JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 95, NO. D1, PAGES 829–850, JANUARY 20, 1990

Regional Validation of Means, Variances, and Spatial Patterns in General Circulation Model Control Runs

B. D. SANTER

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

T. M. L. WIGLEY

Climatic Research Unit, University of East Anglia, Norwich, England

The focus of this study is the control run performance of four general circulation models (GCMs): the Oregon State University (OSU) two-layer atmospheric GCM (AGCM), the OSU coupled ocean-atmosphere model (CGCM), the Goddard Institute for Space Studies (GISS) nine-layer AGCM, and the European Centre for Medium-Range Weather Forecasts (ECMWF) T21 model. The analysis variable is monthly mean sea level pressure (MSLP), and model validation is performed for a limited domain (North America/Atlantic/Europe). The first part of the investigation deals with the magnitude and gross spatial structure of model errors in means and interannual variability (for January and July only). These errors are examined with the aid of maps of time-mean MSLP, difference fields, and local variance ratios. The significance of the local (grid point by grid point) differences in means and variances is then determined by performing univariate *t*- and *F*-tests. This information on the spatial structure of large-scale systematic errors is important for understanding the results of significance tests performed on the overall fields. In the second part of the investigation, the statistics recommended by Wigley and Santer (this issue) for use in model validation are applied to test the overall significance of observed/simulated differences in means, variances, and spatial patterns over the entire annual cycle. Significance levels are determined with the pool permutation procedure (PPP) introduced by Preisendorfer and Barnett (1983). Results indicate that all four models have highly significant errors in the mean field and spatial pattern over the entire annual cycle. Errors in the temporal variance are generally less significant, and significance levels for variance tests can depend critically on the choice of averaging period for observed validation data. The actual test statistic values show that there are considerable differences in model performance. The ECMWF T21 model simulates the spatial pattern and time-mean MSLP field with greater fidelity than the other models considered here.

1. INTRODUCTION

It is generally accepted that there are substantial differences in the regional and seasonal details of the control and perturbed run climates simulated by different general circulation models (GCMs) [e.g., *MacCracken and Luther*, 1985]. Yet these differences have not been fully documented and there are relatively few comparative studies of model performance at the regional scale and over the entire seasonal cycle [e.g., *Reed*, 1986; *Grotch*, 1988; *Santer*, 1988*a*, *b*]. Most control run validation studies test model performance in January and July only, despite the availability of comprehensive observed data bases which provide information on atmospheric behavior throughout the seasonal cycle [e.g., *Oort*, 1983; *Lau*, 1984]. Previous studies have also tended to focus on global- or hemispheric-scale validation of GCM climatologies using visual comparison of observed and simulated fields [e.g., *Schlesinger and Gates*, 1980; *Hansen et al.*, 1983; *Schlesinger and Mitchell*, 1985, 1987].

There is clearly a strong case for validation of control run regional and seasonal details. Increasingly, impact studies based on GCM-derived climate change scenarios are being performed at a regional level, often without any or with only inadequate validation of control run climate [e.g., *Meinl et al.*, 1984; *Cohen*, 1986; *Parry et al.*, 1987]. Without validation appropriate to the spatial and temporal scales of the impact study, the reliability of climate change scenarios, and ultimately of the impact results, is unknown.

There are also strong arguments in support of detailed comparisons of GCM control run performance. Since there are large differences in the structure, physics, and parameterizations of different models, the results from control run intercomparison studies are difficult to interpret unambiguously. But similar control run errors in different models may have common dynamical explanations. Also, as *Mitchell et al.* [1987] point out, control run intercomparison can aid in understanding the causes of intermodel differences in perturbed run results.

Control run validation and model intercomparison are facilitated by the use of rigorous statistical methods, as is the analysis of results from GCM perturbation and predictability experiments. Both climate modelers and impact analysts working with the climate results from perturbation experiments require some objective basis for making decisions. For example, it may be important for a modeler to know whether an error in a simulated time-mean, mean sea level pressure (MSLP) field is sufficiently large to warrant detailed sensitivity studies, or whether the "error" is within the range of natural decadal time-scale variability for MSLP [*Santer*, 1988*a*]. Another common problem is determining whether alterations to a model's resolution or subgrid scale parameterizations have resulted in a significant improvement or deterioration in model performance.

The objective basis necessary for reaching decisions on these and related questions is provided by rigorous signifi-

Paper number 89JD01599.
0148-0227/90/89JD-01599$05.00

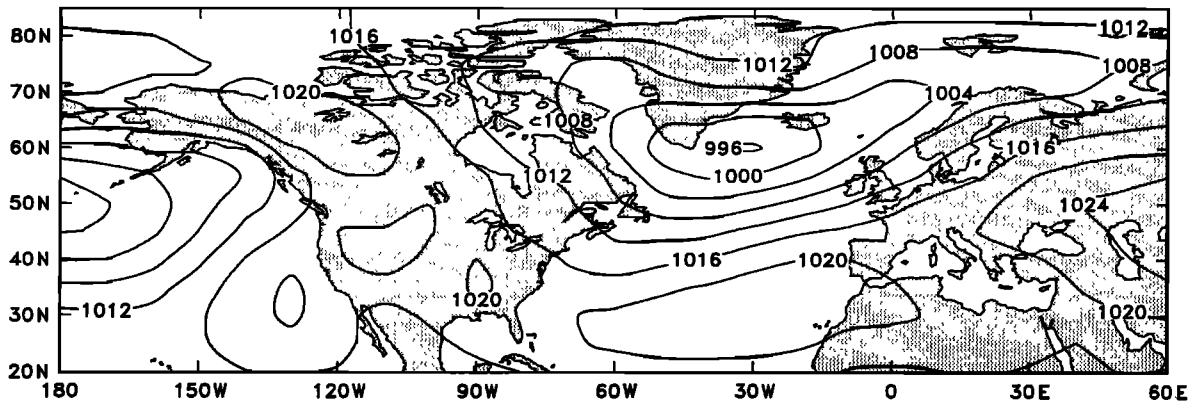

Fig. 1*a*. Observed January MSLP, UKMO data set. Time-mean field for the decade 1971–1980.

cance testing using a variety of different test statistics (measuring differences in means, variances, and spatial patterns). But which significance testing procedures and which test statistics are most suitable for the specific purposes of GCM control run validation and model intercomparison? This is not a simple question to answer. A large number of potentially useful statistics and significance testing procedures are available in the modeling literature and in the wider statistical literature. It is not the intention here to provide a review of the univariate and multivariate test statistics which have been used in the context of studies performed with GCMs and other numerical models. Such reviews have been given by *Laurmann and Gates* [1977], *Livezey* [1985], *von Storch* [1985], *Zwiers* [1987], *Zwiers and Thiébaux* [1987], and *Santer* [1988*a*] for GCMs, by *Daley and Chervin* [1985] and *Stamus* [1985] for numerical weather prediction models, and by *Willmott et al.* [1985] for numerical models in general. Other studies have considered the statistical problems commonly encountered when univariate and multivariate tests are applied to GCM data, notably the problems of multiplicity and spatial autocorrelation for univariate tests [*von Storch*, 1982; *Livezey and Chen*, 1983; *Wigley and Santer*, 1988], and the problem of dimensionality for multivariate tests [*Hasselmann*, 1979; *von Storch and Kruse*, 1985].

Few studies, however, have attempted to evaluate the specific advantages and disadvantages of different statistics when applied for the purposes of validating and intercomparing model climatologies [*von Storch*, 1985; *Santer*, 1988*a*]. Yet this issue is of considerable importance, particularly as the requirement increases for objective methods of assessing and comparing model performance. We need to identify statistics which can be applied operationally, and which provide easily interpretable information of real diagnostic benefit.

Wigley and Santer [this issue] recommended a set of nine statistics for routine use in evaluating the overall significance of data set differences in means, variances, and spatial patterns. These statistics are applied here in order to validate the MSLP fields simulated by four GCMs in extended control integrations. The pool permutation procedure (PPP) introduced by *Preisendorfer and Barnett* [1983] is used to assess the overall significance of test statistic results. Permutation-based methods such as PPP provide a means of circumventing problems commonly encountered in significance testing, notably multiplicity, spatial autocorrelation, and unknown sampling distributions.

The emphasis here is on the practical application of test statistics rather than on detailed power testing using synthetic data [*Preisendorfer and Mobley*, 1982; *Zwiers*, 1987]. A further aim is to show that detailed analysis of the magnitude and gross spatial structure of model errors is essential in order to understand and interpret multivariate significance test results.

2. Observed and Simulated Data Sets

MSLP was selected as the analysis variable for control run validation and model intercomparison. MSLP provides easily interpretable information as to how successfully a model performs in simulating important features of the atmospheric general circulation. If a model has large-scale systematic errors in its simulated MSLP fields, the simulated surface fields of other important variables (e.g., precipitation, zonal and meridional winds) will also be in error. The North American/Atlantic/European study area was identical to that used by *Wigley and Santer* [this issue] for their comparison of MSLP fields for two observed decades.

2.1. *Observed Data*

The United Kingdom Meteorological Office (UKMO) data set was used for validating simulated MSLP. This consists of monthly mean, gridded MSLP data for the period 1873–1980. Data are on a regular 5° × 10° latitude/longitude grid from 15°N to 65°N, and on a 5° × 20° grid from 70°N to 80°N. Pressure data at latitude 85°N are given for four grid points only (180°W, 90°W, 0°, 90°E). The sources and data quality problems of the UKMO data set have been documented by *Williams and van Loon* [1976] and *Jones* [1987].

For significance testing, observed MSLP data for the decade 1971–1980 were selected. The use of decadal data is necessary, since the PPP method requires equal time samples of observed and simulated data. Since only 10 years of data were available for two of the four control runs examined here, 10 years of observed data had to be selected for validation. However, it is not possible to select one "optimum" observed decade for validation purposes, i.e., to define a single most suitable decade in terms of matching boundary conditions in the real world and in the model. It was therefore considered reasonable to use the decade

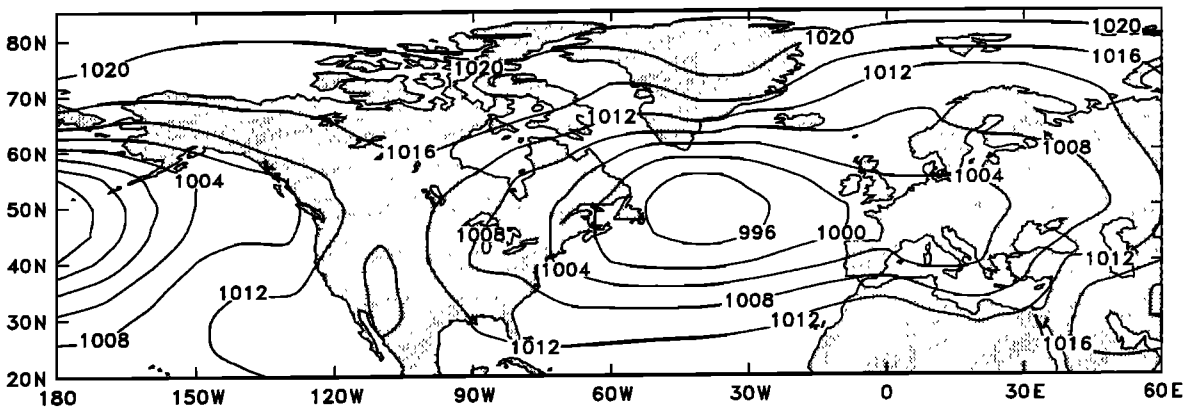

Fig. 1*b*

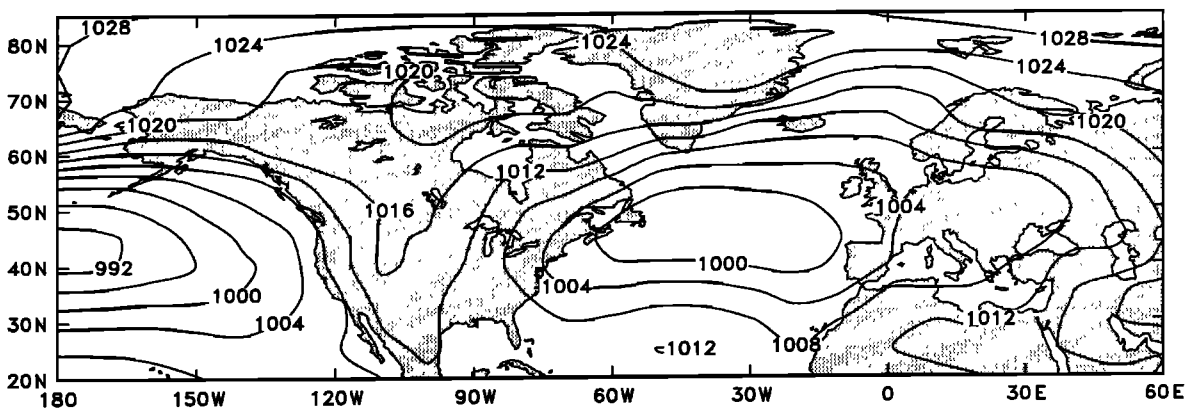

Fig. 1*c*

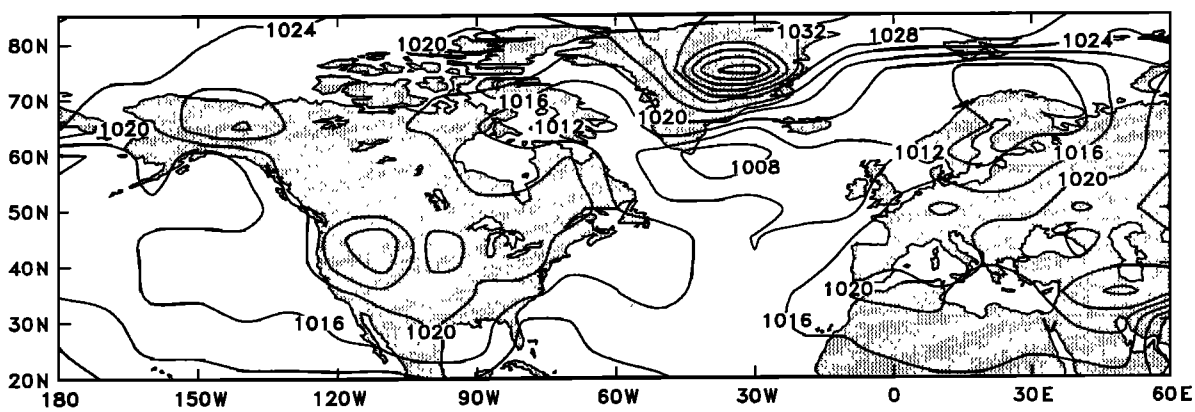

Fig. 1*d*

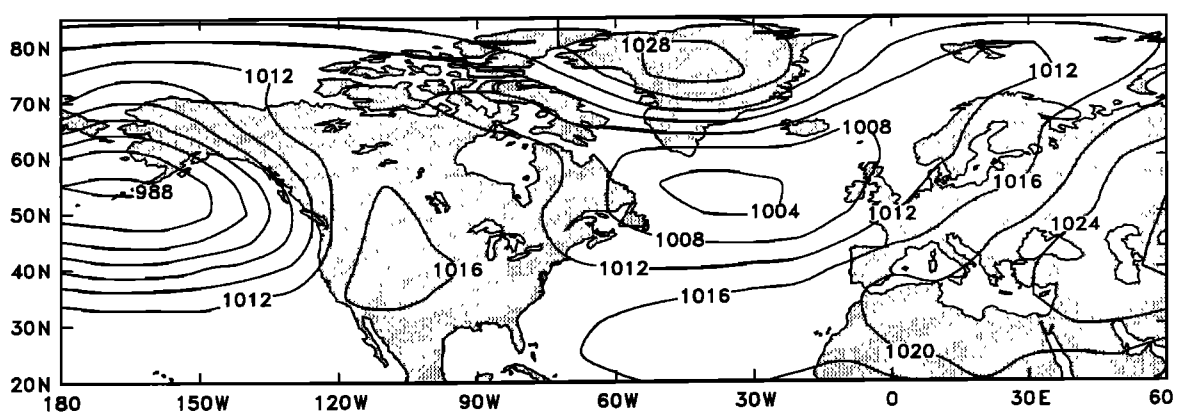

Fig. 1*e*

Fig. 1. (*b–e*) Simulated January MSLP. Time-mean fields for the (*b*) OSU AGCM, (*c*) OSU CGCM, (*d*) GISS AGCM, and (*e*) ECMWF T21 model. For sources and averaging periods of the simulated data, refer to section 2.

