Recurrence Analysis of Climate Sensitivity Experiments

HANS VON STORCH
Max Planck Institut für Meteorologie, Hamburg, Federal Republic of Germany

FRANCIS W. ZWIERS
Canadian Climate Centre, Downsview, Ontario, Canada

(Manuscript received 9 February 1987, in final form 1 September 1987)

ABSTRACT

A difficulty with the statistical techniques which are ordinarily used in the analysis of climate sensitivity experiments is that they do not identify the stable, or recurrent, aspects of the experimental response. Therefore, a new concept called "recurrence" is proposed. With this concept it is possible to identify the parts of the response which are likely to recur with an a priori likelihood each time a new experimental realization is obtained. A variety of statistical tests which may be used to assess an a priori level of recurrence by means of limited samples is suggested.

A recurrence analysis is performed with data simulated by the Canadian Climate Centre general circulation model forced with climatological sea surface temperatures (SSTs) and with several El Niño SST anomalies. All considered SST anomalies, a positive and a negative doubled standard Rasmusson and Carpenter anomaly and the winter 1982/83 anomaly excite a globally significant response in terms of height and temperature. However, only part of the significant response is also recurrent. In the cold SST anomaly experiment, recurrence is confined to a minor part of the tropics. In the warm SST anomaly runs, recurrence is found in most of the tropics and partly over the northeastern Pacific. These results indicate that equatorial Pacific SST anomalies are associated with a rather limited predictive value, even if the anomalies are very strong.

1. Introduction

The effect of boundary conditions (e.g. sea surface temperature, snow coverage, sea ice extent) on the mean atmospheric flow is generally studied by compositing two ensembles: one, denoted by "control," consisting of atmospheric observed or general circulation model (GCM) generated states not affected by anomalous boundary conditions, and the other, "experimental," consisting of atmospheric states observed together with anomalous boundary conditions or simulated under the constraint of modified boundary conditions. These two ensembles are interpreted as statistically independent samples drawn at random from the two populations designated "control" and "experiment" and are eventually intercompared by more or less standard statistical tests.

A difficulty which we have encountered with ordinary statistical techniques is that they often result in the conclusion that the atmosphere responds in a "significant" way to a change in forcing without appreciably improving our physical insight into the nature of the response. The reason for this frustration is that ordinary statistical techniques focus on the "significance" of the outcome: it is often the case that mean states are significantly different but that individual realizations cannot be readily identified as belonging to either the ensemble of control or experimental climate states. We believe that it would be useful not only to determine whether differences are significant, but also to identify aspects of the experimental response which are characteristic of all or at least most individual experimental states and never or seldom with individual control states. Such knowledge would certainly lead to an improved understanding of the effect of anomalous boundary conditions on the mean atmospheric flow as a consequence of the experiment.

The purpose of the present paper is to address this particular problem in hypothesis testing. We intend to focus on statistical tests in which the null hypothesis is rejected only when there is sufficient evidence that classification of individual climate realizations is possible with a predetermined level of reliability. In principle, the problem must be treated in a multivariate manner. However, we limit ourselves to the univariate case in this paper. The extension of these ideas to the multivariate domain will be described in a future paper.

The basic *t*-test concept is reviewed and discussed in section 2 to contrast with our "recurrence analysis" concept given in section 3. Parametric and nonparametric tests are developed in section 4. An example in which some of the tests are applied to a set of El Niño

Corresponding author address: Dr. Francis W. Zwiers, Canadian Climate Centre/CCRN, 4905 Dufferin St., City of North York, Downsview, Ontario M3H5T4 Canada.

experiments conducted with the Canadian Climate Centre (CCC) GCM is described in section 5 and the paper concludes with some discussion and a summary in section 6.

2. Significance Analysis

The problem with ordinary significance tests which was described in the Introduction has its roots in the usual approach to testing for differences of means. Ordinarily, the observed difference of means is expressed in terms of a t-ratio,

$$T = (\bar{Y} - \bar{X})/[S_p(1/n + 1/m)^{1/2}], \quad (1)$$

where n is the size of the control sample, m is the size of the experimental sample, $\bar{X} = \sum_{i=1}^{n} X_i/n$ is the mean of the control sample, $\bar{Y} = \sum_{j=1}^{m} Y_j/m$ is the mean of the experimental sample, and S_p^2 is the pooled estimate of variance. The last is given by

$$S_p^2 = [\sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{j=1}^{m} (Y_j - \bar{Y})^2]/(n + m - 2). \quad (2)$$

The denominator of the t-ratio is an estimate of the standard deviation of the difference of sample means. That is, in the usual setup, the experimental response $(\bar{Y} - \bar{X})$ is expressed in units of the standard deviation of the difference of means. This measure is then referred to the t-distribution in order to ascribe a level of significance to the difference between the mean states. Differences which are larger than approximately two standard deviations of the difference of means (not the individual realizations) are usually identified as being significant.

Difficulties with interpretation may occur because the unit of measurement goes to zero as sample sizes n and m increase. When samples are large, differences do not have to be large, or physically "significant," to be statistically significant. This is illustrated in Fig. 1 where we display the size of the true difference of means which can be detected with probability 0.90 and 0.50 using the t-test conducted at the 5% significance level. It is assumed in this illustration that control and experimental climate samples are equal in size. With samples of size 30 it is possible to detect differences as small as one-half of a standard deviation with a probability of at least 0.5. The size of difference which can be detected with a given level of reliability goes to zero as $1/n^{1/2}$.

Another way to illustrate these ideas is shown in Fig. 2. In schematic form we see the distribution of a control and an experimental ensemble (the solid and dashed curves labeled $n = 1$) and corresponding sampling distributions of the ensemble means for samples of 10 (curves labelled $n = 10$) and 50 (curves labelled $n = 50$).

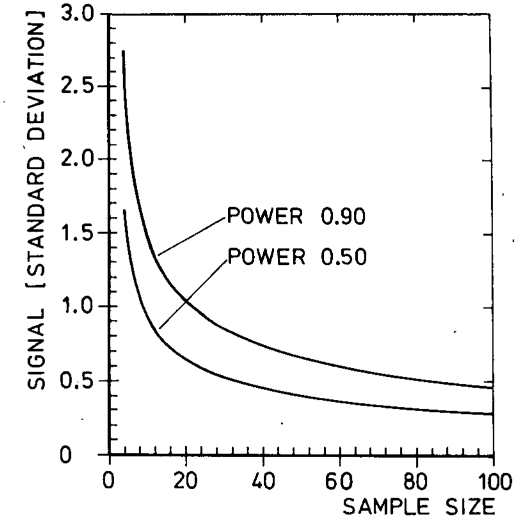

FIG. 1. Departure from null hypothesis of equality of means which is detectable with probability 0.90 and 0.50 using t-test at the 5% level displayed as a function of sample size $n = m$.

In the illustration, the difference between the locations of the two ensembles is one-half of a standard deviation. We see a considerable amount of overlap in the two ensembles. A large proportion of experimental states can occur under control conditions and vice versa. However, as the sample size increases, we see that the distributions of the sample means decrease in spread, and that eventually there is virtually no overlap between the distributions of the control and experimental sample means. Under these conditions we can distinguish control and experimental states with almost perfect reliability. Thus, given a large enough sample we will be able to state with confidence that experimental and control states cluster about different long-term means. However, given a particular realization, we would still be hard pressed to classify it as belonging to one or other of the two climates unless it came from the extreme left-hand tail of the control distribution or the extreme right-hand tail of the experimental distribution.

3. Recurrence analysis

We see then, that for many purposes, the usual hypothesis testing setup does not adequately address the problem at hand. For example, when conducting El Niño experiments with a GCM, we are certainly interested in knowing that the mean climate state is different during El Niño years than during other years. In addition, we are also interested in characterizing differences which are likely to recur during each El Niño episode with a similar SST anomaly, so that useful forecasts of the characteristics of the *next* actual or GCM-simulated El Niño event can perhaps be made. To place this in the context of the previous illustration, we are interested in differences which are relatively large

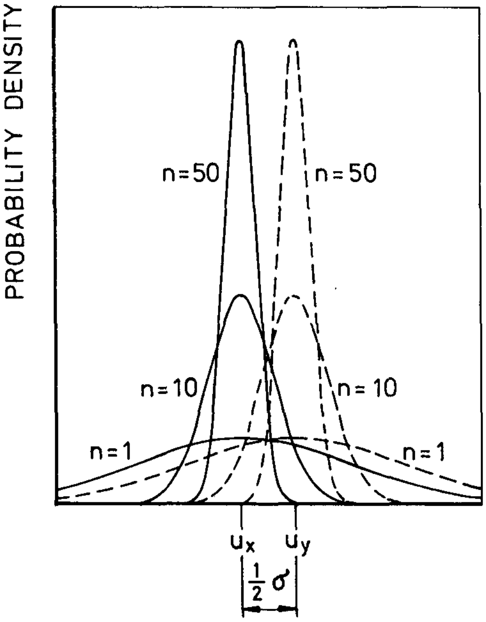

FIG. 2. An example of frequency distributions of control and experimental realizations (solid and dashed curves labelled $n = 1$) and corresponding frequency distribution functions of the means of samples of size 10 (curves labeled $n = 10$) and size 50 (curves labeled $n = 50$).

in the sense that the distributions of control and experiment climate states are well separated so that classification of climate realizations can be done with relatively high reliability.

