Intercomparison of Extended-Range January Simulations with General Circulation Models: Statistical Assessment of Ensemble Properties

Hans von Storch and Erich Roeckner
Meteorologisches Institut der Universität Hamburg. Bundesstr. 55. 2000 Hamburg 13, FRG

Ulrich Cubasch
European Centre for Medium Range Weather Forecasts, Shinfield Park. Reading, Berkshire, U.K.

(Manuscript received 29.10.1984, in revised form 28.06.1985)

Abstract:

With the aid of multivariate test procedures, various 500 mb January geopotential height statistics over the Northern Hemisphere simulated by three general circulation models have been compared with the respective observations of the period1967–1983 analysed by "Deutscher Wetterdienst" (DWD).The simulations were done with two spectral models with different horizontal resolution (T2l, T40) of the "European Centre for Medium Range Weather Forecasts" (ECMWF) and with a 2.8° grid-point model of "Hamburg University" (HU). The T-models were integrated with inclusion of the seasonal cycle while the HU-integrations were performed in the perpetua1 January mode. All three models use climatologically prescribed sea surface temperatures and sea-ice limits.

Though the models are fairly different in most respects, the error structures are remarkably similar pointing to some common deficiencies like the improper parameterisation of land surface processes, for example.

The models behave differently with respect to the simulation of transient processes: The transient eddy variance is reproduced somewhat better by the high-resolution models (HU, T40), however, at the expense of the zonal wind simulation which seems to deteriorate with increasing resolution.

According to a multivariate statistical test, all three investigated model climatologies are significantly different at the 95 % level from the observed climate.

Zusammenfassung: Vergleich von Langzeit-Januar-Simulationen mit Zirkulationsmodellen: Statistische Abschätzung von Ensemble-Eigenschaften.

Mit Hilfe von multivariaten Testmethoden werden verschiedene 500 mb–Statistiken des Geopotentials für den Monat Januar auf der Nordhemisphäre, simuliert mit drei Zirkulationsmodellen, verglichen mit entsprechenden Beobachtungen für den Zeitraum 1967–1983, analysiert vom Deutschen Wetterdienst (DWD). Die Modellsimulationen wurden durchgeführt mit zwei spektralen Modellen des „Europäischen Zentrums für Mittelfristige Wettervorhersage" (EZMW) unterschiedlicher horizontaler Auflösung (T21, T40) sowie mit einem 2.8° -Gitterpunktmodell der „Universität Hamburg" (HU). Die spektralen Modelle wurden unter Einschluß des Jahrganges integriert, das Gittermodell im permanenten Januar-Modus. Alle Modelle verwenden klimatologische Meeresoberflächen-Temperaturen sowie Meereis-Verteilungen.

Obwohl die Modelle sich in vieler Hinsicht unterscheiden, sind die Fehlerstrukturen bemerkenswert ähnlich. Dies legt den Schluß nahe, daß die Modelle gemeinsame Fehlerquellen haben, wie z.B. eine ungenügende Berücksichtigung von Prozessen über Landoberflächen.

Unterschiedliches Verhalten zeigen die Modelle in Bezug auf die Simulation von transienten Prozessen. Die höher auflösenden Modelle (HU, T40) sind hier realistischer, jedoch auf Kosten des mittleren Zonalwindes, dessen Simulation sich mit verbesserter Auflösung zu verschlechtern scheint.

Das Ergebnis eines multivariaten statistischen Tests ist, daß sich alle drei untersuchten Modellklimatologien signifikant (95 %-Niveau) vom beobachteten Klima unterscheiden.

0005-8173/85/04 0477–21 $ 03.00/0

Résumé: Comparaison de la simulation à longue échéance du mois de janvier à l'aide de modèles de cir culation générale: Evaluation statistique des propriétés d'ensembles

On a comparé à l'aide de tests multivarés et avec les observations pertinentes de la période 1967–1983 analysées par le Deutscher Wetterdienst (DWD) l'analyse statistique du géopotentiel à 500 mbar pour janvier sur l'hémisphère nord simulé par trois modèles de circulation générale. Les simulations ont été réalisées avec deux modèles spectraux du Centre Européen pour les Prévisons à Moyen Terme, de résolutions horizontales différentes (T21, T40), et avec le modèle à points de grille distants de 2.8° de l'université de Hambourg (HU). Les modèles T comportaient un cycle saisonnier tandis que le modèle de Hambourg se situait en permanence en janvier. Les trois modèles utilisent les valeurs climatologiques des limites de l'interface océan-glace et des températures de la mer.

Quoique les modèles soient très différents sous bien des rapports, les structures des erreurs sont très semblables et indiquent des déficiences communes comme, par exemple, une paramétrisation inexacte des processus de surface.

Les modèles se comportent différemment quant à la simulation des processus transitoires: la variance des perturbations est un peu meilleure avec le modèle à haute résolution (HU, T40) au détriment toutefois du vent zonal qui semble se détériorer avec l'accroissement de résolution.

D'après le test statistique multivarié, les climatologies des trois modèles diffèrent significativement, au niveau de 95 %, du climat observé.

1 Introduction

In recent years the increasing computer power has made possible numerical experiments with general circulation models (GCM's) in ever longer time ranges. For example, perpetual January and July simulations with a nine-level low-resolution spectral GCM (rhomboidal truncation at wavenumber 15) have been carried out at NCAR (PITCHER et al., 1983; MALONE et al., 1984). A similar model was used at GFDL for the simulation of approximately 15 years including the seasonal cycle (MANABE and HAHN, 1981; LAU, 1981).

Extended-range simulations like those cited above allow to assess the quality of GCM's with more confidence than the traditionally performed integrations of just a few months. This applies not only to the time mean but also to the variability being extended to interannual time scales. As apparent from the studies of MANABE and HAHN (1981) and MALONE et al. (1984), the simulated low-frequency variability is by no means negligible and indeed comparable in magnitude to the observed one, even in models with climatologically prescribed surface conditions. Unless the time averaging interval is not extremely long, the low-frequency variability will noticeably affect the estimate of the time mean state (LEITH, 1973). Thus, the problem of comparing two climate ensembles (e.g. simulated and observed) becomes a statistical one.

In the present paper, similar to a proposal of PREISENDORFER and BARNETT (1983), we have extended the multivariate test procedure of STORCH and ROECKNER (1983a; "SR" in the following) in a way that not only individual (e.g. monthly mean) simulated states can be compared with respect to the observed ensemble, but also first and second moments of the simulated ensemble itself (STORCH and ROECKNER, 1983b). Though the method is not free from subjective elements (namely the choice of: significance level, considered physical parameters and the lowdimensional subspace, the data are projected on) it allows to decide in a fairly objective way whether the simulated and observed states differ significantly, thus being superior to the subjective side-by-side inspection of simulated and observed atmospheric fields.

In this study we use the above test procedure for assessing the quality of two GCM's which differ in nearly every respect (domain, resolution, integration technique, parameterization of physical processes)

so that fairly different test results should be expected. On the other hand, experience in numerical weather prediction has revealed that even rather dissimilar models produce similar error patterns (e.g. TEMPERTON, 1983). The question arises if such a behaviour is typical only for the prediction mode, being subject to initial conditions, or if it can be assigned also to the climate mode, being subject to surface boundary conditions. Of course, by comparing just two models we cannot hope to find proofs but indications at most for any of these alternatives.

In Section 2 we briefly summarize the basic features of the models. The design of the experiments and the evaluation technique is given in Sections 3 and 4, respectively. The results are presented in Sections 5-7, while Section 8 contains a summary of the results and our conclusions. Finally, the multivariate test procedure is described in detail in Appendix B.

2 The Models

The models used for this study are those developed at the European Centre for Medium Range Weather Forecasts (ECMWF) and at Hamburg University (HU), respectively. Their gross features are summarized in Figure 1. The structures of the models and their physical parameterizations are fairly different:

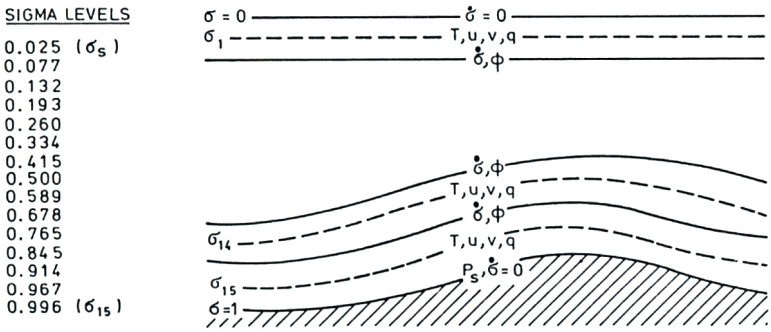
The ECMWF model is a global one with 15 layers including the stratosphere and with a spectral representation up to zonal wavenumber 21 isotropically (T21) in the low-resolution version and up to zonal wavenumber 40 (T40) in the high-resolution version (BAEDE et al., 1979). The physical parameterization packages are equivalent to those used in the operational model (TIEDTKE et al., 1979).