1971–1980 for validation purposes. The data for this decade have few gaps and are more reliable than pre-1941 data. The issue of the sensitivity of multivariate significance test results to decadal time scale variability in the observed MSLP data is treated in section 7.

2.2. *Simulated Data*

The model validation was performed for MSLP fields taken from four separate control runs.

1. A 10-year control run performed with the Oregon State University (OSU) two-layer atmospheric GCM (AGCM) with prescribed sea surface temperatures (SST).
2. A 20-year control run performed with the OSU coupled ocean-atmosphere GCM (CGCM), in which the two-layer AGCM was synchronously coupled with a six-layer OGCM. Both OGCM and AGCM are grid point models with 4° × 5° horizontal solution.
3. A 35-year control run performed with the Goddard Institute for Space Studies (GISS) nine-layer AGCM with prescribed mixed-layer depth and oceanic heat transport. The GISS AGCM is a grid point model with 7.83° × 10° horizontal resolution.
4. A 10-year control run performed with the ECMWF T21 model with prescribed SST. This is a spectral model with 5.625° × 5.625° resolution for nonlinear terms and zonal wave number 21 resolution for linear terms.

Details of model structure and control run parameters for the OSU AGCM and CGCM are given by *Schlesinger and Gates* [1980] and *Gates et al.* [1984], and a complete documentation of the two-layer model has been published by *Ghan et al.* [1982]. For information relating to the structure and control run parameters of the GISS nine-layer AGCM ("model II") and the ECMWF T21 model, refer to *Hansen et al.* [1983, 1984] and *Dümenil and Schlese* [1987], respectively.

In order to ensure equal time samples for all four control runs, 10-year subsets were selected from the 20-year CGCM control run (years 11–20) and the 35-year GISS control run (years 26–35). Since the observed and simulated data sets have different horizontal resolutions, model data were interpolated to the 5° × 10° observed grid (using a two-dimensional Gaussian filter technique) prior to plotting and statistical analyses.

3. ERRORS IN THE TIME-MEAN FIELDS

Examination of the time-mean January and July MSLP maps and corresponding difference fields (Figures 1–4) indicates that model errors can be divided into two general categories: (1) large-scale, spatially coherent pressure biases and (2) errors in the location and intensity of quasi-stationary centers of action (COAs), which are apparent as maxima and minima in the difference fields.

3.1. *January*

3.1.1. *OSU AGCM and CGCM.* In both the OSU AGCM and CGCM, the Iceland Low is displaced southward in January. Its intensity is underestimated in the CGCM (i.e., central pressure too high). Both models substantially underestimate the intensity of the Azores High. These errors are reflected in the difference field maxima off the west coast of Spain (circa 16 mbar for the AGCM and 20 mbar for the CGCM) and the difference field minima over Greenland (circa −8 mbar for the AGCM and −16 mbar for the CGCM). Neither model simulates a discrete North Pacific subtropical high, and the Aleutian Low is too intense. In addition to these COA-related errors, both models have large-scale, spatially coherent pressure biases. Pressure is overestimated over Greenland and most of the Arctic and underestimated over the remainder of the study area.

In a comparison of the MSLP fields simulated by the OSU AGCM and CGCM, *Gates et al.* [1984] noted that "the coupled model has made only small changes relative to the uncoupled atmospheric GCM in the simulation of sea-level pressure". However, for the limited domain examined here, it is evident that errors in the CGCM's time-mean January MSLP field are considerably larger than for the AGCM.

3.1.2. *GISS AGCM.* The time-mean January MSLP field simulated by the GISS AGCM is characterized by an unrealistically large fraction of the variance at high wave numbers, particularly at middle and high latitudes. This behavior is exhibited during all months and is not confined to the northern hemisphere. It is apparently related to the coarse horizontal resolution of the model [*Hansen et al.*, 1983]. Despite this deficiency, the GISS AGCM simulates the correct position of the Iceland Low, North Pacific subtropical high, and the ridge of high pressure over the Canadian Arctic. The model fails, however, to produce a discrete Azores High, underestimates the intensity of the Iceland Low, and simulates numerous small-scale surface ridges and troughs which do not have observed analogs. The most striking of these features is a spurious "Greenland High" (maximum central pressure circa 1048 mbar), which is also generated by the T21 model (maximum central pressure circa 1028 mbar). Both OSU models simulate a comparable feature in July but not in January (Figure 3). *Wigley and Santer* [1988] have shown that the magnitude of these pressure errors over Greenland is not solely due to errors in the reduction of surface pressure to sea level. Part of the error is also related to model deficiencies in simulating strong surface temperature inversions over high-latitude plateaus.

Like both OSU models, the GISS AGCM has large-scale pressure biases, with pressure underestimated in the subtropics and overestimated in middle and high latitudes. Pressure biases are generally smaller than in the OSU AGCM and CGCM, except over the central Pacific and in the vicinity of the Greenland High.

3.1.3. *ECMWF T21 model.* The T21 model simulates the spatial pattern and the absolute magnitude of the time-mean January MSLP field with greater fidelity than the other three models considered here [*Santer*, 1988*b*]. The major errors are in the intensity of the Iceland Low and Azores High (both too weak) and Aleutian Low (too intense). The position of the Azores High is well simulated, but the Aleutian and Iceland lows are displaced to the north and south (respectively). As in the case of both OSU models and the GISS AGCM, there are large, spatially coherent pressure biases. Pressure is higher than observed over Greenland (due to the spurious Greenland High), the European Arctic, and Africa, and lower than observed over the rest of the study area.

One interesting feature of the January simulation is that the T21 model reproduces the Iceland Low's characteristic N.E.-S.W. horizontal axis of orientation. In the observa-

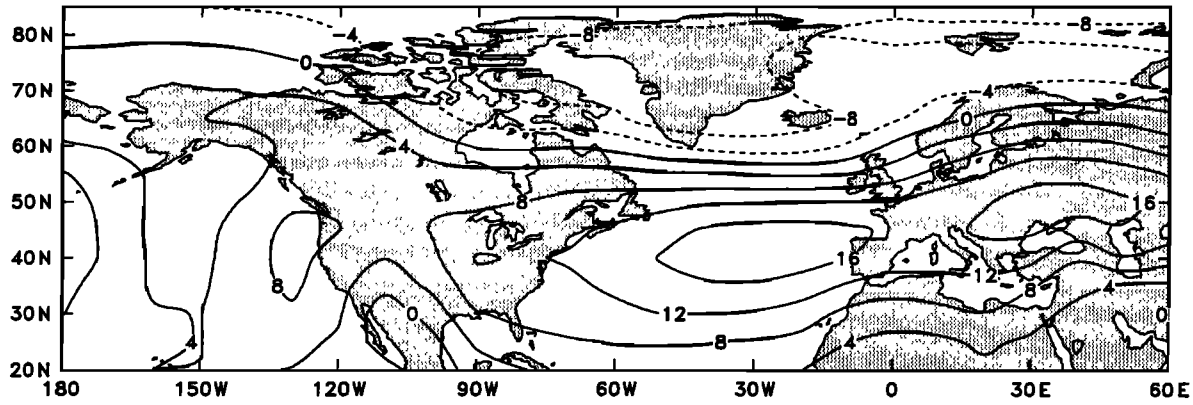

Fig. 2*a*

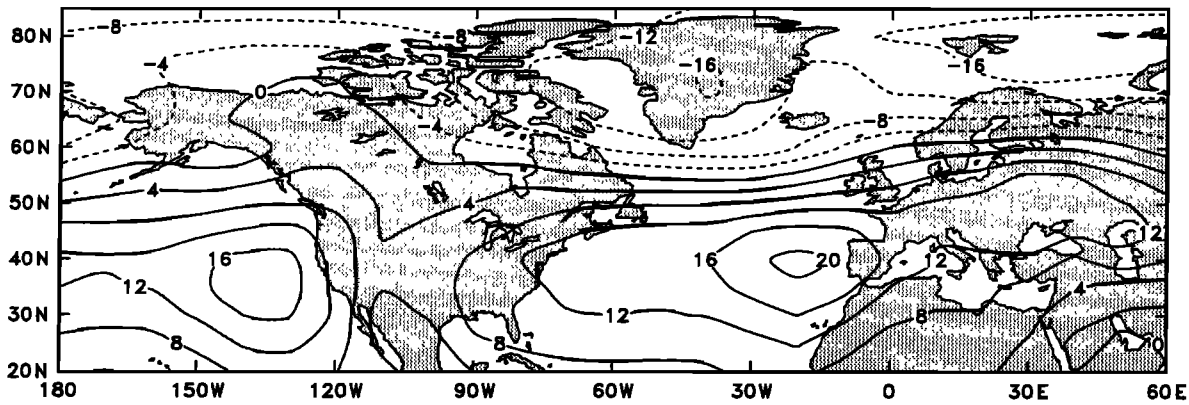

Fig. 2*b*

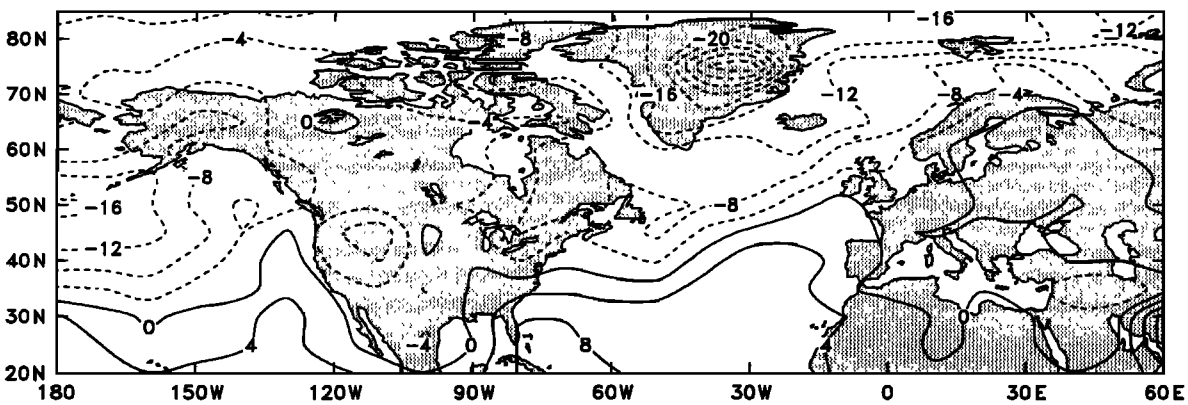

Fig. 2*c*

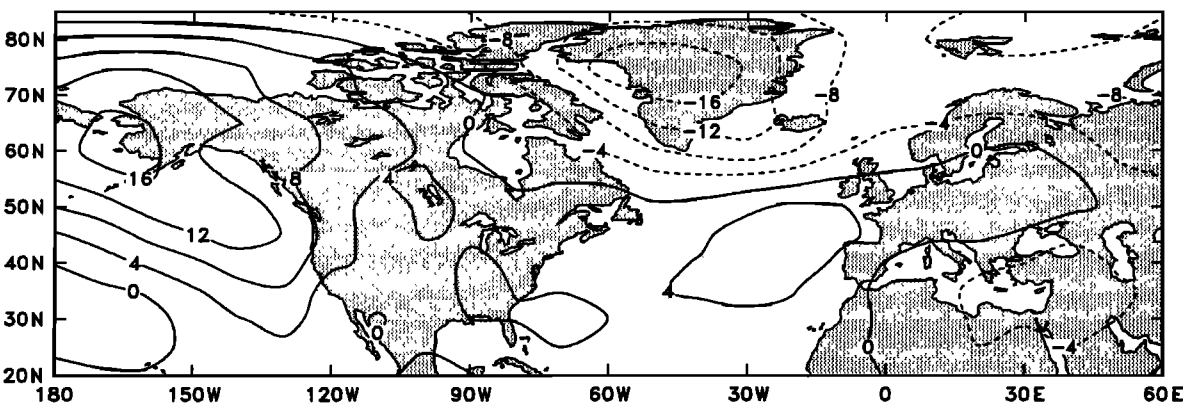

Fig. 2*d*

Fig. 2. Difference fields (observed minus simulated) for January MSLP. UKMO observed data (1971–1980) minus (*a*) OSU AGCM, (*b*) OSU CGCM, (*c*) GISS AGCM, and (*d*) ECMWF T21 model. Dashed isopleths indicate areas where simulated MSLP is greater than observed. For sources and averaging periods of the simulated data, refer to section 2.

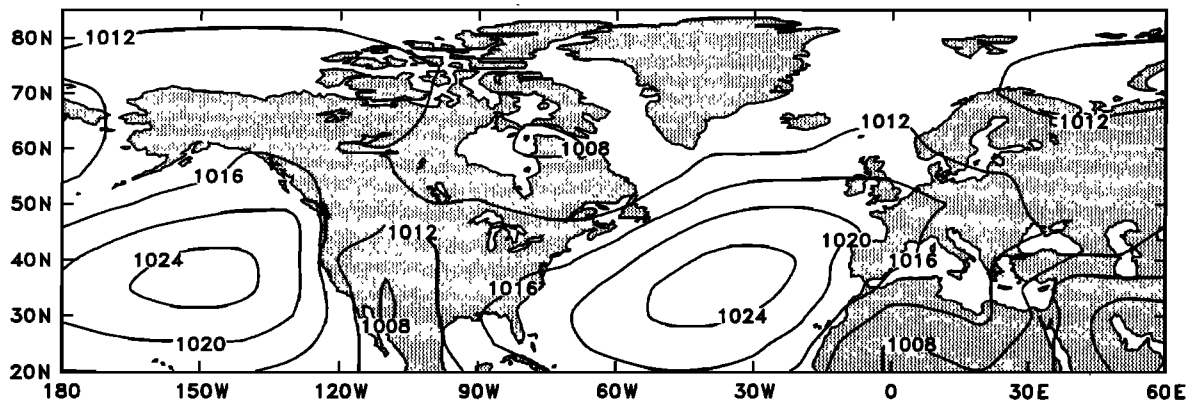

Fig. 3*a*. Observed July MSLP, UKMO data set. Time-mean field for the decade 1971–1980.

tions, thermal forcing is of predominant importance in determining the surface features of the Low [*Wallace*, 1983], and it is probable that this horizontal eccentricity is at least partly explained by the Gulf Stream-induced asymmetry in diabatic heating. However, the prescribing of SST is no guarantee for successful simulation of this feature, as is shown by results for the OSU AGCM.

3.2. *July*

3.2.1. *OSU AGCM and CGCM.* Both OSU models have large errors in their July simulations (Figure 3). Qualitatively, these errors are highly similar. Neither model reproduces the discrete single-cell structure of the observed Azores and North Pacific subtropical highs. Both generate a spurious Greenland High and an unrealistic area of low pressure over the southeastern United States, and both show unrealistic ridging extending south from Greenland and Alaska. Large-scale, spatially coherent biases are also similar: the AGCM and CGCM overestimate pressure north of around 50°N and underestimate MSLP south of this latitude (Figure 4). As for January, errors in the CGCM are generally larger than in the uncoupled model.

3.2.2. *GISS AGCM.* The July circulation simulated by the GISS AGCM is difficult to compare with observations. The model fails to simulate recognizable Azores and North Pacific subtropical highs, and as in the January simulation, high wave numbers explain an unrealistically large fraction of the variance. There are two parallels with the performance of the OSU models: there are similarities in the large-scale pressure biases, and the ridging behavior noted for the OSU AGCM and CGCM is also present here.

3.2.3. *ECMWF T21 model.* The T21 model successfully simulates the spatial pattern of July MSLP. The observed single-cell structure of the Azores and North Pacific subtropical highs is reproduced, along with the positions of the lows over Labrador, Mexico, and the Sahara. However, there are still large errors in the time-mean field, although these are smaller than for the other models considered here (Figure 4). Note that the T21 model underestimates MSLP over virtually the entire study area, with maximum errors (circa 8 mbar) in the subtropics. Thus the large-scale pressure bias differs both qualitatively and quantitatively from the July bias in the other three models.

4. ERRORS IN THE VARIANCE FIELDS

4.1. *January*

Maps of the logarithm of the local variance ratio (observed divided by simulated) provide insights into the spatial structure of model errors in the interannual variability of January MSLP (Figure 5). In the observed data, there are characteristic variance maxima associated with the positions of the Iceland and Aleutian lows [see *Santer*, 1988*a*]. Model errors in the simulation of these variance maxima are reflected in maxima in the log of the variance ratio near the observed positions of both lows.

Large-scale, spatially coherent variance biases can also be identified. In both OSU models and the T21 model, interannual variance is underestimated over most of the study area (i.e., the log of the variance ratio is positive). This result is not surprising, given the fact that important boundary conditions in the OSU AGCM and T21 model are prescribed. In fact, the surprising feature is the similarity of variance results for the coupled and uncoupled OSU models. In the GISS AGCM, the areal extent of positive and negative variance biases is roughly equal. It is notable that all four models overestimate the interannual variability of January MSLP over most of the United States.