This raises the question of how large a difference is of interest when characterizing the response of the atmosphere or a GCM to anomalous boundary forcing. Identifying the undisturbed, "normal" states with a random variable X and the states affected by anomalous boundary conditions with another random variable Y, we propose the use of the following terminology:

Definition: The random variables X and Y are said to be (p, q)-recurrent if

$$\text{Prob}(Y > X_p) > q \qquad (3)$$

where X_p is the $p \cdot 100$th percentile of the X-distribution which is defined by $\text{Prob}(X < X_p) = p$.

Throughout this paper we will state probabilities either as a percentage or as a proportion. The convention used should be clear from the context.

The word recurrence refers to the probability q of observing a response which is greater than the reference level X_p the next time an experimental realization is generated. A large value of q indicates that it is very likely that this will happen. The role of the p-value is essentially to provide a reference value, namely the $p \cdot 100$th percentile of the control ensemble, against which the strength of the response is measured. The probability q is clearly a monotonic function of p for a given degree of separation between the control and experimental ensembles.

The idea of (p, q)-recurrence is illustrated schematically in Fig. 3. In this diagram X_p represents a point on the right-hand tail of the control distribution X. By construction, the proportion of control states less than X_p is p. This point also represents a point on the left-hand tail of the experimental distribution Y and according to (3) the proportion of experimental states greater than X_p is q. Thus, the definition states that a response is (p, q)-recurrent if there is a point between the control and experimental ensembles such that proportion p of the control ensemble lies to one side of that point and proportion q of the experimental ensemble lies to the other side of that point. If p and q are both close to 1, then the two ensembles are almost perfectly separated. On the other hand, if the ensembles are symmetrical, then $p = q = 0.5$ would indicate that the means are exactly equal and that possibly only the variance, or spread, of the two ensembles could be different.

An interesting property may be derived from Fig. 3; if the distributions of X and Y are identical except for location and if they are symmetrical, then (p, q)-recurrence is equivalent to (q, p)-recurrence. This relation will be explored in section 4.

Another way in which to understand the idea of (p, q)-recurrence is to think of a classification problem. Given a realization Z, we are to classify it as belonging to one of two ensembles: the control X or the experimental Y. The classification is performed by choosing a threshold value, X_p, and making the decision "the observed realization belongs to the control ensemble" if $Z < X_p$. If $Z > X_p$, we make the decision "the observed realization belongs to the experimental ensemble." By choosing p we essentially chose the probability $(1 - p)$ of incorrectly determining that Z belongs to the experimental ensemble when in fact it belongs to

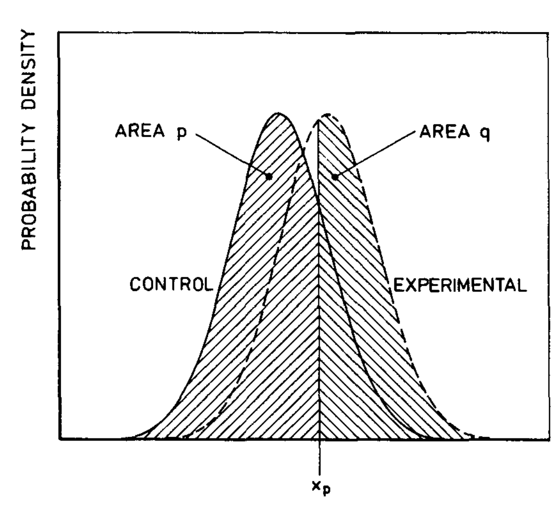

FIG. 3. Schematic diagram illustrating the notion of (p, q)-recurrence.

the control ensemble. Probability q is the probability of correctly deciding that Z belongs to the experimental ensemble when this is in fact the case. Probability q is determined by p and the degree of separation between the two ensembles. This interpretation of (p, q)-recurrence can be summarized in a decision table as in Table 1. When p and q are both close to 1 the ensembles are well enough separated that the next realization can be classified as belonging to either the control or experimental ensemble with only small probabilities of error.

A given degree of overlap between two ensembles can be described by a continuum of (p, q) pairs. As an illustration, consider the example shown in the left half of Fig. 4. In this example both control and experimental ensembles are Gaussian with variance σ^2 and means μ_x and $\mu_y = u_x + \sigma$. Thus, the two ensembles are separated by σ. The control and experimental density functions intersect at $X_p = \mu_x + 0.5\sigma$ which is equal to $\mu_y - 0.5\sigma$. Sixty-nine percent of control climate realizations lie to the left of this point and 69% of experimental climate realizations lie to the right. Thus, the experimental response is (p, p)-recurrent with $p = 69\%$. However, we could choose any point between the two means to describe this degree of recurrence. For example, as illustrated in the right half of Fig. 4, 84% of experimental realizations lie to the right of the control ensemble mean μ_x. Thus the response is also (50%, 84%)-recurrent.

The degree of recurrence can be described uniquely by just one number p by finding the p for which the response is (p, p)-recurrent. This will simply be denoted as p-recurrent. If both ensembles have the same shape and spread (i.e., differ only in location) and if both ensembles are symmetric, then this number is given by the point at which the two ensembles' density functions intersect. This point will have the same proportion p of control states to its left as it has experimental states to its right. Table 2 contains some p and q which describe equivalent (50%, q)- and p-recurrence for Gaussian ensembles with equal variance.

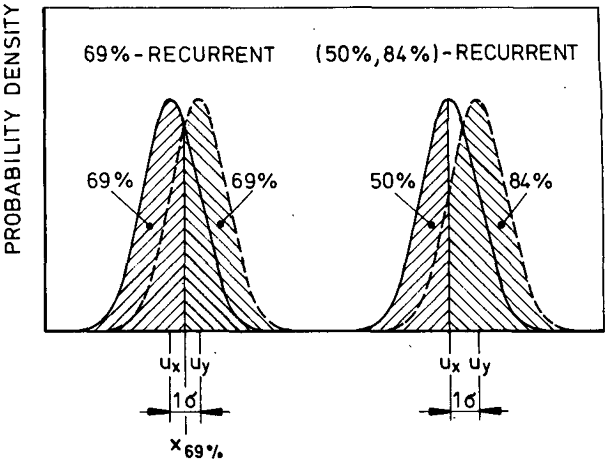

FIG. 4. Schematic diagram illustrating the idea that 69%-recurrence is equivalent to (50%, 84%)-recurrence.

Suppose now that both ensembles X and Y have a Gaussian distribution with variance σ^2. The point of intersection of the control and experimental density functions is $(\mu_x + \mu_y)/2$, which may be written as $\mu_x + z\sigma$ or $\mu_y - z\sigma$ where $z = (\mu_y - \mu_x)/2\sigma$. The point of intersection may also be expressed as $(\mu_x + \mu_y)/2 = \mu_x + Z_p\sigma = \mu_y - Z_p\sigma$, where Z_p is the percentile of the standard Gaussian distribution which has value z. Thus the response is p-recurrent. If the control and the experimental means are separated by approximately two standard deviations, the response is about 84%-recurrent. A response which is 98%-recurrent indicates that the ensembles are separated by about four standard deviations, because the 98th percentile of the standard Gaussian distribution is approximately 2 and suggests that an observed climate state could be almost unequivocally classified as belonging to either the control or experimental ensemble. More generally, the following equivalence holds:

$$X, Y\ p\text{-recurrent} \leftrightarrow \mu_y - \mu_x > (\sigma_x + \sigma_y)\cdot Z_p. \quad (4)$$

The definition given above anticipates that the response to the experimental conditions is to move the ensemble of experimental states to the right of the ensemble of control states. If the opposite response occurs, then the response is (p, q)-recurrent if

$$P(Y < X_{1-p}) > q. \quad (5)$$

4. Statistical tests

To test whether a control and an experimental ensemble are (p, q)-recurrent or p-recurrent, we assume

TABLE 1. A decision table illustrating the idea of (p, q)-recurrence.

