

Verification of General Circulation Models Applied to the Hamburg University GCM. Part I: Test of Individual Climate States

HANS VON STORCH AND ERICH ROECKNER

Meteorologisches Institut der Universität Hamburg, Bundesstrasse 55, 2000 Hamburg 13, FRG

(Manuscript received 12 April 1982, in final form 12 July 1983)

ABSTRACT

One objective of general circulation models is to simulate, e.g., a "January" which is not distinguishable from observed Januaries. A strategy to verify an individual simulated state is proposed. Its main elements are: data compression by means of EOFs, performance of a multivariate parametric test, and a subsequent univariate analysis.

The suggested technique is applied to four January simulations performed with the Hamburg University GCM. The meteorological parameters treated are the zonally averaged January mean of the geopotential itself and of the intensity of transient and stationary eddies of geopotential height at 300, 500 and 850 mb. The comparison is based on daily observations from 15 Januaries (1967–81).

It turns out that the midlatitudinal meridional gradient of geopotential height is significantly overestimated at all levels. The intensity of the transient eddies is significantly overestimated at 850 mb at practically all latitudes and at 300 and 500 mb at midlatitudes.

1. Introduction

This paper is concerned with a comparison between the observed January climate and a few single Januaries simulated with a general circulation model (GCM). Such a comparison is a prerequisite for applying a GCM to climate change experiments.

The traditional way (e.g., Gates, 1975; Blackmon and Lau, 1980) of comparing individual simulated months or seasons with the respective multiyear average gives some insight into the ability of the model to simulate the observed climate, but it may lead to erroneous assessments (see Section 2a). An approach which takes the interannual variability into account was proposed by Chervin (1981). Unfortunately, Chervin's procedure is based on a large number of simultaneously performed univariate tests, which is statistically doubtful (cf. Storch, 1982a). An alternative statistical strategy has been used in the present study. Its background is presented in some detail in Sections 2b–2e.

At the present stage, we test whether the model-generated circulation might be observed. At a future stage of investigation, one has to inquire whether the interannual moments of the model-generated circulation coincide with the respective observed ones.

Sections 3 and 4 are descriptions of the Hamburg University GCM and the experiments on which the proposed strategy is applied, as far as it is useful for understanding the results. It is beyond the scope of this paper to perform a comprehensive comparison of the model Januaries and the observed climate.

Rather, it is intended to apply the strategy of Section 2 exemplarily to an easily accessible variable like the geopotential height. Accordingly, in Section 5, the comparison is confined to the monthly mean and the intensity of the stationary and transient eddies of geopotential height at 300, 500 and 850 mb. In order to check the quality of the proposed scheme, it is applied to an independent case, namely January 1982 (Section 6).

The novel aspect of this paper is the application of a multivariate statistical procedure (cf. Hasselmann, 1979; Storch, 1982a,b; Hayashi, 1982) to GCM verification in order to evaluate the quality of the simulation. This procedure may also be used as an aid to trace the response of SST-anomalies; most likely, however, this will require an *a priori* first guess (Hasselmann, 1979).

2. The technique to perform the comparison

a. The necessity of a statistical approach

The traditional way to verify model-generated distributions of monthly or seasonal climate statistics is the side by side comparison of the model state with the respective long-term observations (e.g. Blackmon and Lau, 1980). Generally, this approach is not very effective and may lead to an incorrect assessment of agreement or disagreement of the simulated and observed states. To see this, let Y be a spatially distributed climate variable, and \bar{Y} the mean of a number of observed Y s. It can easily be shown that generally 1) the

long-term average \bar{Y} is smoother than the individual Y and 2) the maxima (minima) of \bar{Y} are smaller (larger) than those of Y . In fact, the situation may occur that the averaged state \bar{Y} will never be observed as an individual state. Thus, a comparison of \bar{Y} with an observed Y may lead to the conclusion that the observed state Y is "significantly" different from climate. On the other hand, a model-generated state which is identical with \bar{Y} would be interpreted as an ideal agreement of the model and the atmosphere (cf. Fischer and Storch, 1982).

This discrepancy might be understood from the internal and interannual variability of the atmospheric circulation. Obviously, the model-generated state has to be related not to the *point* \bar{Y} but to the *random variable* Y . A proper way to do this is the performance of a statistical test.

b. The statistical frame

In what follows, the random variable itself and the samples drawn from it, are denoted by Y . An individual vector, which is to be tested whether it is drawn from Y or not, is denoted by X . As *null hypothesis* we have

$$N: X \text{ is drawn from } Y$$

and as an *alternative*

$$A: X \text{ is not drawn from } Y.$$

The basic idea is to fix a number $\tilde{p} \in (0; 1)$ and a bounded domain $F(\tilde{p})$ which contains $100 \cdot \tilde{p}\%$ of all samples drawn from Y . The latter is identical to the relation

$$P_Y\{Y \in F(\tilde{p})\} = \tilde{p}, \tag{2.1}$$

where P_Y denotes the probability that the event given in the brackets is true. The complementary of $F(\tilde{p})$ is called the "critical region". The decision rule for $X \in \mathbf{R}$ is

$$\left. \begin{array}{l} \text{reject } N \text{ if } X \notin F(\tilde{p}) \\ \text{do not reject } N \text{ if } X \in F(\tilde{p}) \end{array} \right\} \tag{2.2}$$

The probability of deciding to "reject N " is $1 - \tilde{p}$, if N is true. This quantity is called "risk" in statistical theory. On the other hand, the probability of deciding "no rejection of N ", if A is true, i.e., the "power" of the test, may be quite small, e.g., if X is drawn from a random variable very similar to Y .

The first statement of (2.2) is interpreted as " A is true" or " N is false". It is identical with the disagreement of the model-generated state from the atmospheric states at the \tilde{p} -level. The second statement of (2.2), however, means neither " N is true" nor " A is false," but merely: " N has not turned out to be false".

In what follows, we fix \tilde{p} , the level of significance, to be 95%, which is arbitrarily chosen but usual in meteorological applications. The remaining problem

is how to fix the "region of no rejection" $F(\tilde{p})$. Let f_Y be the probability density function of Y ; then the following definition is compatible with (2.1):

$$F(\tilde{p}) := \{X \in \mathbf{R}^n; f_Y(X) \geq a_{\tilde{p}}, \tag{2.3}$$

where the numbers $a_{\tilde{p}}$ and \tilde{p} are connected through

$$\tilde{p} := \int_{f \geq a_{\tilde{p}}} f = \int_{F(\tilde{p})} f.$$

c. Normality

In most applications the assumption that the probability distributions are normal is permitted. Then, the isoface $f_Y(X) = a_{\tilde{p}}$ is the surface of an ellipsoid in \mathbf{R}^n , the region of no rejection $F(\tilde{p})$ is the interior, and the critical region is the exterior of this ellipsoid. If $\mu = E(Y)$ is the mean vector of the distribution and \mathbf{S} its covariance matrix, the isofaces are given by

$$t(X) := (X - \mu)' \mathbf{S}^{-1} (X - \mu) = a_{\tilde{p}}, \tag{2.4}$$

and the interior of the ellipsoid by $t(X) \leq a_{\tilde{p}}$. From statistical theory it is known that $t(X)$ is χ^2 distributed with n degrees of freedom. Thus, (2.3) changes to

$$F(\tilde{p}) = \{X, t(X) \leq \kappa(\tilde{p}, n)\},$$

where $\kappa(\tilde{p}, n)$ is the \tilde{p} -quantile of the χ^2 distribution with n degrees of freedom. The decision rule (2.2) now becomes

$$\left. \begin{array}{l} \text{reject } N \text{ if } t(X) > \kappa(\tilde{p}, n) \\ \text{do not reject } N \text{ if } t(X) \leq \kappa(\tilde{p}, n) \end{array} \right\} \tag{2.5}$$

In Fig. 1, the situation for $n = 1$ and $n = 2$ is sketched. For $n = 1$, the ellipsoid degenerates to two points and its interior to an interval. For two dimensions, the ellipsoid is a common ellipse. Fig. 1a shows the density function f of the standard normal distribution and the one-dimensional ellipsoid (interval) $F(95\%)$. In Fig. 1b, the isopleths of the density function of the two-dimensional normal distribution centered at the origin with the diagonal covariance matrix with diagonal coefficients 1 and 2 are given. The ellipse $F(95\%)$ is hatched. In both Figs. 1a and 1b, the application of the decision rule (2.5) will result in a rejection of the null hypothesis N for X_1 , but not for X_2 .

