model testing procedure outlined above, we choose the structure matrix **N** that will maximize the test statistic τ^2 given by (31) if the GCM response lies in the space spanned by the guess vectors $\mathbf{h}_\alpha$. In other words, **N** is defined as the structure matrix that will yield the maximum statistical significance for the model parameters if the model provides an exact representation of the GCM response. It is demonstrated in Appendix A that under this condition the test statistic is maximum when

$$N_{ij} = \lambda_i^{-2} \delta_{ij}. \tag{36}$$

Then, one has from (35),

$$\tilde{\varphi}_\alpha{}^N = \sum_{\beta=1}^{p} \sum_{k=1}^{n_c} K^{-1}_{\alpha\beta} s_k h_{\beta,k} \lambda_k^{-2}. \tag{37}$$

Thus the components are the same as before, when

"rotated" guesses were considered. The present interpretation shows clearly how the portions of the signal with low variance are given more emphasis than portions with high variance, since components in the least squares procedure are weighted by the inverse of the sample noise variance.

Now, the premises of the optimizing procedure must be examined. The statistical significance was shown to be maximum if the GCM signal is entirely in the guess space and if rotated guesses were considered according to (33). By continuity arguments one expects that the significance will still be increased by the rotation if the signal is largely, but not entirely, contained in the guess space. However, the significance should *decrease* if the guesses are not good. In addition, sampling errors on the EOFs may be important when the number of degrees of freedom is not large. This should also decrease the power of the optimizing procedure, since the rotation (33) then puts a high emphasis in directions where the *true* noise is not small. Thus, a partial rotation of the guess vectors may be advisable. Clearly, the optimization strategy has not been established for the case where the sample size is small, and further study is needed.

To summarize, the usefulness of the optimal testing procedure is strongly dependent on the available number of degrees of freedom:

1) If the sample size is small, the analysis should be done in the original GCM space, and no optimization of the significance should be attempted.

2) If the sample size is large, the analysis may be done in the truncated EOF subspace where the signal-to-noise ratio can be optimized.

4. Application to the GISS model I response to an SST anomaly

a. The GISS general circulation Model I

The influence of a midlatitude SST anomaly on the winter time circulation is investigated here with the first version of the GCM developed recently at GISS and described as Model I by Hansen *et al.* (1983). The model has a "coarse resolution" (8° × 10° grid) and is designed with computational efficiency which allows long-range climate experiments. Model I has seven layers in the vertical, including one layer in the stratosphere and one in the planetary boundary layer. Cloud cover, snow depth, ground temperature and moisture are computed. Sea surface temperature and sea ice coverage are prescribed.

As described by Hansen *et al.* (1983), Model I has a few deficiencies. In particular, the zonal wind field which is realistic in the troposphere increases with height in the stratosphere, reaching excessive velocities near the model top. Also, the eddy kinetic energy is a little deficient through most of the troposphere, so that the model natural variability is too small. However, the time scales of the fluctuations are in good agreement with the observations; typical decay time is a few days, reaching about a week at the largest spatial scales (see Hannoschöck, 1984 for details). Model I deficiencies, which may have some bearing on the model response to an SST anomaly, have been corrected in the more realistic Model II. However, for this first application of the multivariate statistical method, Model I is satisfactory, and no other GCM could have provided us at the time with

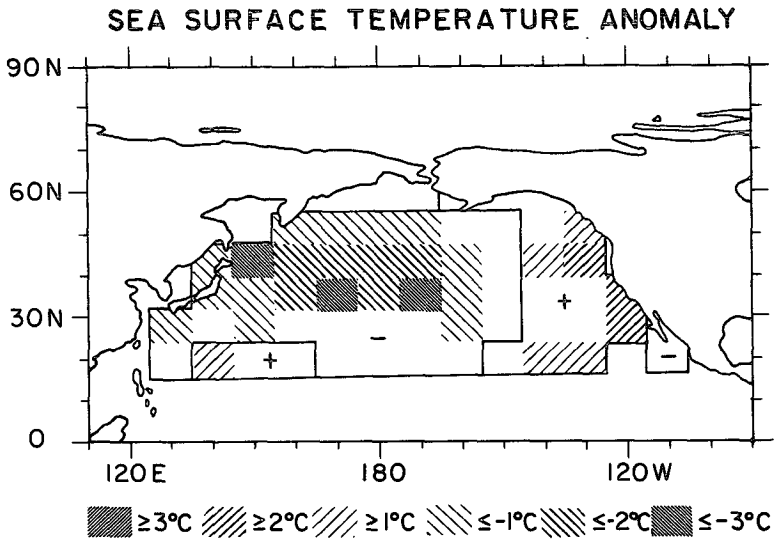
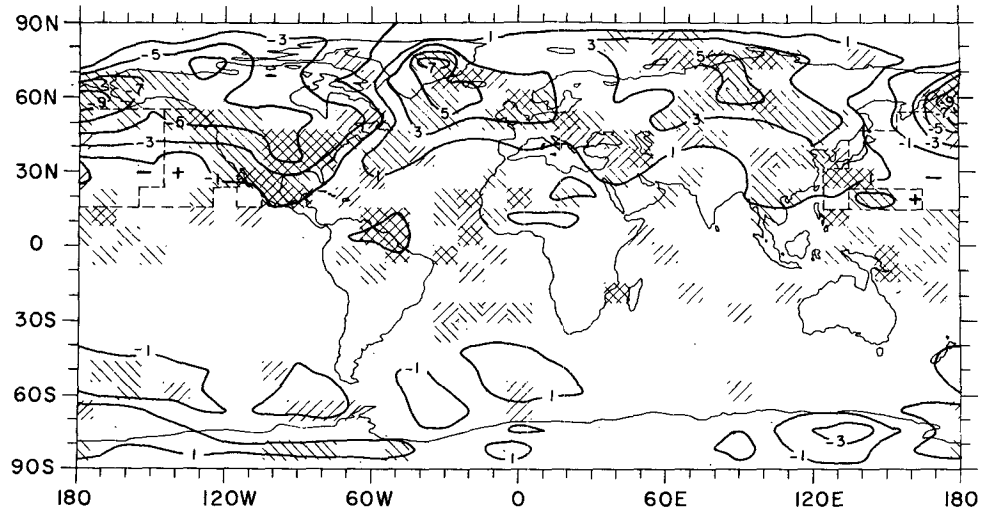
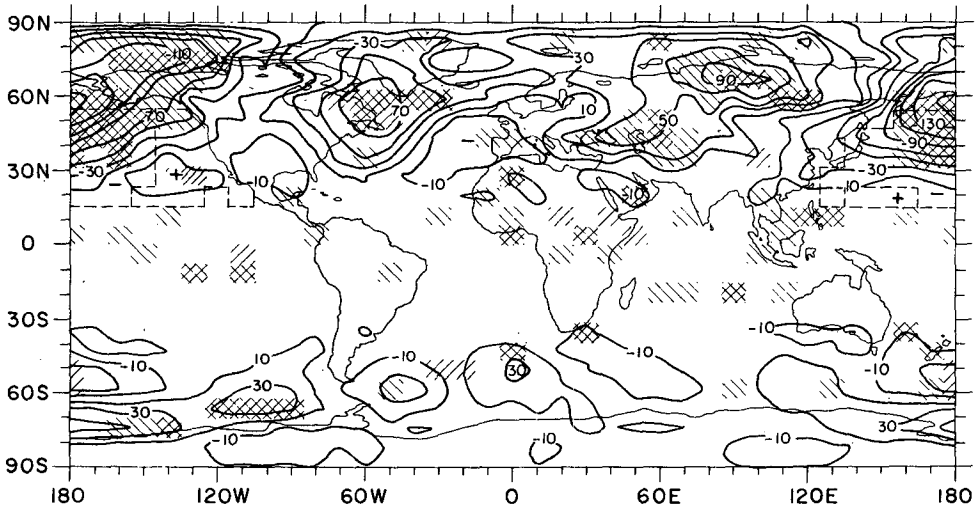


FIG. 1. The prescribed sea surface temperature anomaly (°C).