	Z is actually from	
Decision	Control ensemble	Experimental ensemble
Z belongs to control ensemble	p	$1 - q$
Z belongs to experimental ensemble	$1 - p$	q

TABLE 2. The correspondence between (0.5, q)-recurrence and (p, p)-recurrence assuming that both the control and experimental ensembles are Gaussian with the same variance. The third column expresses the recurrence in terms of the distance between control and experimental means as expressed in standard deviations.

q	p	Separation
.692	.599	0.5
.841	.692	1.0
.933	.773	1.5
.977	.841	2.0
.994	.894	2.5
.9987	.933	3.0
1.0000	.977	4.0

that we have a sample $X_1, X_2, \ldots, X_n$ of control states, and a sample $Y_1, Y_2, \ldots, Y_m$ of experimental states, and that all observations are mutually statistically independent. Seasonal means derived from atmospheric GCM climate simulations, and to a lesser extent from observations, satisfy these assumptions approximately.

a. Parametric tests

To construct a parametric test we adopt a model for the populations of control and experimental states, namely both populations are Gaussian with approximately the same variance, so that the only difference between the two populations lies in their locations.

We want to derive a scheme allowing for a decision "X and Y are p-recurrent" with prescribed, small risk. Therefore, using (4), the following pair of hypotheses is reasonable:

null hypothesis $\quad H_0$: $\mu_y - \mu_x < 2\sigma Z_p$ (6a)

alternative hypothesis $\quad H_a$: $\mu_y - \mu_x > 2\sigma Z_p$. (6b)

Under null hypothesis (6a), the test statistic T given by (1) has a noncentral t-distribution with $n + m - 2$ degrees of freedom (df) and noncentrality parameter Δ no larger than

$$\Delta = 2Z_p/(1/n + 1/m)^{1/2}. \qquad (7)$$

Therefore, in order to test (6) at a significance level no greater than α, we should compare the computed T with the $(1 - \alpha)\cdot 100$th percentile of the noncentral t-distribution with $n + m - 2$ df and noncentrality parameter Δ. Pearson and Hartley (1976, p. 242) provide tables which can be used to determine the critical values of the noncentral t-distribution. Alternately, the IMSL (1982) subroutine MDTN can be used to find the probability of observing a noncentral t-value greater than the observed T for a given noncentrality parameter. For large sample sizes, the critical values $t_{n+m-2,\Delta,1-\alpha}$ of the noncentral t-distribution for testing hypotheses (6) are given by

$$t_{n+m-2,\Delta,1-\alpha} = Z_{1-\alpha} + 2Z_p/(1/n + 1/m)^{1/2}. \qquad (8)$$

This expression illustrates the dependence of the critical value upon the sample size when samples are large: the difference of means, as expressed in units of standard deviations of the difference of means, must also be large in order to reject the null hypothesis (6).

As an example, the El Niño experiments described below resulted in a sample of $n = 30$ control climate realizations and $m = 5$ experimental climate realizations. Assuming that the response is positive, testing that the response is at least 84%-recurrent is equivalent to testing that the difference of means is at least 2σ. Using Pearson and Hartley's table we see that the null hypothesis can only be rejected at the 5% significance level if $T > 6.25$. That is, we demand considerably stronger evidence to conclude that the response is at least 84%-recurrent than to conclude that the means are significantly different (i.e., the response is more than 50%-recurrent). Rejection of the null hypothesis in the former case is much more informative about the nature of the response. The asymptotic critical value derived from (8) is 5.785, indicating that the use of (8) makes the test somewhat too liberal. Testing that the response is at least 98%-recurrent is equivalent to testing for a difference of means of at least 4σ. In this case the null hypothesis can only be rejected if $T > 11.1$.

The discussion to this point has dealt only with one-sided tests. A two-sided test of the hypotheses

$$H_0\colon |\mu_y - \mu_x| < 2\sigma Z_p \quad \text{vs} \quad H_a\colon |\mu_y - \mu_x| > 2\sigma Z_p \qquad (9)$$

with significance level no greater than α can be conducted by comparing the absolute value of the computed T with the $\alpha/2$ critical value of the noncentral t-distribution.

b. Nonparametric tests

A nonparametric test can be constructed with less constraining assumptions than the parametric test described above at the expense of some power. We will discuss three such approaches in this section. To do this we will need to make some assumptions. For the most part we will assume that the two samples are taken from ensembles which are identical except for their locations.

A simple approach makes use of the ideas on which the "sign" test is based. To test (p, q)-recurrence,

$$H_0\colon P(Y > X_p) < q \quad \text{vs} \quad H_a\colon P(Y > X_p) > q \qquad (10)$$

is an appropriate pair of hypotheses. We will use

$$T = \text{number of } Y \text{ greater than } X_p \qquad (11)$$

as the test statistic. If H_0 holds, T has a binomial distribution with m trials and probability of success q. The strongest evidence that the null hypothesis should be rejected is obtained when we observe $T = m$, and a test which rejects the null hypothesis only when $T = m$ will have significance level q^m.

Since any (p, q)-recurrence is equivalent to $(0.5, q')$-recurrence for some uniquely determined q', it is reasonable to limit ourselves to $(0.5, q')$-recurrence. Unfortunately the desired large value of q' is in conflict with the desired small significance level q'^m: if the number m of experimental samples is small ($m = 5$ in our case), the smallest possible significance level of this test is dauntingly large unless q' is close to 0.5. That is, $T = m$ does not provide sufficient evidence to conclude anything other than the means of the two ensembles are different.

An improvement is possible if one makes use of the equivalence of (p, q)-recurrence and (q, p)-recurrence noted in section 3. By doing so one implicitly makes the additional assumptions that control and experimental distributions are symmetric and differ only in the mean. With these assumptions, it is reasonable to test for (q', 0.5)-recurrence, that is to use (11) with q' instead of 0.5. In this case the test statistic T is the number of Y larger than the $q' \cdot 100$th percentile of X. Thus, in the case studied in section 5 with $m = 5$, it is possible to derive the result $P(Y > \mu_x) > q'$ (i.e. (0.5, q')-recurrence) with risk less than $0.5^m = 3\%$ if all Y are larger than the q'-quantile of X.

Even if (0.5, q')-recurrence and (q', 0.5)-recurrence are equivalent, the proposed nonparametric tests differ. The form of the test statistic, and hence its power, depends upon the form of the null hypothesis. Suppose, for example, that the samples are taken from Gaussian distributions, that the alternate hypothesis is that the response is at least (0.5, 0.84)-recurrence, and that in fact the response is (0.5, 0.98)-recurrent. The significance level of the test for (0.5, 0.84)-recurrence is 0.84^m (approximately 0.05 if $m = 17$) if the null hypothesis is rejected only when $T = m$. In this case the power of the test (the probability of rejecting the null hypothesis) is 0.98^m (=0.71 if $m = 17$). When testing for (0.84, 0.5)-recurrence the significance level of the test will be approximately 0.025 if we reject the null hypothesis whenever $T > 12$ (if $m = 17$) and the power will be approximately 0.89. Thus we see that although the significance levels of the two tests are comparable, their sensitivity is considerably different. In particular, the test for (q', 0.5)-recurrence is more powerful that the test for (0.5, q')-recurrence.

As noted above, the test for (0.84, 0.5)-recurrence has a significance level of approximately 0.03 when $m = 5$ and if the null hypothesis is rejected when $T = m$. In this case the probability of rejecting the null hypothesis when the response is in fact (0.5, 0.98)-recurrent is approximately 42%, indicating that the proposed test is not very powerful for these small sample sizes.

A practical consideration is that the threshold value $X_{q'}$ in (11) is not known and must be estimated from the sample of control climate states. This introduces a source of variation which the binomial distribution does not take into account and hence will alter the significance level of the test from the nominal level which is determined from this distribution. To minimize the effects of this source of variation it may be appropriate to fit a parametric model, such as the Gaussian distribution, to the sample of control climate states and to use the fitted model to estimate $X_{q'}$. The resulting test is no longer precisely nonparametric, but still has some of the advantages of a nonparametric test. The parametric model is fitted only to the larger of the two samples (which we assume to be the control sample), and the information in this sample can be used to assess the goodness of the fit of the parametric model prior to any recurrence testing.

The results of the test for (q', 0.5)-recurrence should be interpreted with care. It only tests explicitly for (q', 0.5)-recurrence. In that context, fitting the Gaussian distribution to the control sample has no implications for the experimental sample. However, the test for (q', 0.5)-recurrence was motivated in the context of a test for (0.5, q')- or q-recurrence. If we interpret the results in terms of (0.5, q')- or q-recurrence then the assumptions that both ensembles are symmetric and differ only in mean are made implicitly. In this case fitting the Gaussian distribution to the control sample implies that the experimental ensemble is also approximately Gaussian and that both ensembles have the same variance.