4.2. *July*

The GISS AGCM and T21 model overestimate the interannual variability of July MSLP over virtually the entire study area (Figure 6). The overall variance bias of the T21 model is thus reversed relative to January, a result which has been noted previously for results of multivariate variance tests with the T21 model [*Santer*, 1988*b*]. As in January, the OSU AGCM generally underestimates the variance, while the areas of positive and negative variance biases are approximately equal in the OSU CGCM.

5. UNIVARIATE SIGNIFICANCE TEST RESULTS

5.1. *Grid Point t-Tests*

The grid point *t*-test results for January and July (Figures 7 and 8, respectively) clearly reflect the large errors in the simulated time-mean fields (Figures 2 and 4). Results show the probability ("*p* value") of obtaining the observed local *t*

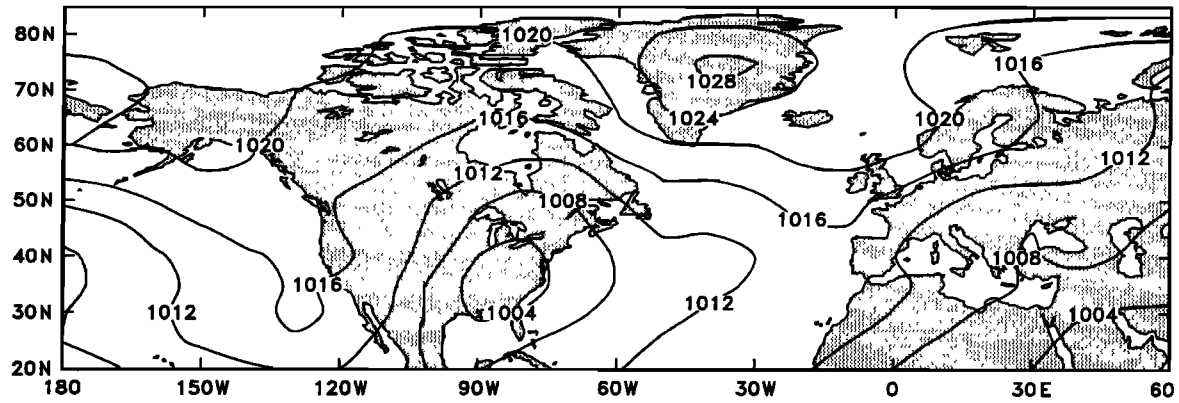

Fig. 3*b*

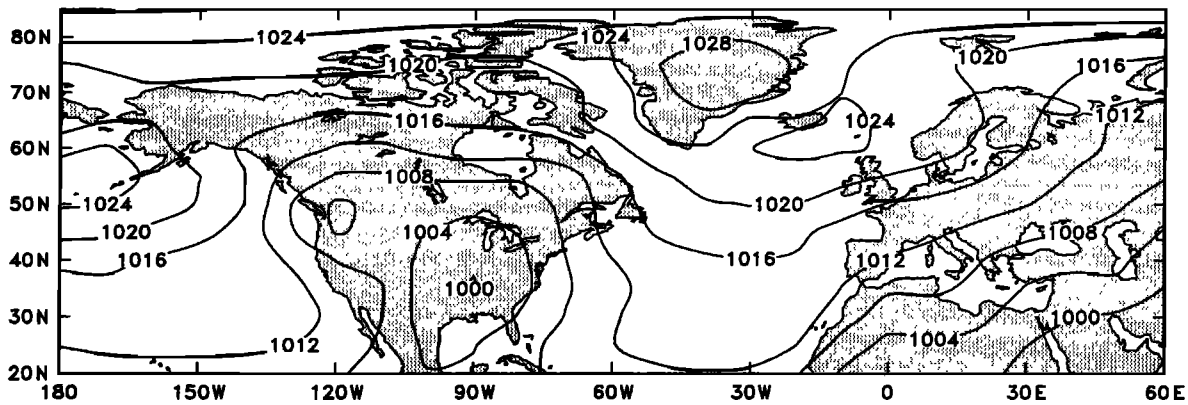

Fig. 3*c*

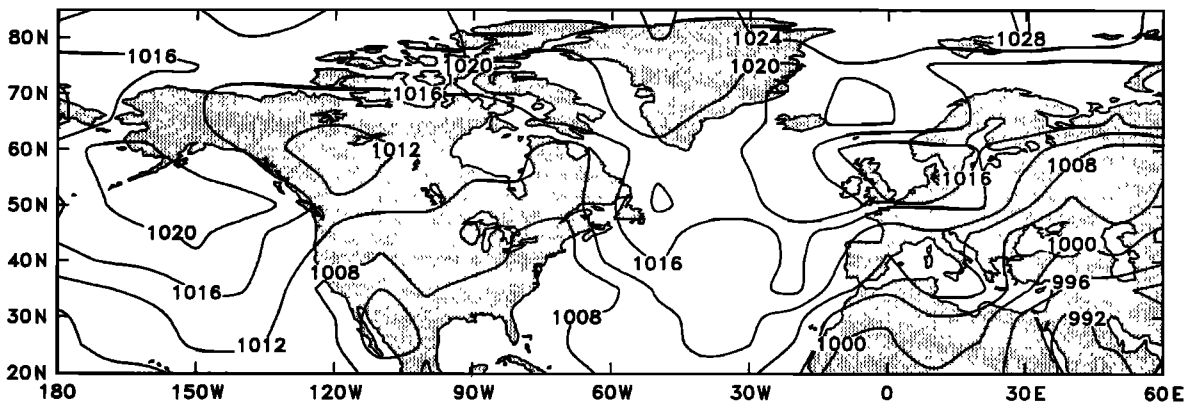

Fig. 3*d*

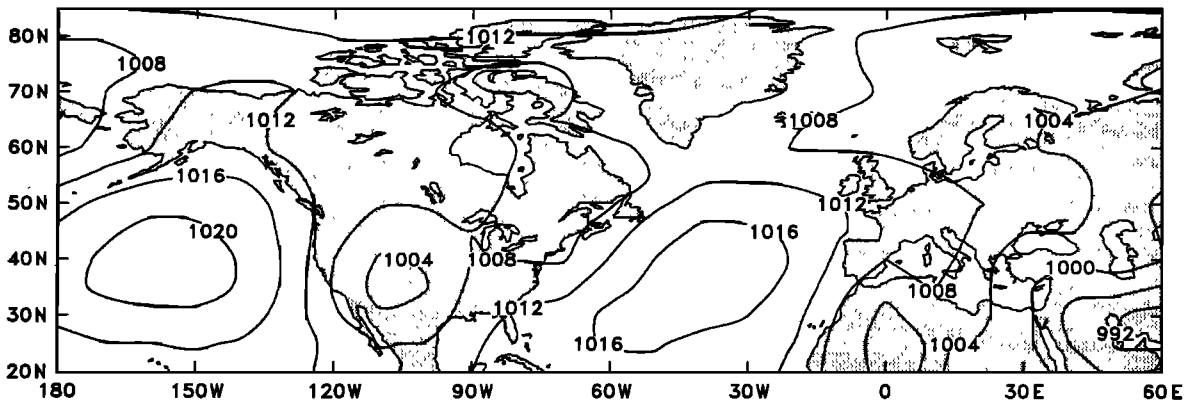

Fig. 3*e*

Fig. 3. (*b–e*) Simulated July MSLP. Time-mean fields for the (*b*) OSU AGCM, (*c*) OSU CGCM, (*d*) GISS AGCM, and (*e*) ECMWF T21 model. For sources and averaging periods of the simulated data, refer to section 2.

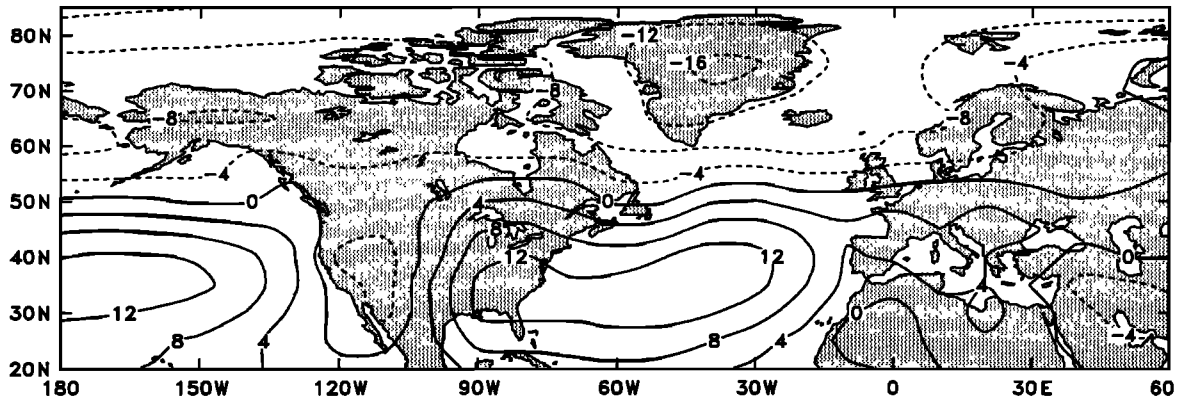

Fig. 4*a*

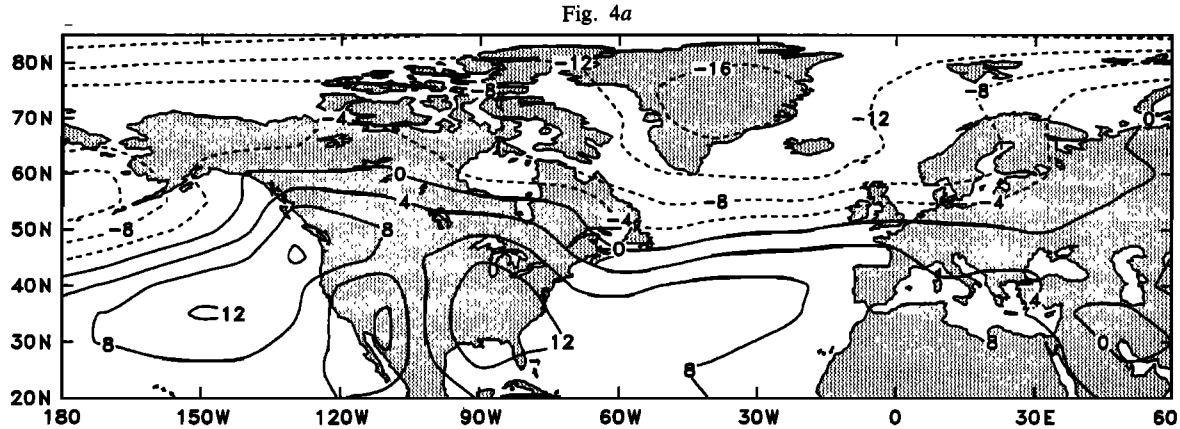

Fig. 4*b*

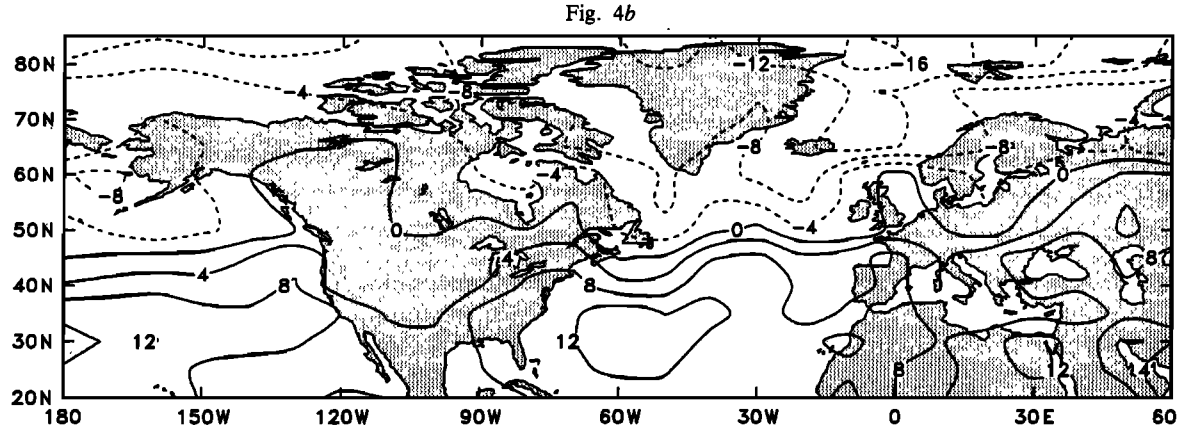

Fig. 4*c*

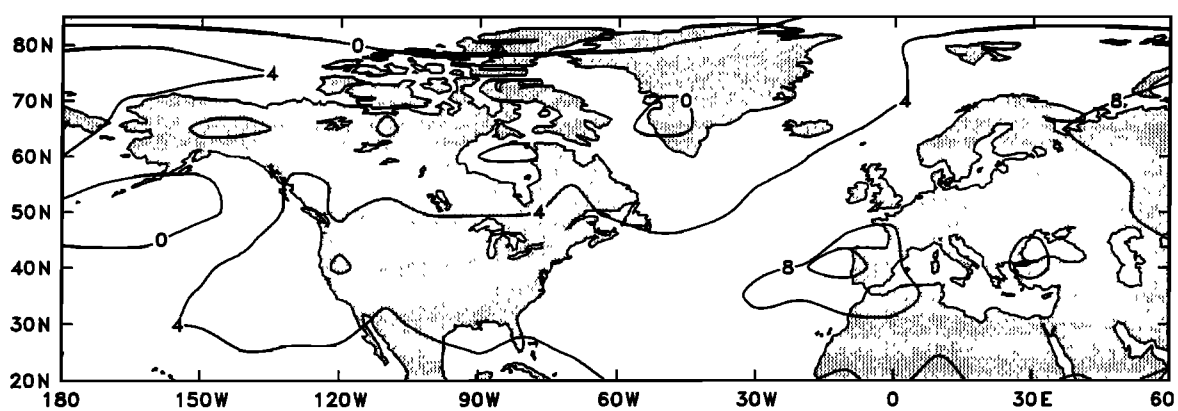

Fig. 4*d*

Fig. 4. Difference fields (observed minus simulated) for July MSLP. UKMO observed data (1971–1980) minus (*a*) OSU AGCM (*b*) OSU CGCM, (*c*) GISS AGCM, and (*d*) ECMWF T21 model. Dashed isopleths indicate areas where simulated MSLP is greater than observed. For sources and averaging periods of the simulated data, refer to section 2.