The HU model (ROECKNER, 1979) is a hemispheric one with 3 layers vertically excluding the stratosphere and with a regular latitude-longitude grid of 2.8125° which corresponds roughly to the Gaussian-grid resolution of the T40 model. The model is identical to that used by STORCH and ROECKNER (1983a), except for improvements in the PBL parameterization (cf. Appendix A) and for the use of more reliable surface boundary data being mostly interpolated from the same data base used in the ECMWF models (i.e. orography, sea surface temperature, sea ice limits, soil moisture, surface roughness length depending on subgrid-scale orography).

3 The Experiments

The Hamburg University model (HUM) was integrated for an extended January up to 300 days starting with initial conditions of January 2, 1974. The total range of 300 days has been subdivided into ten 30 day intervals each of which representing an individual January. The spin-up time has been assumed to be 30 days so that nine Januaries are left for evaluation.

The spectral models were integrated for a number of annual cycles. For that purpose, the sea surface temperature, the deep soil moisture, the deep soil temperature and the sea-ice have been adjusted to their climatological values every fourth day. The climatological values for the SST have been obtained by ALEXANDER and MOBLEY (1976), while the deep soil moisture and deep soil temperature have been interpolated to the T21 /T40 model grid from the values used in the operational ECMWF forecasting model (LOUIS, 1984). The initial data were obtained by means of a spin-up experiment starting with the FGGE analysis of November 11, 1979 and run until December 31, 1979. This day represents the start of the 10 (2)-year integration with the T21 (T40) model. The elapsed time of 6 weeks was thought to be long enough to erase any information which might be contained in the initial field and to let the model find its own climate (CUBASCH, 1981).



Vertical dispositions of variables in the ECMWF models
Vertical coordinate: $\sigma = p/p_s$

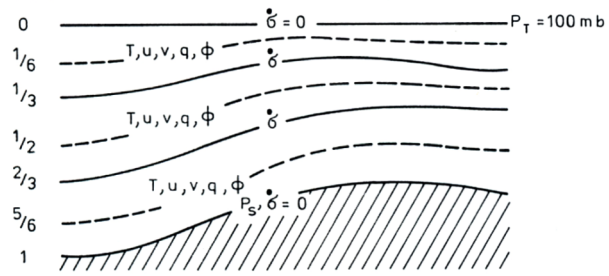
Independent variables	λ, φ, σ, t
Dependent variables	T, u, v, q, p_s
Grid	Gaussian, resolution: T21 max. resolved zonal wave nr. 21, T40: max. resolved zonal wave nr. 40: non-uniform spacing of vertical levels.
Finite difference scheme	Second order accuracy.
Time-integration	Leapfrog, semi-implicit (Δt: T21–2400 sec; T40–1600 sec)
Horizontal diffusion	Linear, fourth order (diffusion coefficient: T21 = 2.10^{15} m^4/s T40 = 7.10^{14} m^4/s)
Earth surface	Albedo, roughness, soil moisture, snow and ice specified geographically. Albedo, soil moisture and snow time dependent.
Orography	Averaged from high resolution data set.
Vertical boundary conditions	$\dot{\sigma} = 0$ at $p = p_s$ and $p = 0$.
Physical parameterisation	(i) Boundary eddy fluxes dependent on roughness length and local stability (Monin Obukov)
	(ii) Free-atmosphere turbulent fluxes dependent on mixing length and Richardson number
	(iii) Kuo convection scheme
	(iv) Full interaction between radiation and clouds
	(v) Full hydrological cycle
	(vi) Computed land temperature, no diurnal cycle
	(vii) Climatological sea-surface temperature adjusted every 4th day.

● Figure 1a Characteristics of the ECMWF model

4 Evaluation Technique

The results presented in the next Sections have been obtained with different evaluation techniques.

In order to get a first insight into the performance of the models, we compare long-term averages of the simulated and observed January mean flow in terms of a few representative variables like 500 mb height, 850 mb temperature and 300 mb wind (Section 5). The observed height and wind field (referred to as NCAR climatology) have been made available by CRUTCHER and JENNE (1970). The temperature field has been compared with a climate derived from DWD data.



Vertical disposition of variables in the HU model

Vertical coordinate: $\sigma = \frac{p - p_T}{p_s - p_T}$

Independent variables	$\lambda, \varphi, \sigma, t$
Dependent variables	$\pi T, \pi u, \pi v, \pi q, \pi = p_s - p_T$
Grid	Regular latitude-longitude grid with $\Delta\lambda = \Delta\varphi = 2.8125°$ Uniform spacing of vertical levels.
Finite difference	Second order accuracy. Energy- and quasienstrophy conservation for non-divergent vorticity advection on B-grid.
Horizontal diffusion	Non-linear, fourth order for θ, q; high order numerical filter for u, v and SLP
Earth surface	SST, roughness length, soil moisture specified geographically
Orography	Averaged from high resolution data set
Vertical boundary conditions	$\dot{\sigma} = 0$ at $p = p_s$ and at $p_T = 100$ mb
Horizontal boundary conditions	Symmetry with respect to the Equator
Physical parameterizations	(i) Surface boundary fluxes calculated from similarity theory for the whole PBL.
	(ii) Free-atmosphere vertical turbulent fluxes calculated from K-theory with constant Austausch coefficient.
	(iii) Kuo convection scheme
	(iv) Radiative fluxes calculated from empirical relations. No explicit cloud treatment in the radiation scheme.
	(v) Full hydrological cycle.
	(vi) Computed land temperature, no diurnal cycle.

● Figure 1b Characteristics of the Hamburg University grid-point model (HUM).

Secondly in Section 6, a multivariate test [The necessity to use multivariate techniques instead of the widely used univariate t-test approach was discussed in detail by SR.] is applied to a set of 500 mb-height statistics allowing to objectively decide if some simulated climate state is significantly different from the respective observed one or not. The testing is done

– for each simulated individual January sample
– for the mean of the January states ("multiyear mean") and
– for the standard deviation of the January states ("interannual variability").

Details of the method are given in Appendix B. The following 500 mb height statistics have been tested:

a) $[\bar{z}]$ "zonal mean height minus hemispheric mean"

b) $[\bar{z}^{*2}]\cos\varphi$ and $\{\bar{z}^{*}\}$ "quasi-stationary eddies"

c) $[\overline{z'^{*2}}]\cos\varphi$ and $\{\overline{z'^{2}}\}$ "transient eddies"

d) Wavenumber-frequency spectra along 50 °N estimated with the aid of the maximum-entropy method (cf., STORCH and FISCHER, 1983).

The following mostly standard notations have been used:

z	height of the 500 mb surface minus hemispheric mean
—, '	time average, deviation
[ ],*	zonal average, deviation
{}	meridional average between 30° and 60°N.

Finally in Section 7, the above statistics are examined in a univariate way in order to quantify and localize differences between the simulated and observed climate states.

The observational basis for the statistical tests are daily 500 mb geopotential height analyses of Januaries 1967-1983 produced by the "Deutscher Wetterdienst" (DWD) on a 381-km stereogrpahic grid north of 15 °N.