Finally, when testing for (q', 0.5)-recurrence the occurrence of $T = m$ is extremely unlikely if $\mu_x = \mu_y$, namely $(1 - q')^m$. Even if we perform the analysis at, say, 500 statistically independent locations, the probability of observing this result by chance at just a single point is less than 0.02% if $q' = 95\%$ and $m > 2$. Thus, the presence of just one (0.5, 0.95)-recurrent point identified by $T = m$ (for $m > 2$), is almost certain proof that the whole signal is significant in a multivariate sense.

The tests can be made two sided by using as the test statistic

$$T = \max(T_1, T_2) \qquad (12)$$

where T_1 is the number of Y less than X_{1-q} and T_2 is the number of Y greater than X_q. To have a test with significance level at most α, this statistic is then compared with the critical value for the binomial distribution with m trials and probability p of success which leads to a one-sided test with significance level at most $\alpha/2$.

The nonparametric tests described above are constructed by counting the number of experimental outcomes in excess of a certain threshold. This threshold is derived from the control outcomes, either directly or by fitting a parametric model. This approach uses the information about the magnitudes of the control and experimental outcomes, and this information leads to some difficulty and the necessity of adopting a parametric model for the control climate states. A way to avoid this difficulty is to replace the observations with their relative ranks in the combined pool of control and experimental outcomes. One could conduct a test by rejecting the null hypothesis of (0.5, q)-recurrence if the smallest (largest) experimental outcome was greater (less) than the largest (smallest) control outcome. Such an outcome represents very strong evidence that the control and experimental distributions are well separated. The probability of such a result at a single point is $2 \cdot n!m!/(n + m)!$ when both distributions are symmetric and have the same mean. In our example, in which $n = 30$ and $m = 5$, this probability is approximately 6×10^{-6}. Thus again, we would argue that in the example the presence of just one such result in a field is almost certain proof that the whole signal is significant in a multivariate sense. The derivation of the significance level of this test is given in the Appendix. Table 3 contains the significance level of this test as a function of q for the case in which both ensembles have Gaussian distributions with the same variance and in which $n = 30$ and $m = 5$. Under these circumstances we see that this test has significance level 0.05 under the null hypothesis that the response is at most (0.5, 0.97)-recurrent, or equivalently, that the means of the control and experimental ensembles are separated by at most 1.88 standard deviations.

A third approach to constructing a suitable test would be to attempt to duplicate the noncentral difference of means test (6) of the previous section, using a nonparametric test. Assuming that the ensembles are in fact not far from being Gaussian, then testing hypotheses (6) is approximately equivalent to testing hypotheses (10) according to (4). If we knew σ, then we could test (6) using a one-sided Mann–Whitney test (e.g. Conover, 1980) by first subtracting $2\sigma Z_p$ from the experimental realizations and then proceeding in the usual way.

5. Application—El Niño sensitivity GCM experiments

To describe the merits of the proposed recurrence analysis, we analyze data from sensitivity experiments performed with the Canadian Climate Centre GCM (Boer et al., 1984a,b). This model was integrated under regular boundary conditions for a total of 30 winters. A series of runs was also performed in which three different El Niño-type SST anomalies were added to the climatological annual cycle of the SST.

a. The GCM experiments

The imposed SST anomalies were

(i) Two times the Rasmusson and Carpenter (1982) standard El Niño anomaly. This anomaly is confined to the equatorial Pacific and is positive everywhere. We will we refer to it in this paper as "2RC".

(ii) The same as 2RC but with reversed sign. We refer to this anomaly as "−2RC".

(iii) The 1982–83 global SST anomaly as analyzed by the European Centre for Medium Range Weather Forecasts (ECMWF). We refer to this anomaly as "82/83".

TABLE 3. The significance level of the nonparametric test for (0.5, q)-recurrence which rejects the null hypothesis when $Y_{(1)} > X_{(n)}$. The sample sizes are $n = 30$ for the control ensemble and $m = 5$ for the anomaly ensemble. The significance level is computed assuming that both control and experimental ensembles are Gaussian and have the same variance.

q	Separation	Significance level
.692	0.5	.0001
.841	1.0	.0016
.933	1.5	.0133
.977	2.0	.0652
.994	2.5	.2001
.9987	3.0	.4186
1.0000	4.0	.8324

A number of results obtained from the 82/83-experiment are given by Boer (1985).

The positive equatorial Pacific SST anomaly of 82/83 is similar to 2RC. However, the 82/83 SST regime also contains extratropical SST anomalies, negative SST anomalies in the western part of the equatorial Pacific and a large-scale anomaly of about +1°C in the equatorial Indian Ocean. Both equatorial areas are relevant with respect to the model's response, because normal SSTs are quite high in the Indian Ocean and the western equatorial Pacific, namely about 28°C, and, more importantly, because of the existence of low-level convergence associated with heavy convective activity in these areas.

For each SST anomaly a total of five different, statistically independent winter seasons were simulated. Thus, for each experiment, a control ensemble of 30 and an experimental ensemble of 5 was available. Considered are the temperature and the geopotential height at 500 mb.

From sensitivity experiments performed with other GCMs we may anticipate the following results. The tropical response is strong and clearly statistically significant in contrast to the extratropical response, which, in terms of geopotential height, is often identified as a Pacific–North America (PNA) pattern or at least as being reminiscent of such a pattern. GCMs have turned out to be quite sensitive to SST anomalies in regions with normally high SST and low-level convergence (e.g. Palmer and Mansfield, 1984; Storch et al., 1987). Thus we expect that the 82/83 response will be different from the 2RC response not only in the extratropics but also in the tropics. With respect to −2RC, we may expect a tropical signal, which is similar to that of 2RC but somewhat weaker and with reversed sign (Blackmon et al., 1983; Cubasch, 1985). In the extratropics, the −2RC signal will most likely be weaker and less statistically significant than that of 2RC. Also, the more remote response will have a different mean pattern, which is associated with a considerable intrasample variability (Storch and Kruse, 1985).

b. Statistical significance of the responses

A disadvantage of the proposed recurrence analysis is the fact that it is based solely on univariate considerations. It could happen that we find erroneously significant recurrence simply because of the large number of analyses performed at the various grid points (Storch, 1982; Livezey and Chen, 1983). To lessen the likelihood of the occurrence of such unwanted events we perform a regular multivariate test procedure to determine whether the mean of the control and the experimental ensembles are statistically significantly different from zero prior to the pointwise recurrence analysis.

We used two different two-step techniques. The first step of the first procedure (Livezey and Chen, 1983; Zwiers, 1987) is to count the number of locally significant differences. In the second step, the probability of obtaining this number from two samples taken from ensembles with identical means is assessed by means of a permutation procedure. The advantage of this approach for the recurrence analysis is that one need only calculate the t-statistic at all grid points. These t-statistics may then be used for both the multivariate test above and the parametric recurrence test [hypotheses (6) and test statistic (1)]. The computed t-statistics are displayed in Figs. 5 and 6 for 500 mb height and temperature respectively. Regions in which the local test for equality of means can be rejected at the 5% significance level are hatched. According to the multivariate test, the response which is shown in these diagrams is highly significant with a risk less than 1% in all three experiments.

The second procedure was proposed by Storch and Kruse (1985). Its first step is to project the whole fields on a small number of fixed "guess patterns" which are determined a priori from scale reasoning, or problem-related experience derived from observations, or from similar but statistically independent GCM experiments (Storch, 1987).

Here, we used a series of winter mean 500 mb height fields derived from National Meteorological Center (NMC) analyses covering 1949 to 1985. The major cold and warm events were extracted from the whole dataset and their 500 mb height anomalies calculated. Because the NMC analyses are almost all confined to the Northern Hemisphere extratropics (north of 20°N), the tropics, and thus the strongest part of the multicomponent signal, is excluded. Nevertheless, the response in all three experiments was significant according to this test. As a by-product, we found for 500 mb height that

(i) The 2RC response was parallel to the 1965/66 warm event anomaly.

(ii) The −2RC response was parallel to the 1955/56 cold event anomaly and antiparallel to the 1982/83 warm event anomaly.

(iii) The 82/83 response was parallel to the warm event winter mean anomalies of 1957/58 and 1982/83 and antiparallel to the 1963/64 warm event anomaly.

c. Recurrence analysis

We used the parametric hypotheses (6) with test statistic (1) to screen the data for p-recurrence and the nonparametric hypotheses (10) with test statistic (11) to study formally (q, 0.50)-recurrence which we assume to be approximately equivalent to the more relevant (0.50, q)-recurrence.