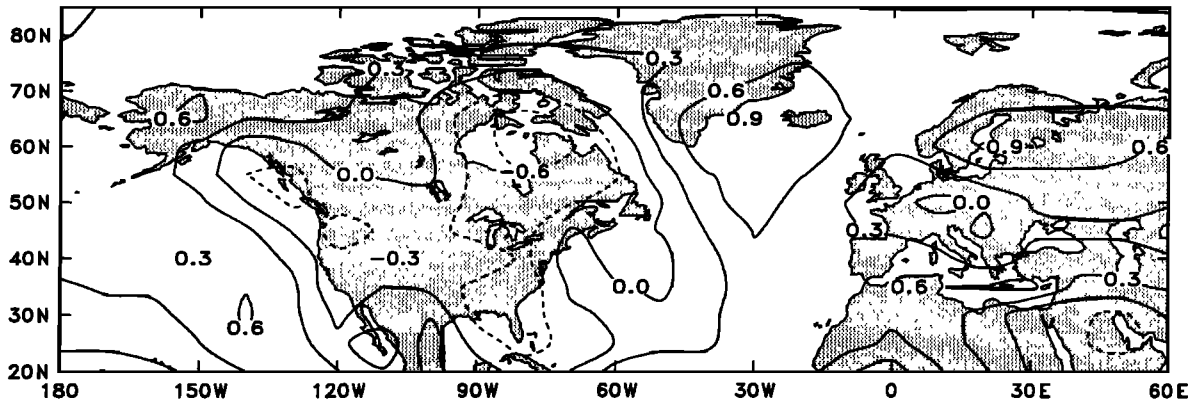

Fig. 5a

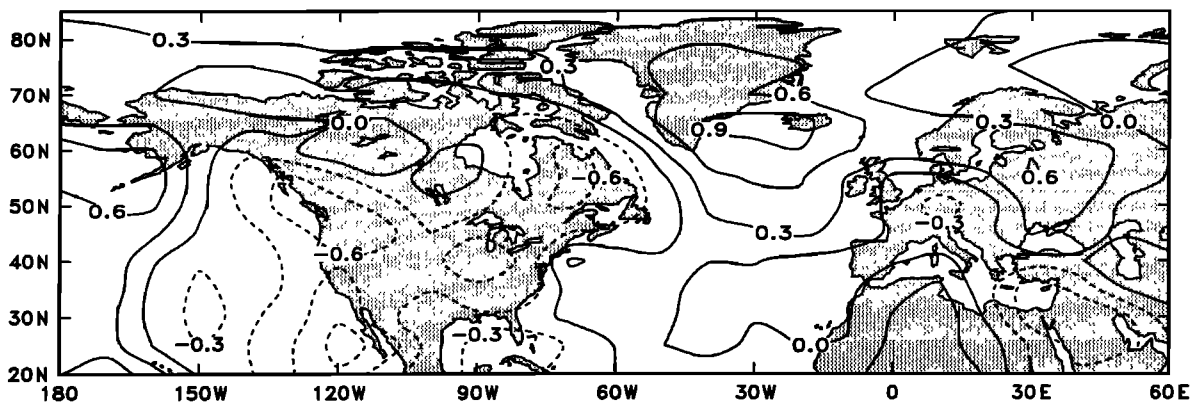

Fig. 5b

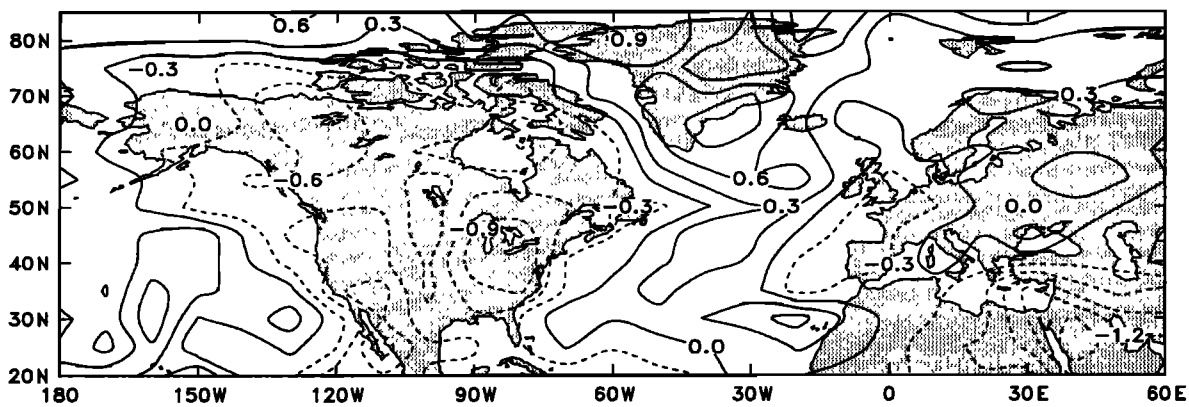

Fig. 5c

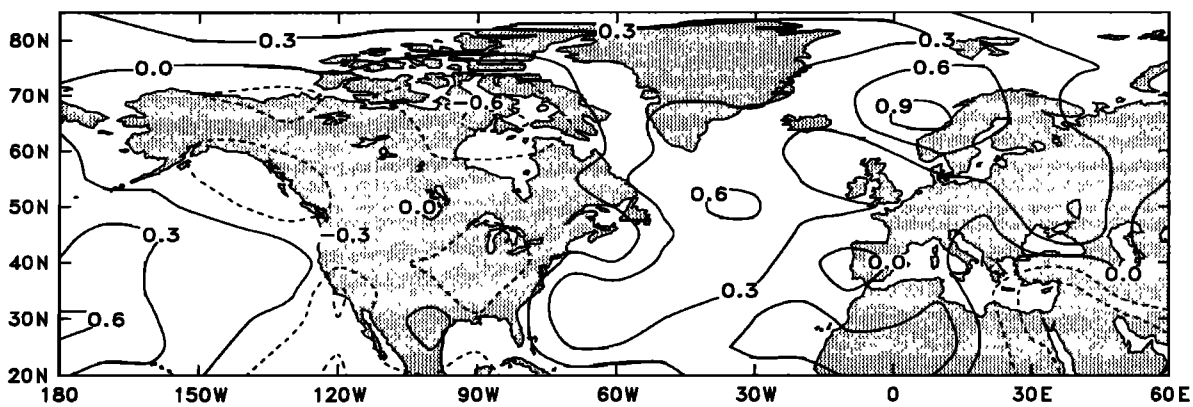

Fig. 5d

Fig. 5. Variance ratios for January MSLP, UKMO observed (1971–1980) divided by (*a*) OSU AGCM, (*b*) OSU CGCM, (*c*) GISS AGCM, and (*d*) ECMWF T21 model. The isopleths show the logarithm of the variance ratio in order to identify unusually high or low ratios. Dashed isopleths indicate areas where the model variance is greater than observed. Note that all four models overestimate the variance over most of the United States.

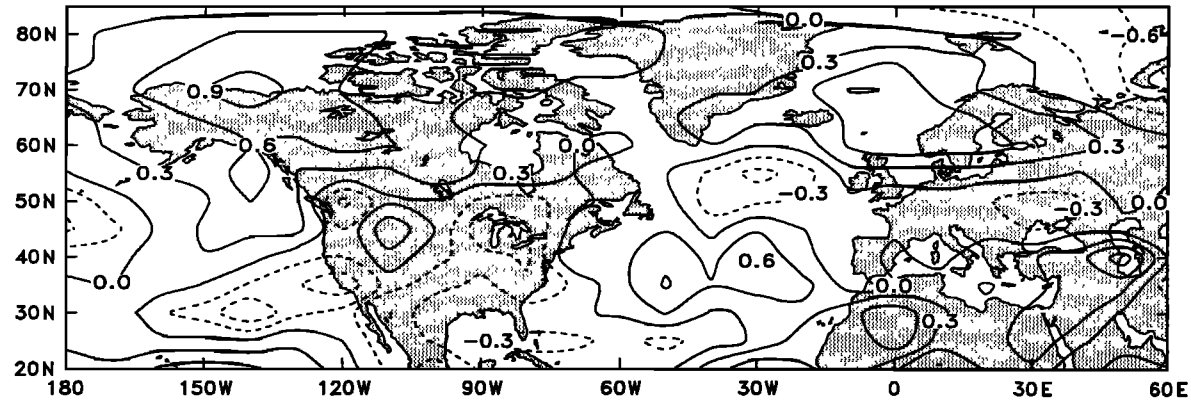

Fig. 6*a*

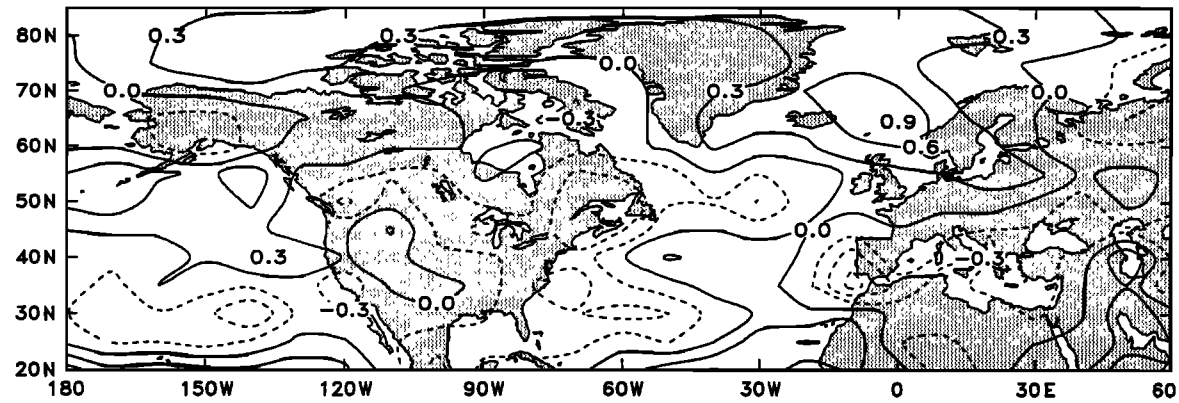

Fig. 6*b*

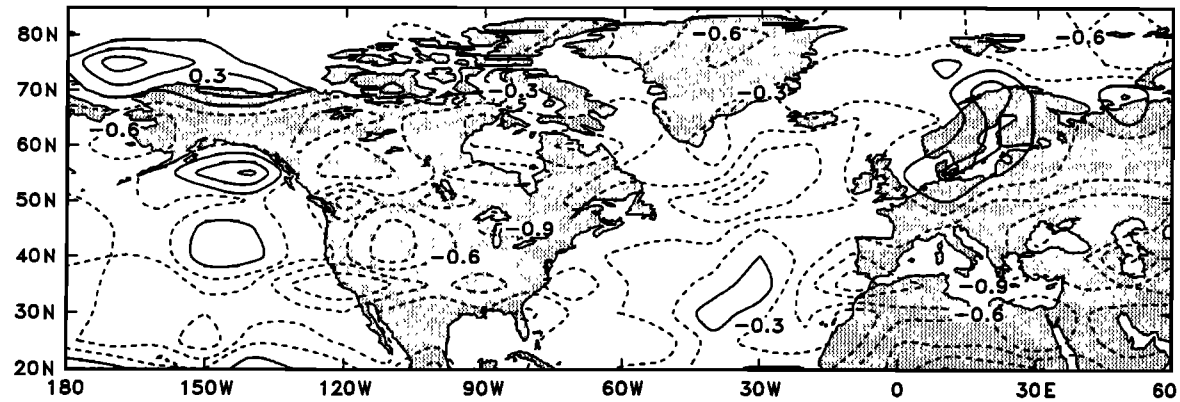

Fig. 6*c*

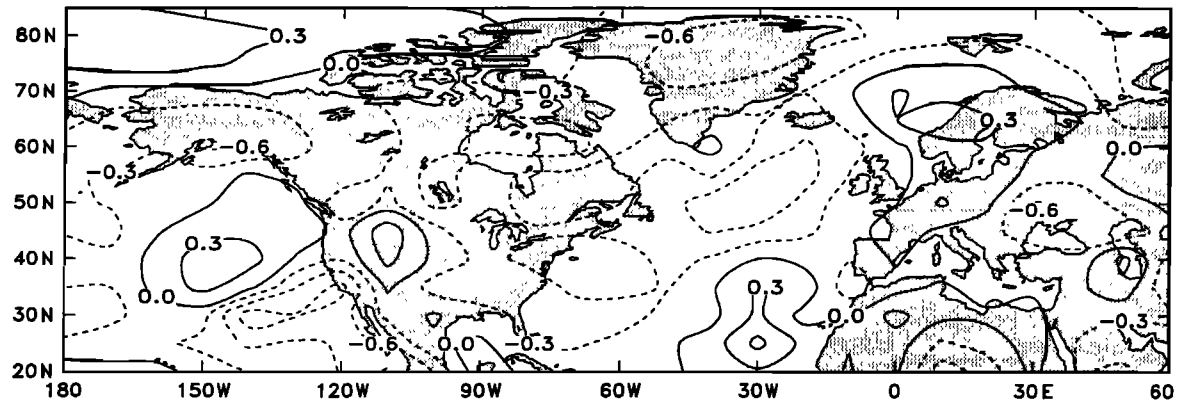

Fig. 6*d*

Fig. 6. Variance ratios for July MSLP, UKMO observed (1971–1980) divided by (*a*) OSU AGCM, (*b*) OSU CGCM, (*c*) GISS AGCM, and (*d*) ECMWF T21 model. The isopleths show the logarithm of the variance ratio in order to identify unusually high or low ratios. Dashed isopleths indicate areas where the model variance is greater than observed.

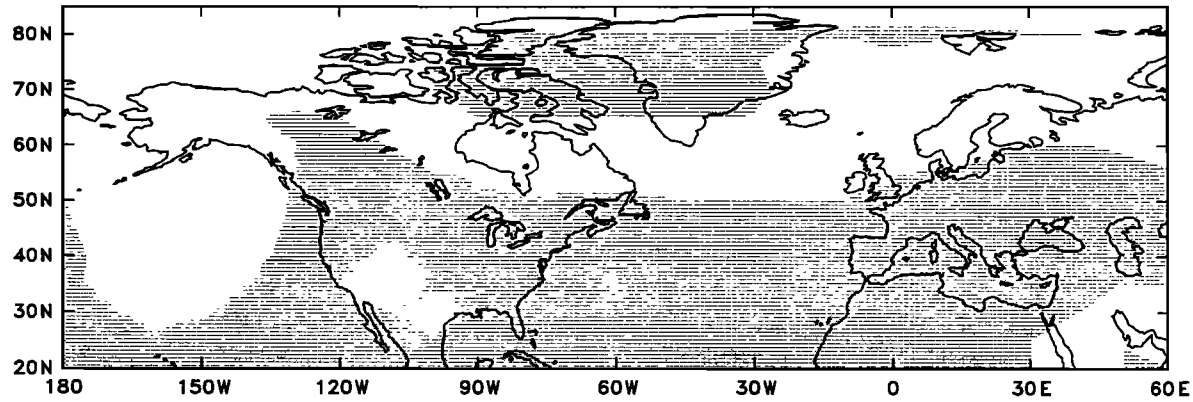

Fig. 7*a*

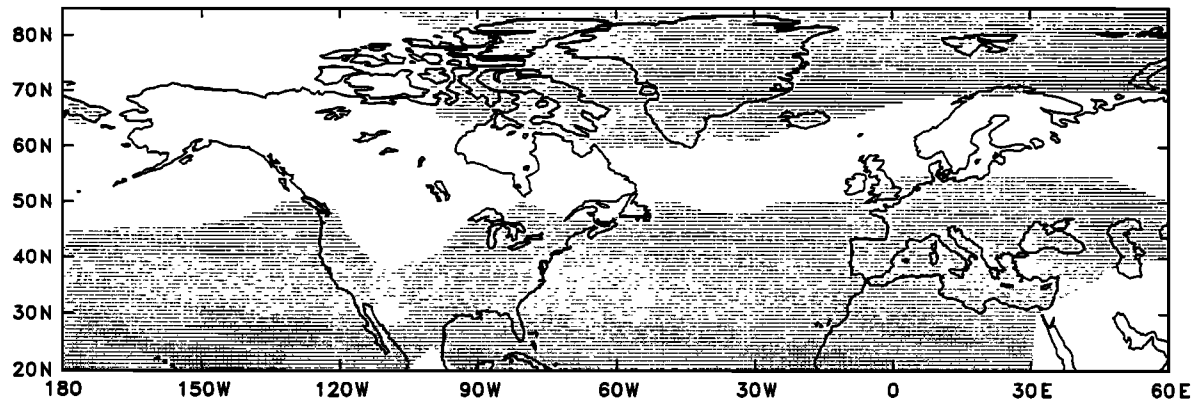

Fig. 7*b*

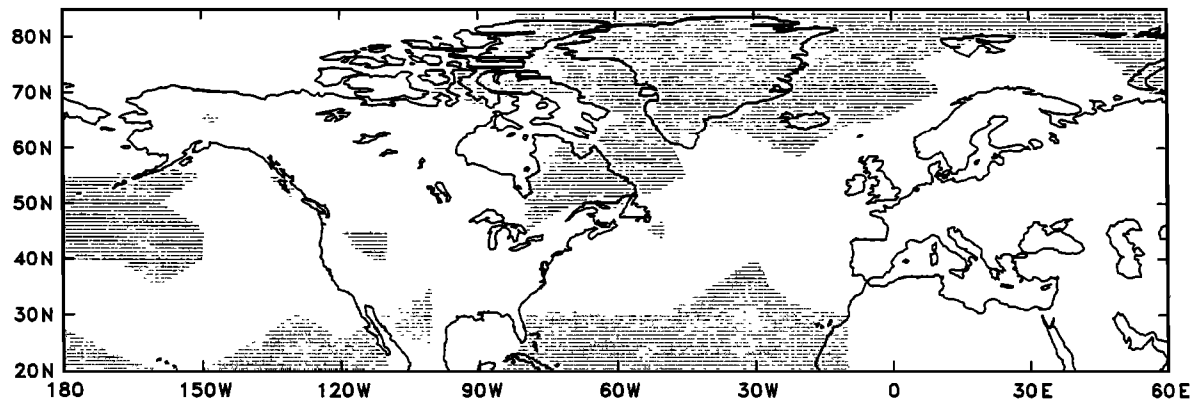

Fig. 7*c*

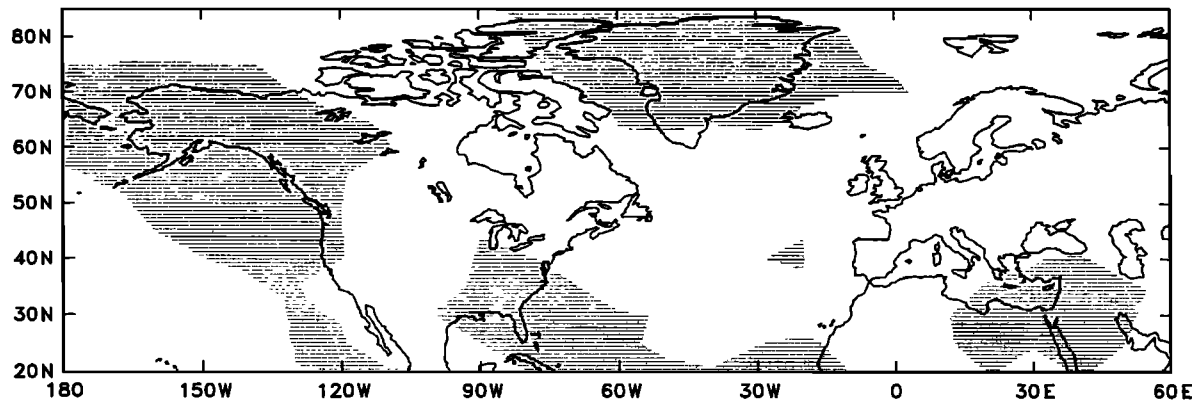

Fig. 7*d*

Fig. 7. Local *t*-test results for January MSLP. Results are for UKMO observed data (1971–1980) versus the (*a*) OSU AGCM, (*b*) OSU CGCM, (*c*) GISS AGCM, and (*d*) ECMWF T21 model. Shading indicates areas where differences in means are significant at or greater than the 1% level. All local tests are two-tailed. Note that in certain cases, shaded areas cross the zero difference line (Figure 2) due to the coarse resolution of the data and the use of an objective contouring routine.