In order to apply (2.5), one has to compute the expression $t(X)$. For this purpose, the mean vector and the covariance matrix \mathbf{S} of the random variable Y are needed, which are unknown. Therefore, both the mean vector and the covariance matrix must be estimated. We have done this by means of daily analyses of the Deutscher Wetterdienst (DWD) for the Januaries 1967-81.

As a consequence of the application of the estimated μ and \mathbf{S} in Eq. (2.4), $t(X)$ is actually only asymptotically χ^2 distributed. Thus, the quantiles

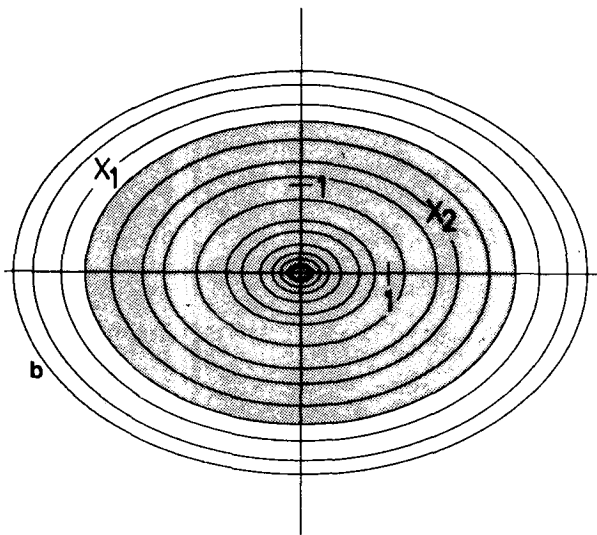
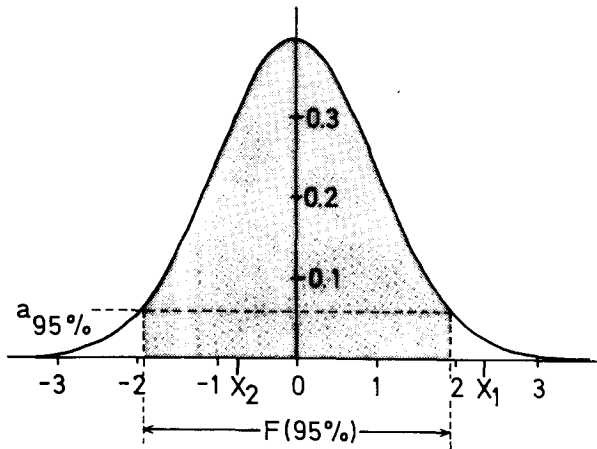


FIG. 1. (a) Standard normal distribution. The event X_1 will be called significantly different from climate at the 95% level, the event X_2 , undistinguishable from climate. (b) Two-dimensional normal distribution centered in the origin with diagonal covariance matrix and variances 1 and 2. Isopleths with $\chi(X) = \chi(p, 2)$ with $p = 99.5\%, 99\%, 95\%, 90\%, 80\%, 70\%, 50\%, 30\%, 20\%, 10\%, 5\%, 2.5\%, 1\%, 0.5\%$. Stippled area: $F(95\%)$. The event X_1 will be rejected, event X_2 not.

$\chi(\vec{p}, n)$ are valid only asymptotically, strictly speaking. But, in order to assure that these values do apply, Section 6 has been included demonstrating the fairness of the proposed scheme and the $\chi(\vec{p}, n)$ quantiles. When the covariance matrix must be estimated from few samples, Hotelling's T^2 test might be more appropriate.

d. EOF expansion

The variables which will be investigated in Section 5 form vectors with 32 and more components. It is not reasonable to apply the algorithm given in Section 2c to vectors with such dimensions because:

- 1) Numerical problems may occur.
- 2) The chance to detect model deficiencies decreases if the number of degrees of freedom increases (cf. HasseImann, 1979). In order to demonstrate this, assume \mathbf{X} to be a n -variate normal random variable with vanishing expectation $E(\mathbf{X}) = 0$ and the identity covariance matrix $\text{diag}(1, \dots, 1)$. Let \mathbf{Y} be another random variable identical to \mathbf{X} besides of $E(\mathbf{Y}) = (a, 0, \dots, 0)$. The application of (2.4) yields: $E[t(\mathbf{Y})] = a^2 + n$. If we assume, for example, that $a = 2$, then, the chance to detect $a \neq 0$ is quite good if $n \leq 2$, but bad if $n \geq 4$.

Therefore, the fields and profiles are expanded into a finite series of empirical orthogonal functions

$$\mathbf{X} = \sum_{i=1}^K a_i \mathbf{y}_i + r, \tag{2.6}$$

with the EOF coefficients a_i defined as the inner product of the vector \mathbf{X} and the i th EOF $a_i := \langle \mathbf{X}, \mathbf{y}_i \rangle$. r denotes the residuum and K the number of available EOFs. The EOFs are ordered such that a low index is connected with large variance. They were obtained without an *a priori* subtraction of the first moment such that the first EOF shows the structure of the multiyear mean. This approach is permitted because no theoretical properties of EOFs besides of their effectiveness to compress data are utilized; in our study, EOFs are applied because of the fast convergence of (2.6).

Instead of the vector \mathbf{X} itself, the vector

$$\mathbf{A} := (a_1, \dots, a_I, r)$$

in the case of hemispheric fields and

$$\mathbf{A} := (a_1, \dots, a_I)$$

in the case of meridional profiles are utilized. The number I will be specified below. The number n of degrees of freedom used for $\chi(\vec{p}, n)$ is $I + 1$ and I , respectively.

Unfortunately, there are some difficulties involved in the estimation of the covariance matrix of a vector consisting of EOF coefficients if the two data sets used to compute the EOFs, and to estimate the second moments of the EOF coefficients, are identical. The second of the low-indexed EOF coefficients are overestimated and those with high indices underestimated (see Appendix). Therefore, from the available total data set of 15 Januaries only the first seven Januaries (i.e., the Januaries 1967-73) are applied for EOF estimation. The choice of the first seven Januaries is irrelevant; any other subset of about the same magnitude could be employed. The moments are computed from the EOF coefficients which are projections

of the total data record of 15 Januaries onto the EOFs estimated from the first seven Januaries.

Because the EOFs connected with small eigenvalues generally are bad estimators of the corresponding principal vectors, the last two EOFs were not taken into account, i.e., $I = 5$.