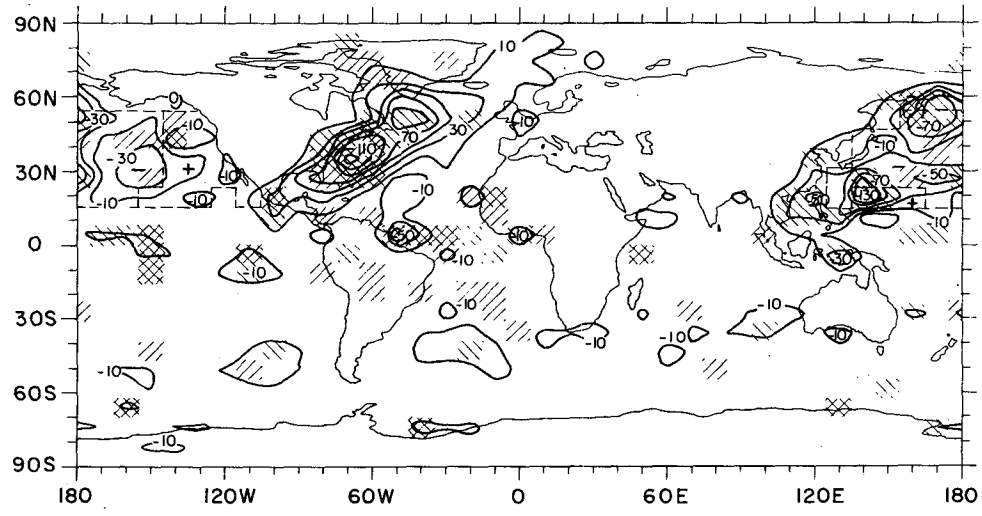
SEA LEVEL PRESSURE (mb)



500 mb HEIGHT (m)



SURFACE HEAT FLUX (Wm^{-2})



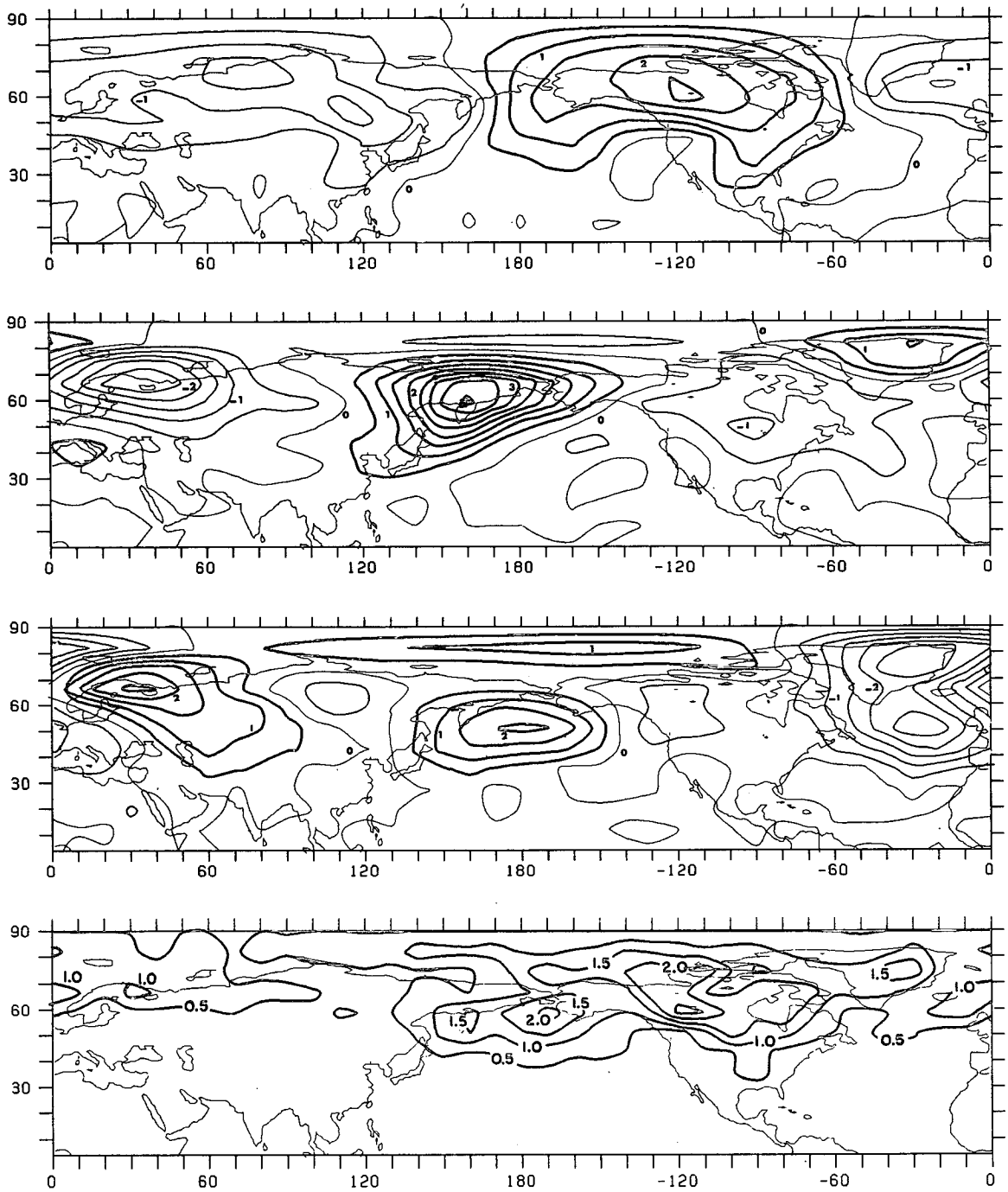


FIG. 3. First three panels from top, successively: First, second and third empirical orthogonal functions, respectively, of the Northern Hemisphere sea level pressure at low frequencies. They represent 37, 15 and 12% of the variance, respectively. Lower panel: Isolines of the power density estimate at zero frequency (mb^2/cpd).

FIG. 2. Upper panel: Mean sea level pressure difference (in mb) between anomaly and control runs. Middle panel: Same for 500 mb height (in m). Bottom panel: same for the net surface heat flux (W m^{-2}). The regions of positive and negative SST anomalies are delineated. Regions with forward slashes indicate grid points where the null hypothesis is rejected at the 5% level of significance using the standard two-sample *t*-test. Regions with backward slashes indicate the corresponding areas using the test for autocorrelated variables.

the lengthy simulations needed for development and testing of the statistical procedures.

b. The sea surface temperature anomaly experiment

In all experiments described here, Model I was run in the perpetual January mode with incoming solar radiation, sea ice and sea surface temperature held constant at the January climatological values. In the anomaly experiments, a fixed SST anomaly was added in the North Pacific (Fig. 1). The SST anomaly is primarly negative in the central and west Pacific, and positive in the eastern Pacific. It is similar to the large SST anomaly observed in the North Pacific during the 1976/77 winter (Namias, 1978), except that it is multiplied by a factor of 1.8 and does not quite vanish in the subtropics near 20°N.

Three long runs were made. After five months of spin up, the model appeared to be in a statistically steady state, i.e., no trend was seen in zonal averages of the model variables (there was, however, a very slow increase in snow depth, but not in snow cover). The experiments were then continued, yielding one 8-month control run, one 8-month anomaly run and one additional 15-month control run. For the analysis, we obtained daily averaged values of a number of variables at each grid point; among others, sea level pressure, 500 mb geopotential height and temperature, surface heat fluxes (positive downwards) and precipitation. In the 15-month control run, geopotential height and temperature were saved at 700 mb instead of 500 mb so that the data base for the 500 mb variables is smaller.

The mean changes between anomaly and control runs are illustrated for sea level pressure, 500 mb height and surface heat flux in Fig. 2. The changes in sea level pressure and 500 mb height are similar, suggesting a primarily barotropic behavior, with large negative values above the North Pacific and positive values above the North Atlantic and Siberia. However, these are regions with the largest "natural" variability in the model (see Fig. 3) and statistical tests must be performed before it can be concluded that the changes are caused by the SST anomaly. This also holds for the change in net surface heat flux which peaks in the Kuroshio and the Gulf Stream regions.

As mentioned previously, the natural variability in the troposphere is somewhat too small in Model I. This, combined with the near doubling of the SST anomaly, should yield signal-to-noise ratios that will be substantially larger than in the real world.

c. Univariate analysis

1) THE TWO-SAMPLE *t*-TEST

For test (4), we have considered mean values over 6-month durations, which lead to one anomaly sample and three control samples. For the 500 mb variables, 2-month means were considered instead, yielding a sample size of three in both anomaly and control conditions. Mean values calculated from the same run were separated by an interval of one month to avoid statistical dependence. However, one month between samples may not be sufficiently long in "perpetual conditions" to guarantee independence, and there is, indeed, some low frequency variability in the model (see Appendix B). For simplicity, statistical independence will nonetheless be assumed in this section.

The results of the test are illustrated in Fig. 2, where forward slashes (/ / /) indicate areas corresponding to grid points where the null hypothesis is rejected at the 5% level of significance; no smoothing was applied. For sea level pressure, these areas are rather scattered and the null hypothesis is rejected in only part of the regions, where there is a large change between anomaly and control runs. On the other hand, the largest mean changes in 500 mb height seem significant according to this test, including in particular the three centers of action above the North Pacific, the eastern North Atlantic and Siberia. The results for the surface heat flux also show considerable scatter; there seem to be significant changes above the SST anomaly and the Gulf Stream, and in the tropics. The apparent nonlocal character of the surface heat flux anomalies is not surprising, since these are a function of the computed lower air and ground conditions which are themselves affected by the atmospheric response to the SST anomaly.

The percentage of grid points for which the null hypothesis is rejected at the 5% level is given in Table 1 for different variables in different regions. Bold characters indicate global significance at the 5% level, assuming that all grid points are independent. As discussed above, this is clearly a liberal indicator of statistical significance. Table 1 suggests that there might be a significant response to the SST anomaly in the Northern Hemisphere and possibly also in the Southern Hemisphere tropics. Farther south, the atmospheric fields seem unperturbed.

As a measure of the local character of the mean changes we have considered the oceanic grid points separately. There seems to be a local intensification of the changes of the near-surface variables above the SST anomaly itself, with precipitation correlated mostly positively with the SST anomaly, and low-level cloudiness negatively correlated with the SST anomaly. The surface heat flux does not seem related to the SST anomaly in a simple fashion.