1) PARAMETRIC APPROACH

We tested the data for a separation of control and experimental means of at least two standard deviations, which, under relatively weak assumptions (see above), is equivalent to 84%-recurrence or to a probability of at least 98% that a random experimental sample is larger than the control mean [that is (50%, 98%)-recurrence; see Table 2]. As already mentioned in section 3, the critical t-value for $n = 30$, $m = 5$ and a significance level of 5% is 6.25.

Even if 84%-recurrence is a rather strong requirement, the criterion is fulfilled at a number of points, which are almost all located in the tropical belt. The areas in which the response is found to be at least 84%-recurrent at the 5% significance level are indicated by the contours in Figs. 5 and 6. These areas are, of course, small relative to the regions in which the response is locally significant. The latter regions cover a majority of grid points in all three experiments.

In the "positive" experiments, namely 2RC and 82/83 (Figs. 5a, b) we see increased height between 20°S and 20°N, over northeastern Canada, part of central Europe and the tip of South America, and decreased height along the North American west coast, over most of the midlatitude Southern Hemisphere and part of central Siberia. In 2RC the response is roughly 84%-recurrent everywhere in the zonal belt between, say, 15°N and 15°S, and over the southwestern part of North America. However, the same magnitude of response occurs in a much smaller equatorial area in the 82/83 experiment. The reduced height of the 500 mb topography over the northeast Pacific in the 82/82 experiment is connected with a minimum $t = -5.72$, which is smaller in magnitude than the critical value of -6.25 and much smaller in magnitude than the minimum $t = -8.36$ which is found in this area in 2RC. Also, in the equatorial belt, the 2RC t-statistics are much larger in magnitude than in the 82/83 experiment. In this region t is less than -8 everywhere on the equator and has a minimum of -13.18. In the 82/83 experiment, no more than about 5% of all points are associated with $t < -8$ and the minimum value is -9.4.

The locally significant temperature response (Figs. 6a, b) is most pronounced in the tropics in the 20°N–20°S belt, with gaps or minima over the Atlantic and at about 140°E. The temperature is reduced significantly in 2RC in a local sense over most of Siberia and the northern Pacific to the North American coast, and over Australia. The portion of this area associated with 84%-recurrence is restricted in the tropics to the longitudes 160°E–40°W (2RC) and 150°−100°W (82/83). However, in 2RC, there are three extratropical regions with recurrently reduced temperatures: western Siberia, the northeastern Pacific, and Australia. The Siberian and Pacific recurrent temperature responses are similar to the height responses, but the Australian response is not present in the height fields.

In the "negative" experiment −2RC (Fig. 5c), the 500 mb height topography is reduced significantly in a local sense all around the equatorial belt, over northern Canada and in part of Siberia. An increase in height

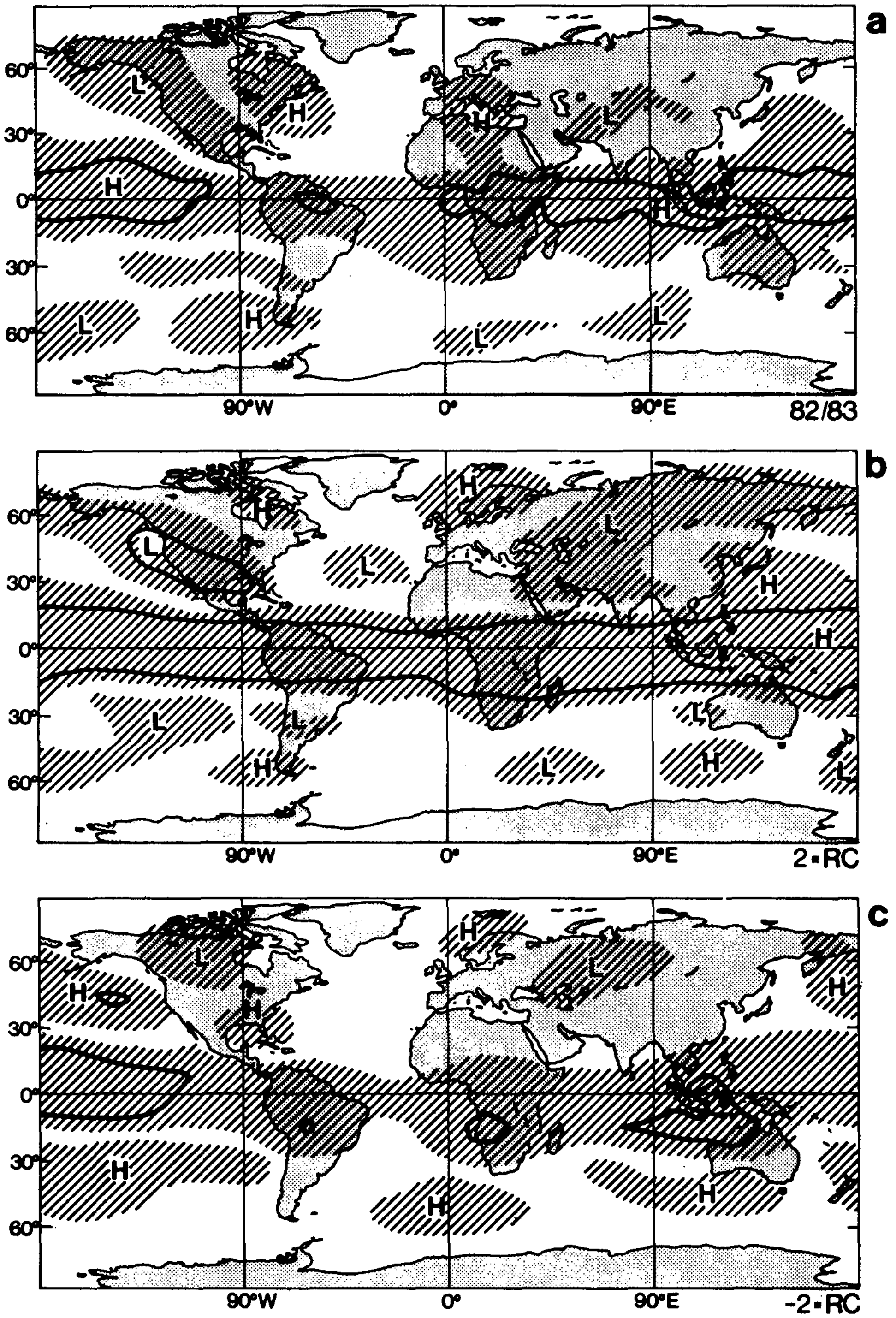

FIG. 5. Locally significant (hatched) and 84%-recurrent (according to parametric test; closed contours) geopotential height at 500 mb simulated in (a) the 82/83 experiment, (b) the 2RC experiment, and (c) the −2RC experiment.

associated with t-values greater than two is found over the northern Pacific and southeastern United States, and at about 40°S latitude. In this experiment a clear PNA pattern has evolved. With respect to the tropics and the midlatitude Southern Hemisphere, the −2RC and 2RC responses have similar patterns but are reversed in sign.

Only a small fraction of the locally significant height response in −2RC is associated with t-values large enough to indicate 84%-recurrence: a 70 degree lon-