It turned out that in the case of meridional profiles, the first five EOFs explain a sufficient amount of variance. This is different when hemispheric fields are expanded; therefore, the residuum is included in the "hemispheric" vector A.

e. Univariate analysis

After having performed the ultimate multivariate test described above to a certain vector X , the eventual decision "reject N " may be followed by a closer univariate analysis. This is done by applying the one-dimensional version of the above test (2.5) to 1) grid-points of the hemisphere and the meridional profiles, respectively, and 2) to the EOF coefficients a ($i = 1, \dots, I$) and the residuum r , individually. In this way, the model generated values are inspected whether they are outstandingly small or large.

The objective of the multivariate approach is to emphasize the significance of a disagreement. The objective of the series of univariate considerations, on the other hand, is to find those components which may have caused the significant difference. The univariate approach is not thought of as a test giving rise to the statement "significant change of the field at a certain point or of a certain EOF coefficient".

3. The model

A detailed description of the GCM developed at the Meteorological Institute of Hamburg University (MIHU) was given by Roeckner (1979). Therefore, a summary of its main aspects should be sufficient here.

a. Structure and parameterization

The model equations are discretized on the so-called B-grid (Arakawa and Lamb, 1977) with a horizontal grid size of $\Delta = 2.8125^\circ$, and $3\Delta\sigma$ layers of equal depth.

Surface fluxes of heat, moisture and momentum are calculated from generalized similarity theory with empirical constants derived by Yamada (1976) for momentum and heat and by Brutsaert and Chang (1978) for water vapor. Condensation is assumed to take place if the relative humidity within a grid volume exceeds a threshold value of 100%. Convective fluxes of heat and moisture are parameterized according to a scheme proposed by Kuo (1965) for deep convection. A dry adiabatic adjustment scheme prevents thermal instability.

Solar radiation is calculated in a rather crude way:

20% of the incoming radiation is assumed to be absorbed by the atmosphere and 50% by the ground surface. The remaining 30% is reflected back to space. The sun is fixed at a mean January declination angle. Thermal radiative heating is calculated from empirical formulas prepared by Moeller (1954) for the 400 and 1000 mb levels. The net emissions at these levels are given as a function of temperature and humidity. Surface heating is calculated from a heat transfer equation assuming a constant heat capacity of the soil. The temperatures of the oceans and the sea ice are held fixed.

Vertical turbulent fluxes above the boundary layer are calculated only for momentum. Horizontal diffusion of heat (potential temperature) and moisture is parameterized according to a nonlinear fourth-order scheme with a proportionality constant of 0.1 (Roeckner and Storch, 1980). Instead of horizontal momentum diffusion, a high order numerical filter is applied intermittently depending on the small-scale noise production of the model.

b. Boundaries

The model covers the Northern Hemisphere with symmetric boundary conditions at the equator. The top of the model is at $p_T = 100$ mb. The "vertical velocity" $\dot{\sigma} = d\sigma/dt$ vanishes at the top $p = p_T$ and at the bottom $p = p_s$ of the model atmosphere. Mountains are included in a realistic but slightly smoothed way. The initial surface temperature distribution is interpolated from mean January data prepared by Schutz and Gates (1971). The sea surface temperature (SST) and the surface relative humidity are fixed in time. The surface relative humidity is assumed to be 100% over sea and 70% over land. The surface roughness which is measured by a "roughness length" z_o is calculated from friction velocity over sea and held constant at $z_o = 20$ cm over land.

To summarize, the spatial distribution of the surface conditions is realistic only with respect to orography and surface temperature. All other surface conditions are specified in a highly idealized way.

TABLE 1. List of experiments. Numbers added to the letters denote the respective 30-day intervals of each experiment, e.g. A1: 0–30 days of experiment A, B2: 30–60 days of experiment B, and so on.

| Experiment | Simulated time (days) | Change with respect to experiment A |
|------------|-----------------------|--|
| A | 90 | — |
| B | 60 | Pacific SST anomaly in middle latitudes |
| C | 60 | No horizontal diffusion of heat and moisture |

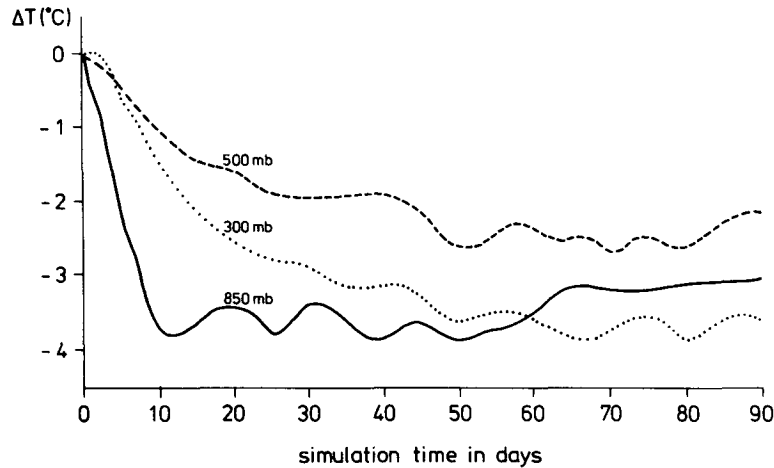


FIG. 2. Difference $T(t) - T(0)$ of the hemispheric mean temperature as a function of time for experiment A.

4. Experiments

a. Description

Three experiments have been performed so far, initialized with observed data at 0000 GMT 2 January 1974. The total simulated time is either 60 or 90 days. The experiments differ with respect to parameterizations or boundary conditions (Table 1). The global response of the model to the midlatitude SST anomaly in experiment B is weak so that the "climates" of experiments A and B can barely be distinguished (see Section 5). Similarly, the slight

model change in experiment C does not produce a notable response in the circulation.

b. Data and analysis

The model variables originally defined at σ -levels were interpolated vertically to three pressure levels: 300, 500 and 850 mb. They are available every 12 h. An interval of 30 simulated days is treated as an individual "model January". Thus, according to Table 1, three model Januaries were calculated during experiment A and two during both experiments B

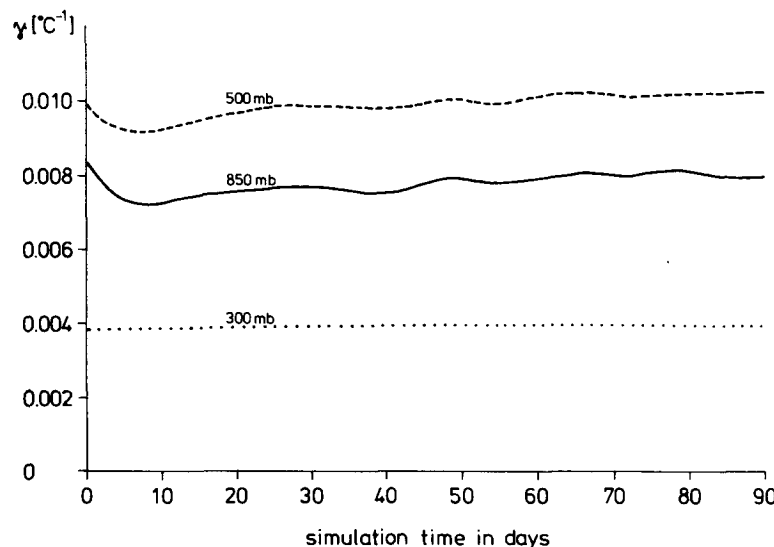


FIG. 3. Hemispheric mean static stability parameter

$$\gamma = -\frac{\Theta}{T} \frac{\kappa}{p} \left(\frac{\partial[\Theta]}{\partial p} \right)^{-1}$$

as a function of time for experiment A.

and C. The first January of A is denoted by A1, the second of B by B2, and so on.

c. Quasi-stationary state

A prerequisite of comparing simulated and observed climate states is that both are free from a noticeable temporal trend. Such a state will be called "quasi-stationary". This requirement is most likely fulfilled for the observed January climate and after some "response time" for the simulated climate.