2) THE *t*-TEST FOR AUTOCORRELATED VARIABLES

Here we have applied the test (10). To estimate the spectral density $f_x(0)$, we have Fast-Fourier trans-

TABLE 1. Univariate test: Percentage of rejections of the null hypothesis at the 5% level according to the standard t-test (top number) and the red noise t-test (bottom number) at the model grid points (no area weighting). Numbers in parentheses give a measure of the bias of the latter method (global percentage of rejections for two control runs, minus 5%). Bold characters indicate overall rejection of the null hypothesis if the grid points are treated as independent.

Variable (bias)	Globe	30–90°N	0–30°N	0–30°S	30–90°S
Sea level pressure (4)	**10**	**13**	**16**	**10**	4
	17	**29**	**28**	**10**	4
Surface heat flux (2)	**8**	**9**	**12**	**10**	3
	8	**8**	**17**	7	4
Precipitation (3)	7	**8**	8	7	4
	12	**15**	**16**	**11**	7
500 mb height	**8**	**15**	7	3	5
	15	**29**	**13**	8	6
500 mb temperature	**9**	**11**	**15**	6	4
	15	**21**	**22**	8	**9**
200 mb zonal wind (3)	6	6	6	7	6
	11	**11**	**17**	8	**11**
200 mb meridional wind (4)	7	7	8	**10**	4
	11	**18**	**15**	**12**	2

formed the entire time series for the two control runs and the anomaly run. A smoothed spectral estimator was obtained by frequency averaging, as described in Section 2b. To select the frequency cutoff in (11), a compromise had to be made between significance and bias, since GCM spectra are white only at very low frequencies. It is shown in Appendix B that the bias remains small (a few percent) when periods larger than about two months are considered in (11), but that it becomes rapidly unacceptable if smaller periods are included. For the univariate tests, there was no need for a large number of degrees of freedom and the cut-off period was taken to be 200 days, yielding $\nu = 8$ ($\nu = 4$ at 500 mb).

The rejection of the null hypothesis at the 5% level by the test (10) is indicated in Fig. 2 by the backward slashes (\ \ \). It is seen that in the Southern Hemisphere the null hypothesis is frequently rejected over areas different than those with the test (4), while the regions of rejection coincide more often in the Northern Hemisphere. Since the variance estimates are nearly independent in the two tests, this confirms that the SST anomaly influence, if any, is largest in the Northern Hemisphere. Table 1 indicates that the averaged rejection rate of the null hypothesis is larger with (10) than with (4) by about 5%, which is consistent with the 3% bias estimated in Appendix B.

d. Multivariate statistical analysis

In the multivariate approach discussed in Section 3, the *a priori* guesses can be the predictions of some simplified dynamical model or a sequence of nondynamical guesses, and this choice is largely subjective. The main usefulness of the method lies in the testing of dynamical models. However, it was decided to first use nondynamical guesses, which are less restrictive (a large ensemble of spherical harmonics), allowing us to develop and test the new methodology more safely.

1) A HIERARCHY OF NONDYNAMICAL GUESSES

Theoretical studies suggest that, in the absence of resonance, the transfer function for the atmospheric response to diabatic heating is largest at the largest scales. Thus, we decided to use as guess vectors a sequence of spherical harmonics of decreasing spatial scales. Assuming that the response to the North Pacific SST anomaly primarily occurs in the Northern Hemisphere, we have in this hemisphere restricted *a priori* our analysis to one variable, the sea level pressure field. The guesses selected are even spherical harmonic functions $Y_n^m(\lambda, \varphi)$, where λ denotes longitude, φ latitude, m the zonal wavenumber and n the "total" wavenumber (one has $0 \leqslant m \leqslant n$, with $n - m$ even). The even spherical harmonics form a complete orthonormal set on a hemisphere. The hierarchy among the functions was chosen *a priori*, without special reference to the SST anomaly, as the following (n,m) sequence: (1,1), (3,1), (2,2), (4,2), (3,3), (5,1), (5,3), (4,4), (6,2), (6,4), (5,5), (6,6). Each (n,m) pair stands for a pair of real functions, the cosine and sine modes, which will be considered together in the analysis. In all, the sequence contains 24 guesses (the number of degrees of freedom of the error covariance matrix estimate). The guesses are orthonormal, although this is not a prerequisite of the method.

2) SEA LEVEL PRESSURE VARIABILITY

Since the experiment was run in the perpetual January mode, a power spectrum approach has been used to estimate the error covariance matrix. The procedure is a straightforward generalization of the one-dimensional case. To eliminate grid scale noise, the daily values of sea level pressure in the Northern Hemisphere were expanded into even spherical harmonics and the expansion limited by a triangular truncation at wavenumber $n = 18$. Excluding the zonally symmetric functions ($m = 0$), this led to a representation of the daily pressure field in terms of 179 even spherical harmonics. This does not alter the original data on the scales resolved realistically by the GCM (compare Fig. 2 with Fig. 6, upper panel). The entire time series for each spherical harmonic in the two control runs and in the anomaly run were then Fast-Fourier transformed and the cross-spectral properties of the spherical harmonics at zero frequency estimated by frequency averaging as described for the univariate test, using a cutoff period of 75 days. This should lead to a small but still acceptable bias (Appendix B), while providing a relatively high number of degrees of freedom $\nu = 24$[1] for the covariance matrix estimate. The latter property is particularly desirable for this first application of the method [ν should be used instead of $N + M - 2$ in (23)].

The 24 nonzero eigenvalues of the matrix **S** can be distinguished from white noise according to the criteria of Preisendorfer and Barnett (1977), and there are 24 corresponding nondegenerate EOFs or principal components. The first three EOFs, which represent 64% of the variance, are very large scale (Fig. 3). Thus, the zero frequency variability of the sea level pressure field (Fig. 3, lower panel) primarly reflects changes on the planetary scale. This stresses the inadequacy of the univariate significance tests.

It should be pointed out that the dominance of very large spatial scales in the low-frequency SLP field of Model I is not realistic and the EOFs in Fig. 3 have little resemblance to the main EOF patterns of the observed January sea level pressure anomalies (Kutzbach, 1970). This is partly due to the deficient eddy activity on the baroclinic scale, which was remedied in Model II. However, results from the forced variability of Model I may not be applicable to the real atmosphere since it does not satisfactorily reproduce its natural variability.

3) THE MULTIVARIATE ANALYSIS

The validity of the *a priori* sequence of guesses was first tested in the original (spherical harmonics) space, using the expansion (17) for the sea level pressure signal. The Hotelling test (23) is applied in order to estimate whether the amplitudes of the guesses are significantly different from zero, including successively more functions from the *a priori* sequence. Since we had no strong confidence in the *a priori* ordering of the guesses, we have used a rather mild selection criterion for the optimal model, which was taken as the maximum-order model satisfying the significance test (rule C: nonsequential fixed significance-level selection criterion in Barnett *et al.*, 1981). Thus, at the 95% level, the sequence of guesses must be interrupted at $p = 14$, (first seven spherical harmonics), since adding further guesses does not bring the value of T^2 above the significance level (Fig. 4).

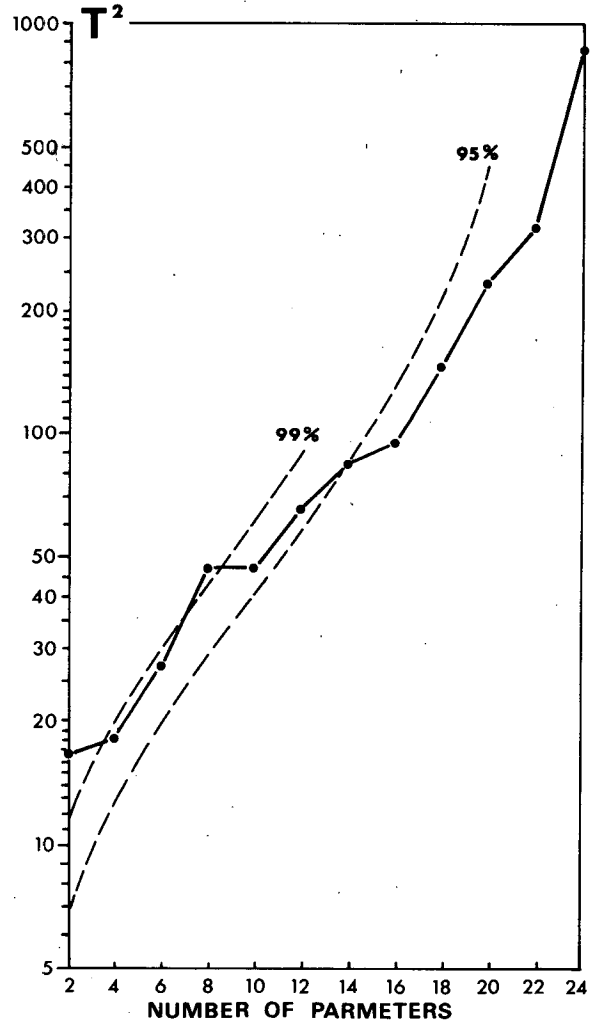


FIG. 4. Test statistic T^2 as a function of the number of guesses for sea level pressure changes between anomaly and control runs. The guesses are even spherical harmonics Y_n^m in the following (n,m) sequence: (1,1), (3,1), (2,2), (4,2), (3,3), (5,1), ... The 95% and the 99% significance bounds are given for the null hypothesis (Hotelling distribution).