5 The Time-Mean State

5.1 The 500 mb Height Field

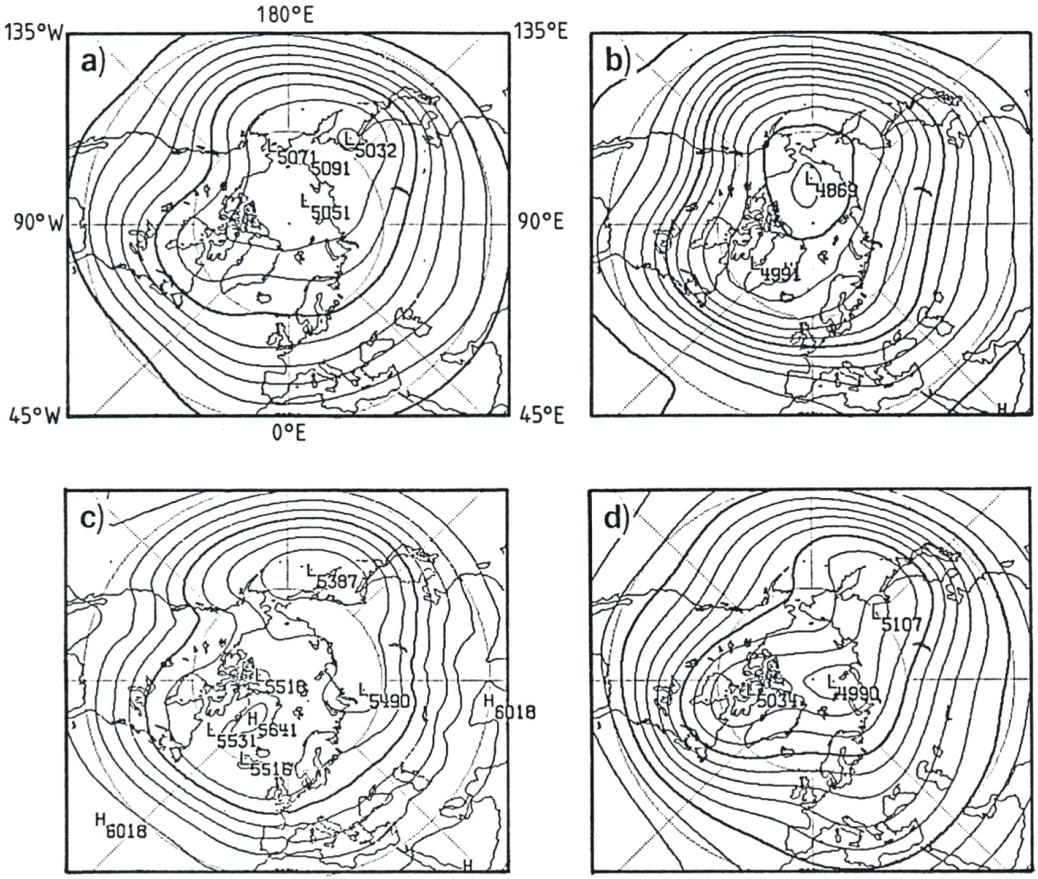
In Figure 2 the geographical distributions of the simulated 500 mb height fields (a-c) are shown together with the respective observations (d). A number of model deficiencies that will be discussed in Sections 6 and 7 in a statistical context may already be inferred from Figure 2: All 3 models tend to broaden the Pacific and especially the Atlantic trough too far to the east so that the circulation becomes too zonal over large parts of the oceans. This tendency can also be found in the ECMWF operational model (ARPE, 1983). Moreover, in the T40 model the meridional height gradient at mid-latitudes is clearly too strong while it is too weak at high latitudes in the HU model. The systematic raise of the height in the latter is due to the choice of the finite difference form of the hydrostatic equation being different from that used in the other models. Therefore, when discussing mean height profiles in Section 6 and 7, the hemispheric mean of z will always be removed.

By and large, the low resolution T21 model seems to be closer to the observed 500 mb height field than the other models.

5.2 The 850 mb Temperature Field

In Figure 3 we compare the Northern Hemisphere distribution of the January 850 mb temperature error of the 3 models (a-c). The observed temperature distribution is shown in Figure 3d. The error structure for the 3 models is remarkably similar with a tendency of warming over the continents. The largest errors up to 20 °C over the Himalaya are produced by the HU model.

Over the oceans the situation is different. While the T-models reveal a general cooling of up to –5 °C, the HUM errors are smaller and mostly positive. The rather cold PBL in the T-models points to an insufficient vertical heat transfer from the ocean surface. The reasonable simulation in the HU-model is related to the assumption of a countergradient vertical heat flux which has to be introduced into

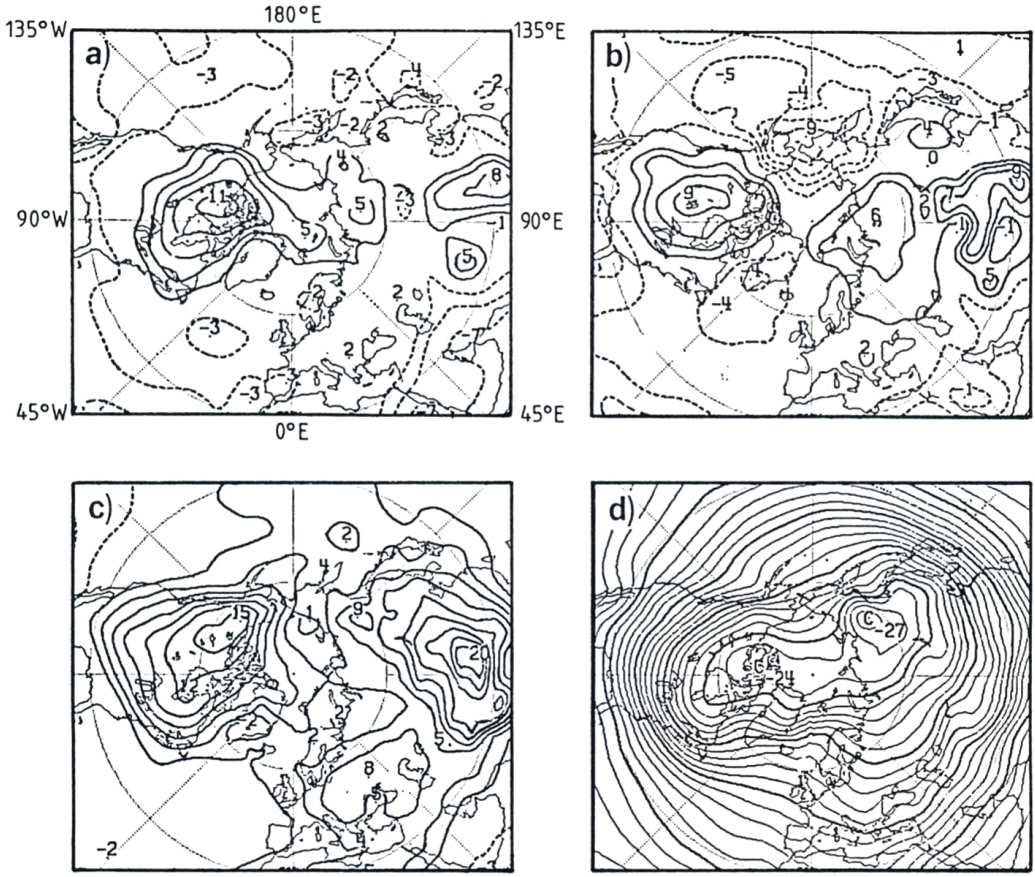


● Figure 2 The time-mean 500 mb height field in January for a) T21, b) T40, c) HUM and d) the observed climate (source: NCAR). Contour spacing: 8 dam

a model with a coarse vertical resolution in order to approximately maintain the observed static stability in the PBL (cf. Appendix A).

The reason for the systematic warming of the air over the continents is unknown. A number of model defects may have produced these errors, e.g. insufficient radiative cooling (this is known for HUM), or insufficient heat exchange between the air and the surface, or horizontal heat fluxes by East Pacific depressions moving wrongly too far inland.

The dynamical implications of the continental warming during winter are that the diminshed zonal temperature gradient affects the quasi-stationary eddies by a reduction of the diabatic heat source and also the transient eddies by a reduction of baroclinicity over the western parts of the oceans mainly.