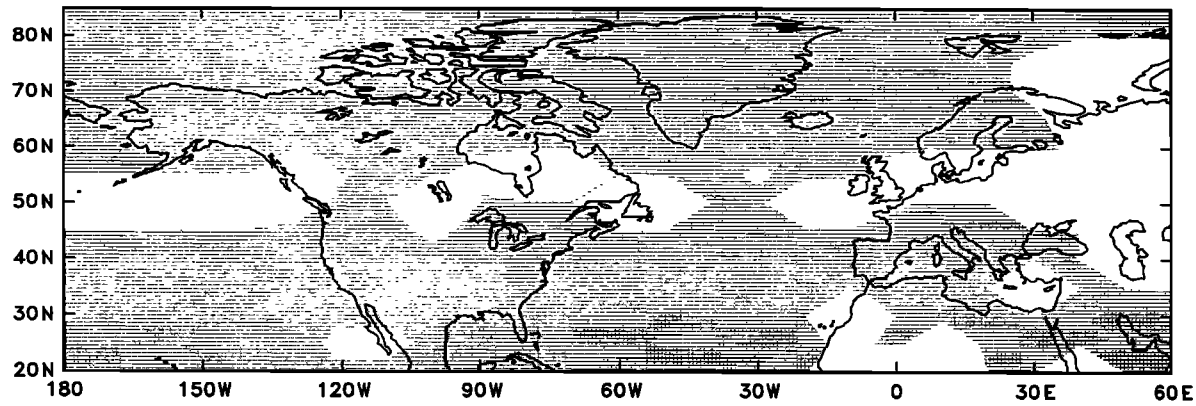

Fig. 8*a*

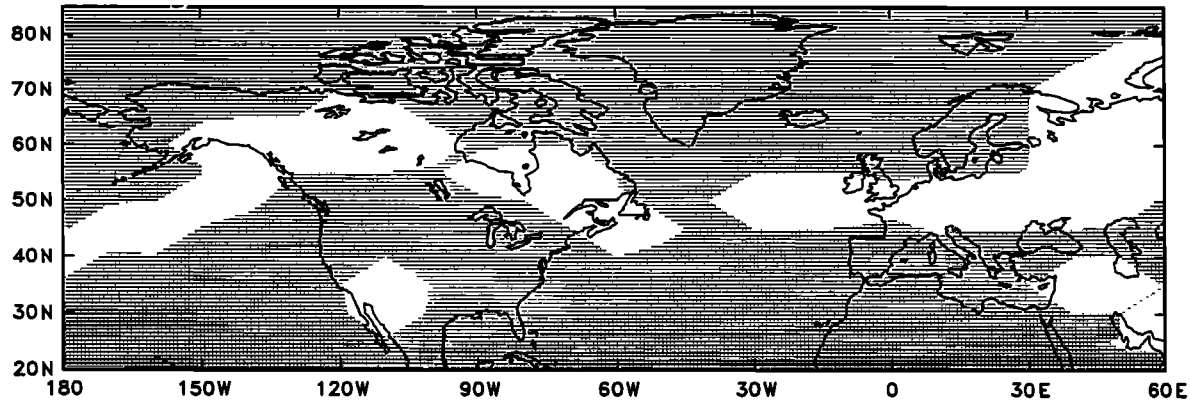

Fig. 8*b*

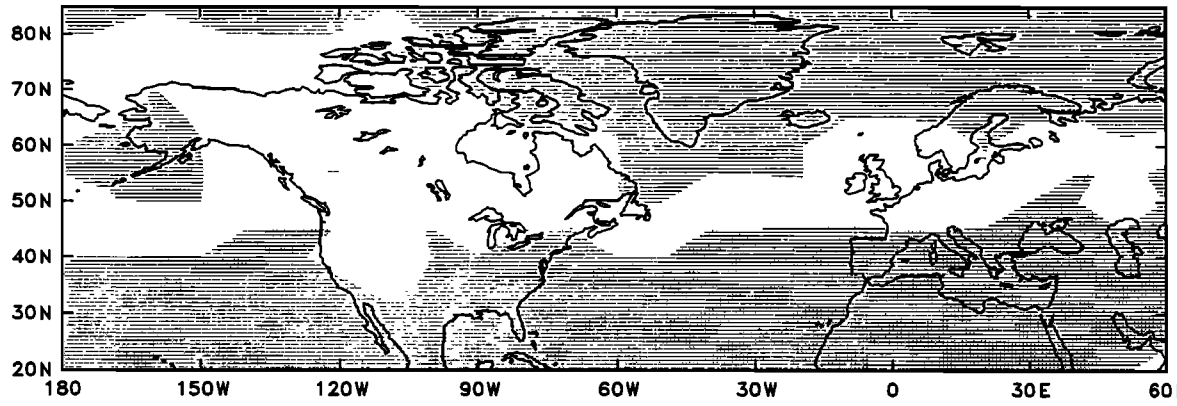

Fig. 8*c*

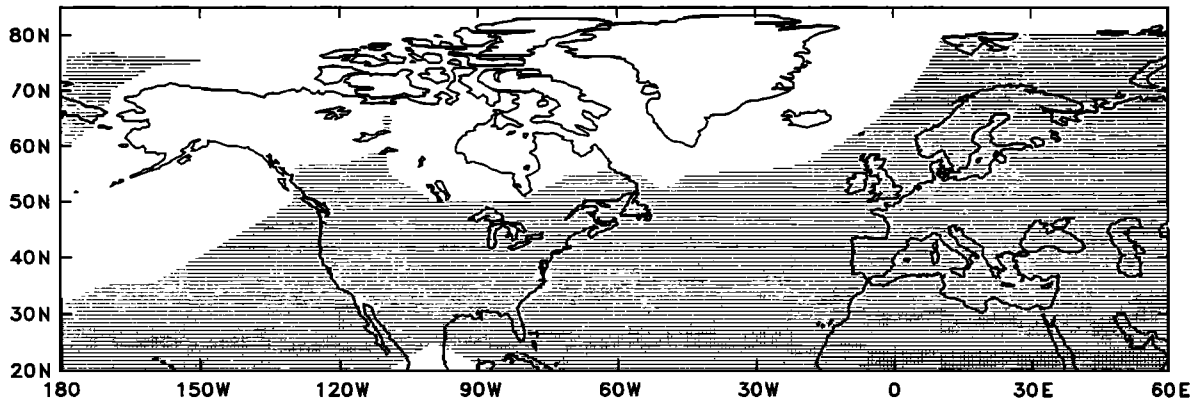

Fig. 8*d*

Fig. 8. Local *t*-test results for July MSLP. Results are for UKMO observed data (1971–1980) versus the (*a*) OSU AGCM, (*b*) OSU CGCM, (*c*) GISS AGCM, and (*d*) ECMWF T21 model. Shading indicates areas where differences in means are significant at or greater than the 1% level. All local tests are two-tailed.

value by chance. All local tests were two-tailed. Areas with local differences significant at the 1% level (denoted by shading in Figures 7 and 8) are largest for the OSU coupled and uncoupled models in July, and cover most of the study area. The previously noted similarities in the OSU AGCM and CGCM difference fields (in both months) are evident in the close correspondence between their respective patterns of local significance levels.

Areas with locally significant differences (1% level) are smaller for the GISS AGCM and T21 model, although still sufficiently large to ensure significant results for multivariate tests of the mean (section 6). For all four models, the univariate test results indicate that errors in the time-mean field are larger in July than in January.

5.2. *Grid Point F-Tests*

Results for the grid point F-tests indicate that, for the four models examined here, errors in the interannual variability of January and July MSLP are consistently less significant than errors in the time-mean field (Figures 9 and 10). This result is at least partly due to the lower power of the F-test and the small time sample used here [see *Zwiers and Thiébaux*, 1987]. For the test situation pertaining here, namely, two-tailed F-tests with nine ($n_t - 1$) degrees of freedom in the numerator and denominator, an observed variance ratio greater than 4.03 (6.54) is required in order to achieve significance at the 5% (1%) level.

In July the GISS AGCM has large areas of locally significant variance ratios. In the T21 model and the OSU AGCM and CGCM (in both January and July), local variance differences significant at the 5% level generally occur at isolated grid points only. There are, however, a few clusters of locally significant points. One such cluster is located between Greenland and Iceland in the January results for the coupled and uncoupled OSU models and is a function of model errors in simulating the variance maximum associated with the Iceland Low (see Figure 5).

In contrast to the case of the univariate t-test results, we cannot reach a field decision (i.e., whether the overall model and real world variances are significantly different at some prescribed level) simply by visual inspection of the local F-test p values, at least not for the T21 model or either OSU model. Could these variance test results have been obtained by chance alone? In order to answer this question it is necessary to account for the twin effects of multiplicity and spatial autocorrelation, e.g., by applying the test statistics and significance testing procedures used by *Wigley and Santer* [this issue].

6. MULTIVARIATE SIGNIFICANCE TESTS

Wigley and Santer [this issue] applied a number of different statistics in order to test the significance of overall differences in means, variances, and spatial patterns, using MSLP data for two observed decades. They recommended the use of a reduced set of nine test statistics for routine quantitative evaluation of data set differences: total number of "successful" local 1% and 5% t-tests (NT1, NT5) (where success is defined as rejection of the local null hypothesis at the stipulated level of significance), overall difference in means (SITES), differences between grand means (T1), total number of successful local 1% and 5% F-tests (NF1, NF5), overall difference in temporal variances (SPRET1), overall difference in spatial variances (SPREX1), and differences in spatial patterns of time mean fields (r). These statistics were applied here for validation of the simulated MSLP fields. Significance testing was performed using the PPP introduced by *Preisendorfer and Barnett* [1983], with 1000 randomizations of the **D** (data) and **M** (model) time ordering. For further details of the test statistics and PPP, refer to *Wigley and Santer* [this issue]. Details of the PPP-generated reference distributions are given by *Santer* [1988*a*].

Significance tests were conducted over the entire seasonal cycle using the data sets described in section 2. Thus there are 48 test statistic values and p values for each test statistic (12 months × 4 models). (The test statistic values are calculated for unrandomized **D** and **M** data sets; the one-tailed Monte Carlo probabilities are determined with PPP.) Note that all grid points in **D** with missing data and corresponding points in **M** were excluded from the analysis, so that n_x (the number of valid grid points) varies from 288 to 292 over the seasonal cycle. All grid point t- and F-tests were two-tailed, and results for NT1, NT5, NF1, and NF5 are expressed as the fraction of locally significant test results (relative to n_x).

6.1. *Results for NT1 and NT5*

The p values for NT1 and NT5 indicate that the 48 results for each statistic are all highly field significant (Table 1). In fact, all p values except one are zero, indicating that (except for the T21 model in July) the actual NT1-NT5 test statistic values are always larger than every value in the PPP-generated NT1-NT5 reference distributions. For each month and model, the null hypothesis that **D** and **M** are drawn from populations with identical time-mean fields must be rejected. This result is not unexpected in view of the large model errors in the time-mean fields for January and July (see sections 3 and 5).

While the p values indicate that all four models have highly significant errors in their time-mean fields, the actual test statistic values for NT1-NT5 reveal considerable differences in model performance (Figure 11). These actual values effectively summarize the univariate information presented in Figures 7 and 8. The T21 model generally has the lowest NT1-NT5 values throughout the seasonal cycle, implying smaller errors in its simulation of the time-mean field. Errors are largest for the OSU CGCM. We conclude that the prescription of important boundary conditions in the T21 model and OSU and GISS AGCMs acts as a constraint on the magnitude of the overall error in the time-mean field. This constraint is much less severe in the OSU CGCM, where only surface salinity is prescribed. In the control run analyzed here, the two-layer AGCM and six-layer OGCM were synchronously coupled without the use of any flux corrections (e.g., as used by *Sausen et al.* [1988]). Although physically realistic, this coupling strategy allows large errors to develop as a result of feedbacks between any errors which exist in the separate (uncoupled) atmospheric and oceanic models.

Figure 11 shows that model errors in the simulation of MSLP vary over the seasonal cycle. For all four models, errors in the time-mean field are largest in July, August, and September.