[1] We use spectral estimates for periods of 240, 120 and 80 days for the 8-month control and anomaly runs, and additionally, 450, 225, 150, 112.5, 90 and 75 days for the 15-month control run, i.e., $=2 \times (2 \times 3 + 6)$.

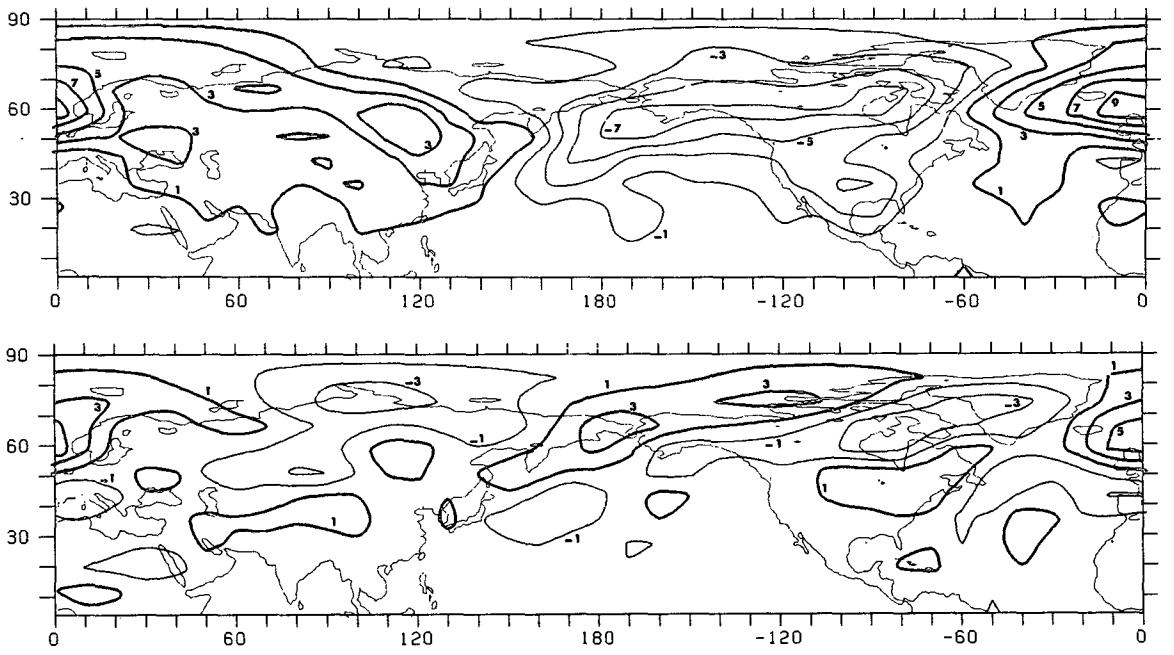


FIG. 5. Upper panel: Projection of the significant spherical harmonics of the mean sea level pressure signal on the original grid, at the 99% level of confidence. Lower panel: Difference between the mean sea level pressure signal and the significant part above.

At the 99% level, the first 4 spherical harmonics have nonzero amplitudes. This demonstrates unambiguously that the North Pacific SST anomaly had a significant large-scale effect on the wintertime circulation of the model. For visualization purposes, the "significant" spherical harmonics have been projected onto the original grid, and the results contoured (Fig. 5, upper panel). One must bear in mind that this presentation refers to, and to some extent depends on, the particular choice of *a priori* guesses. A com-

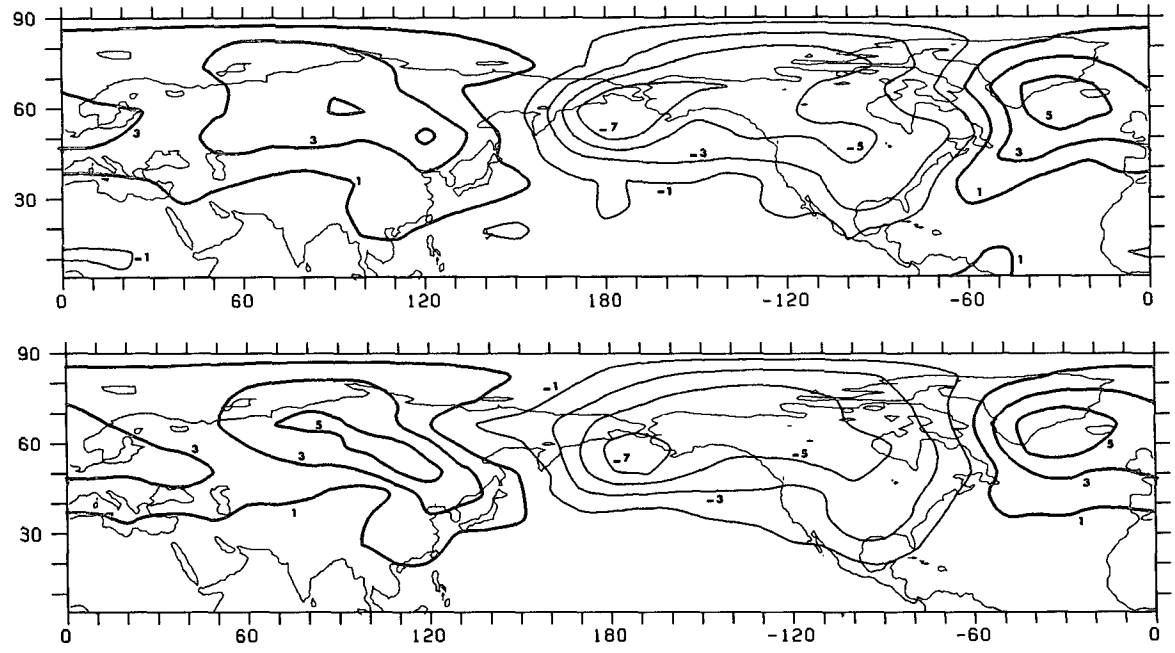


FIG. 6. Upper panel: Mean change in sea level pressure after reduction to 179 even spherical harmonics (in mb). Bottom panel: Signal after projection onto the 24 EOF. (The results are represented on the original grid.)

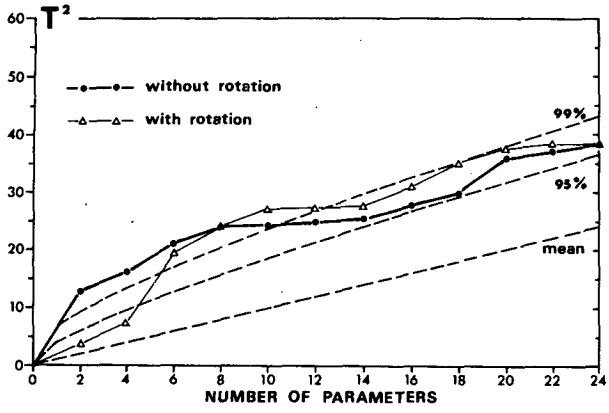


FIG. 7. As in Fig. 4 for the mean sea level pressure changes in the EOF subspace. Full circles and thick line represent the results of the test without optimization, open triangles and thin line the results with optimization. The 95% and 99% significance bounds are given for the null hypothesis using the χ^2 distribution.

parison with the sample mean difference (Fig. 6, upper panel) shows that many features of the GCM signal are significant at this level. The "nonsignificant part" of the signal has smaller amplitude and, as expected from the hierarchy, smaller scales (Fig. 5, lower panel).

4) THE MULTIVARIATE ANALYSIS IN THE EOF SUBSPACE

Because of the relatively large number of degrees of freedom achieved, the multivariate analysis could also be attempted in the EOF subspace, and the optimization procedure could be applied. The EOF subspace is 24-dimensional, and it can therefore be expected that signal and guesses are well-represented. Fig. 6 (lower panel) shows for comparison the mean sea level pressure change on the original grid before and after EOF projection. It can be verified that the main features of the signal are retained with little loss of variance.

In the truncated space, the sea level pressure signal is expanded in terms of the guess vectors, using (27). The τ^2 values for increasing number of guesses are shown in Fig. 7 (full circles and thick line), with the 95% and 99% significance bounds given by the χ^2 distribution. Although the statistical significance decreases for $p > 8$, τ^2 remains above the 95% level for all values of p. Since the signal significance should be lower in the EOF subspace, where the sample noise is maximal, it is likely that the available number of degree of freedom is not large enough for using this asymptotic form of the test of significance (Section 3), especially for large p. Lacking a more precise probability model for τ^2, we decided, nonetheless, to choose the optimal model in this preliminary analysis by comparing the incremental changes in the test statistic to the χ^2 significance bounds. This gives $p = 8$. The "significant" part of the response, according to our *a priori* sequence of guess is shown in Fig. 8 on the original grid.