Figure 2 shows an initial cooling trend of the model atmosphere at all levels for experiment A. However, the cooling seems to be initiated in the lowest layer which is probably caused by deficiencies in the boundary layer treatment. The surface heat fluxes depend crucially on the differences between the virtual potential temperatures at the surface and at the mixed layer height which is no model variable, however. For convenience, therefore, the mixed layer height is assumed to be given by the height of the lowest model level, i.e., at about 850 mb which corresponds to a height of ~ 1500 m. This level, however, lies in many cases (e.g., over the tropical oceans) in the stable free atmosphere. Consequently, the surface heat fluxes during the first ten days are directed predominantly downward, resulting in a heat loss over the oceans whose temperature is not allowed to change. This heat loss is probably the main cause for the overall cooling of ~ 3 K with respect to the initial value which is close to the climatological mean. Accordingly, the mean static stability parameter drops sharply at the 850 and 500 mb levels during the first eight days and then recovers gradually to approximately the initial value (Fig. 3). The stability parameter at the highest level shows little variation. Its rather low value is caused by the upper boundary value which is calculated from the assumption of an isothermal atmosphere above 300 mb.

Figure 4 shows the time evolution of the total kinetic energy for experiment A (upper curve) which is split additionally into its zonal and eddy parts (lower curves). The increase of the total kinetic energy is produced mainly by an increase of the zonal part whereas the eddy kinetic energy shows no systematic growth with time, although there is a vacillation-like behavior beyond day ~ 40 . While the amplitude of these fluctuations is realistic for winter conditions, the time scale of approximately 2 weeks is less than the observed 3–4 weeks in winter (McGuirk and Reiter, 1977).

The results shown in Figs. 2–4 do not differ essentially from the results obtained in experiments B and C which were run until day 60 only.

To sum up, the choice of 30 days as the response time the model needs to reach a quasi-stationary state seems to be reasonable, which is in agreement with

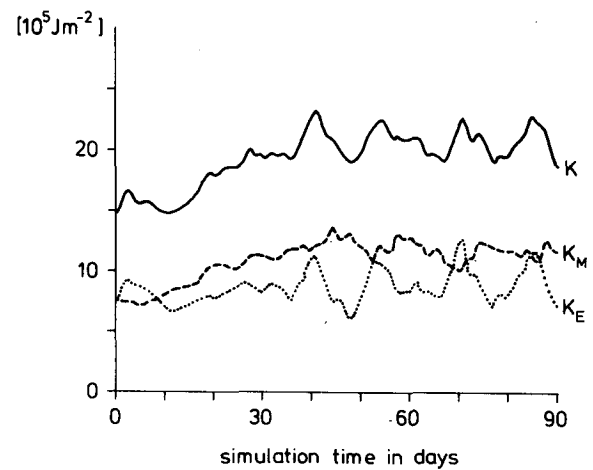


FIG. 4. Total hemispheric and vertical mean kinetic energy $K = K_M + K_E$ together with its zonal part K_M and eddy part K_E .

a suggestion of Washington and Chervin (1980). Therefore, the first 30 days of each experiment, i.e., the model Januaries A1, B1 and C1, are disregarded.

5. Accomplishment of the comparison

The geopotential height at 300, 500 and 850 mb is analyzed. For each January, observed and simulated, the fields are decomposed according to

$$\left. \begin{aligned} \Phi(\varphi, \lambda, t) &= \Phi'(\varphi, \lambda, t) + \bar{\Phi}(\varphi, \lambda) \\ \Phi'(\varphi, \lambda, t) &= \Phi'^*(\varphi, \lambda, t) + [\Phi'](\varphi, \lambda) \\ \bar{\Phi}(\varphi, \lambda) &= \bar{\Phi}^*(\varphi, \lambda) + [\bar{\Phi}] \end{aligned} \right\}, \quad (5.1)$$

where a bar denotes a temporal average and a bracket a zonal average. The deviation from the temporal (zonal) average is marked by a prime (asterisk); Φ' and $\bar{\Phi}$ are called the transient and the stationary fields, Φ'^* and $\bar{\Phi}^*$ the transient and stationary eddies and $[\Phi']$ and $[\bar{\Phi}]$ the transient and stationary cells, respectively.

TABLE 2. Values of $t(X)$ at different levels for the four experiments for the stationary features. The quantiles are $\kappa = 11.1$ for $[\bar{\Phi}]$ and $[\Phi'^*]$ and $\kappa = 12.6$ for $\bar{\Phi}$.

| Level | Statistic | B2 | A2 | A3 | C2 |
|--------|--------------------|--------|--------|--------|--------|
| 300 mb | $X = [\bar{\Phi}]$ | 160.07 | 136.32 | 99.52 | 103.57 |
| | $X = [\Phi'^*]$ | 6.34 | 7.93 | 5.31 | 5.96 |
| | $X = \bar{\Phi}$ | 51.39 | 46.06 | 46.15 | 25.50 |
| 500 mb | $X = [\bar{\Phi}]$ | 147.77 | 134.61 | 94.36 | 101.11 |
| | $X = [\Phi'^*]$ | 4.41 | 6.63 | 4.17 | 2.98 |
| | $X = \bar{\Phi}$ | 20.72 | 16.28 | 17.58 | 9.67 |
| 850 mb | $X = [\bar{\Phi}]$ | 101.55 | 85.99 | 113.51 | 84.16 |
| | $X = [\Phi'^*]$ | 5.46 | 4.26 | 9.38 | 9.17 |
| | $X = \bar{\Phi}$ | 19.88 | 10.04 | 22.74 | 10.08 |

In Section 5a we report on the stationary features, i.e. the zonally averaged statistics $[\bar{\Phi}]$ and $[\bar{\Phi}^{*2}]$ and the hemispheric distribution $\bar{\Phi}$. In 5b the transient features are treated, i.e., the zonally averaged variance of transient eddies $[\bar{\Phi}'^{*2}]$ and the hemispherically distributed $\bar{\Phi}'^2$. The variance $[\bar{\Phi}']^2$ due to the transient cell is small and therefore neglected.

a. Stationary features

1) MULTIVARIATE DECISIONS

In Table 2, the numbers $t(X)$ are given for the zonally averaged values of $X = [\bar{\Phi}]$ and $X = [\bar{\Phi}^{*2}]$ and for the hemispheric distribution $X = \bar{\Phi}$ as obtained by the model simulations. For the zonal means, the number of degrees of freedom is 5, while that of the hemispheric distribution is 6 (see Section 2d). The corresponding 95% quantiles are $\kappa = 11.1$ and $\kappa = 12.6$. A model-generated state is defined to be significantly different from what can be observed, if the $t(X)$ value is larger than the respective 95% quantile (risk: 5%).