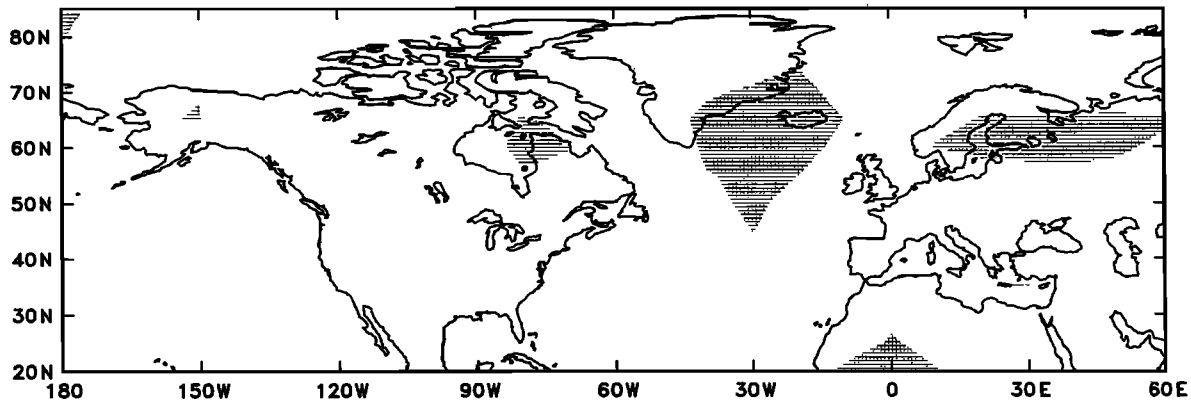

Fig. 9*a*

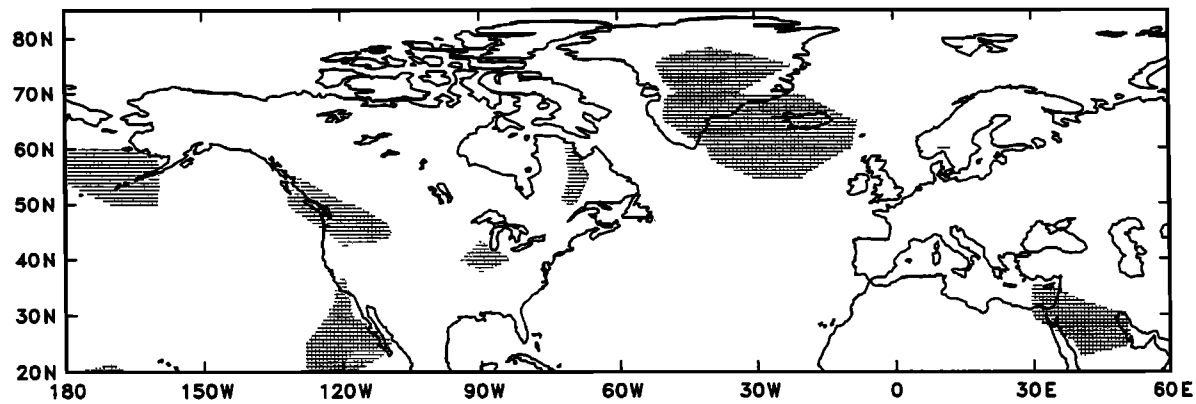

Fig. 9*b*

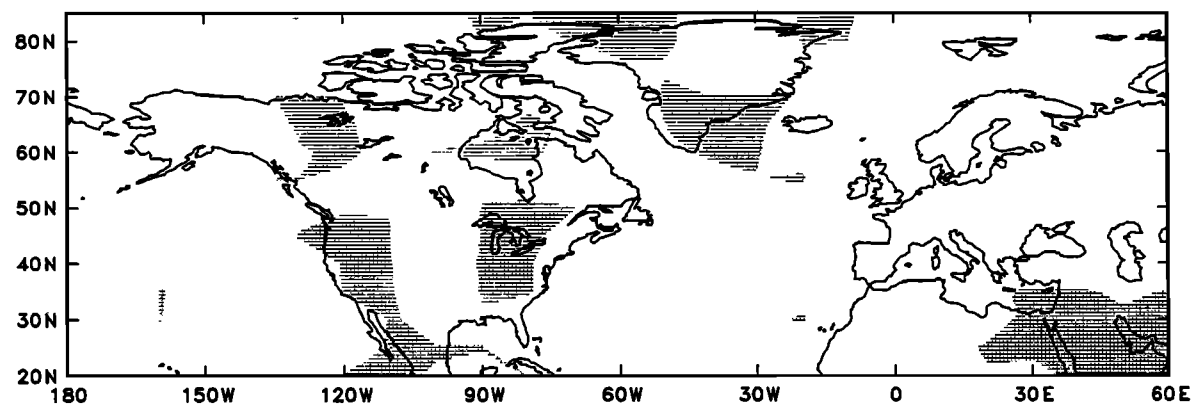

Fig. 9*c*

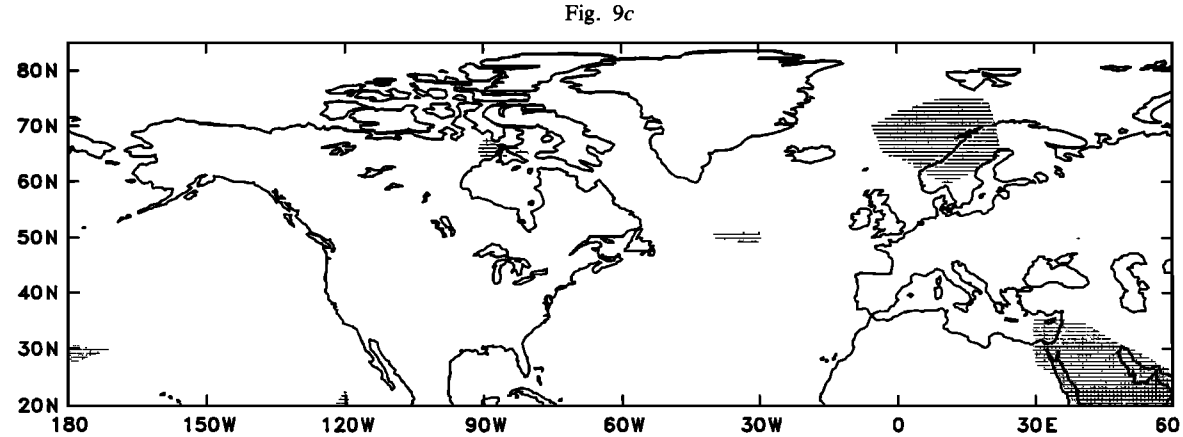

Fig. 9*d*

Fig. 9. Local *F*-test results for January MSLP. Results are for UKMO observed data (1971–1980) versus the (*a*) OSU AGCM, (*b*) OSU CGCM, (*c*) GISS AGCM, and (*d*) ECMWF T21 model. Shading indicates areas where differences in variances are significant at or greater than the 5% level.

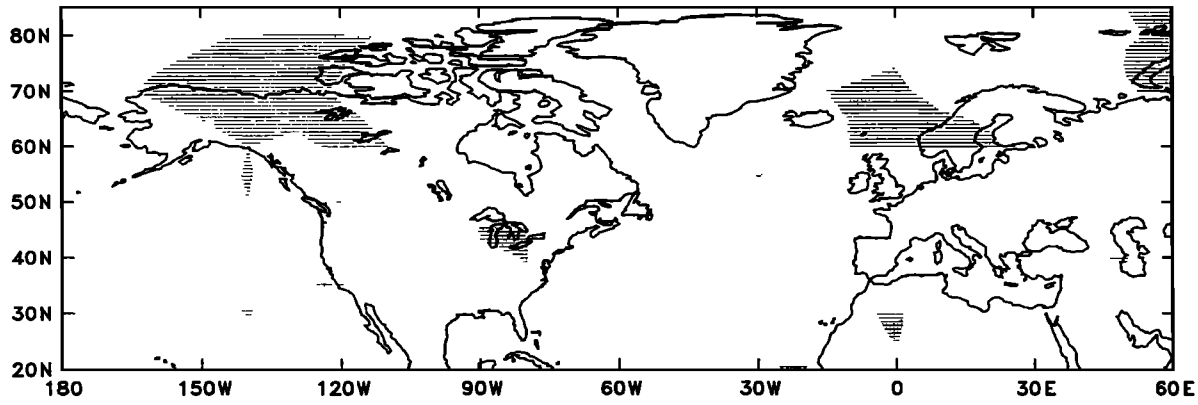

Fig. 10*a*

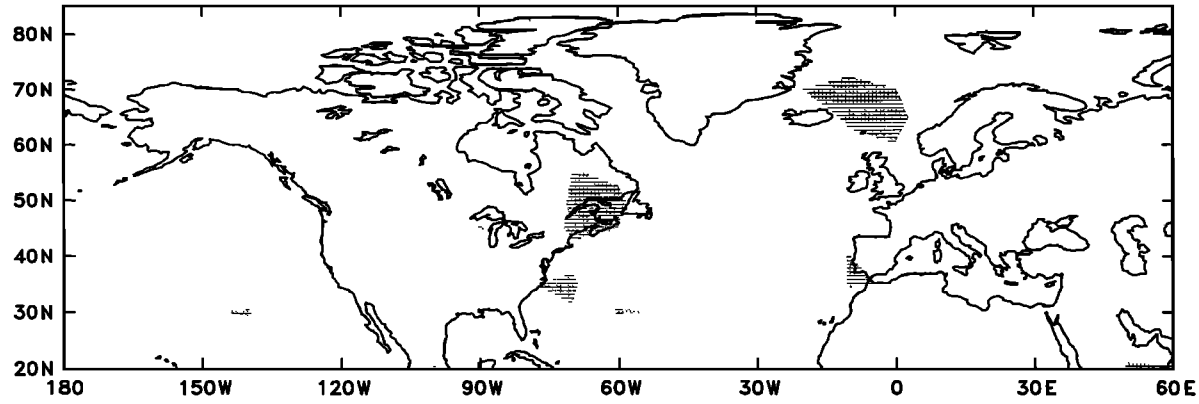

Fig. 10*b*

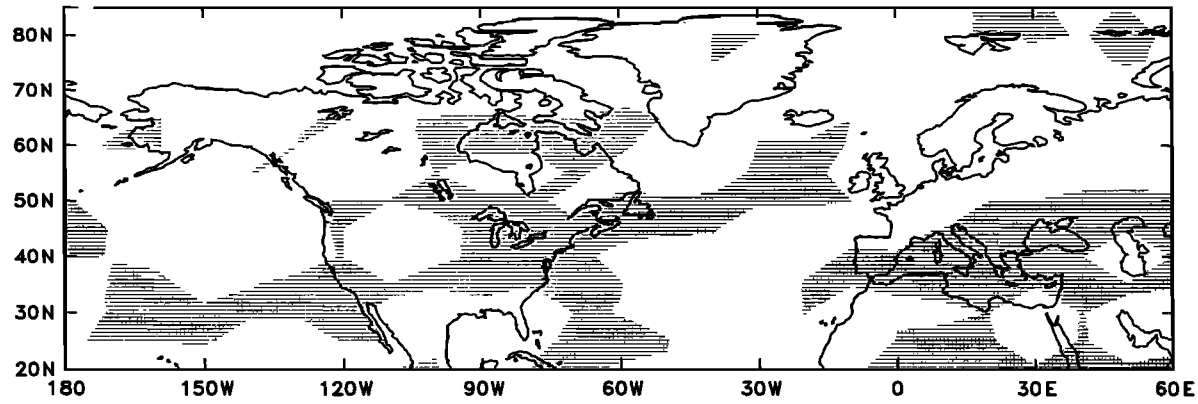

Fig. 10*c*

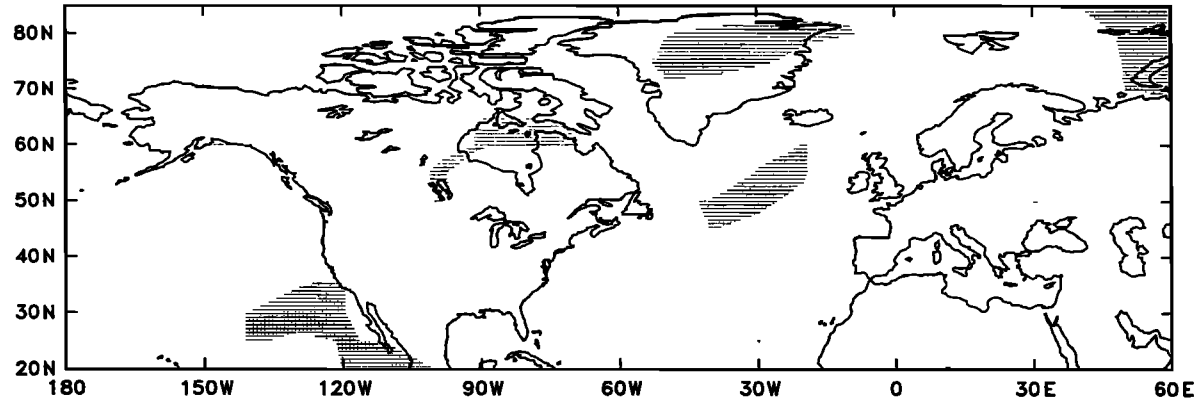

Fig. 10*d*

Fig. 10. Local *F*-test results for July MSLP. Results are for UKMO observed data (1971–1980) versus the (*a*) OSU AGCM (*b*) OSU CGCM, (*c*) GISS AGCM, and (*d*) ECMWF T21 model. Shading indicates areas where differences in variances are significant at or greater than the 5% level. Note that errors in the temporal variance are largest for the GISS AGCM.

TABLE 1. Seasonal Cycle Validation, p Values for 48 Observed Versus Simulated Comparisons

File	NT1	NT5	SITES	T1	NF1	NF5	SPRET1	SPREX1	r
FSJAN	0	0	0	0	0.059	0.002	0	0.097	0
FSFEB	0	0	0	0	0.313	0.018	0.001	0.044	0
FSMAR	0	0	0	0	0	0	0.073	0.669	0
FSAPR	0	0	0	0	0	0	0	0.806	0
FSMAY	0	0	0	0	0	0	0.042	0.126	0
FSJUN	0	0	0	0	0	0	0	0.218	0
FSJUL	0	0	0	0.214	0	0	0	0.498	0
FSAUG	0	0	0	0.008	0	0	0	0.029	0
FSSEP	0	0	0	0	0.007	0	0	0	0
FSOCT	0	0	0	0	0.177	0.009	0.007	0.003	0
FSNOV	0	0	0	0	0.083	0.051	0.208	0.112	0
FSDEC	0	0	0	0	0.008	0	0.005	0.546	0
CGJAN	0	0	0	0	0.002	0	0.040	0.953	0
CGFEB	0	0	0	0	0	0	0.042	0.940	0
CGMAR	0	0	0	0	0	0	0.827	1.000	0
CGAPR	0	0	0	0	0	0	0.058	1.000	0
CGMAY	0	0	0	0	0	0	0.001	1.000	0
CGJUN	0	0	0	0	0	0	0.006	1.000	0
CGJUL	0	0	0	0.036	0	0	0.402	1.000	0
CGAUG	0	0	0	0.095	0	0	0	1.000	0
CGSEP	0	0	0	0	0	0	0	1.000	0
CGOCT	0	0	0	0	0	0	0.006	0.933	0
CGNOV	0	0	0	0	0	0	0.659	0.977	0
CGDEC	0	0	0	0	0.066	0.001	0.001	0.902	0
GSJAN	0	0	0	1.000	0	0	0.547	0.074	0
GSFEB	0	0	0	1.000	0	0	0.760	0.555	0
GSMAR	0	0	0	1.000	0	0	1.000	1.000	0
GSAPR	0	0	0	0.177	0	0	1.000	1.000	0
GSMAY	0	0	0	0.153	0	0	1.000	1.000	0
GSJUN	0	0	0	0.024	0	0	1.000	1.000	0
GSJUL	0	0	0	0.001	0	0	1.000	1.000	0
GSAUG	0	0	0	0	0	0	1.000	1.000	0
GSSEP	0	0	0	1.000	0	0	1.000	1.000	0
GSOCT	0	0	0	1.000	0	0	1.000	1.000	0
GSNOV	0	0	0	0.998	0	0	1.000	0.988	0
GSDEC	0	0	0	0.999	0	0	0.860	0.168	0
ECJAN	0	0	0	0.013	0.139	0.062	0.031	0.910	0
ECFEB	0	0	0	0.001	0.091	0.036	0.168	0.998	0
ECMAR	0	0	0	0	0.197	0.171	0.984	0.864	0
ECAPR	0	0	0	0	0.037	0.011	0.909	0.717	0
ECMAY	0	0	0	0	0	0	1.000	0.224	0
ECJUN	0	0	0	0	0	0	0.975	0.996	0
ECJUL	0	0.001	0	0	0.115	0.011	0.992	0.991	0
ECAUG	0	0	0	0	0.030	0.002	0.773	0.942	0
ECSEP	0	0	0	0.620	0.060	0.002	0.591	0.425	0
ECOCT	0	0	0	0.953	0.097	0.052	0.747	0.442	0
ECNOV	0	0	0	0.089	0.141	0.082	0.926	0.998	0
ECDEC	0	0	0	0.068	0.011	0.003	0.638	0.999	0

Statistics used are those recommended by *Wigley and Santer* [this issue] for comparison of data set means (NT1, NT5, SITES, T1), variances (NF1, NF5, SPRET1, SPREX1) and spatial patterns. The p values were calculated by testing actual test statistic values against reference distributions generated with PPP (with 1000 randomizations of **D** and **M**). The prefixes FS, CG, GS, and EC denote tests involving the OSU AGCM, OSU CGCM, GISS AGCM, and ECMWF T21 model, respectively. For sources of the observed and simulated MSLP data used in the tests, refer to section 2. Note that a p value of zero strictly indicates that $p < 1/N$, where $N = 1000$ is the number of randomizations.

6.2. *Results for SITES*

As for NT1-NT5, all 48 p values are highly significant (Table 1). In each case, we can reject the null hypothesis that **D** and **M** are drawn from populations with identical time-mean fields and accept the alternate hypothesis that the time-mean fields are dissimilar.

The actual values for SITES (Figure 12, left) and NT1-NT5 (Figure 11) clearly identify similar seasonal cycles in the model errors and also identify the same order of model performance. But there are also differences between the actual values of these statistics. SITES and NT1-NT5 sometimes identify different months with time-mean fields most unlike and least unlike the real world (e.g., for the OSU CGCM). In the present case, model errors in the time-mean field are so large that the p values for SITES and NT1-NT5 are always zero or close to zero. In cases where differences in the **D** and **M** time-mean fields are smaller, the p values for NT1-NT5 are almost always lower than for SITES [*Wigley and Santer*, this issue]. This result suggests that SITES has lower power than NT1-NT5 (i.e., higher probability of erroneously accepting the null hypothesis). However, the two statistics do provide different types of information, and there is a need for including both SITES and NT1-NT5 tests in model validation studies.