Although the use of "rotated guesses," or of weighted least-squares estimates, is designed to increase the statistical significance (Section 3), the results

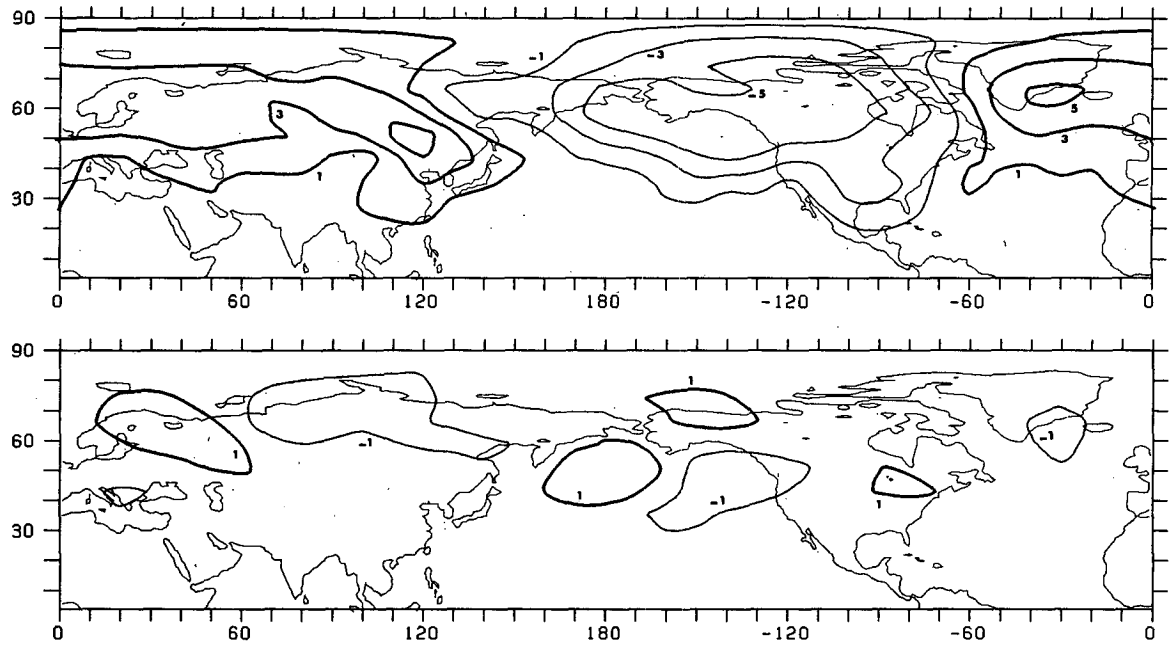


FIG. 8. As in Fig. 5 after projection onto the EOF subspace (no optimization).

of the optimized test proved disappointing (Fig. 7, open triangles). Indeed, rather than increasing τ^2, the optimization decreases τ^2 for the first few guesses down to the point where significance is lost. For larger p, τ^2 increases and even exceeds the unrotated τ^2 values. Since the significance keeps decreasing for $p \geqslant 10$, the "optimal" model has 5 spherical harmonics. As expected, a smaller part of the signal variance is found to be significant (Fig. 9), but the statistical significance of the pattern has not increased. This failure is not unexpected, since the guesses are not very good (a rather arbitrary sequence of spherical harmonics) and sampling errors on the EOFs are large. In our application, the range of eigenvalues covers 2 orders of magnitude for the sea level pressure field. Thus (33) can lead to a strong loss of significance.

5. Testing a dynamical model

Linear wave models have been rather successful at explaining qualitatively the mean asymmetries of the atmospheric circulation as perturbations superimposed on a zonal basic state and caused by topography and the observed diabatic heating contrasts between land and sea (e.g., Egger, 1976; Hoskins and Karoly, 1981). It may therefore be expected that linear wave models are appropriate to represent the atmospheric response to a SST anomaly. In linear wave models, one usually investigates the influence of an anomaly in the diabatic heating (e.g. Hoskins and Karoly, 1981), unless some simple relation is assumed between a prescribed SST anomaly and the resulting anomaly in diabatic heating (e.g., Egger, 1977; Webster, 1981). Although such procedure is justified in basic studies, it is not directly applicable to the GCM case where a SST anomaly is prescribed, but only a noisy estimate can be obtained for the three-dimensional structure of the corresponding heating anomaly. Thus, to interpret the GCM response to a SST anomaly, the relation between SST and heating must be clarified. This is not an easy task, particularly in the midlatitudes.

Unfortunately, insufficient information had been saved in the experiments with the GISS model to describe precisely the diabatic heating field or to establish parameterization schemes. So we shall make only a preliminary attempt at interpreting dynamically the response of the GISS model. More efforts will be devoted to further experiments with Model II, to be discussed elsewhere.

To construct simple dynamical guesses for the Model I experiments, we have considered a linear quasi-nondivergent model in steady zonal motion (e.g., Wiin-Nielsen, 1971), where the streamfunction ψ for the nondivergent part of the velocity field and the geopotential ϕ obey the following equations, in pressure coordinates

$$\bar{U}\frac{\partial}{\partial x}\left(\nabla^2\psi + \frac{f}{\sigma}\frac{\partial^2\phi}{\partial p}\right) + \beta\frac{\partial\psi}{\partial x} = \frac{fR}{C_p\sigma}\frac{\partial}{\partial p}\left(\frac{\dot{q}}{p}\right),$$

$$\nabla^2\phi - \nabla\cdot(f\nabla\psi) = 0. \qquad (45)$$

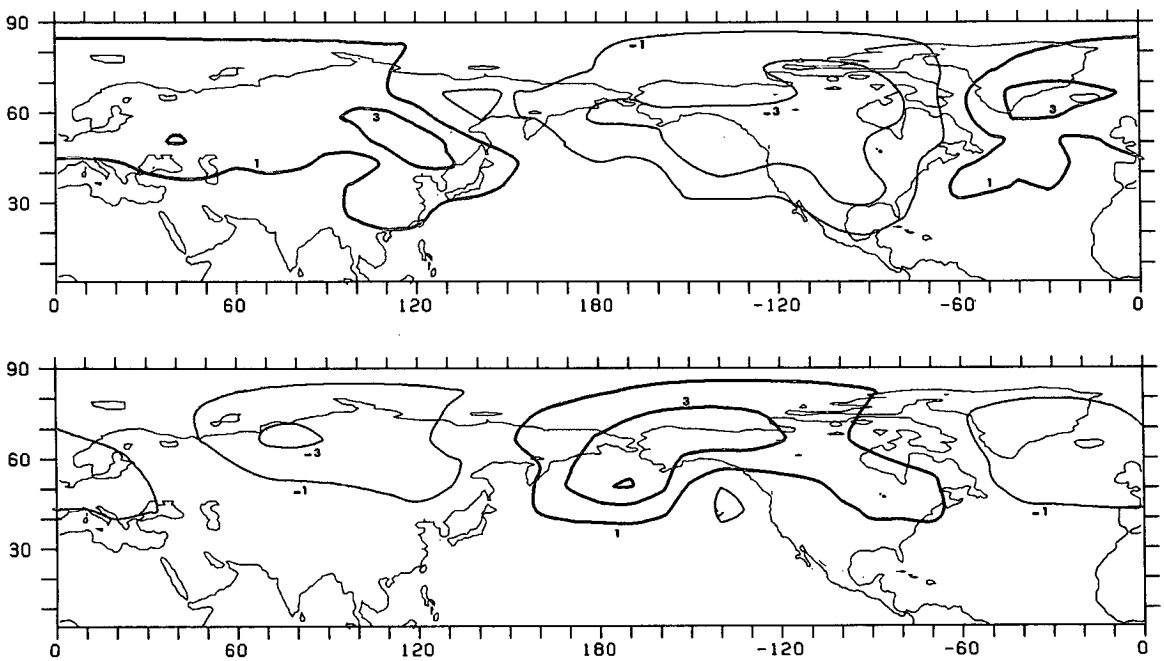


FIG. 9. As in Fig. 8 with optimization.

Here $\bar{U}$ is the mean zonal velocity, p the pressure, f the Coriolis parameter, R the constant for perfect gas, C_p the heat capacity, σ the static stability (assumed constant) and $\dot{q}$ the diabatic heating rate. When the basic zonal flow is independent of pressure, the solutions of the homogeneous problem are separable in the vertical, and the vertical normal modes $F_l(p)$ can be treated separately. The eigenfunctions obey

$$\sigma^{-1}\frac{\partial^2}{\partial p}F_l = -(gh_l)^{-1}F_l, \tag{46}$$

where h_l is the equivalent height. Rigid lid and flat bottom are assumed. To solve the forced problem by a normal mode expansion, it was assumed that the (unknown) vertical distribution of the diabatic heating anomaly is independent of longitude and latitude. Therefore, the model response for each vertical mode depends linearly on the (unknown) projection of the diabatic heating profile on the mode. Rather than specify a profile, this projection was left as a free parameter and estimated by a least-squares fit. The model significance was then tested by establishing whether the estimated parameter was significantly different from zero, at the 95% level.

To solve Eq. (45) in the Northern Hemisphere, the geopotential height and the heating were expanded into even spherical harmonics while the streamfunction was expanded into odd harmonics. The basic state was taken as the mean velocity at 300 mb and estimated geostrophically from the 500 mb data, using 1.8 as a scaling factor; the scale height is 11 km. A linear friction is used, with a damping time of 14.7 days (see Hannoschöck, 1984, for more details).