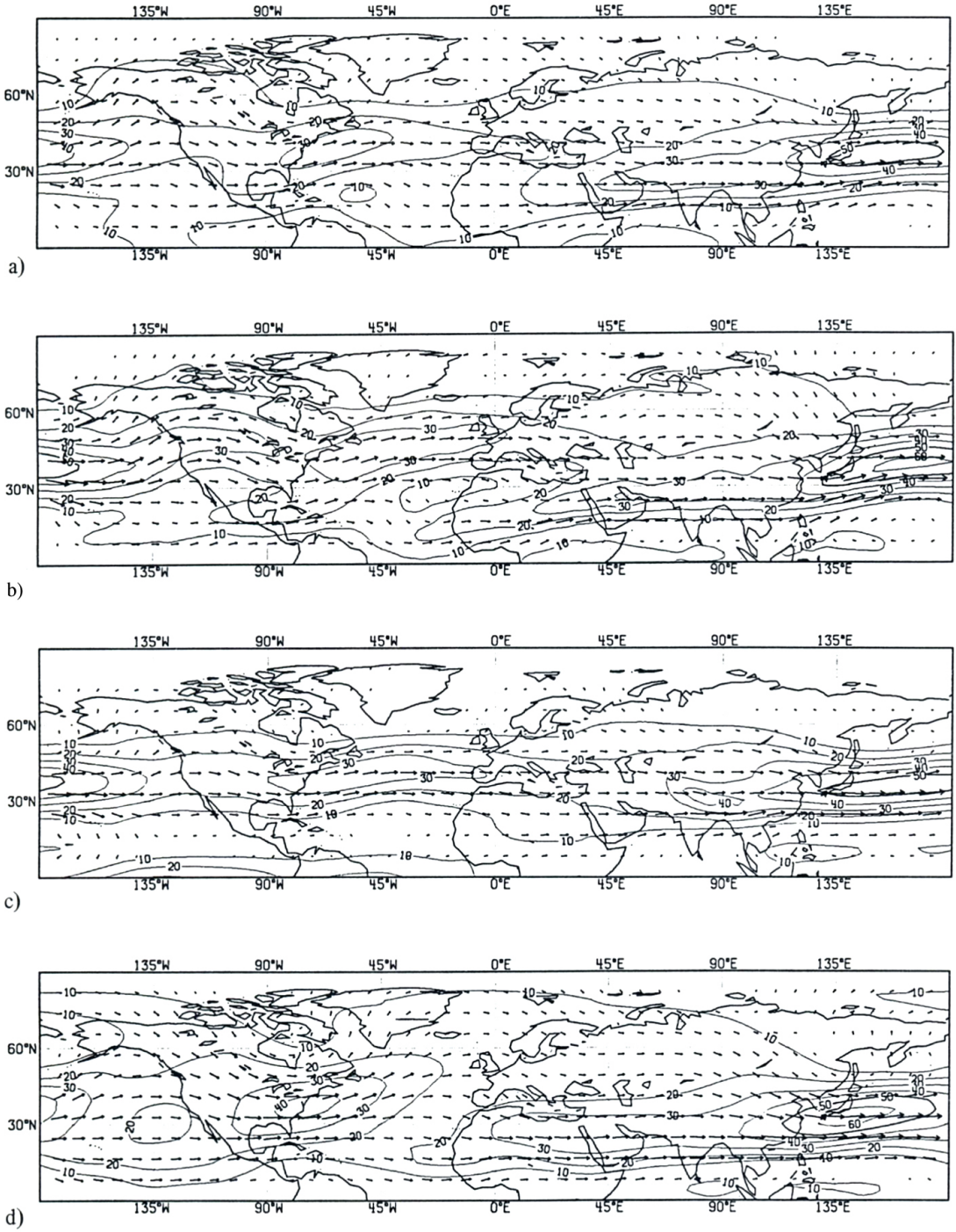


● Figure 3 The time-mean 850 mb temperature error fields in January for a) T21, b) T40, c) HUM. d) represents the observed climate (source: DWD). Contour spacing: 2 °C

5.3 The 300 mb Wind Field

As one might expect from the pressure gradients (Figure 2) the wind is weaker in T21 than in T40 (Figure 4a, b). The HU model (Figure 4c) produces a confined and narrow windband around 40 °N with hardly any north-south structure. This might well be connected with the artificial wall at the equator: While a split of the subtropical jet over the Pacific is bent polewards over Central America thus increasing the jet in the West Atlantic in the global models, it clinges to the equator in the hemispheric HUM integration and creates an abnormal westerly flow over Brazil.

All models simulate the increase of the windspeed at the east side of the continents, but they all underestimate its strenght, The T40 model seems to produce the most realistic flow pattern.



● Figure 4 The time-mean 300 mb wind field in January for a) T21, b) T40, c) HUM. d) the observed climate (source: NCAR). Contour spacing: 10 m/s

6 Multivariate Tests

With the aid of the statistical tests described in detail in SR and Appendix B, we examine if the simulated 500 mb-height statistics (cf. Section 4) are significantly different from the respective observed quantities.

The tests are performed for:

(1) Each of the January states generated by the models (t-statistic according to (2.4) of SR).

(2) The longyear (= interannual) January mean (generalized Mann-Whitney test, see Appendix B).

(3) The interannual variability (generalized Siegel-Tukey test, see Appendix B).

Principally, the data have to be statistically independent for the application of the tests (2) and (3). In case of HUM, this assumption does not hold. However, we found essentially unchanged results when we subdivided the total range of the HUM integration such that each individual January is separated by a 10d interval from the next. Thus, the violation of the independence is seemingly unimportant.

Due to the small sample size of 2, the T40-results could only be tested by the first procedure. The results obtained by applying (1)–(3) with a significance level of 95 % are summarized in Tables 1–3.

Table 1 contains the number of rejections of the nullhypothesis "the simulated variable cannot be distinguished from the ensemble of the respective observed quantity" obtained by (1) for each simulated January. Apparently, all 3 models exhibit problems in simulating amplitude and/or phase of the mid-latitude stationary eddies (i.e. $\{\bar{z}^*\}$ correctly: The rejection rate is 100 % for T40 and HUM and 60 % for T21. Similarly, all $[\bar{z}]$-samples of T40 and HUM are rejected, pointing to a relationship between the mean zonal flow and the stationary eddies. As a whole, the rejection rate is smallest for the low resolution model T21.

In case of the interannual mean (Table 2), the nullhypothesis "the expectation vectors of the observed and simulated ensembles are identical" is rejected for all quantities, except for the HUM wavenumber frequency spectrum. This means that the models T21 and HUM generate a long term mean January climate significantly different from the observed climate with respect to nearly all considered parameters.

■Table 1 Number of rejections by the t-test for zonal averages of variance of transient eddies (column 2), of stationary eddies (column 3), of monthly mean (column 4) and meridional average (30 °–60 °N) of monthly mean (column 5), of transient eddies (column 6) and the frequency wavenumber spectrum at 50 °N (column 7).

model	[z'*2Icos'f'	[z*2l cos 'f'	[z]	{z*}	{z'2}	r	total sampie size
ECMWF							
T21	1	3	2	6	3	0	10
T40	1	0	2	2	0	0	2
HUM	3	6	9	9	2	1	9

■Table 2 Result of the generalized Mann-Whitney test (B 3) applied to zonal averages of variance of transient eddies (column 2), of stationary eddies (column 3), of monthly mean (column 4) and meridional average (30°–60 °N) of monthly mean (column 5), of transient eddies (column 6) and the frequency-wavenumber spectrum at 50 °N (column 7). A significant difference "simulated – observed data "is marked by a star; an insignificant one with a bar.

model	[z'*2) cos'f'	[z*2) cos φp	[z)	[z*}	{z'2}	r
ECMWP						
T21	*	*	*	*	*	*
HUM	*	*	*	*	*	⌴ ⌴

■Table 3 Result of the generalized Siegel-Tukey test (B4) applied to zonal monthly mean (column 2), to the zonally averaged variance of stationary eddies (column 3) and to the meridional average (30 °–60 °N) of monthly mean (column 4). A significant difference "simulated – observed data" is marked by a star, an insignificant one with a bar.

mode!	[zj	[z*2j cos'f'	[z*}
ECMWF			
T21	–	*	*
HUM	–	*	–

Since all experiments were run with climatological bounditions conditions such as SST, sea-ice extent, deep soil moisture and deep soil temperature, it is reasonable to expect a-priori an underestimation of the models interannual variability which is defined to be the variability of monthly means. Therefore, we use the one-sided nullhypothesis "the variance of the simulated ensemble is not smaller than that of the respective observed quantity", i.e. the anticipated result coincides just with the alternative hypothesis of our test. Since it is difficult to interprete 2nd moments of parameters measuring transients (i.e. interannual variability of mean daily variability), we restrict ourselves to the discussion of the quasistationary parameters $[\bar{z}]$, $[\bar{z}^{*2}]\cos\varphi$ and $\{\bar{z}^*\}$. The result of the testing with method (3) is given in Table 3: The decreased interannual variability is insignificant for [Z], but significant in both experiments, HUM and T21, for [Z*2] COSlp and for (Z*} of T21.

After having objectively found a number of significant model deviations from the real atmosphere, a detailed discussion of the features that might have caused these deviations will be given in the next Section. We often do this by means of an "univariate analysis" (cf. SR), i.e. by comparing the set of simulated January states with the respective observed 95 % band which is defined to be the set of univariately estimated intervals containing about 95 % of all observed states. This representation is independent of the performance of the multivariate test. Additionally, this a-posteriori analysis yields arguments of plausibility and stability: If, for example, a number of individual states leave the band at the same location, this may be seen as a hint that the model's climatology is incorrect at that particular location at least.

7 Discussion of 500 mb Height Statistics

7.1 Zonal Mean Geopotential Height

Figure 5a shows the latitudinal distribution of $[\bar{z}]$ for the ensemble means of T21, HUM and observations, respectively. In Figure 5b both T40 samples are compared with the observed univariately established "95 %-band" containing roughly 95 % of all available observed states.

The reasons for the rejection of the nullhypothesis (Tables 1 and 2) are different: The rejection of the mean HUM-profile is caused by an underestimated height at low latitudes and a severely overestimated height at high latitudes. The slope at mid-latitudes, i.e. the mean zonal geostrophic wind, seems to be correct. On the other hand, both T40-profiles are too steep at mid-latitudes which is also apparent from Figures 2b, d. Moreover, in contrast to T21 and HUM, there is a tendency of underestimating the heigh at high latitudes.