It turns out that at all levels the circulation of the model atmosphere is significantly distinct from the

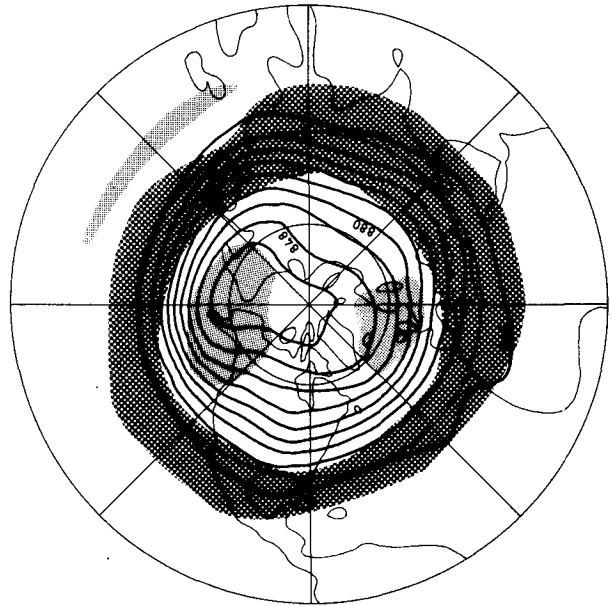


FIG. 6. $\bar{\Phi}$ at 300 mb generated by experiment B2. Areas with (univariately) too high geopotential heights are coarsely stippled, those with too low values are finely stippled (units: gpdam).

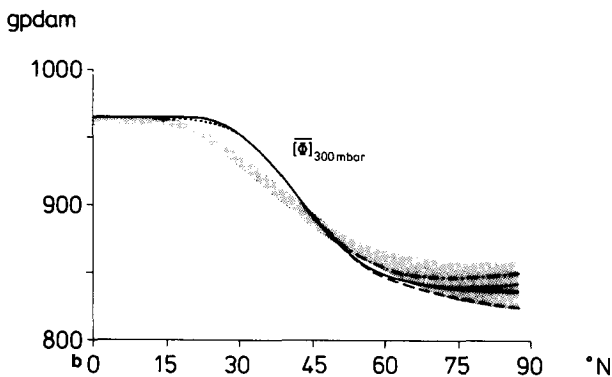
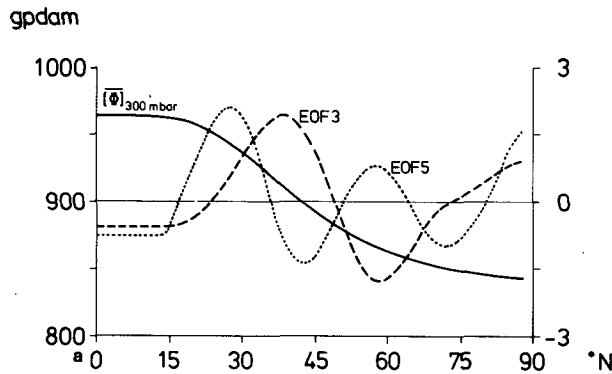


FIG. 5. (a) Multiyear mean of $[\bar{\Phi}]$ at 300 mb and EOFs 3 and 5, the coefficients of which are overestimated by all four experiments. (b) Observed 95% tube and simulated $[\bar{\Phi}]$ profiles at 300 mb (A2, solid line; A3, dashed; B2, dotted; and C2, dashed-dotted).

real atmosphere for the temporally and spatially averaged fields. The result is significant for $[\bar{\Phi}]$ at all levels and all experiments, and also for the hemispheric $\bar{\Phi}$, except for experiment C2 at 500 mb and 850 mb, and experiment A2 at 850 mb. For the stationary eddies $[\bar{\Phi}^{*2}]$, no significant disagreement can be detected.

2) UNIVARIATE ANALYSIS

For inspection of which structures might have caused those significant disagreements mentioned above, the behavior of the EOF coefficients of $[\bar{\Phi}]$ at all levels and of $\bar{\Phi}$ at 300 mb are checked. Furthermore, $[\bar{\Phi}^{*2}]$ is examined whether it is within its 95% tube

TABLE 3. Values of $t(X)$ at different levels for the four experiments for the transient features. The quantiles are $\kappa = 11.1$ for $[\bar{\Phi}^{*2}]$ and $\kappa = 12.6$ for $\bar{\Phi}'^2$.

| Level | Statistic | B2 | A2 | A3 | C2 |
|--------|-------------------------|-------|-------|--------|-------|
| 300 mb | $X = [\bar{\Phi}^{*2}]$ | 32.43 | 44.20 | 21.94 | 46.18 |
| | $X = \bar{\Phi}'^2$ | 3.42 | 2.45 | 3.39 | 5.40 |
| 500 mb | $X = [\bar{\Phi}^{*2}]$ | 11.34 | 20.62 | 4.63 | 8.18 |
| | $X = \bar{\Phi}'^2$ | 1.82 | 3.96 | 4.37 | 6.52 |
| 850 mb | $X = [\bar{\Phi}^{*2}]$ | 21.80 | 23.89 | 116.33 | 29.92 |
| | $X = \bar{\Phi}'^2$ | 10.95 | 29.84 | 32.14 | 4.54 |

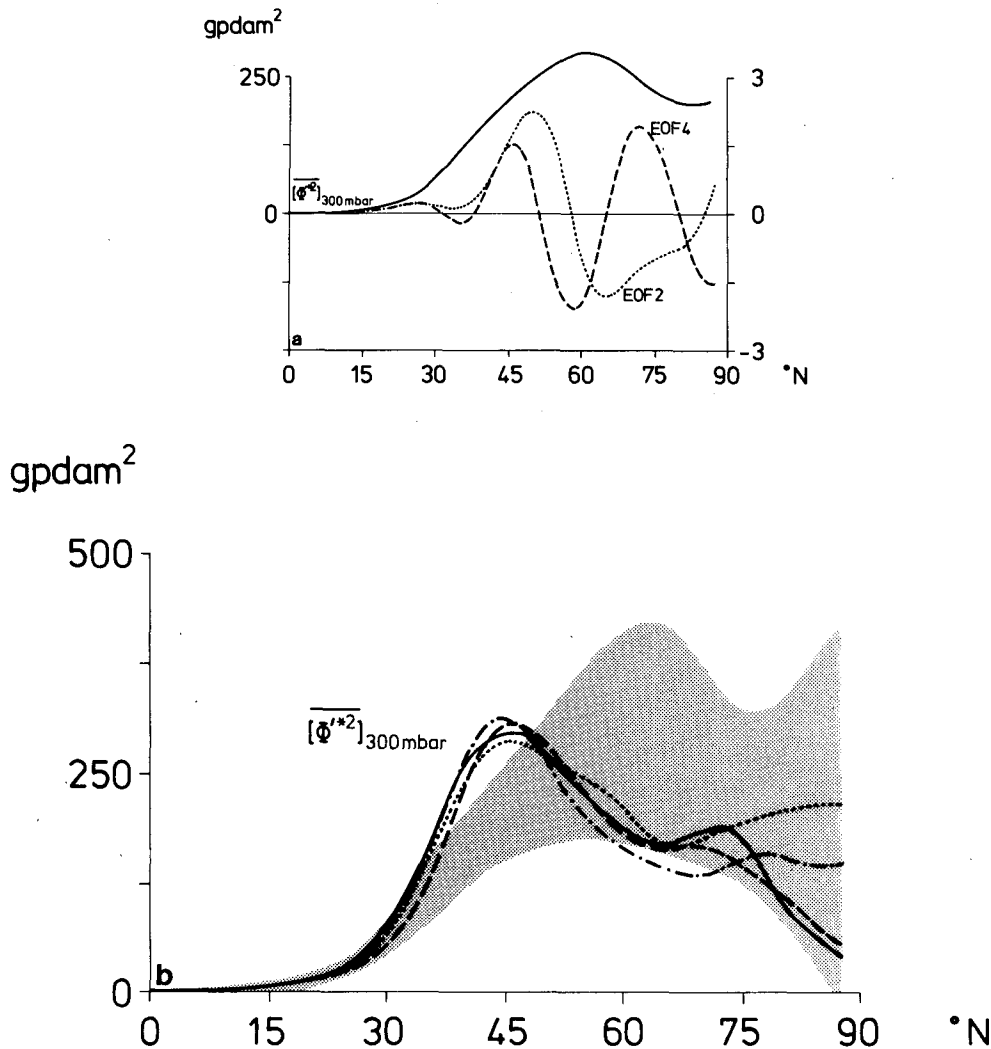


FIG. 7. (a) Multiyear mean of the intensity of the transient eddies $[\overline{\Phi''^2}]$ at 300 mb and EOFs 2 and 4, the coefficients of which are overestimated by all four experiments. (b) As in Fig. 5b, except for $[\overline{\Phi''^2}]$ at 300 mb.