The horizontal distribution of the diabatic heating

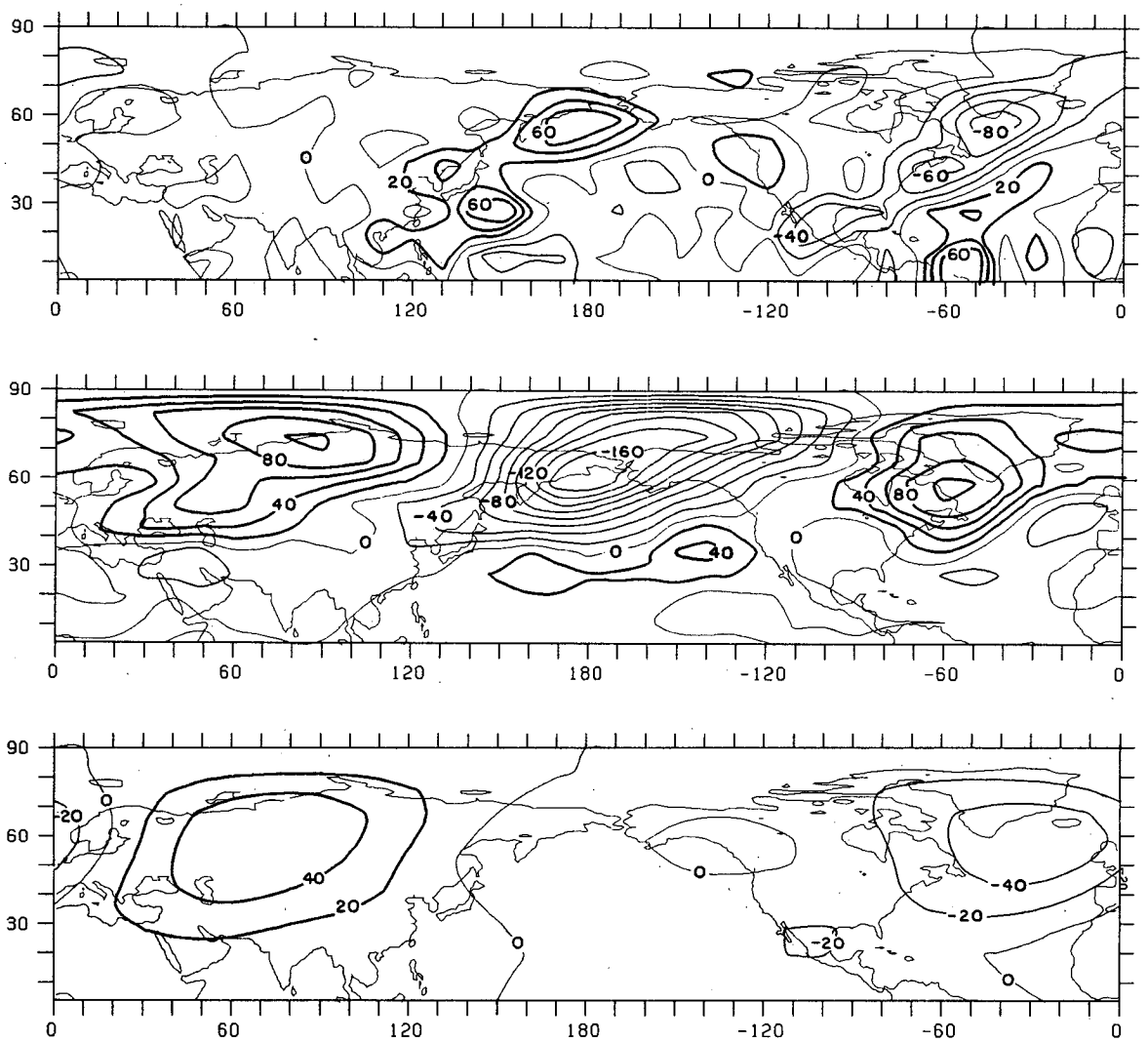


FIG. 10. Upper panel: Estimated heating anomaly (in W m^{-2}) for the linear wave model (see text). Middle panel: Mean change in the 500 mb height (in m) between anomaly and control runs. Bottom panel: Estimated barotropic model response for the 500 mb geopotential height (the fit is statistically significant at the 95% level).

anomaly was estimated from the surface heat flux (minus latent heat flux) and the precipitation data. Figure 10 (upper panel) indicates very noisy changes, with small scale features associated with zones of precipitation. In view of the space–time intermittency of the latter process, the sampling errors on the model forcing are particularly large. This noisy estimate was used "deterministically" as the *true* model forcing (we also used the SST anomaly to define the forcing, but the model response turned out to be inconsistent with the GCM data).

The analysis was conducted for the 500 mb geopotential height for which there is only one 8-month control run and one 8-month anomaly run, so that $\nu = 12$. As noted before, the mean change (Fig. 10, middle panel) resembles that of sea level pressure. The guesses were chosen to be the linear wave model response for successive vertical normal modes, in order of increasing mode number. Only the barotropic and the first baroclinic modes were considered. The procedure is the same as for the nondynamical guesses in the original (spherical harmonics) space, where the model significance is evaluated with the T^2 test.

The amplitude of the barotropic model was found to be different from zero at the 95% level of confidence. The barotropic model prediction is represented in Fig. 10 (lower panel), and it shows only little resemblance with the mean GCM signal. This illustrates that a single pattern does not necessarily represent a large fraction of the signal variance, even if it is statistically significant. Using the 500 mb level to define the basic flow did not bring significant improvements.

Adding as a second guess the first baroclinic mode (with an equivalent height of 100 m) lead to a complete loss of significance. This failure is not unexpected in view of the model simplicity (no vertical shear), indeterminacies in the estimated forcing field, and the arbitrariness of our assumptions for the forcing (e.g., no surface heating), which are critical for the baroclinic modes.

It is also possible that the GISS Model I is very sensitive to changes in the equatorial regions. Indeed, we looked at the reverse problem of deducing the heat sources that would be consistent with the observed GCM changes, if the response were purely linear and barotropic. To simplify the calculations, the 500 mb height signal was first smoothed spatially, keeping only 6 spherical harmonics. Fig. 11 shows that the hypothetical heat source has strong maxima in the equatorial regions. Thus, a quasi-geostrophic model which is not valid at very low latitudes may not be very appropriate.

6. Discussion

The experiments discussed in this paper demonstrate unambiguously that a midlatitude SST anomaly can have a significant large-scale influence on the wintertime circulation in a general circulation model of the atmosphere. This result could be established on a firm statistical basis by developing multivariate statistical tests based on the hypothesis testing method of Hasselmann (1979). Using a sequence of large-scale spherical harmonics as *a priori* guesses of the anticipated response of the Northern Hemisphere sea level pressure to the SST anomaly, we found a significant signal at the planetary scales (zonal wavenumbers 1–3). The 500 mb geopotential height appears to behave similarly, and thus the GISS Model I response to the North Pacific SST anomaly seems primarily large-scale barotropic. Such behavior differs from the sensitivity exhibited by other GCMs. For instance, in the "superanomaly" experiments of Kutzbach *et al.* (1977) and Chervin *et al.* (1980), the 6-layer NCAR model (5° × 5° resolution) shows some sensitivity to very large SST anomalies in the North Pacific (12°C maximum amplitude), but the response is dominated by a direct thermal circulation and a primarily zonal wavenumber 3 response for the pressure field (plus changes in the cyclonic activity). Shukla and Bangaru (1979) have briefly discussed the response of the 9-layer GLAS model (4° × 5° resolution) to a SST anomaly very similar to that in Fig. 1, but the response, if any, seems also largest near zonal wavenumber 3 or 4. Finally, preliminary results on the GISS Model II response to the SST anomaly considered in this paper suggest a predominantly wavenumber 3–5 response (Frankignoul,

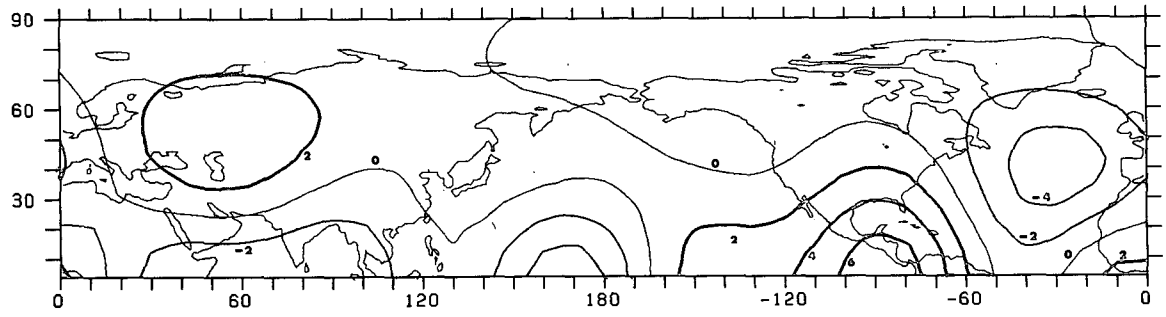


FIG. 11. Heat sources consistent with the 500 mb geopotential signal as obtained by inverting the barotropic model. The geopotential data were smoothed by keeping only six spherical harmonics. Contour is arbitrary.

1984). Thus, the response of GCMs to midlatitude SST anomalies seems to be very model-dependent.