The smallest errors are clearly produced by the low resolution model T21 with a rejection rate of only 20 % (Table 1). Nevertheless, the interannual mean is rejected in the respective test (Table 2), like that of HUM. The fact that the mid-latitudinal westerlies in Northern Hemisphere winter become stronger

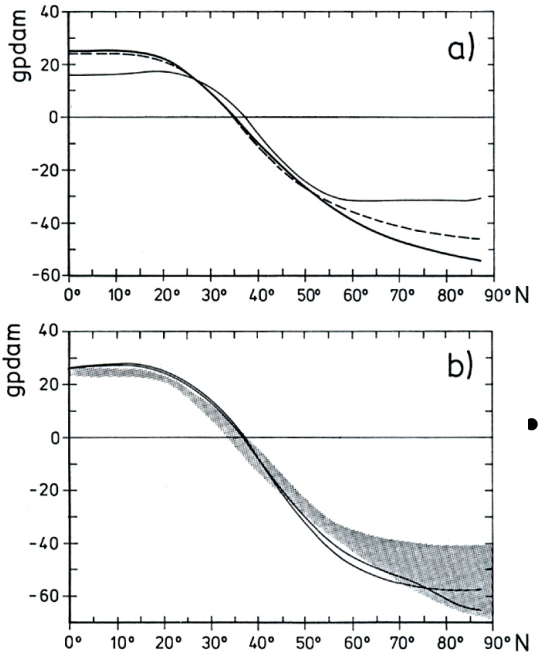


Figure 5

Latitudinal profiles of zonally averaged January 500 mb height. [$\bar{z}$], centered at the hemispheric mean.

a) Multiyear mean of T21 (dashed), HUM (thin) and DWD analyses 19671983 (solid).

b) 95 %-band based on DWD analyses 19671983 together with two T40 samples (30 d averages).

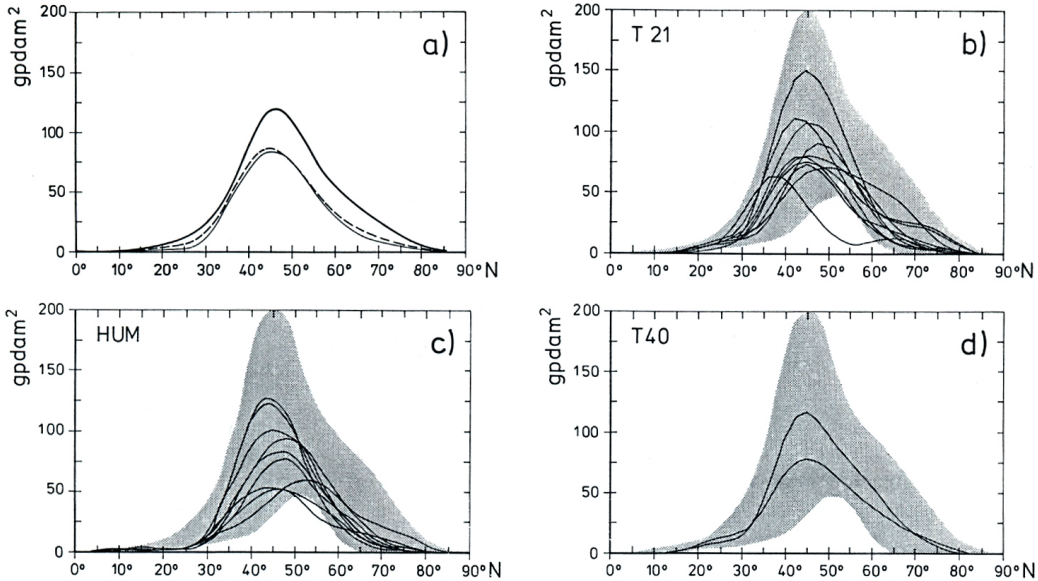
and more unrealistic with increasing horizontal resolution was already noted by MANABE et al. (1979) when comparing the climates of the spectral GFDL model with varying wavenumber truncation.

7.2 Quasi-Stationary Eddies

Figure 6 shows the latitudinal distribution of [$\bar{z}^{*2}$] cos φ for the ensemble means (a) and for the individual samples together with the observed 95 %-band (b–d). Though most of the individual profiles fit nicely into the observed band, the mean and also the interannual variability of both, T21 and HUM, differ significantly from the respective observations (Tables 2, 3 and Figure 8a). Obviously, the T21 and HUM models are close together.

As noted above, all models discussed in the present paper have problems in simulating amplitudes and phases of the stationary waves correctly (Tables 1 and 2, column 5). This is documented in Figure 7 showing the longitudinal distribution of { $\bar{z}^*$}. Analogous to Figure 6, Figure 7 shows the mean distributions (a) and the individual samples of the three models (b–d). The dashed region again represents the observed 95 %-band. While the amplitude of the Pacific trough-ridge system is simulated satisfactorily by all 3 models the phases in HUM and T40 are shifted by some 30° to the east. Moreover, in the T-models there is a tendency of developing a trough over the eastern Pacific. In the Atlantic the situation is different: Here, the amplitude errors dominate which is striking for the East Atlantic ridge being not well represented, especially in HUM.

AccordingtoTable3, the interannual variability of { $\bar{z}^*$} is significantly underestimated for T21 but not for HUM. The difference between both simulations is, however, noticeable only in the East Pacific between 180° and 120 °W (Figures 7b, c and 8b). In fact, the actual level of significance is just below 95 % for HUM and a little above this threshold value for T21.



● Figure 6 Latitudinal profiles of zonally averaged January 500 mb height stationary eddy variance. $[\overline{z}^{*2}] \cos \varphi$.

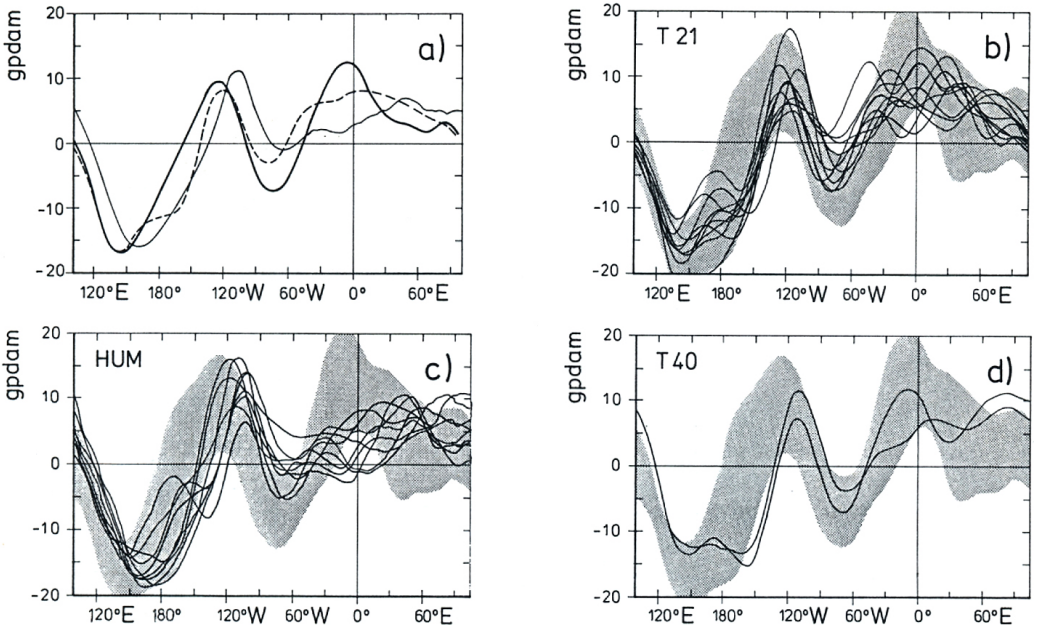
a) Multiyear mean of T21(dashed), HUM (thin) and DWD analyses 1967–1983 (solid).

b) –d) 95-bandbased on DWD analyses 1967–1983 together with the individual January samples of T21, T40 and HUM, respectively.

7.3 Transient Eddies

As compared to the stationary eddies, the rate of rejection of the transient eddy profiles is relatively small (Table 1). On the other hand, the ensemble mean profiles are rejected without exception (Table2). The reason is obvious from Figure 9 showing the mean profiles (a) together with the individual samples simulated by the 3 models (b–d). For example, 90 % of all T2l-profiles are not rejected according to the statistical test (Table1, column2), however, all profiles are located at the lower boundary of the observed 95%-band (Figure 9b) resulting in a significant reduction of the ensemble mean (Figure 9a). Both high resolution models, HUM and T40, tend to produce larger transient eddy variance than the low resolution model T21, but still significantly less than the real atmosphere. All models tend to shift the maximum of transient activity by some 15° southwards. Thus, the errors become largest at about 60 °N which is probably due to the reduced baroclinicity at the east coasts of North America and Asia (cf. Section S.2).