(i.e., the sum of the 95% probability intervals estimated for each latitude independently) or not.

Meridional profile $[\overline{\Phi}]$. For $[\overline{\Phi}]$ at 300 mb, the coefficients of EOF 3 and 5 are noteworthy: the model-generated coefficients vary between 5.6 and 7.0 (EOF3) and between 2.3 and 3.0 (EOF5), while the atmospheric 95% intervals are $[-3.59; 2.6]$ (EOF3) and $[-1.4; 1.2]$ (EOF5).

In the lower part of Fig. 5, the four model-generated profiles at 300 mb are shown together with 95% intervals at each latitude (stippled area) as estimated from the 15 Januaries data record. In the upper part, EOFs 3 and 5 are shown together with the 15-January average. (These two EOFs describe deviations from the observed multiyear mean.)

Taking into account that the model-generated coefficients a_3 and a_5 are positive and that a_3 is about twice as large as a_5 , these two EOFs describe an amplification in $[20^\circ\text{N}; 45^\circ\text{N}]$ and a somewhat weaker reduction in $[45^\circ\text{N}; 75^\circ\text{N}]$. This significant feature— increase of geopotential height south of 45°N and decrease of it north of 45°N —is displayed by the fact that the model-generated profiles lie partly outside of the univariately estimated 95% tube, which is shown in the lower part of Fig. 5. Similar results hold for the 500 and 850 mb levels (not shown).

Hemispheric distribution $\overline{\Phi}$. The univariate analysis of $\overline{\Phi}$ again yields an overestimated meridional gradient. Moreover, the midlatitude troughs of $\overline{\Phi}$ are broadened, flattened and smeared out to the east at

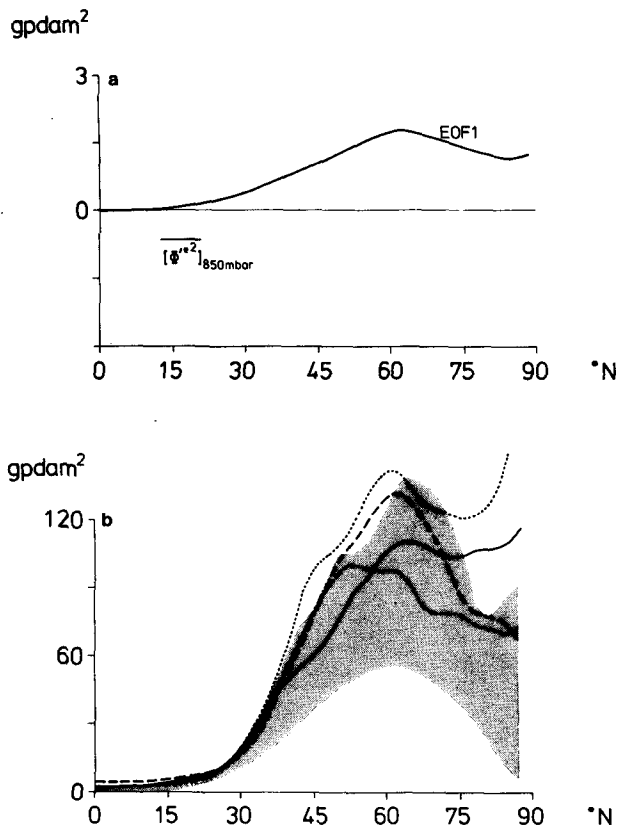


FIG. 8. (a) EOF1 of intensity of the transient eddies $[\overline{\Phi'^*2}]$ at 850 mb showing the structure of the multiyear mean. The model generated profiles have a too large EOF1 coefficient. (b) As in Fig. 5b, except for $[\overline{\Phi'^*2}]$ at 850 mb.

all levels: Fig. 6 shows $\overline{\Phi}$ of experiment B2 at 300 mb together with coarsely (finely) stippled areas of univariately too high (too low) geopotential heights.

Intensity of stationary eddies. After we detected no significant differences between the model generated $[\overline{\Phi'^*2}]$ and the real ones, a check was done to see to what extent the meridional profiles leave their respective 95% tubes. This happens neither at 850 mb nor at 500 mb (not shown). Only at 300 mb, at about 40–50°N, is the intensity of the model-generated stationary eddies smaller than 95% of the observed states (not shown). The fact that this was found for all four experiments points to the possibility of a systematic underestimation of stationary features at these latitudes, an impression confirmed by Fig. 6.

b. Transient features

1) MULTIVARIATE DECISIONS

The numbers $t(X)$ obtained for $[\overline{\Phi'^*2}]$ and for $\overline{\Phi'^2}$ are listed in Table 3. As in Section 5a1), the quantiles α are 11.1 and 12.6, respectively.

At 300 mb and 850 mb, the meridional profiles are significantly different from atmospheric ones for all four experiments. At 500 mb, this is valid only for experiments A2 and B2.

The model generated $\overline{\Phi'^2}$ are not distinguishable from atmospheric ones at the two higher levels. But, at the lowest level, 850 mb, the probability that the states