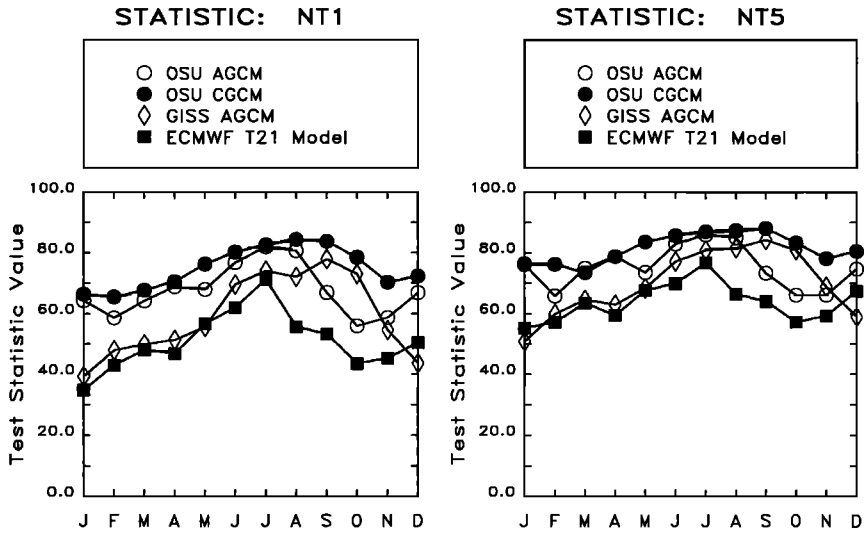

Fig. 11. Actual test statistic values for (left) NT1 and (right) NT5 for the seasonal cycle validation. NT1 and NT5 are the total number of "successful" local two-tailed 1% and 5% t-tests, where success is defined as a locally significant result at the stipulated local significance level. Results are expressed as percentages of the total number of tests performed. Both statistics show that errors in the time-mean field are generally largest for the OSU CGCM and smallest for the ECMWF T21 model.

6.3. *Results for T1*

The p values for T1 provide information about the direction of the overall bias in the simulated time-mean field. Values close to zero (**D** overall mean greater than **M**) or close to 1 (**M** overall mean greater than **D**) are significant. In total, 41 out of 48 results for T1 are significant at the 5% level (i.e., either $p \leq 0.05$ or $1 - p \leq 0.05$; see Table 1). Thirty-three p values show that the observed overall mean is significantly greater than in the models. This type of result occurs most frequently in both OSU models and the T21 model. For the remaining eight significant results, the bias is reversed. Seven of the eight cases in which the **M** overall mean is significantly greater than in **D** occur for the GISS AGCM.

The seven nonsignificant results for T1 are due to large but compensating errors in the simulated time-mean fields. This is clearly shown using the example of the OSU AGCM in July ($p = 0.214$). The AGCM's difference field for this month has large but approximately compensating positive and negative biases (Figure 4). This illustrates the principal disadvantage of the T1 statistic: nonsignificant results do not necessarily indicate overall similarity in the **D** and **M** time-mean fields [*Wigley and Santer*, this issue]. Despite this deficiency, the directional information supplied by T1 is useful. This information cannot be provided by SITES and can only be provided by NTα if one-tailed local tests are performed.

The actual statistic values for T1 are given in Figure 12 (right). Overall biases in the time-mean MSLP field are consistently positive for both OSU models and (with the exception of September and October) the T21 model and are both positive and negative for the GISS AGCM.

6.4. *Results for NF1 and NF5*

The p values for the grid point variance tests indicate that 36 NF1 and 43 NF5 results achieve overall significance at the 5% level (Table 1). In these cases, the null hypothesis that **D** and **M** are drawn from populations with identical temporal variances can be rejected. Results for NF1-NF5 are generally less significant than for NT1-NT5, as expected on the basis of the univariate test results in section 5. This is partly due to the lower power of the F-test and partly due to the fact that errors in the simulation of the interannual variability of MSLP are generally smaller than errors in the simulation of the time-mean field (for the models examined here). (Note that the p values calculated with PPP for NFα and SPRET1 are sensitive to the differences in overall **D** and **M** means. Here, the overall means were not subtracted prior to performing variance tests. Subtraction of the overall means makes the variance test results more significant, since the actual test statistic values remain unchanged but the numerical values of NFα (SPRET1) reference distribution means decrease (become closer to 1.0). Therefore the NFα and SPRET1 p values presented here are conservative estimates of the true significance levels.)

There are considerable intermodel differences in p values. For the GISS AGCM, all NFα results are highly significant, while only five (eight) of the NF1 (NF5) results for the T21 model are significant at the 5% level. Actual statistic values for NF1 and NF5 show that the GISS AGCM's temporal variance errors are consistently larger than those in the other three models (Figure 13). These results are in accord with the univariate F-test results for January and July.

In contrast to the NT1-NT5 actual values, results for NF1-NF5 do not show pronounced seasonal cycles in the model errors, except for the GISS AGCM.

6.5. *Results for SPRET1*

As for T1, p values for SPRET1 have a directional interpretation. Values close to zero (overall **D** temporal variance greater than **M**) and close to 1 (overall **M** temporal variance greater than **D**) are significant. In total, 32 out of 48 SPRET1 results are significant at the 5% level. In 19 cases, the overall variance in **D** is significantly greater than in **M**; in the remaining 13 cases the bias is reversed (Table 1). As

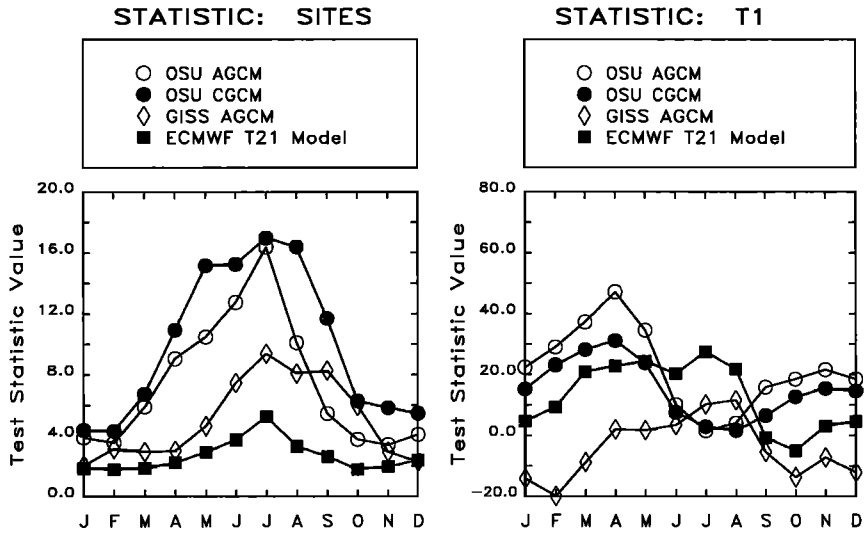

Fig. 12. Actual test statistic values for (left) SITES and (right) T1 for the seasonal cycle validation. The Preisendorfer and Barnett SITES statistic is a measure of the difference in the **D** and **M** time-mean fields. T1 is a measure of the difference in the **D** and **M** overall means. SITES indicates that errors in the time-mean field are largest for the OSU CGCM and smallest for the ECMWF T21 model, while T1 shows that overall MSLP biases are both positive and negative for the GISS AGCM, and are almost always positive for the other three models.

expected on the basis of the variance ratio plots for January and July (Figures 5 and 6), the GISS AGCM has the most cases (nine) of overall **M** variance significantly greater than **D**. The T21 model has both types of significant result. In the OSU AGCM and CGCM, only significant variance underestimates occur.

The actual test statistic values for SPRET1 show an interesting result: the T21 model underestimates the interannual variability in January and February but overestimates the temporal variance in all other months (Figure 14, left). The T21 model's underestimation of the January variance has also been shown by *von Storch et al.* [1985] for the northern hemisphere 500-mbar height field. The model's overestimation of the interannual variability of MSLP in all months except January and February is puzzling in view of the constraints on boundary condition variability imposed by prescribing SST and the temperature and moisture content of the lowest soil layer in a three-layer model. There are several possible explanations.

1. The T21 model's internal dynamical variability ("weather noise") for March–December is higher than in the real world.

2. Certain of the model's nonprescribed components of boundary condition variability (for these 10 months) are unrealistically high, e.g., snow cover and soil moisture content of the upper soil layers.

3. The result is partially fortuitous and depends on the selection of observed data for the decade 1971–1980.

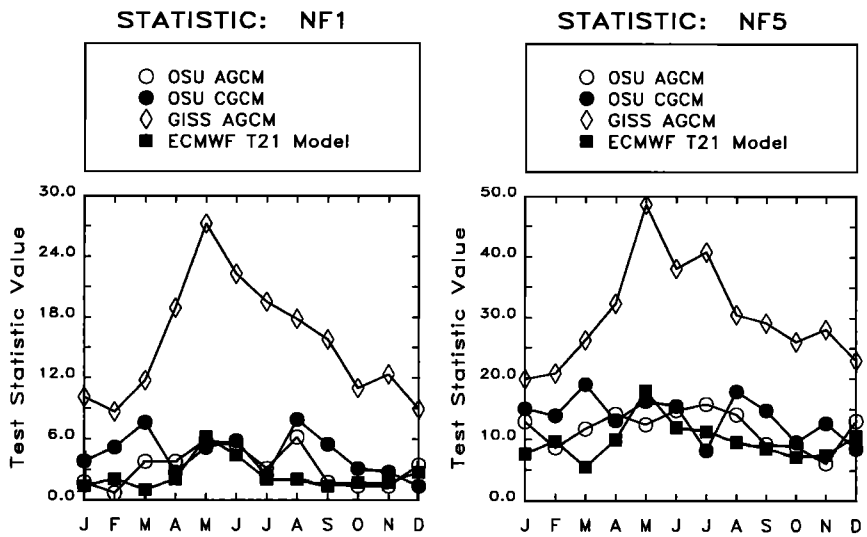

Fig. 13. Actual test statistic values for (left) NF1 and (right) NF5 for the seasonal cycle validation. NF1 and NF5 are the total number of successful local two-tailed 1% and 5% *F*-tests. Results are expressed as percentages of the total number of tests performed. Both statistics show that errors in the interannual variability of MSLP are largest in the GISS AGCM.

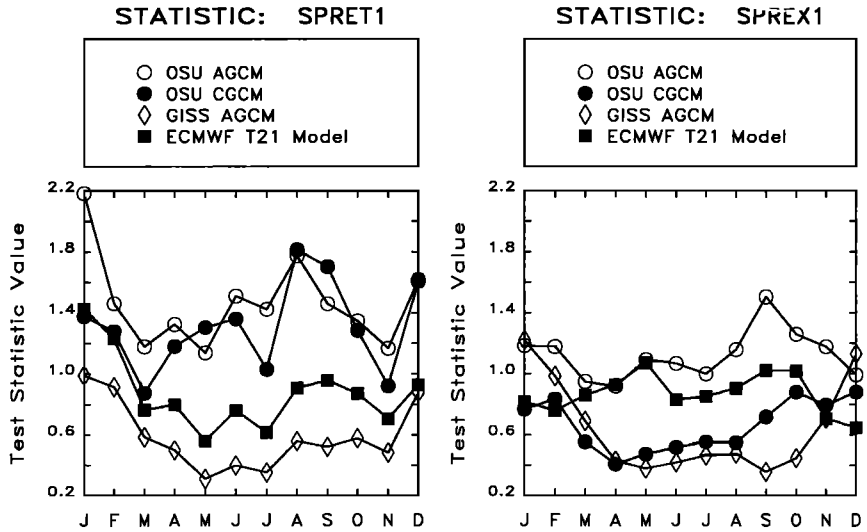

Fig. 14. Actual test statistic values for (left) SPRET1 and (right) SPREX1 for the seasonal cycle validation. SPRET1 is the ratio of the spatially averaged time variances in **D** and **M**, while SPREX1 is the ratio of the time-averaged spatial variances in **D** and **M**. Both statistics provide information on the direction of overall variance biases. Ratios less than 1.0 indicate that the overall **M** variances exceed the overall **D** variances.

6.6. *Results for SPREX1*

Twenty-seven out of 48 SPREX1 results are significant at the 5% level (Table 1). In 23 cases, the overall spatial variance in **M** is significantly greater than in **D** (p value close to 1); in the remaining four cases, the bias is reversed (p value close to zero). As in the case of the other variance ratio statistics, there are considerable intermodel differences in p values. All cases of significant spatial variance underestimates are for the OSU AGCM. Significant spatial variance overestimates occur for the GISS AGCM and OSU CGCM (nine each) and the T21 model (five).

The multivariate SPRET1 results can be readily interpreted in terms of the univariate F-test results. In the case of SPREX1, however, it is more difficult to interpret p values and actual test statistic values by simple visual examination of the time-mean fields and difference fields (Figures 1–4). Clearly, SPREX1 is sensitive to "outliers" such as the spurious Greenland High, and to large underestimates or overestimates of COA intensity, which tend to inflate the spatial variance relative to observations. The poor performance of the GISS AGCM and OSU CGCM in terms of SPREX1 (Figure 14, right) is thus easier to understand, since the Greenland High is most intense in the GISS AGCM and OSU CGCM (maxima of around 1050 and 1031 mbar, respectively; see Figures 1 and 3).

6.7. *Results for r*

While the p values for r indicate that the differences between **D** and **M** time-mean spatial patterns are highly significant in all 48 cases (Table 1), analysis of the actual test statistic values (Figure 15) reveals considerable differences in model performance. As expected from simple visual examination of January and July time-mean MSLP maps (section 3), the T21 model performs consistently better than the other three models in simulating the time-mean spatial pattern. Note that all four models show strong seasonal cycles in the spatial field correlation, despite the prescribing of important boundary conditions in the three AGCMs.

January and July results for the T21 model indicate why r is a useful complement to the standard NTα tests. Although errors in the time-mean field are larger in July than in January (as indicated by NTα and SITES, Figures 11 and 12, left), the T21 model simulates the July spatial pattern with greater fidelity ($r_{JUL} = 0.89$; $r_{JAN} = 0.64$). This result is due to the previously noted January versus July differences in the sign, magnitude, and spatial coherence of the T21 model's MSLP biases (section 3). A significant result for NTα or SITES therefore does not necessarily preclude a nonsignificant result for r.

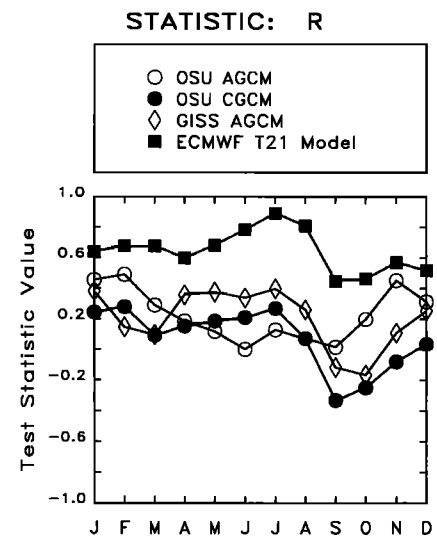

Fig. 15. Actual test statistic values for r (the correlation between observed and simulated time-mean fields) for the seasonal cycle validation. The ECMWF T21 model simulates the time-mean spatial field with the greatest fidelity. All four models have large errors in September.

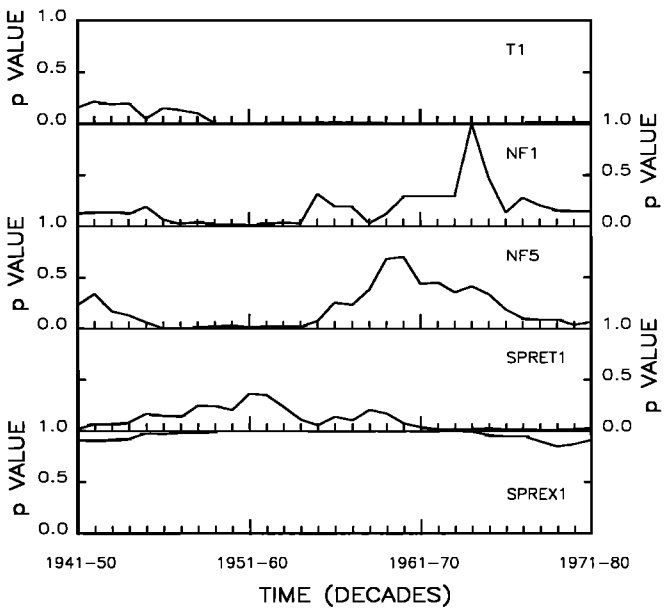

Fig. 16. Sensitivity of p values to decadal time-scale variability in the observed MSLP data. Results are for the ECMWF T21 model January MSLP versus observed January MSLP data for the 31 overlapping decade times from 1941–1950 to 1971–1980. For T1 and the variance ratio statistics NF1, NF5, SPRET1, and SPREX1, decisions on overall significance can depend on the chosen period of observed validation data. In the case of NT1, NT5, SITES, and r (not shown here), results for all 31 decadal tests are significant at the 1% level.