Similar conclusions may be drawn from Figure 10 which shows the longitudinal distribution of transient eddy variance at mid-latitudes. Analogous to the stationary eddies (Figure 7) the transient eddies are underestimated especially in the North Atlantic sector. This deficiency of the T2l-model was already found by VOLMER et al. (1984) who compared the EOF structure of the ensembles of observed and simulated daily 500 mb-height fields.



● Figure 7 Zonal profiles of latitudinally averaged (300° –60 °N) January 500 mb height stationary eddies, $\{\bar{z}^*\}$.

a) Multiyear mean of T21 (dashed), HUM (thin) and DWD analyses (solid).

b) –d) 95 % band based on DWD analyses 1967–83 together with the individual January samples of T21, T40 and HUM, respectively

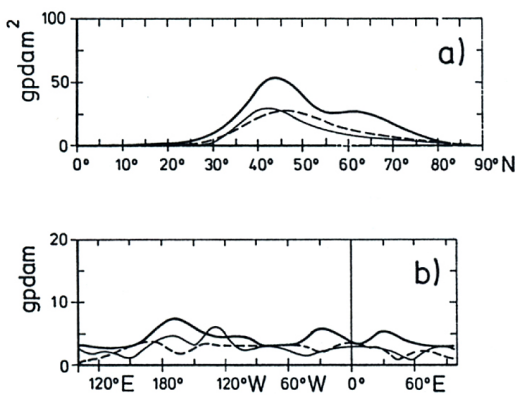


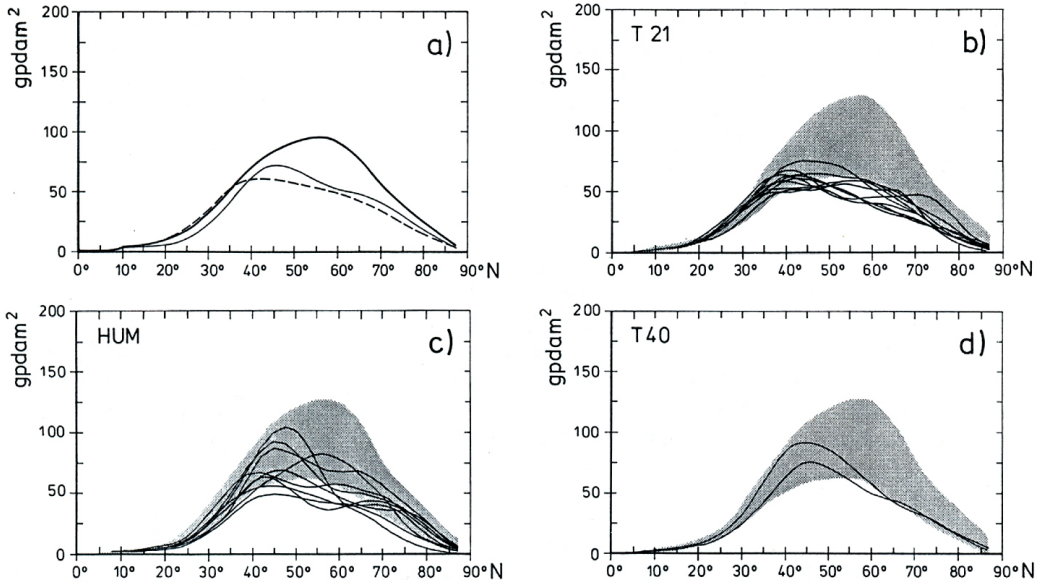
Figure 8

Interannual variability of January 500 mb height stationary eddies measured in terms of standard deviation of

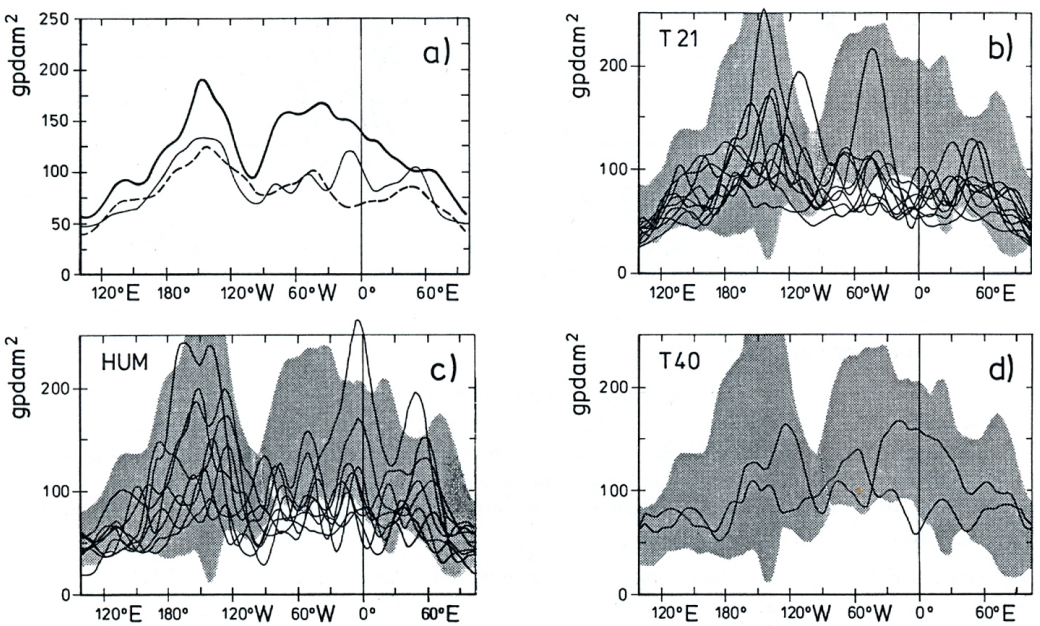
zonally averaged $[\bar{z}^{*2}]\cos\varphi$

latitudinally averaged (30° –60 °N)$\{\bar{z}^*\}$

for T21(dashed), HUM(thin) and DWD analyses (solid)



● Figure 9 As Figure 6, except for transient eddy variance, $\left[\overline{z'^{*2}}\right]$ cos φ

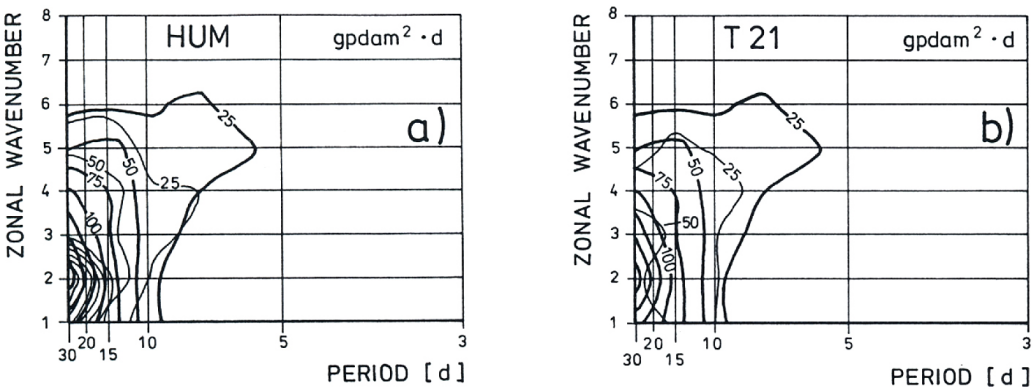


● Figure 10 As Figure 7, except for transient eddy variance, $\left\{\overline{z'^{2}}\right\}$

An overall reduction of the daily variability in the T21-model was already found by CAO (1984) who studied the T21 and HUM daily northern hemisphere January 500 mb-height fields with respect to the occurrence of climatologically rare events. With the aid of a non-parametric statistical method, CAO found 5.3 % of all daily HUM fields rare at the 95 % level. Since this is close to expectation, it points to an acceptable daily variability of the model. For T21, however, the rate of rare events is zero (CAO, pers. comm.), indicating an insufficiently developed transient activity in the T21-model.

7.4 Wavenumber-Frequency Spectra

According to Table 1 (column 7), only 1 of 21 spectra totally are rejected. However, the ensemble mean of T21 was found to be significantly different from the respective observations (Table 2). The reason is apparent from Figure11b: The mean T21 variance is smaller than the observed one everywhere in the space-time domain accounting, for example, only for 50 % of the maximum observed variance. The ensemble mean HUM spectrum is not rejected, revealing larger variance than T2l, especially in the low frequency–low wavenumber part where the observed and simulated intensities are close together.