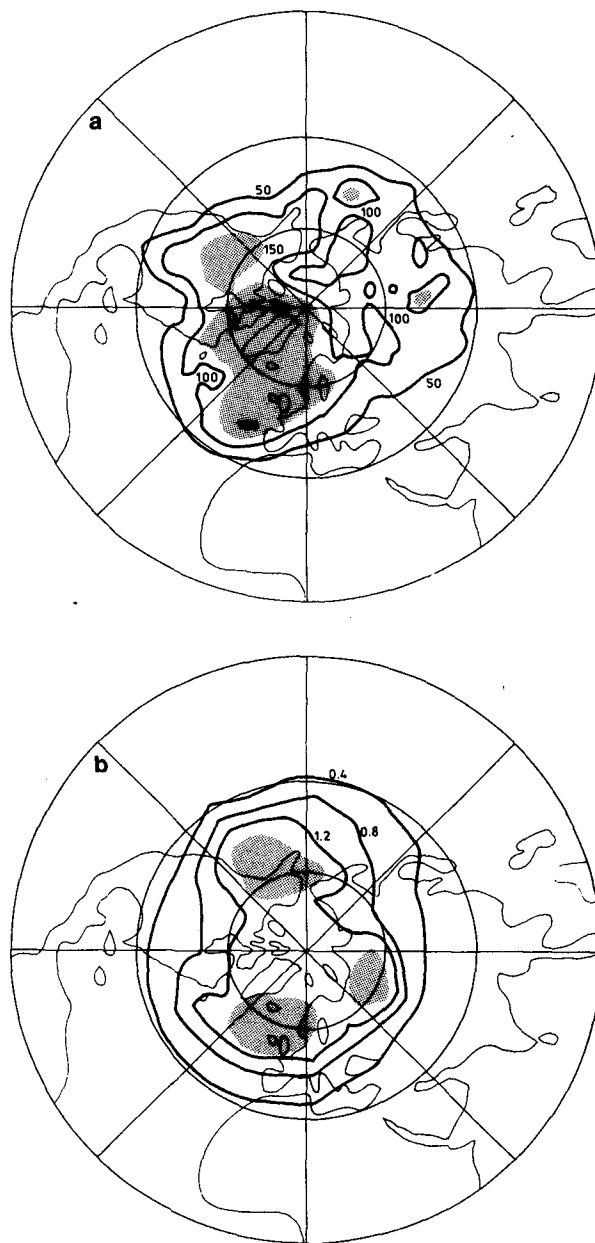


FIG. 9. (a) Hemispherically distributed variance of transient eddies at 850 mb as simulated during experiment A3. Spacing: 50 gpdam², stippled area: ≥ 150 gpdam². (b) EOF1 of $\overline{\Phi'^2}$ at 850 mb showing the structure of the multiyear average. Dimensionless units, spacing: 0.4, stippled area: ≥ 1.6 .

generated by experiments A2 and A3 will be observed in nature, is smaller than 5%. The field $\overline{\Phi^2}$ obtained by experiment C2 looks quite different as those obtained in the other three experiments (not shown), a fact which is reflected by Fig. 8. The reason for this is unknown.

2) UNIVARIATE ANALYSIS

*The meridional profile $[\overline{\Phi^{*2}}]$.* At 300 mb, the rejection of the corresponding null hypothesis is caused by too large coefficients of EOF2 and (minor) EOF4 (see Fig. 7a), a fact which is reflected quite clearly by Fig. 7b, showing the model-generated states. The EOFs have maxima where the simulated $[\overline{\Phi^{*2}}]$ are outside the 95% tube. Thus, the difference between observed and model-generated transient eddy variance is mainly due to the unrealistic maximum at about 45°N, where the overestimation of the latitudinal gradient of $\overline{\Phi}$ is maximum (Fig. 5b).

At 850 mb, the rejection is caused mostly by EOF1 for experiments A2, A3 and B2. Thus, the reason for the rejection of the model-generated $[\overline{\Phi^{*2}}]$ at 850 mb is different from that of the rejection at 300 mb—at 850 mb, the transient activity everywhere is too strong. The rejection of experiment C2 cannot be explained by an individual EOF. Fig. 8 shows the model-generated states and EOF1.

Hemispheric distribution $\overline{\Phi^2}$ at 850 mb. In the foregoing cases, the univariate analysis of EOF coefficients

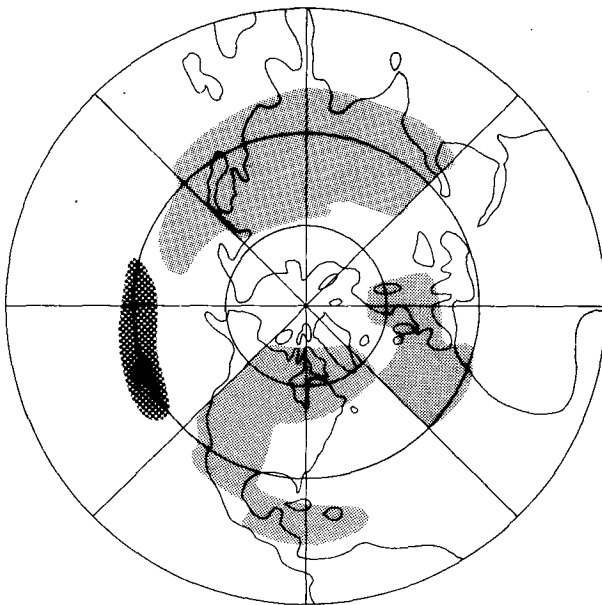


FIG. 10. Result of the point by point performance of the univariate analysis of $\overline{\Phi^2}$ at 850 mb, experiment A3, [finely (coarsely) stippled: stimulated variance too small (large) compared with observations].

TABLE 4. Values of $t(X)$ at different levels for the data record observed during January 1982. The quantile is $\alpha = 11.1$.

| | Pressure level (mb) | | |
|--------------------------|---------------------|-----|-----|
| | 850 | 500 | 300 |
| $[\overline{\Phi^{*2}}]$ | 12.8 | 4.1 | 4.1 |
| $[\overline{\Phi^{*2}}]$ | 1.8 | 0.5 | 0.3 |
| $[\overline{\Phi}]$ | 0.8 | 1.2 | 1.4 |

and of gridpoint values yielded coinciding results. Thus, one might argue that an analysis of gridpoint values would suffice. The following case shows that this is generally not true.

In Fig. 9, the hemispheric distribution of $\overline{\Phi^2}$ at 850 mb, obtained from experiment A3 is shown together with EOF1 which was found to be too large. The maxima of EOF1 are at 60°N, consistent with the meridional profiles of the transient eddy variance (cf. Fig. 8). The maxima of EOF1 (Fig. 9b) are centered over the North Atlantic, Gulf of Alaska and northern Siberia, pointing to an overestimation of $\overline{\Phi^2}$ in these regions. A univariate comparison performed simultaneously at gridpoints, shows a different structure (Fig. 10). Only the Siberian maximum is a common feature of both analyses. The reason for this discrepancy is that large variances can be observed over the North Atlantic and the Gulf of Alaska, however, not at the same time. This can, of course, not be noticed by a point by point analysis.

For the sake of completeness we mention that an overestimated EOF1 is a common property of experiments A2, A3 and B2. Furthermore, sometimes the residual and the coefficients of EOF3 and 5 are too high.

6. Quality control

The proposed strategy may be checked to determine whether it is radical (i.e., that the probability of an erroneous rejection of the null hypothesis is clearly larger than \hat{p}) by applying it to data fulfilling the null hypothesis, i.e., to some independent and observed data. To do so, the zonally averaged geopotential height and its transient and stationary eddies of January 1982 were computed and tested. The result was (see Table 4) that only the intensity of the transient eddies at 850 mb were erroneously rejected as conspicuous. Thus, one of a total of nine decisions was incorrect; this is an acceptable rate.

7. Conclusion

A strategy to compare the GCM simulation of one individual month with the respective ensemble of observations was presented. Its main elements are an

EOF expansion in order to reduce the number of degrees of freedom drastically, a multivariate parametric test, and a final series of univariate point by point comparisons. The test may result in the decision that the simulated state is not to be expected to become observed (rejection of the null hypothesis). If the null hypothesis is not rejected, nothing can be said.

The strategy is performed exemplarily to the January geopotential height. The estimation of the moments of the January climate is done by means of the observations from 1967 to 1981. The application to four January simulations performed with the Hamburg University GCM yields a series of rejections, i.e., a series of significant differences from the observed climate. Thus, the technique is able to detect discrepancies.

To assure that the test was not a radical one, it was finally applied to the observations of January 1982. This check was also successful.