7. Sensitivity of p Values

How sensitive are the p values obtained in the previous section to decadal time-scale variability in the observed MSLP data? This question was addressed using MSLP data for the T21 model only, since section 6 showed that errors in the means, variances, and spatial patterns are generally smaller for this model (and thus closer to the significance threshold) than in the other models considered here. Model MSLP data for one selected month (January) were tested against observed January data for the 31 overlapping decades from 1941–1950 to 1971–1980, using the test statistics and significance testing procedures applied in section 6.

In the case of NT1, NT5, SITES, and r, results for all 31 decadal tests are significant at the 1% level. We conclude that errors in the time-mean field and spatial pattern are so large that the significance levels for these statistics are insensitive to decadal time scale variability in the observed MSLP data. The January p values for T1 are, however, sensitive to the choice of observed validation data. Eight of the 31 values fail to achieve significance at the 5% level (Figure 16), indicating that large but compensating pressure biases must exist for these eight decades (since NTα and SITES are always significant).

For the variance ratio statistics NF1, NF5, SPRET1, and SPREX1, decisions on the overall significance of **D**, **M** differences in temporal and spatial variances are critically dependent on decade-to-decade variations in the observed MSLP data (Figure 16). In the case of NF1 (NF5), nine (ten) out of 31 results indicate that the observed and simulated temporal variances are significantly different at the 5% level (p value close to 0). (Note that one NF1 result (for the decade 1964–1973) shows **D** and **M** temporal variances that are significantly similar (p value close to 1).) For SPRET1 (SPREX1), 12 (21) results achieve significance. Although the decision of whether or not a result is significant depends on the chosen period for the validation data, it is at least reassuring that the direction of the overall variance bias is unaffected by this choice (i.e., all significant SPRET1 and SPREX1 results show **D** temporal variance greater than **M**, and **D** spatial variance less than **M**, respectively). The sensitivity of significance test results to decadal time scale variability in the observed data will become greater as the fidelity of the models improves.

8. Discussion and Conclusions

The set of nine multivariate statistics which were recommended by *Wigley and Santer* [this issue] for routine use in comparison of data set means, variances, and spatial patterns was applied here in a model validation/intercomparison context. These statistics provide complementary information which is easy to interpret and of real diagnostic benefit. Significance levels for all multivariate statistics were determined using the PPP method, which provides a means of circumventing such problems as multiplicity, spatial autocorrelation, unknown reference distributions, and small time samples of model data. In order to fully understand the multivariate significance test results, it is first necessary to examine the magnitude and spatial structure of model errors in means and variances. This was done here with the aid of maps of time mean MSLP, difference fields, local variance ratios, and univariate t- and F-test results.

For all four GCMs considered in this study, errors in the mean field and spatial pattern of MSLP are highly significant throughout the entire annual cycle, and are large enough to ensure that the p values for NTα, SITES, and r are insensitive to decadal time scale variability in the observed MSLP data (at least in January). The fact that some of the results for tests of the **D** and **M** grand means (with T1) are nonsignificant is attributable to the existence of large but compensating errors in the simulated time-mean field.

Errors in the temporal and spatial variance are generally smaller and less significant than errors in the mean field and spatial patterns (except in the case of the GISS AGCM's temporal variance). For the NFα statistics, this result is also related to the lower power of the F-test relative to the t-test. Sensitivity studies with January MSLP data for the T21 model show that p values for NFα, SPRET1, and SPREX1 are critically dependent on the choice of observed MSLP data for validation.

The actual test statistic values reveal considerable differences in model performance. Errors in the time-mean field are largest for the OSU CGCM, which has fully synchronous coupling of atmospheric and oceanic models (without flux corrections) and in which only surface salinity is prescribed. This coupling strategy should allow the CGCM to simulate important atmosphere/ocean interactions. However, it also permits large errors to develop as a result of feedback between errors in the separate (uncoupled) atmospheric and oceanic models. Errors in the time-mean field and spatial pattern are smallest in the T21 model, in which important boundary conditions (SST, deep-soil moisture) are prescribed.

Actual test statistic values (and p values) for SPRET1 and

SPREX1 provide information about the direction of overall biases in temporal and spatial variance. Both OSU models generally underestimate the interannual variability of MSLP, while the GISS AGCM consistently overestimates this property. In the T21 model, temporal variance is underestimated in January and February but overestimated during the rest of the year. This result requires further investigation, as does the similarity of temporal variances in the coupled and uncoupled OSU models (despite fundamental differences in their treatment of boundary condition variability). Differences between the model and observed spatial variances are largely dictated by the magnitude of the spurious Greenland High, simulated by all four models, and by errors in the intensity and location of major COAs.

Model intercomparisons need to be treated cautiously. The four GCMs examined here have different horizontal and vertical resolution and different levels of atmosphere-ocean interaction. More meaningful intercomparisons should involve models of similar resolution and with similar levels of atmosphere-ocean interaction. It is also important to investigate the relation of the statistics and significance testing procedure used here to other statistics (e.g., Hotelling's T^2, Mahalanobis D^2) and validation methods, such as parametric time series modeling [*Katz*, 1982] and univariate and multivariate recurrence analysis [*von Storch and Zwiers*, 1988; *Zwiers and von Storch*, 1988].

Statistical results for tests of means, variances, and spatial patterns cannot positively identify the dynamical cause or causes of the large-scale systematic errors identified here. Only detailed sensitivity studies can provide such insights. The role of rigorous, objective model validation and intercomparison studies is to provide the information which is necessary to design useful and efficient sensitivity studies.

In model intercomparison, we want to determine and compare the significance levels of errors in the means, variances, and spatial patterns which are related to real differences in model physics, resolution, and parameterizations. Unfortunately, intermodel differences which are unrelated to these factors complicate the task of model intercomparison, e.g., the use of different observed data sets for initialization or prescribing boundary conditions. This situation could easily be rectified if standard observed data sets were used by the various modeling groups.

A further difficulty in model intercomparison relates to the conservation of mass. Both spectral and grid point models generally have small changes in the total atmospheric mass at each time step as a result of purely numerical errors (truncation and rounding errors). In most models these mass changes are corrected, e.g., in the Canadian Climate Center (CCC) T20 model (F. W. Zwiers, personal communication, 1988) and in the coupled and uncoupled OSU models [*Ghan et al.*, 1982]. Such corrections were not performed for the T21 model, which therefore fails to conserve mass. Unfortunately, it is difficult to perform a posteriori correction of the T21 model's mass changes, since these are not consistent (either in direction or magnitude) over the entire spatial field. (Note that there is also a strong annual cycle in the T21 model's total atmospheric mass, which further complicates the correction of numerically induced mass changes. This is introduced by temperature errors (particularly over Antarctica) and consequent errors in the reduction of pressure from the lowest sigma level to mean sea level [see *Wigley and Santer*, 1988].) Consistent treatment of mass conservation among the various modeling groups would remove a further reason for ambiguous or misleading results in model intercomparison.

It has been shown that the statistics and the significance testing procedure applied here are useful for validation of simulated MSLP for a limited study area. Future studies should examine the utility of these methods for addressing other problems where rigorous significance testing is appropriate, e.g., in the intercomparison of GCM equilibrium response results for doubled CO_2 and in evaluating the significance of paleoclimate experiments. These methods should also be extended to include other variables, such as surface temperature and precipitation rate, and global-scale fields.

Acknowledgments. We would like to acknowledge the assistance of Larry Gates and Michael Schlesinger at the Climatic Research Institute (CRI), Oregon State University, for providing model results from the OSU AGCM and CGCM and for encouragement in the earlier stages of this work. Data from the GISS AGCM were supplied by Jim Hansen, David Rind, and Gary Russell at the Goddard Institute of Space Studies, New York. The grid transformation programs used were written by Bill McKie at CRI and by Edilbert Kirk at the Max-Planck-Institut für Meteorologie, Hamburg. Rick Katz, Dan Wilks, and Hans von Storch provided useful suggestions and criticisms concerning statistical aspects of this work. Part of this work was funded by the Carbon Dioxide Research Division of the U.S. Department of Energy under grant DE-FG02-86-ER60397 and under contract W-7405-ENG-48 with the Lawrence Livermore National Laboratory.

REFERENCES

Cohen, S. J., Impacts of CO_2-induced climatic change on water resources in the Great Lakes Basin, *Climatic Change*, *8*, 135–153, 1986.

Daley, R., and R. M. Chervin, Statistical significance testing in numerical weather prediction, *Mon. Weather Rev.*, *113*, 814–826, 1985.

Dümenil, L., and U. Schlese, Description of the general circulation model, Climate simulations with the ECMWF T21-model in Hamburg, *Rep. 1*, pp. 3–11, Meteorol. Inst. der Univ. Hamburg, Federal Republic of Germany, 1987.

Gates, W. L., Y.-J. Han, and M. E. Schlesinger, The global climate simulated by a coupled atmosphere-ocean general circulation model: Preliminary results, *CRI Rep. 57*, 31 pp., Clim. Res. Inst., Oreg. State Univ., Corvallis, 1984.

Ghan, S. J., J. W. Lingaas, M. E. Schlesinger, R. L. Mobley, and W. L. Gates, A documentation of the OSU two-level atmospheric general circulation model, *CRI Rep. 35*, 395 pp., Clim. Res. Inst., Oreg. State Univ., Corvallis, 1982.

Grotch, S. L., Regional intercomparisons of general circulation model predictions and historical climate data, *Rep. TR041*, 291 pp., Carbon Dioxide Res. Div., U.S. Dep. of Energy, Washington, D. C., 1988.

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

Hansen, J., A. Lacis, D. Rind, and G. Russell, Climate sensitivity: Analysis of feedback mechanisms, in *Climate Processes and Climate Sensitivity*, *Maurice Ewing Ser. 5*, edited by J. Hansen and T. Takahasi, pp. 130–163, AGU, Washington, D. C., 1984.

Hasselmann, K., On the signal-to-noise problem in atmospheric response studies, in *Meteorology of Tropical Oceans*, pp. 251–259, edited by D. B. Shaw, Royal Meteorological Society, London, 1979.

Jones, P. D., The early twentieth century Arctic high—Fact or fiction?, *Clim. Dyn.*, *1*, 63–75, 1987.

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

Lau, N.-C., Circulation statistics based on FGGE level III-B analyses produced by GFDL, *NOAA Data Rep. ERL GFDL-5*, 427 pp., Princeton, N. J., 1984.

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

Livezey, R. E., Statistical analysis of general circulation model climate simulation, sensitivity and prediction experiments, *J. Atmos. Sci.*, *42*, 1139–1149, 1985.

Livezey, R. E., and W. Y. Chen, Statistical field significance and its determination by Monte Carlo techniques, *Mon. Weather Rev.*, *111*, 46–59, 1983.

MacCracken, M. C., and F. M. Luther (Eds.), *Projecting the Climatic Effects of Increasing Carbon Dioxide*, 381 pp., Carbon Dioxide Research Division, U.S. Department of Energy, Washington, D. C., 1985.

Meinl, H., W. Bach, J. Jäger, H.-J. Jung, H. Knottenberg, G. Marr, B. D. Santer, and G. Schwieren, The socio-economic impacts of climatic changes due to a doubling of atmospheric CO_2 content, contract CLI-063D, Comm. of the Eur. Community, Brussels, and contract V30501-0004/81, Deut. Forsch. und Versuchanst. für Luft- und Raumfahrt, Cologne, 1984.

Mitchell, J. F. B., C. A. Wilson, and W. M. Cunnington, On CO_2 climate sensitivity and model dependence of results, *Q. J. R. Meteorol. Soc.*, *113*, 293–322, 1987.

Oort, A. H., Global atmospheric circulation statistics, 1958–1973, *Prof. Pap. 14*, 180 pp., Geophys. Fluid Dyn. Lab., Natl. Oceanic and Atmos. Admin., Princeton, N. J., 1983.

Parry, M. L., T. R. Carter, and N. T. Konijn (Eds.), *Assessment of Climate Impacts on Agriculture*, vol. 1, *High Latitude Regions*, D. Reidel, Hingham, Mass., 1987.

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

Preisendorfer, R. W., and C. D. Mobley, Data intercomparison theory, II, Trinity statistics for location, spread and pattern differences, *Tech. Memo. ERL PMEL-39*, 91 pp., Pac. Mar. Environ. Lab., NOAA, Seattle, Wash., 1982.

Reed, D. N., Simulation of time series of temperature and precipitation over eastern England, *J. Climatol.*, *6*, 233–253, 1986.

Santer, B. D., Regional validation of general circulation models, *Clim. Res. Unit Publ. 9*, 375 pp., Univ. of East Anglia, Norwich, England, 1988*a*.

Santer, B. D., Validation of sea-level pressure simulated by the ECMWF T21 model for the northern hemisphere, Climate simulations with the ECMWF T21 model in Hamburg, *Large-Scale Atmos. Modelling Rep. 4*, pp. 65–98, Meteorol. Inst. der Univ. Hamburg, Federal Republic of Germany, 1988*b*.

Sausen, R., K. Barthel, and K. Hasselmann, Coupled ocean-atmosphere models with flux correction, *Clim. Dyn.*, *2*, 145–163, 1988.

Schlesinger, M. E., and W. L. Gates, The January and July performance of the OSU two-level atmospheric general circulation model, *J. Atmos. Sci.*, *37*, 1914–1943, 1980.

Schlesinger, M. E., and J. F. B. Mitchell, Model projections of the equilibrium climatic response to increased carbon dioxide, in *Projecting the Climatic Effects of Increasing Carbon Dioxide*, edited by M. C. MacCracken and F. M. Luther, pp. 81–147, Carbon Dioxide Research Division, U.S. Department of Energy, Washington, D. C., 1985.

Schlesinger, M. E., and J. F. B. Mitchell, Climate model simulations of the equilibrium climatic response to increased carbon dioxide, *Rev. Geophys.*, *25*, 760–798, 1987.

Stamus, P. A., Applications of a new verification methodology for regional-scale numerical models, *Coop. Thesis 94*, 147 pp., Univ. of Okla. and Natl. Cent. for Atmos. Res., Boulder, Colo., 1985.

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

von Storch, H., Über die Verifikation atmosphärischer Zirkulationsexperimente, internal report, 112 pp., Meteorol. Inst. der Univ. Hamburg, Federal Republic of Germany, 1985.

von Storch, H., and H. A. Kruse, The extra-tropical atmospheric response to El Niño events—A multivariate statistical analysis, *Tellus*, *37A*, 361–377, 1985.

von Storch, H., and F. W. Zwiers, Recurrence analysis of climate sensitivity experiments, *J. Clim.*, *1*, 151–171, 1988.

von Storch, H., E. Roeckner, and U. Cubasch, Intercomparison of extended-range January simulations with general circulation models: Statistical assessment of ensemble properties, *Beitr. Phys. Atmos.*, *58*, 477–497, 1985.

Wallace, J. M., The climatological mean stationary waves: Observational evidence, in *Large-Scale Dynamical Processes in the Atmosphere*, edited by B. J. Hoskins and R. P. Pearce, pp. 27–53, Academic, San Diego, Calif., 1983.

Wigley, T. M. L., and B. D. Santer, Validation of general circulation climate models, in *Physically-Based Modelling and Simulation of Climate and Climatic Change, Part 2*, edited by M. E. Schlesinger, pp. 841–879, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

Wigley, T. M. L., and B. D. Santer, Statistical comparison of spatial fields in model validation, perturbation, and predictability experiments, *J. Geophys. Res.*, this issue.

Williams, J., and H. H. van Loon, An examination of the northern hemisphere sea-level pressure data set, *Mon. Weather Rev.*, *104*, 1354–1361, 1976.

Willmott, C. J., S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell, and C. M. Rowe, Statistics for the evaluation and comparison of models, *J. Geophys. Res.*, *90*, 8995–9005, 1985.

Zwiers, F. W., Statistical considerations for climate experiments, II, Multivariate tests, *J. Clim. Appl. Meteorol.*, *26*, 477–487, 1987.

Zwiers, F. W., and H. J. Thiébaux, Statistical considerations for climate experiments, I, Scalar tests, *J. Clim. Appl. Meteorol.*, *26*, 464–476, 1987.

Zwiers, F. W., and H. von Storch, Multivariate recurrence analysis, *Rep. 17*, 49 pp., Max-Planck-Inst. für Meteorol., Hamburg, Federal Republic of Germany, 1988.

B. D. Santer, Max-Planck-Institut für Meteorologie, Bundesstrasse 55, 2 Hamburg 13, Federal Republic of Germany.

T. M. L. Wigley, Climatic Research Unit, University of East Anglia, Norwich NR4 7TJ, England.

(Received July 22, 1988;
revised April 14, 1989;
accepted April 28, 1989.)