● Figure 11 Multiyear mean of January 500 mb height wavenumber-frequency spectra at 50 °N for a) HUM and b) T21. The solid lines represent observed spectra calculated from DWD analyses 1967–1983.

8 Summary and Conclusions

In the present study we considered the ability of three different GCM store produce the January climatology of the Northern Hemisphere in terms of some statistics of the daily 500 mb-height. Although two of the models (T21 and HUM) differ in many respects (spectral versus grid-point representation; low versus high horizontal resolution and vice versa vertically; global versus hemispheric domain; seasonal cycle integration versus perpetual January mode; different physical parameterizations), the main error characteristics are remarkably similar, e.g.

— Overestimation of mean height at high latitudes (more pronounced in HUM than in T21).

— Underestimation of stationary eddy variance everywhere.

— Eastward shift of mid-latitude stationary waves in the Pacific/North America sector and too small amplitudes in the Atlantic/Europe sector.

— Underestimation of transient eddy variance especially over the North Atlantic (more pronounced in T21) with a southward shift of the maximum from 60 ~ (observed) to 45 ~ (simulated).

— Underestimation of interannual stationary eddy variability.

The most apparent differences between the models (HUM, T21, T40) are:

— The mean zonal geostrophic wind is underestimated at high latitudes in T2l (slightly) and HUM (strikingly) but overestimated, especially at mid-latitudes, in the T40 simulations.

— The transient eddy variance is smallest in the low resolution model T21. This is true not only for the respective zonal average but also for the wavenumber-frequency spectra at 50 °N in the whole space-time domain.

From the results presented above, we draw the following conclusions:

(l) In terms of various 500 mb-geopotential height statistics, the three investigated model climatologies are significantly different from the respective observed Northern Hemisphere January climatology. With a few exceptions, the deviations point into the same direction, i.e. underestimation of variance of transient eddies (timescale: 1 day–1 month) and of stationary eddies as well. These deficiencies are probably connected with the spurious lower-troposphere heating over the continents produced by all 3 models. Since the structure of the models is remarkably different, the main error sources are probably due to the incomplete model physics and/or inaccurate surface boundary forcing being similar in all models. A comparison with previously published results (e.g., MANABE and HAHN, 1981; MALONE et al., 1984) is difficult because in those studies the differences between simulations and observations were discussed only subjectively, mainly in terms of geographical distributions of the respective quantities. There are, however, indications that the NCAR model (MALONE et al., 1984) simulates the transient and stationary eddy variances better than the models discussed in this study.

(2) With respect to most parameters the model climatologies are closer together than any of the considered model climatologies and the observed one. An analogous result was already found for the prediction mode (i.e. for integrations of a few days; e.g. TEMPERTON, 1983) but has not been explicitly established for the climate mode (i.e. integration time of several months or years).

(3) The increase of horizontal resolution improves the simulation of transient processes, but the reverse is true for the simulation of the mean zonal flow in the T-models (T21, T40).

(4) The "seasonal cycle model" (T21) does not reveal a larger interannual variability than the "perpetual Janury model" (HUM). Though both models differ also in other respects it seems unlikely that the inclusion of the seasonal cycle will considerably increase the interannual variability. A corresponding conclusion was drawn by MALONE et al. (l984) when comparing the interannual variabilities produced by the NCAR and GFDL model, respectively. The reduction of interannual variability of the stationary eddies may be understood from the fact that the surface boundary conditions are partly prescribed from climatological conditions (e.g. MANABE and HAHN, 1981).

Appendix A

Modification of the Bulk Stability Parameter in the PBL

Over low-latitude oceans the virtual potential temperature quite often shows a vertical distribution similar to that sketched in Figure 12: an unstable surface layer, a well-mixed neutral layer and stable stratification above the top h of the mixed layer, typically at 500 m.

In the 3-layer model the lowest model level at H≈ 1500 m is found generally above the top of the mixed layer. Thus, the basic assumption H ≈ h used in the PBL parameterization (Ekman similarity theory according to VAMADA, 1976) is violated. In Figure 12, the bulk stability parameter measured in the model $\Delta\theta = \theta_2 - \theta_0$ is positive and causes a downward directed surface heat flux and thus an energy loss of the lowest layer. In a model integration this situation would lead to a permanent cooling (at least at low latitudes) until $\theta_2 < \theta_0$. Obviously, the correct procedure for calculating the energy input from the surface is the use of $\Delta\theta^* = \theta_1 - \theta_0$ leading to an upward directed surface heat flux and hence to an energy gain of the PBL. The problem is how to estimate $\Delta\theta^*$in a model with a coarse vertical resolution. For the present study we use a "countergradient" approach according to

$$\Delta\theta^* = \Delta\theta - \gamma\,(H - h) \qquad \text{for } h < H \tag{AI}$$

with $\gamma = \partial\theta/\partial z = 0.004$ K/m which is a value typically measured above the top of the mixed layer over low-latitude oceans (e.g. FITZJARRALD and GARSTANG, 1981, Figures 8a and 10). The mixed-layer height h has been estimated consistently with the length scale used in the surface flux parameterization (VAMADA, 1976):

$$h = 0.3\,|v_*| / |f| \tag{A2}$$

where v_* is the friction velocity and f the Coriolisparameter which is not allowed to drop below a value of $f = 5\ 10^{-5}\ s^{-1}$. Furthermore, we assume h = Max (h, 300 m). The introduction of γ in (Al) has some formal resemblance to the countergradient derived by DEARDORFF (l966) to account for an upward directed heat flux being measured during slightly stable conditions. However, while DEARDORFF has shown that a countergradient heat flux can be understood from physical arguments, the countergradient in (A1) is motivated only by the coarse vertical resolution of the GCM. The countergradient heat flux within the mixed-layer has been neglected because the (A1)-correction is nearly an order of magnitude larger than the countergradient proposed by DEARDORFF.

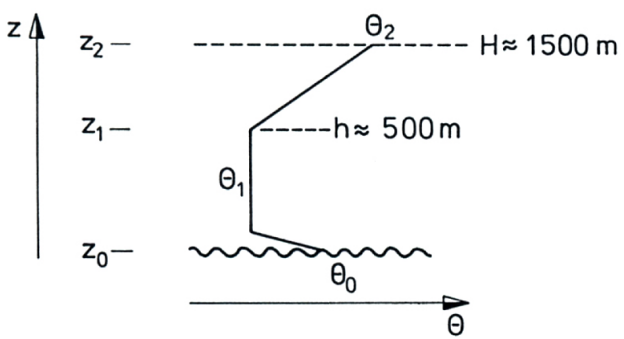


Figure 12
Schematic vertical distribution of virtual potential temperature 6 over low-latitude oceans.

Appendix B

A Class of Nonparametric Two-Sample Tests

Our procedure is build up by the elements "projection onto an appropriate subspace = EOF expansion", "multivariate test" and "univariate analysis". This frame is identical to that used by SR. Because the element "multivariate test" is different, we describe it in the following in some detail.

The test is a generalization of a permutation technique originally proposed by R. A. FISCHER and further developed by BOYETT and SCHUSTER (1977). A similar approach was presented by PREISENDORFER and BARNETT (l983).

a) Univariate approach

Let x and y be two univariate random variables (r.v.) and $x_1, \ldots x_n$ and $y_1, \ldots y_m$ mutually independent samples of x and y, respectively. z denotes the (n + m)-dimensional vector composed by all samples of x and y:

$$z=(x_1, x_2, \ldots x_n, y_1, y_2, \ldots y_m)$$

We want to compare a parameter P of the two r.v., e.g. expectation P = E or variance P = Var. In this paper, y denotes observed quantities and x GCM simulated numbers.

There are two pairs of hypotheses which can be used (N denotes the nullhypothesis, A the alternative), namely either

"one sided": N: $P(x \geq P(y)$ against A: $P(x) < P(y)$

or

"two sided": N: $P(x) = P(y)$ against A: $P(x) \neq P(y)$

We use the two-sided test for the comparison of means. If a GCM is integrated with climatologically fixed boundary conditions, a reduction of its low-frequent variability may be expected a-priori. Thus, the usage of the one-sided test is appropriate, if GCM generated low-frequent variances are compared with observed ones.

For the test, we need a function t (z) taking small values if N holds and large values otherwise. Strictly spoken, t has to fulfil the following condition:

(*) If S is the set of all permutations σ of {1, ..., n + m} and if N is true, then the r.v. $t(z_\sigma) := t(Z_{\sigma(1)}, \ldots Z_{\sigma(n+m)})$ is distributed as t(z) itself for all permutations $\sigma \varepsilon S$.