In the present paper, we have also discussed the multivariate hypothesis testing method. Hasselmann's (1979) method has been extended to the more realistic case where the sample size (number of degrees of freedom) is limited, using the Hotelling T^2 statistic to evaluate the significance of the anticipated response patterns. It was shown that, unless the available number of degrees of freedom is very large, it is preferable to apply the hypothesis testing method directly in the original GCM space, without attempting to optimize the signal-to-noise ratio. It may be useful to work in the EOF subspace when the sample size is large, but more work needs to be done on the best method for optimizing the signal-to-noise ratio for the guessed response. Furthermore, the statistical distribution of the test statistic, which behaves asymptotically as a χ^2 variable, should be established more precisely.

Besides the need for a sufficiently large integration time, the only real difficulty of the multivariate method lies in the choice of the *a priori* guesses of the anticipated model response. Here, we have first considered a sequence of spherical harmonics of decreasing spatial scales, which is a rather trivial, if noncommital choice. Another reasonable sequence could have been based on the EOFs, thereby assuming that free and forced responses are dominated by the same spatial patterns. Such *a priori* guesses are helpful in establishing the statistical significance of a model response, but clearly the great potential of the method lies in the use of dynamical guesses. Linear wave models or other simple dynamical models should provide the most pertinent guesses of the GCM behaviors and if they do not, one should seriously question their applicability to the real atmosphere. From this point of view, GCMs can be used to test theories and tune simpler models.

Some attempts were made here at using a quasi-geostrophic linear wave model to analyse the GCM data, but only some barotropic behavior could be established. This lack of success may have arisen from using a too simplified model, and from missing knowledge regarding the diabatic heating which is associated with the SST anomaly. Although more careful planning of the experiments could have decreased these uncertainties, the difficulty can only be remedied by modeling explicitly the mechanisms that relate the diabatic heating to the SST changes. This aspect has often been overlooked in linear wave models and requires further study.

Acknowledgments. We would like to thank Prof. K. Hasselmann for help and stimulating discussions, and Drs. J. Hansen, G. Russel and D. Rind for generously providing the GCM data. Thanks are also due Drs. D. Olbers and E. Maier-Reimer. Mr. M. Terman assisted in some of the calculations and is gratefully acknowledged. For C.F., this research was supported by the National Science Foundation, Climate Dynamics Program, under Grant ATM-8116047 with the Massachusetts Institute of Technology, and by a research Grant from the CNEXO.

APPENDIX A

Optimal Weight for the Least-Squares Fit

Here we demonstrate that in the EOF subspace, the structure matrix **N** that maximizes the test statistics

$$\tau^2 = \left(\frac{1}{N} + \frac{1}{M}\right)^{-1} \tilde{\boldsymbol{\varphi}}^{N\prime} \boldsymbol{\Phi}^{N-1} \boldsymbol{\varphi}^N \tag{A1}$$

is given by $N_{ij} = \lambda_i^{-2}\delta_{ij}$ when the GCM signal **s** is an exact combination of the guess vectors $\mathbf{h}_\alpha$.

To simplify the writing, we define the p-dimensional random vector ψ by

$$\tilde{\boldsymbol{\varphi}}^N = \mathbf{K}^{-1}\psi, \tag{A2}$$

and consider as measure of significance

$$P^2 = \psi' \boldsymbol{\Psi}^{-1} \psi, \tag{A3}$$

where the error covariance matrix $\Psi_{\alpha\beta} = \langle \psi_\alpha \psi_\beta' \rangle$ is computed from $\boldsymbol{\Lambda}$ as in (31), using the projectors $\mathbf{b}_\alpha$ defined by

$$\psi_\alpha = \mathbf{b}_\alpha' \mathbf{s} = \sum_{k=1}^{n_c} b_{\alpha,k} s_k. \tag{A4}$$

Since τ^2 does not change if $\tilde{\boldsymbol{\varphi}}^N$ is multiplied by the nonsingular matrix $\mathbf{K}^{-1}$, we have to find the projectors $\mathbf{b}_\alpha$ which maximize P^2 when the signal is contained in the guess space; for example if

$$s_k = \sum_{\alpha=1}^{p} \sigma_\alpha h_{\alpha,k}. \tag{A5}$$

(Note incidently that in this hypothetical case the residual error vanishes and the coefficients φ_α are independent of the norm.)

If the guess vectors $\mathbf{h}_\alpha$ are p elements of the n_c-dimensional euclidean space R^{n_c}, the value of P^2 depends on the signal **s** and on the regression norm or, equivalently, the p projectors $\mathbf{b}_\alpha$. These projectors $\mathbf{b}_\alpha$ form a basis of a p-dimensional linear subspace B of the dual space to R^{n_c}. Since the quadratic form P^2 is invariant with respect to linear transformations of coordinates, one has

$$P^2 = P^2(B, \mathbf{s}).$$

First we establish that for arbitrary B one has

$$P^2(B, \mathbf{s}) \leqslant \sum_{k=1}^{n_c} s_k{}^2 \lambda_k{}^{-2}. \tag{A6}$$

To do this, we consider the $(p-1)$-dimensional subspace B_{p-1} that contains all linear forms projecting **s** onto zero. Within B_{p-1}, we choose statistically

orthonormal basis vectors $\mathbf{b}_2, \mathbf{b}_3, \ldots, \mathbf{b}_p$ which verify

$$\langle(\mathbf{b}_\alpha'\mathbf{s})(\mathbf{b}_\beta'\mathbf{s})\rangle = \delta_{\alpha\beta} = \sum_{k=1}^{n_c} b_{\alpha,k}b_{\beta,k}\lambda_k^2. \quad \text{(A7)}$$

Then we complete the basis of B by a vector $\mathbf{b}_1$ which also satisfies (A7). From (A3) and (A4), one has

$$P^2 = \psi_1^2 = (\sum_{k=1}^{n_c} b_{1,k}s_k)^2 = (\sum_{k=1}^{n_c} b_{1,k}\lambda_k s_k \lambda_k^{-1})^2 \quad \text{(A8)}$$

and, using the Cauchy-Schwartz inequality,

$$P^2 \leq (\sum_{k=1}^{n_c} b_{1,k}^2\lambda_k^2)(\sum_{k=1}^{n_c} s_k^2\lambda_k^{-2}), \quad \text{(A9)}$$

which reduces to (A6) in view of (A7).

Now we show that if (A5) holds, the upper bound (A6) is reached for the projectors obeying

$$b_{\alpha,k} = h_{\alpha,k}\lambda_k^{-2} \quad \text{(A10)}$$

for all α and k. Indeed, consider a projector subspace B^* with basis vectors $\mathbf{b}_2^*, \mathbf{b}_3^*, \ldots, \mathbf{b}_p^*$ obeying (A10) and projecting the signal $\mathbf{s}$ onto zero, and $\mathbf{b}_1^*$ which obeys

$$b_{1,k}^* = Cs_k\lambda_k^{-2}, \quad \text{(A11)}$$

where C is a normalizing constant. The projector $\mathbf{b}_1^*$ is contained in the previous subspace B, for if $\mathbf{s}$ is in the guess space, (A5) and (A10) yield

$$s_k\lambda_k^{-2} = \sum_{\alpha=1}^{p} \sigma_k b_{\alpha,k}; \quad \text{(A12)}$$

thus $\mathbf{b}_1^*$ is a linear combination of the basis vectors $\mathbf{b}_\alpha$. Moreover, the orthonormality condition (A7) holds since one has

$$\langle(\mathbf{b}_\alpha^{*\prime}\mathbf{s})(\mathbf{b}_1^{*\prime}\mathbf{s})\rangle = C\sum_{k=1}^{n_c} b_{\alpha,k}^*\lambda_k^2 s_k\lambda_k^{-2} = C\mathbf{b}_\alpha^{*\prime}\mathbf{s}, \quad \text{(A13)}$$

which is equal to zero by construction for $\alpha > 1$. The measure of significance becomes

$$\tau^2 = (\sum_{k=1}^{n_c} -b_{1,k}^*s_k)^2 = C^2(\sum_{k=1}^{n_c} s_k^2\lambda_k^{-2})^2 = \sum_{k=1}^{n_c} s_k^2\lambda_k^{-2}, \quad \text{(A14)}$$

where we have used $|\mathbf{b}_1^*| = 1$. The Cauchy-Schwartz inequality has become an equation for the choice (A10).

APPENDIX B

Bias of the Spectral Density Estimates at Zero Frequency

GCM variables, like their atmospheric counterparts, have a correlation time of a few days. It is therefore difficult to estimate the spectral density near zero frequency in typical GCM experiments. Since GCM time series can often be reasonably well modeled as a discrete first-order Markov process, the bias associated with the use of a frequency ω to estimate the zero frequency energy $f(0)$ can be computed from (Madden, 1976)

$$\frac{F(0)}{F(\omega)} = \frac{1 + \alpha^2 - 2\alpha\cos\omega}{1 + \alpha^2 - 2\alpha}, \quad \text{(B1)}$$

where α is the lag one autocorrelation and the frequency ω is in radians per unit time. The ratio (B1) is greater than one, hence $F(0)$ is always underestimated. Using $\alpha = 0.7$ as a typical value (three day decay time), one finds that the bias in $F(0)$ is only 4% for a 100-day period, but as large as 34% for a 30-day period. Thus, spectral densities near zero frequency cannot be estimated satisfactorily if the GCM run length is only one or two months.