APPENDIX

Estimation of Variances of EOF Coefficients

In the following, the problem of estimating second moments of the coefficients \hat{a}_i of a fixed set of estimated EOFs $\{\hat{y}_1, \dots, \hat{y}_M\}$ is discussed. The coefficient \hat{a}_i is defined as inner product $\langle \hat{y}_i, \mathbf{x} \rangle$ of the i th EOF \hat{y}_i with the random vector \mathbf{x} under concern. Without loss of generality is assumed that the expectation $E(\mathbf{x})$ of \mathbf{x} vanishes and that the covariance matrix \mathbf{X} of \mathbf{x} is diagonal with λ_1, \dots as diagonal elements. For the true EOFs $\{y_1, \dots\}$ of \mathbf{x} with eigenvalues λ_i and coefficients $a_i = \langle y_i, \mathbf{x} \rangle$ holds

$$\left. \begin{aligned} E(a_i) &= 0 \\ \text{Var}(a_i) &= E(a_i^2) = \lambda_i \end{aligned} \right\} \quad (A1)$$

where E denotes expectation. If N samples \mathbf{x}^i of \mathbf{x} are available, by means of M , $M \leq N$, samples \mathbf{x}^i , EOFs $\{\hat{y}_1^M, \dots, \hat{y}_M^M\}$ with eigenvalues $\hat{\lambda}_i^M$ may be estimated. The resulting coefficients are denoted by \hat{a}_i^M .

The moments of \hat{a}_i^M are

$$\begin{aligned} E(\hat{a}_i^M) &= \langle \hat{y}_i^M, E(\mathbf{x}) \rangle = 0 \\ \text{Var}(\hat{a}_i^M) &= \hat{y}_i^{M'} \mathbf{X} \hat{y}_i^M = \sum_k \hat{y}_{ik}^2 \lambda_k. \end{aligned} \quad (A2)$$

The question is how to estimate $\text{Var}(\hat{a}_i^M)$. There are three different approaches possible:

Set $M = N$, i.e., use all available data to estimate the EOFs, and

$$\widehat{\text{Var}}(\hat{a}_i^M) := \hat{\lambda}_i^M \quad (A3)$$

according to (A1).

Set $M = N$ as in (A3) and compute conventionally:

$$\widehat{\text{Var}}(\hat{a}_i^N) := \frac{1}{N} \sum_{k=1}^N \left(\hat{a}_i^N - \frac{1}{N} \sum_{k=1}^N \hat{a}_i^N \right)^2. \quad (A4)$$

Eqs. (A3) and (A4) yield the optimal estimation of the EOFs at the given data set. The difference between (A3) and (A4) is due to the fact that the sample mean is not equal to zero though $E(\mathbf{x}) = 0$. Eq. (A5) is identical to (A4) except that $M < N$ i.e.

$$\widehat{\text{Var}}(\hat{a}_i^M) := \frac{1}{N} \sum_{k=1}^N \left(\hat{a}_i^M - \frac{1}{N} \sum_{k=1}^N \hat{a}_i^M \right)^2. \quad (A5)$$

Eq. (A5) yields EOFs of minor quality compared to the two other possibilities but estimates the variances better. This is demonstrated by Monte Carlo simulations with two 20-variate random vectors, $N = 15$ and $M = 7$ (as in the preceding text). The biases of the three different variance estimators (A3)–(A5) of the first five EOF coefficients (as in the text) (sample means of 100 Monte Carlo trials) are listed in the following two tables.

TABLE A1.

| i | λ_i | (A2)–(A3) | (A2)–(A4) | (A2)–(A5) |
|-----|-------------|-----------|-----------|-----------|
| 1 | 5.0 | -1.25 | -0.90 | -0.75 |
| 2 | 2.5 | -0.05 | -0.10 | -0.01 |
| 3 | 1.3 | 0.07 | 0.16 | 0.03 |
| 4 | 0.6 | -0.02 | 0.03 | 0.06 |
| 5 | 0.3 | -0.07 | -0.05 | -0.01 |

TABLE A2.

| i | λ_i | (A2)–(A3) | (A2)–(A4) | (A2)–(A5) |
|-----|-------------|-----------|-----------|-----------|
| 1 | 3.0 | -1.56 | -1.33 | -1.02 |
| 2 | 2.0 | -0.61 | -0.45 | -0.33 |
| 3 | 1.4 | -0.14 | -0.03 | -0.01 |
| 4 | 1.2 | 0.14 | 0.22 | 0.07 |
| 5 | 1.0 | 0.28 | 0.32 | 0.12 |

The results given in the two tables are stable, i.e., are essentially reproduced if different random samples are used. A negative (positive) sign means that a statistical test based on this estimate would be conservative (radical). As can be deduced from the tables, (A3) is the worst estimator and (A5) the best.

REFERENCES

Arakawa, A., and V. R. Lamb, 1977: Computational design of the basic dynamical processes of the UCLA general circulation model. *Methods Comput. Phys.*, **17**, 174–265.
 Blackmon, M. L., and N. C. Lau, 1980: Regional characteristics of the Northern Hemisphere wintertime circulation: A comparison of the GFDL general circulation model with observations. *J. Atmos. Sci.*, **37**, 497–514.
 Brutsaert, W., and F. K. F. Chang, 1978: Similarity functions D for water vapor in the unstable atmospheric boundary layer. *Bound. Layer Meteor.*, **14**, 441–456.

- Chervin, R. M., 1981: On the comparison of observed and GCM simulated climate ensembles. *J. Atmos. Sci.*, **38**, 885-901.
- Fischer, G., and H. von Storch, 1982: Klima = langjähriges Mittel?. *Meteor. Rundsch.*, **35**, 152-158.
- Gates, W. L., 1975: The January global climate simulated by a two-level general circulation model: A comparison with observation. *J. Atmos. Sci.*, **32**, 449-476.
- Hasselmann, K., 1979: On the signal-to-noise problem in atmospheric response studies, *Meteorology of Tropical Oceans*. Roy. Meteor. Soc., 251-259.
- Hayashi, Y., 1982: Confidence intervals of a climatic signal. *J. Atmos. Sci.*, **39**, 1895-1905.
- Kuo, H. L., 1965: On formation and intensification of tropical cyclones trough latent heat release by cumulus convection. *J. Atmos. Sci.*, **22**, 40-63.
- McGuirk, J. P., and E. R. Reiter, 1977: Non-random fluctuations in atmospheric energy parameters. *Beitr. Phys. Atmos.*, **50**, 239-246.
- Moeller, F., 1954: Ein Kurzverfahren zur Bestimmung der langwelligen Ausstrahlung dicker Atmosphärenschichten. *Arch. Meteor., Geophys. Bioklim.*, **A7**, 158-169.
- Roeckner, E., 1979: A hemispheric model for short range numerical weather prediction and general circulation studies. *Beitr. Phys. Atmos.*, **52**, 262-286.
- , and H. von Storch, 1980: On the efficiency of horizontal diffusion and numerical filtering in an Arakawa-type model. *Atmos.-Ocean*, **18**, 239-253.
- Schutz, C., and W. L. Gates, 1971: Global climatic data for surface, 800 mb, 400 mb: January, Rep. R-915-ARPA, The Rand Corporation, 173 pp.
- Storch, H. von, 1982a: A remark on Chervin/Schneider's algorithm to test significance of climate experiments with GCMs. *J. Atmos. Sci.*, **39**, 187-189.
- , 1982b: Comparison of a sequence of model generated 500 mb topographies with climate. *Tellus*, **34**, 89-91.
- Washington, W. M., and R. M. Chervin, 1980: Response time of an atmospheric general circulation model to changes in ocean surface temperature: Implications for interactive large-scale atmosphere and ocean models. *Tellus*, **32**, 119-132.
- Yamada, T., 1976: On the similarity functions *A*, *B* and *C* of the planetary boundary layer. *J. Atmos. Sci.*, **33**, 781-793.