The decision of the test is based on the comparison of the number t (z), where z is the combined sample vector with orginal ordering, and all numbers t (z_σ). For that, think of a arbitrary (n + m)-dimensional combined vector w which fulfils the nullhypothesis N. The actual ranking of the components of w is irrelevant for the distribution of t (w) according to (*). Consequently the probability that the actual ranking of w yields a t (w) larger than 95 % of t (w_σ) obtained for all permutation $\sigma \varepsilon S$ is just 5 %.

Thus, an acceptance of the alternative A is associated with a risk of less than 5 %, if the actual ranking's t (z) is larger than 95 % of t (z_σ) with $\sigma \varepsilon S$. If i | denotes the magnitude of a set, i.e. the number of its elements, an appropriate test statistic is

$$H := \frac{\left|\{\sigma \varepsilon S;\ t(z_\sigma) > t(z)\}\right|}{|S|} \qquad \text{(B1)}$$

and N is rejected (A is accepted) at the 95 % level if H <5 %.

If the sample sizes n and m are large, it may become impossible to compute t (z_σ) for all permutations. To overcome this difficulty, one may restrict oneself to a randomly formed subset of S. We found subsets of 2000 permutations to be sufficient.

b) Multivariate generalization

If instead of a univariate quantity a p-dimensional vector is studied, a multivariate test statistic T may be defined by means of the univariate test statistics t for each of the components, j= 1, ... p, as follows:

$$T := \sum_{j=1}^{p} t^{(j)} \qquad \text{or} \qquad T := \sum_{j=1}^{p} \left| t^{(j)} \right| \tag{B2}$$

The superscript (j) is mean t as an index and not as an exponent. In contrast to the components of p-dimensional vetor, samples of random variables are numbered by subscripts. The same procedure as outlined above is applied to T instead of t. (If (*) holds for t, then it is also true for T.). The version using absolute values is suitable for two-sided tests, the other for one-sided tests.

c) Choice of t

If the coincidence of the expectations of the (multivariate) random variables x and y is to be tested (i.e. P (x) = E (x)), the rank statistic proposed by MANN and WHITNEY (cf. CONOVER, 1971) is appropriate.

$$t := \sum_{i=1}^{n} R_i - \frac{n\,(n+m+1)}{2} \tag{B3}$$

Note that the sum is formed just for the first n components. The rank $R_i = k$ is assigned to the i. th observation, if k –1 observations are greater than z_i and n +m–1 + k are smaller. Thus, the rank 1 is given to the largest z_i, rank 2 to the second largest and n + m to the smallest value.

If the traces of the covariance matrices of x and y are considered (i.e. P (x) = Var (x)), the nonparametric rank statistic proposed by SIEGEL and TUKEY (cf. CONOVER, 1971) may be used. This statistic t, denoted by (B4) is formally identical to (B3), but the ranking is different. The assignment to the components of z is such that rank $R_i = 1$ is given to the largest component, rank $R_i = 2$ to the smallest, rank $R_i = 3$ to the second largest, rank $R_i = 4$ to the second smallest and so on.

We name a test based on (B1) and (B3) or (B4) a "generalized MANN-WHITNEY" or "SIEGEL-TUKEY" test and add the term "randomized" if instead of all permutations only a randomly selected subset of S is used. Both tests are fully nonparametric: no estimation of the parameters which are to be tested is necessary. This aspect is favorable in case of small samples.

The condition (*) is strictly spoken fulfilled only if x and y are identically distributed. By means of some Monte Carlo studies we found that in case of (B3) a violation of (*) is not crucial. For (B4), the problem may be solved practically by the replacement of the original data x and y by "sample centered data" $x_i - \bar{x}$ and $y_i - \bar{y}$, where $\bar{x}$ and $\bar{y}$ denote the sample means.

References

ALEXANDER, R. C. and R. L. MOBLEY, 1976: Monthly average sea surface temperatures and ice pack limits on a 1 degree global grid. Mon. Wea. Rev. **104**, 143–148.

ARPE, K., 1983: Diagnostic evaluation of analysis and forecast climate of the ECMWF model. Proceedings to the ECMWF seminar on "Interpretation of numerical weather prediction products", ECMWF, Shinfield Park, Reading, U.K.

BAEDE, A., M.JARRAUD and U. CUBASCH, 1979: Adiabatic formulation and organization of ECMWF's spectral model. Techn. Rep. No. 15, ECMWF, Shinfield Park, Reading, U. K.

BOYETT, J. M. and J. 1.SHUSTER, 1977: Nonparametrie one-sided tests in multivariate analysis with medical applications. J. Amer. Stat. Ass. **72**, 665–668.

CAO, H., 1984: Fuzzy verification of the simulated 500 mb topographies. Submitted to Tellus.

CONOVER, W. J., 1971: Practical nonparametrie statistics. John Wiley & Sons Inc., New York, London, Sidney, Toronto – 462 pp.

CRUTCHER, H. L. and R. J. JENNE, 1970: An interim note on Northern Hemisphere climatological grid data tape. NOAA Environmental Data Service, NWRC, Asheville.

CUBASCH, U., 1981: The performance of the ECMWF model in 50 day integrations. ECMWF Techn. Mem. No. 32, ECMWF, Shinfield Park, Reading, U. K.

DEARDORFF, J. W., 1966: The counter-gradient heat flux in the lower atmosphere and in the laboratory. J. Atmos. Sci. **23**, 503–506.

FITZJARRALD, D. R. and M. GARSTANG, 1981: Vertical structure of the tropical boundary layer. Mon.Wea.Rev. **109**, 1512–1526.

LAU, N. C., 1981: A diagnostic study of recurrent meteorological anomalies appearing in a 15 year simulation with the GFDL general circulation model. Mon. Wea. Rev. **109**, 2287–2296.

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

LOUIS, J. F. (ed.), 1984: The ECMWF forecasting model. ECMWF, Shinfield Park, Reading, U.K.

MALONE, R. C., E. J. PITCHER, M. L. BLACKMON, K. PURI and W. BOURKE, 1984: The simulation of stationary and transient geopotential height eddies in January and July with a spectral General Circulation model. J. Atmos. Sci. **41**, 1394–1419.

MANABE, S., D. G. HAHN and J. L. HOLLOWAY Jr., 1979: Climate simulations with GFDL spectral models of the atmosphere: Effect of spectral truncation. GARP Publication Series No. **22**, WMO, Geneva.

MANABE, S. and D. G. HAHN, 1981: Simulation of atmospheric variability. Mon. Wea. Rev. **109**, 2260–2286.

P1TCHER, E. J., R. C. MALONE, V. RAMANATHAN, M. L. BLACKMON, K. PURI and W. BOURKE, 1983: January and July simulations with a spectral General Circulation model. J. Atmos. Sci. **40**, 580–604.

PREISENDORFER, R. W. and T. P. BARNETT, 1983: Numerical model-reality intercomparison tests using small-sample statistics. J. Atmos. Sci. **40**, 1884–1896.

ROECKNER, E., 1979: A hemispheric model for short range numerical weather prodiction and general circulation studies. Beitr. Phys. Atmosph. **52**, 262–286.

STORCH, H. v. and I. FISCHER, 1983: An analysis of geopotential height at 300 mbar in the frequency wavenumber domain along 50 °N in observed and modelled January climate. Beitr. Phys. Atmosph. **56**, 199–212.

STORCH, H. v. and E. ROECKNER, 1983a: Methods for the verification of general circulation models applied to the Hamburg University GCM. Part I: Test of individual climate states. Mon. Wea. Rev. **111**, 1965–1976.

STORCH, H. v. and E. ROECKNER, 1983b: On the verification of January GCM simulations. -II International Meeting on Statistical Climatology Sep. 26–30, 1983, Lisboa, Portugal, Instituto Nacional de Meteorologica e Geofisica, Rua. C. Aeroporto, 1700 Lisboa, 14.7.1–8.

TEMPERTON, C., 1983: Survey of forecasts starting from a particular SOP-I initial state. WGNE Forecast Comparison Experiments, WCRP Report No. **6**, 9–71.

TIEDTKE, M., J. F. GELEYN, A. HOLLINGSWORTH and J. F. LOUIS, 1979: ECMWF-model parameterisation of subgrid processes. Techn. Rep. No. 10, ECMWF, Shinfield Park, Reading, U. K.

VOLMER, J. P., M. DEQUE and D. ROUSSELET, 1984: EOF analysis of 500 mb geopotential: A comparison between simulation and reality. Tellus **36**, 336–347.

YAMADA, T., 1976: On the similarity functions A, Band C of the planetary boundary layer. J. Atmos. Sci. **33**, 781793.