In the t-test (10) or in the corresponding multivariate analysis, the underestimation of the zero frequency spectral density leads to an overestimation of the rejection rate of the null hypothesis. This is illustrated for the GISS data by applying (10) to the sea level

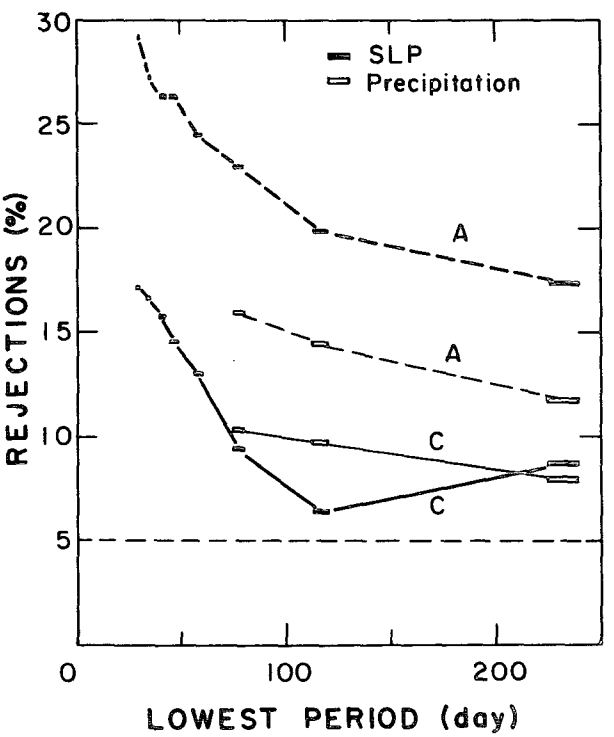


FIG. B1. Percentage of rejection of the null hypothesis at the 5% level of significance in the test, Eq. (13), as a function of the cutoff period (see text) for the estimation of $f(0)$. Results are presented for sea level pressure (thick line) and precipitation (thin line), and for changes between anomaly and control runs (dashed line), and between the two control runs (continuous line).

pressure in two cases. First, we consider the changes between the anomaly and the two control runs, then we consider the changes between the two control runs. In the latter case, the expected rejection rate is 5% at the 5% level. In both cases, $F(0)$ was calculated as described in Section 4c from the three runs. The global percentage of rejection of the null hypothesis is represented in Fig. B1 as a function of the cutoff period (since the runs have unequal length, 240 and 450 days, we have represented the range of minimum period). The bias appears clearly in the two cases: it is large when short periods are included in the estimate of $f(0)$, but decreases rapidly with increasing cut-off period.

Similar tests made with precipitation (Fig. B1) indicate that these results are applicable to other variables. Note incidently that if the bias is comparable (see the control cases), the mean response seems smaller for precipitation than for sea level pressure. The bias with a 200 day cutoff has been estimated for several variables by considering the differences between the two control runs. The results are reported in Table 1.

REFERENCES

Anderson, T. W., 1971: *The Statistical Analysis of Time Series.* Wiley & Sons, 704 pp.

Barnett, T. P., R. W. Preisendorfer, L. M. Goldstein and K. Hasselmann, 1981: Significance tests for regression model hierarchies. *J. Phys. Oceanogr.*, **11,** 1150–1154.

Bell, T. L., 1982: Optimal weighting of data to detect climatic change: Application to the carbon dioxide problem. *J. Geophys. Res.*, **87,** 11 161–11 170.

Chervin, R. M., and S. H. Schneider, 1976: On determining the statistical significance of climate experiments with general circulation models. *J. Atmos. Sci.*, **33,** 405–412.

——, J. E. Kutzbach, D. D. Houghton and R. G. Gallimore, 1980: Response of the NCAR general circulation model to prescribed changes in ocean surface temperature. Part II: Midlatitude and subtropical changes. *J. Atmos. Sci.*, **37,** 308–332.

Egger, J., 1976: On the theory of steady perturbations in the troposphere. *Tellus,* **28,** 381–389.

——, 1977: On the linear theory of the atmospheric response to sea surface temperature anomalies. *J. Atmos. Sci.*, **34,** 603–614.

Frankignoul, C., 1985: Multivariate analysis of sensitivity studies with atmospheric GCM's. *Coupled Atmosphere–Ocean Models,* Chap. 15, J. C. J. Nihoul, Ed., *Elsevier Oceanogr. Ser.*, **40** (in press).

Hannoschöck, G., 1984: A multivariate signal-to-noise analysis of the response of an atmospheric circulation model to sea surface temperature anomalies. *Hamburger Geophysikalische Einzelschriften,* **67,** 100 pp.

Hansen, J., G. Russel, D. Rind, P. Stone, A. Lacis, S. Lebedeff, R. Ruedy and L. Traves, 1983: Efficient three-dimensional global models for climate studies: Models I and II. *Mon. Wea. Rev.*, **111,** 609–662.

Hasselmann, K., 1979: On the signal-to-noise problem in atmospheric response studies. *Meteorology of the Tropical Oceans,* D. B. Shaw, Ed., Roy. Meteor. Soc., 251–259.

Hayashi, Y., 1982: Confidence intervals of a climatic signal. *J. Atmos. Sci.*, **39,** 1895–1905.

Hoskins, B. J., and D. J. Karoly, 1981: The steady linear response of a spherical atmosphere to thermal and orographic forcing. *J. Atmos. Sci.*, **38,** 1179–1196.

Katz, R. W., 1982: Statistical evaluation of climate experiments with general circulation models: A parameter time series modelling approach. *J. Atmos. Sci.*, **39,** 1446–1455.

——, 1983: Statistical procedures for making inferences about precipitation changes simulated by an atmospheric general circulation model. *J. Atmos. Sci.*, **40,** 2193–2201.

Kutzbach, J. E., 1970: Large-scale features of monthly mean Northern Hemisphere anomaly maps of sea level pressure. *Mon. Wea. Rev.*, **98,** 708–716.

——, R. M. Chervin and D. D. Houghton, 1977: Response of the NCAR general circulation model to prescribed changes in ocean surface temperature. Part I: Midlatitude changes. *J. Atmos. Sci.*, **34,** 1200–1213.

Jones, R. H., 1976: On estimating the variance of time averages. *J. Appl. Meteor.*, **15,** 514–515.

Laurmann, J. A., and W. L. Gates, 1977: Statistical considerations in the evaluation of climatic experiments with atmospheric general circulation models. *J. Atmos. Sci.*, **34,** 1187–1199.

Leith, C. E., 1973: The standard error of time-average estimates of climatic means. *J. Appl. Meteor.*, **12,** 1066–1069.

Livezey, R. E., 1983: Statistical analysis of general circulation model climate simulation, sensitivity and prediction experiments. *Preprints Second Int. Meeting on Statistical Climatology,* Lisbon, 8 pp.

——, and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111,** 46–59.

Madden, R. A., 1976: Estimates of the natural variability of time-averaged sea level pressure. *Mon. Wea. Rev.*, **104,** 942–952.

Morrison, D. F., 1976: *Multivariate statistical methods.* McGraw-Hill, 415 pp.

Moura, A. D., and J. Shukla, 1981: On the dynamics of droughts in Northeast Brazil: Observations, theory and numerical experiments with a general circulation model. *J. Atmos. Sci.*, **38,** 2653–2675.

Namias, J., 1978: Multiple causes of the North American abnormal winter 1976–77. *Mon. Wea. Rev.*, **106,** 279–295.

North, G. R., T. L. Bell and R. F. Cahalan, 1982: Sampling errors in the estimation of empirical orthogonal functions. *J. Atmos. Sci.*, **110,** 699–706.

Preisendorfer, R. W., and T. P. Barnett, 1977: Significance tests for empirical orthogonal functions. *Preprints 5th Conf. on Probability and Statistics in Atmospheric Sciences,* Las Vegas, Amer. Meteor. Soc., 169–172.

——, and ——, 1983: Numerical model-reality intercomparison tests using small-sample statistics. *J. Atmos. Sci.*, **40,** 1884–1896.

Rowntree, P. R., 1972: The influence of tropical East Pacific Ocean temperature on the atmosphere. *Quart. J. Roy. Meteor. Soc.*, **98,** 290–321.

——, 1979: The effects of changes in ocean temperature on the atmosphere. *Dyn. Atmos. Oceans,* **3,** 373–390.

Shukla, J., 1975: Effect of Arabian sea surface temperature anomaly on Indian summer monsoon: A numerical experiment with the GFDL model. *J. Atmos. Sci.*, **32,** 503–521.

——, and B. Bangaru, 1979: Effect of a Pacific sea surface temperature anomaly on the circulation over North America. *Fourth NASA Weather and Climate Program Science Review,* The NASA/Goddard Space Flight Center, Greenbelt, MD, 171–176.

Storch, H. V., 1982: A remark on Chervin-Schneider's algorithm to test significance of climate experiments with GCMs. *J. Atmos. Sci.*, **39,** 187–189.

——, and H. Kruse, 1985: The significant tropospheric midlatitudinal El Niño response patterns observed in January 1983 and simulated by a GCM. *Coupled Atmospheric-Ocean Models,* J. C. J. Nihoul, Ed., *Elsevier Oceanogr. Ser.*, **40** (in press).

Webster, P. J., 1981: Mechanisms determining the atmospheric response to sea surface temperature anomalies. *J. Atmos. Sci.*, **38,** 554–571.

Wiin-Nielsen, A., 1971: On the motion of various vertical modes of transient, very long waves. *Tellus,* **23,** 87–98.