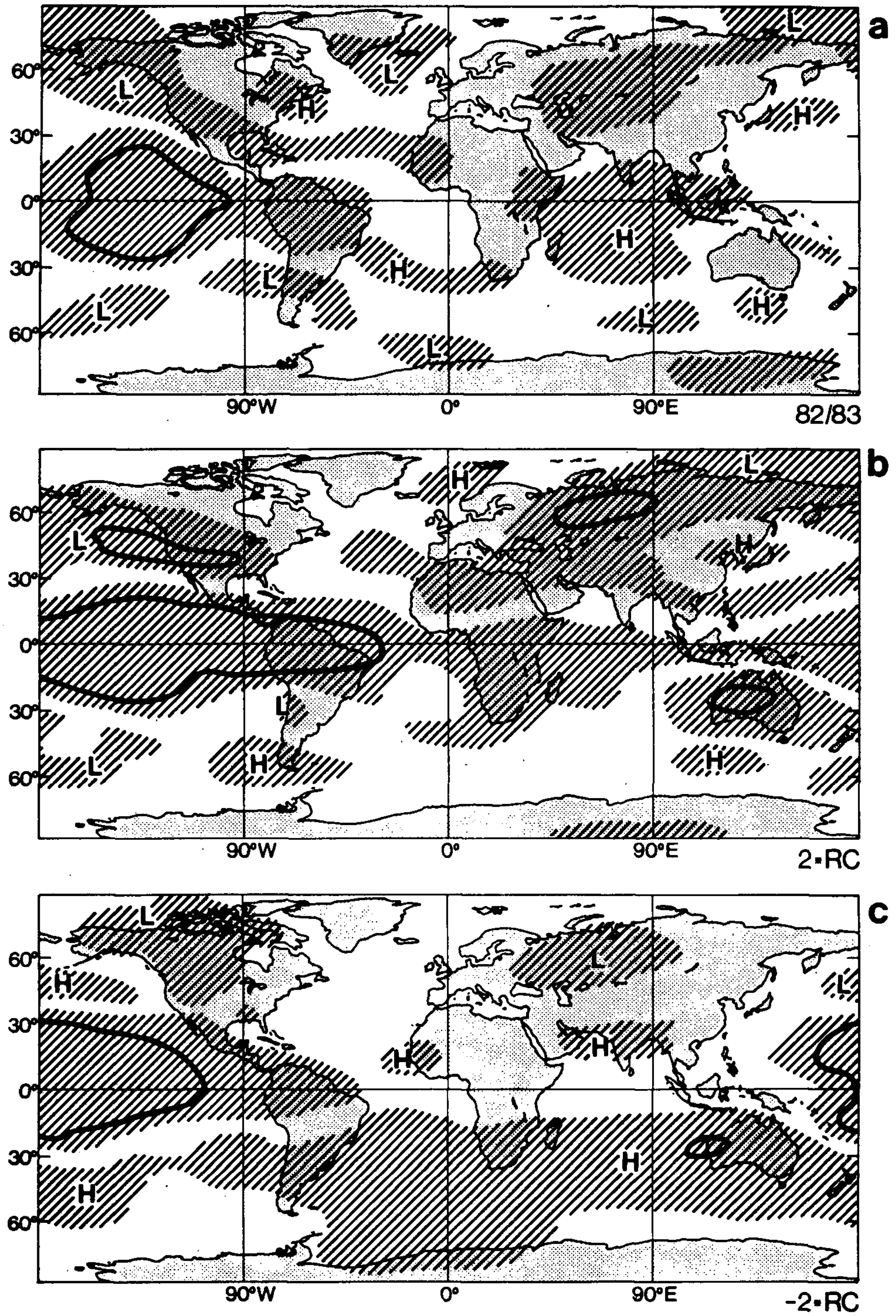

FIG. 6. Locally significant (hatched) and 84%-recurrent (according to parametric test; closed contours) temperature at 500 mb simulated in (a) the 82/83 experiment, (b) the 2RC experiment, and (c) the −2RC experiment.

gitude sector from about the dateline eastward (minimum of $t = -11.8$), an area covering Indonesia (minimum of $t = -7.7$), and two small spots over South America and southern Africa.

The tropical temperature response of −2RC (Fig. 6c) is confined to the Pacific. Over most of the midlatitude Southern Hemisphere including Australia, locally significant positive temperature anomalies have evolved. The −2RC 84%-recurrent portion is larger than in the 82/83 experiment covering the equatorial

Pacific from 150°E to 100°W and also a small spot over Australia.

2) Nonparametric approach

We also tested the temperature and height data for $(q', 0.50)$-recurrence with $q' = 98\%$ by counting the number T of experimental outcomes larger (smaller) than $X_{q'}$ ($X_{1-q'}$). As mentioned above, the probability of obtaining $T = m = 5$ is at most 3% if the response is less than $(q', 0.50)$-recurrent. The 98th quantile of the control distribution was estimated by fitting the Gaussian distribution to the control sample and deriving the corresponding quantile from the fitted distribution.

The $T = m = 5$ contours analyzed in the three experiments 82/83, 2RC and −2RC are shown in Figs. 7 and 8 for 500 mb height and temperature respectively. The results are similar in character but do not coincide closely with those of the parametric analysis technique. With respect to geopotential height, 2RC appears to have the weakest recurrent response in the tropics, while −2RC and 82/83 generate recurrence patterns of similar extent, which contradicts the findings with the parametric test.

Apart from small areas in 2RC and −2RC over western Siberia, which are connected with the same sign of anomaly, the recurrent 500 mb temperature response is confined entirely to the tropics in all three experiments. The 82/83 and 2RC responses differ over the maritime continent and the India Ocean, which is reasonable because of the large-scale, positive SST anomaly in this area mentioned above. The −2RC response is associated with the smallest recurrence area which lies primarily in the Pacific. Thus, with respect to temperature, the nonparametric analysis yields results which are physically more reasonable than those obtained with the parametric one.

Differences in test results are most likely the result of small experimental samples and the fact that not the same information is used in both tests. The parametric test is based primarily on the distance between sample means while the nonparametric test is essentially a comparison of the extremes of the two samples. In the case of the parametric test the null hypothesis is rejected when $\bar{y} > \bar{x} + 6.25 \cdot S_p \cdot (1/n + 1/m)^{1/2}$ or when $\bar{y}$ is more than approximately $3 \cdot S_x$ from the control sample mean (because S_p is dominated by the standard deviation S_x of the control sample unless the control and experimental variances are vastly different). On the other hand, the null hypothesis is rejected by the nonparametric test if the smallest experimental observation $y_{(1)}$ is more than $2 \cdot S_x$ from the control mean. It is easy to see that both conditions will not be satisfied simultaneously at all points. Also, because the tested fields have strong large-scale spatial correlation structure, we can expect that there will be discrepancies over large areas. These difficulties will be less pronounced when sample sizes are larger because the nonparametric test will then involve a less variable order statistic than the first order statistic $y_{(1)}$. As will be shown below, discrepancies between test results are not due to differences in control and experimental variances.

Differences in control and experimental variances affect the interpretation of test results as has been noted in section 4b. The variances of the 500 mb height fields in the experimental and control samples (not shown) are not generally significantly different in the positive SST anomaly experiments but in both cases the ratio of variances is less than 1 over most of the globe, suggesting that ensemble variance has been reduced by imposing positive SST anomalies. On the other hand, variances have been significantly increased in the tropics in the negative SST anomaly experiment (−2RC). Variances are reduced in the Southern Hemisphere in this case, but not significantly.

To understand some of the effects of unequal variance consider the case in which both ensembles are Gaussian but have different variances. In this situation it can be shown that q-recurrence is equivalent to $(q', 0.5)$-recurrence where q and q' are related by the equation $Z_{q'} = (1 + \sigma_y/\sigma_x) \cdot Z_q$. When q' is fixed, as in the nonparametric test, the degree of equivalent q-recurrence depends on the ratio of experimental to control variance. When variances are equal, evidence for (98%, 50%)-recurrence is also evidence for 84%-recurrence. When the experimental variance is the greater, evidence for (98%, 50%)-recurrence is evidence for something less than 84%-recurrence. That is, the test for (98%, 50%)-recurrence becomes somewhat liberal when it is interpreted as a test for 84%-recurrence. The opposite occurs when the control variance is the greater of the two.

The analysis is more difficult in the case of the parametric test because development of a parametric test which takes differences in variance into account is not mathematically tractable. However, we can make some observations about the asymptotic version of the parametric test. First we note that the denominator of t-statistic (1) is an estimate of $\sigma_x \cdot [1/m + (\sigma_y/\sigma_x)^2/n]^{1/2}$. Then, if we assume that the variances are known and replace the denominator of (1) with this expression, the resulting statistic has a Gaussian distribution with mean $Z_q \cdot (1 + \sigma_y/\sigma_x) \cdot \{mn/[n + m(\sigma_y/\sigma_x)^2]\}^{1/2}$ and variance $[m + n(\sigma_y/\sigma_x)^2]/[n + m(\sigma_y/\sigma_x)^2]$ when the response is q-recurrent. Thus, the critical value for the asymptotic test of q-recurrence is $Z_q \cdot (1 + \sigma_y/\sigma_x) \cdot \{mn/[n + m(\sigma_y/\sigma_x)^2]\}^{1/2} + Z_{1-\alpha} \cdot \{[m + n(\sigma_y/\sigma_x)^2]/[n + m(\sigma_y/\sigma_x)^2]\}^{1/2}$. This reduces to (8) when the variances are equal. With $n = 30$ and $m = 5$ the critical value is greater than given by (8) when the experimental variance is greater than the control variance, and hence the asymptotic test using (8) is liberal. The opposite occurs when the control variance is greater than the experimental variance.

It is difficult to say whether this behavior is also characteristic of the nonasymptotic test, but the suggestion is that both the parametric and nonparametric

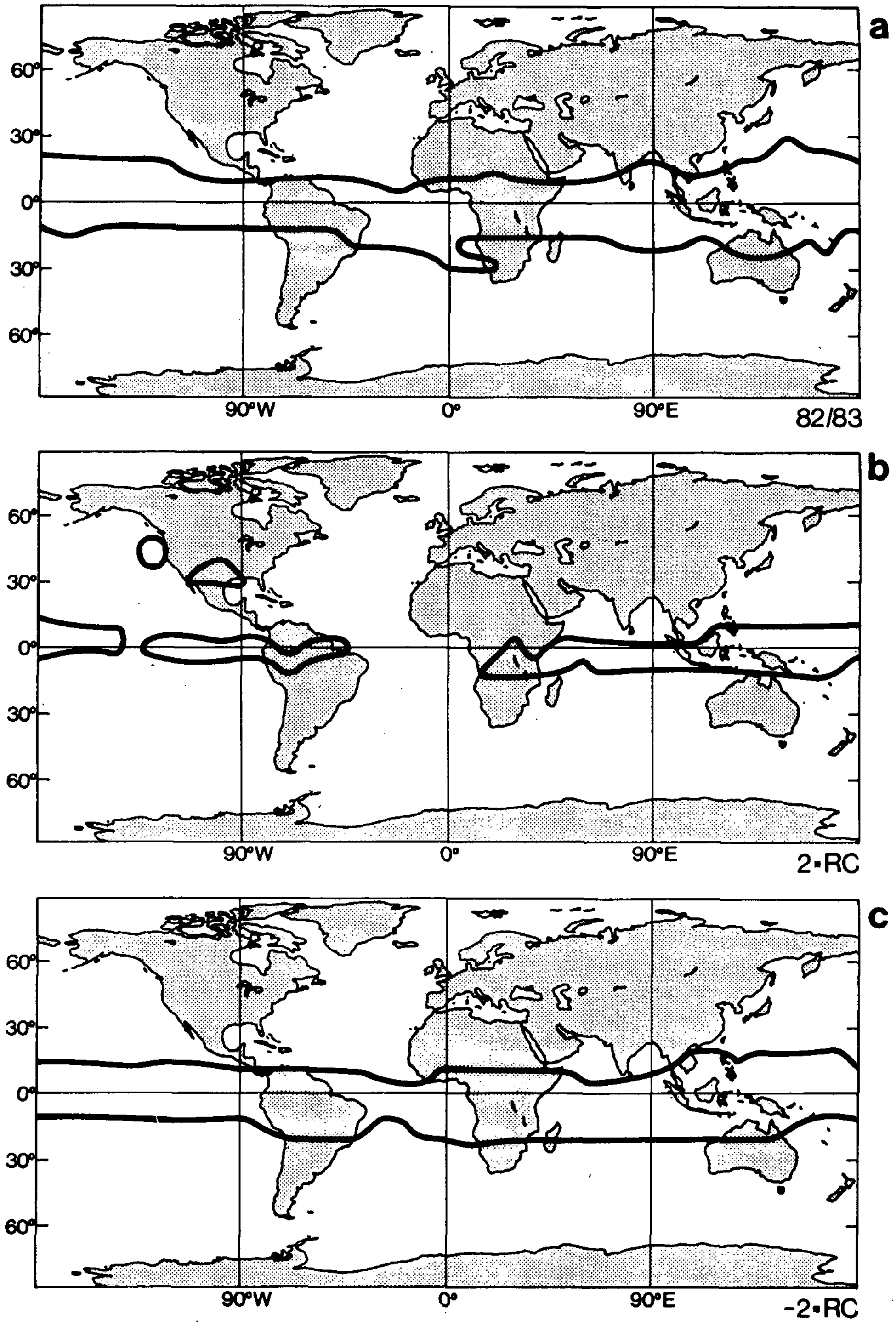

FIG. 7. (98%, 50%)-recurrent response of 500 mb height simulated in (a) the 82/83 experiment, (b) the 2RC experiment and (c) the −2RC experiment as determined by the nonparametric test.

tests are affected by unequal variance in the same way when they are thought of as tests for q-recurrence. The same considerations can be made if we think of the tests as tests for (q', 0.5)-recurrence. As noted above, the nonparametric test is not affected by unequal variance in this case, and as we will see, the asymptotic version of the parametric test is only mildly affected. When the response is (q', 0.5)-recurrent, the asymptotic test statistic has a Gaussian distribution with mean $Z_{q'} \cdot \{mn/[n + m(\sigma_y/\sigma_x)^2]\}^{1/2}$ and variance $[m + n(\sigma_y/$

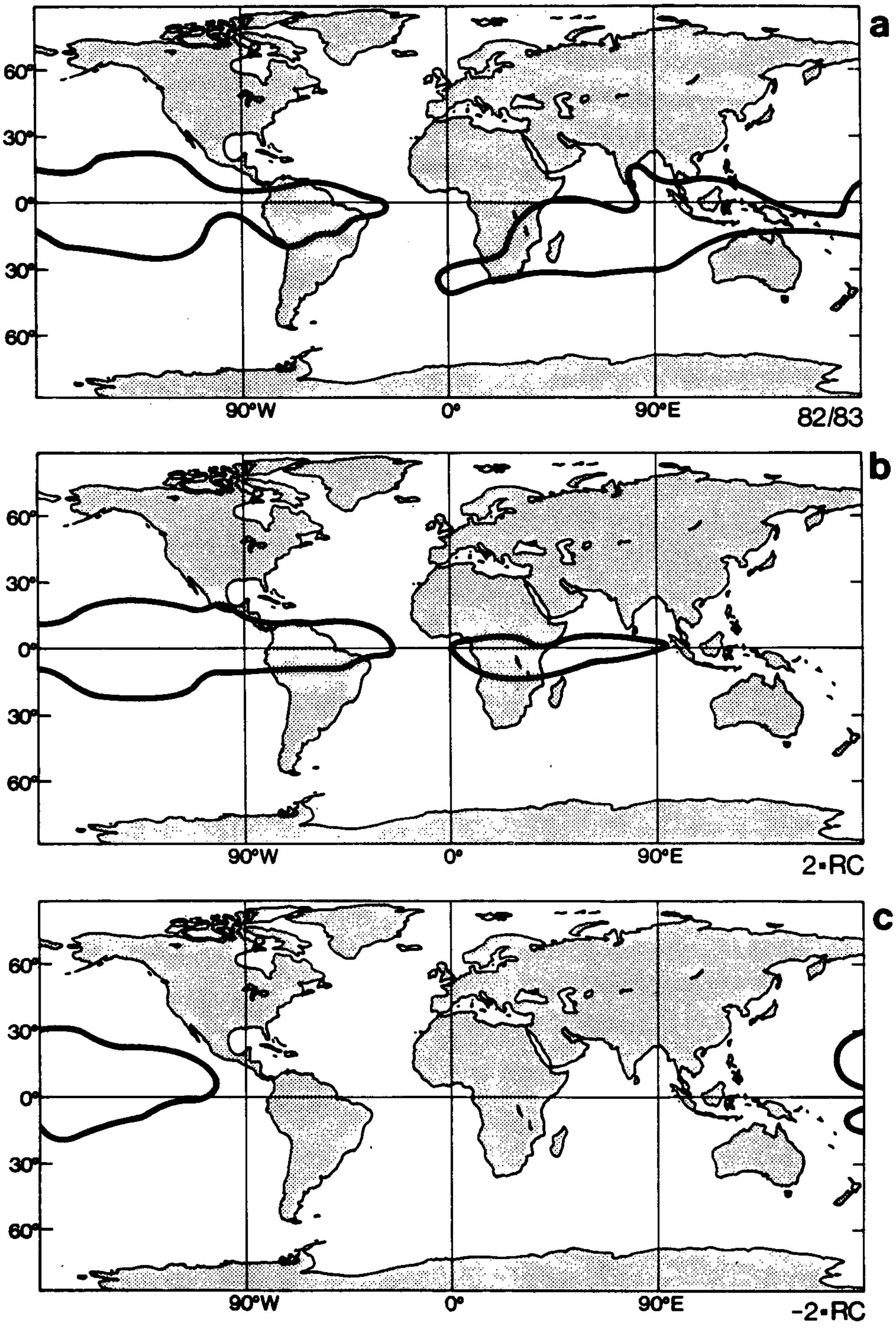

FIG. 8. (98%, 50%)-recurrent response of 500 mb temperature simulated in (a) the 82/83 experiment, (b) the 2RC experiment, and (c) the −2RC experiment as determined by the nonparametric test.

$\sigma_x)^2]/[n + m(\sigma_y/\sigma_x)^2]$. Thus the critical value when this test is regarded as a test of $(q', 0.5)$-recurrence is $Z_{q'} \cdot \{mn/[n + m(\sigma_y/\sigma_x)^2]\}^{1/2} + Z_{1-\alpha} \cdot \{[m + n(\sigma_y/\sigma_x)^2]/[n + m(\sigma_y/\sigma_x)^2]\}^{1/2}$, which again reduces to (8) when the variances are equal. This latter expression is almost constant over a large range of variance ratios. Thus both tests appear to be unaffected by differences in variance when they are regarded as tests for $(q', 0.5)$-recurrence. Because the experimental and control variances are apparently different, the tests, particularly in the case of the negative SST anomaly experiment, should only be interpreted as tests for (98%, 50%)-recurrence.

6. Conclusions

In the ordinary testing setup one asks if the mean of the observed variable under experimental conditions is significantly different from the mean under control conditions. Equivalently, one might ask whether the experimental treatment, such as altering the SSTs, has an effect upon a model's climate. Unfortunately, addressing the results of an experiment in this way does not necessarily lead to the identification of the predictable, or stable, part of the response. We have attempted to address this problem by describing a number of tests which focus on whether the ensemble of control climate states is far enough from the ensemble of experimental climate states to enable identification of recurrent aspects of the experimental response. That is, does the response have certain attributes which one would be likely to observe each time a new realization is observed under experimental conditions?

To develop statistical tests which might be used to address the results of climate experiments in this way, we introduced the idea of recurrence. A response to experimental conditions is recurrent if there is little overlap between the ensembles of control and experimental climate states. Then several tests were described which attempt to diagnose the degree of recurrence of a response. One approach was based on a parametric model, in which the amount of overlap between ensembles was related to the difference of ensemble means. The particular test which was developed requires that both ensembles are Gaussian and that they both have the same variance. A family of nonparametric tests based on the sign test was also developed. They operate by counting the number of experimental outcomes that are greater than a given threshold. This threshold is determined from the sample of control outcomes. These tests are more rugged but perhaps less powerful than the parametric test, even when it is necessary to use parametric models to determine the threshold. We also described a direct nonparametric analogue of the parametric test which is based on the Mann–Whitney statistic, and we described a third type of nonparametric test which compared the order statistics of the control and experimental samples.

The parametric test and the sign-test-based nonparametric test were applied to a set of climate experiments that were conducted with the Canadian Climate Centre GCM. The example illustrates several things about recurrence analysis. Features of the response to the experimental conditions which are recurrent cover much smaller regions than the regions in which it is possible to say that experimental and control climates are significantly different. However, the features which were identified as being recurrent were physically reasonable, supporting the notion that the use of these techniques in climate experiments may allow for a clearer understanding of the form of the expected response. We were able to identify aspects of the response in all three experiments which are likely to recur each time a new realization of the experimental climate is observed. There were differences between the results of the two tests which were attributed to small sample size and the fact that the two tests use different aspects of the information in the experimental sample. Also, it was pointed out that care is required in the interpretation of test results because of evidence that the assumption of equality of variance was violated.

There are still many aspects of this work which are not yet complete. The tests which have been developed are univariate tests which can only be applied at individual grid points or to individual expansion function coefficients. It has been argued that a recurrent response at a single grid point may well be strong enough evidence to indicate that control and experimental mean fields are significantly different from each other. A test which can indicate whether an entire pattern of response is recurrent is presently being developed. Also, we know very little about the power of the tests which we have described. This problem will also be addressed in future work.

Acknowledgments. The research by H. von Storch was supported in part by the Canadian Climate Centre. We thank Mike Lazare for his assistance with preparing the data and we thank Dr. Thomas Bell for several useful suggestions.

APPENDIX

Derivation of the Significance Level of a Rank Band Test for Recurrent Differences

In this appendix we derive the significance level of the test of

$$\mathrm{H}_0\colon P(Y > X_{0.5}) < q \quad \text{vs} \quad \mathrm{H}_a\colon P(Y > X_{0.5}) > q$$

which is conducted by rejecting H_0 when $Y_{(1)} > X_{(n)}$. Here we use the notation $X_{(n)}$ to denote the largest observation in the control sample and $Y_{(1)}$ to denote the smallest observation in the experimental sample. The significance level α is given by

$$\alpha = \mathrm{Prob}(Y_{(1)} > X_{(n)} | \text{ response is at most } (0.5, q)\text{-recurrent})$$

which is given by

$$\alpha = \int_{-\infty}^{\infty} \int_{X_{(n)}}^{\infty} f[X_{(n)}, Y_{(1)}] dY_{(1)} dX_{(n)} \qquad \text{(A1)}$$

where $f[X_{(n)}, Y_{(1)}]$ is the joint probability density function of $X_{(n)}$ and $Y_{(1)}$ under the null hypothesis. Because of the independence of the two samples, the joint density function is given by

$$f[X_{(n)}, Y_{(1)}] = g[X_{(n)}] h[Y_{(1)}]$$

where g and h are the density functions of $X_{(n)}$ and $Y_{(1)}$ respectively. As a consequence of the independence of observations within samples, the density functions g and h are given by

$$g[X_{(n)}] = n\{F_x[X_{(n)}]\}^{n-1}f_x[X_{(n)}]$$
$$h[Y_{(1)}] = m\{1 - F_y[Y_{(1)}]\}^{m-1}f_y[Y_{(1)}]$$

where f_x and f_y are the density functions of the control and experimental climate ensembles respectively, and F_x and F_y are the corresponding distribution functions. By substituting these expressions into (A1) and simplifying the integral it can be shown that the significance level is given by

$$\alpha = \int_{-\infty}^{\infty} nF_x(x)^{n-1}[1 - F_y(x)]^{m-1}f_x(x)dx.$$

This expression may be evaluated numerically after models have been adopted for the control and experimental climate ensembles. Table 3 was derived by using Gaussian models with equal variance and a difference of means $\mu_y - \mu_x = \sigma Z_q$. The integral was evaluated using an adaptive Rhomberg extrapolation algorithm.

REFERENCES

Blackmon, M. L., J. L. Geisler and E. J. Pitcher, 1983: A general circulation model study of January climate anomaly patterns associated with interannual variation of equatorial Pacific sea surface temperatures, *J. Atmos. Sci.*, **40,** 1410–1425.

Boer, G. J., 1985: Modeling the atmospheric response to the 1982/83 El Niño. *Coupled Ocean-Atmosphere Models.* J. C. J. Nihoul, Ed., Elsevier, 7–17.

——, N. A. McFarlane, R. Laprise, J. D. Henderson and J.-P. Blanchet, 1984a: The Canadian Climate Centre spectral atmospheric general circulation model. *Atmos. Ocean,* **22,** 397–429.

——, —— and ——, 1984b: The climatology of the Canadian Climate Centre general circulation model as obtained from a five-year simulation. *Atmos. Ocean,* **22,** 430–475.

Conover, W. J., 1980: *Practical Nonparametric Statistics.* 2nd ed., Wiley, 493 pp.

Cubasch, U., 1985: The mean response of the ECMWF global model to the El Niño anomaly in extended range prediction experiments. *Atmos. Ocean,* **23,** 43–66.

IMSL, 1982: *IMSL Reference Manual.* IMSL Inc., 7500 Bellaire Blvd., Houston, Texas 77036-5085.

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

Palmer, T., and D. A. Mansfield, 1984: Response of two general circulation models to sea surface temperature anomalies in the tropical East and West Pacific. *Nature,* **310,** 483–485.

Pearson, E. S., and H. O. Hartley, Eds., 1976: *Biometrika Tables for Statisticians.* Vol. 2, Biometrika Trust, University College, London, 385 pp.

Rasmusson, E., and Carpenter, 1982: Variations in tropical SST and surface wind fields associated with the Southern oscillation/El Niño. *Mon. Wea. Rev.,* **110,** 354–384.

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

——, 1987: A statistical comparison with observations of control and El Niño simulations using the NCAR CCM. *Beitr. Phys. Atmos.,* **60,** 464–477.

——, and H. A. Kruse, 1985: The extratropical atmospheric response to El Niño events—a multivariate significance analysis. *Tellus,* **37,** 361–377.

——, H. van Loon and G. N. Kiladis, 1987: The Southern oscillation. Part VIII: Model sensitivity to SST anomalies in the tropical and subtropical regions of the South Pacific Convergence Zone. *J. Climate,* in press.

Zwiers, F. W., 1987: Statistical considerations for climate experiments. Part II: Mulitivariate tests. *J. Climate Appl. Meteor.,* **26,** 477–487.