Statistical Field Significance and its Determination by Monte Carlo Techniques

ROBERT E. LIVEZEY AND W. Y. CHEN

Climate Analysis Center, NMC, NWS, NOAA, Washington, DC 20233

(Manuscript received 6 March 1982, in final form 5 October 1982)

ABSTRACT

The effects of number and interdependence in evaluating the collective significance of finite sets of statistics are frequently non-trivial, especially for spatial networks of time-averaged meteorological data. These effects can be taken into account in two steps: By first prescreening for significance assuming data independence and then, if necessary, by taking into consideration dependence through the use of estimated effective degrees of freedom and the binomial distribution or, failing that, Monte Carlo simulation. Seasonal averages of 700 mb height data are used to illustrate the problem and to demonstrate how the data set properties are taken into account. Papers by Hancock and Yarger (1979), Nastrom and Belmont (1980) and Williams (1980) are critically examined in light of these considerations and Monte Carlo strategies for clarification of ambiguities suggested.

1. Introduction

A neglected aspect of statistical testing in a large number of geophysical studies has been the evaluation of the collective significance of a finite set of individual significance tests. This neglect has stemmed not only from deficiencies in sample sizes or computational power, but also from a lack of understanding of the combined effects of number and interdependence of set numbers. Here, we will try to clarify the problem and propose viable solutions within present data and computer constraints. The sets most frequently considered will be geographic arrays or fields where the two set properties mentioned above become number of grid points and spatial correlation, respectively. The ideas presented below, however, will be applicable to many other collections of tests.

Typically in studies, a statistic is estimated from a time series at each point on a grid and tested for statistical significance at some level, say 95%. This involves taking into account the finite length of the time series and, when needed, the serial correlation or higher-order time dependence (through a degree of freedom adjustment like those suggested by Mitchell *et al.*, 1966, or Davis, 1976). A recent study by Chen (1981) (which will be discussed in detail in Section 3) provides an example of such a situation in Fig. 1. Approximately 11.4% of the experimental domain is represented by correlations statistically significant at the 95% level. The question is what is the probability this could have occurred by accident? Or, in terms of a specific confidence level, what percent of area represented by significant correlations would be equalled or exceeded one out of twenty times (5%) by accident?

To answer these questions, a description of the probability distribution of percent of area is needed. From probability theory, the expected value or mean of this distribution is 5%. If the distribution was a vanishingly narrow spike, 5% would be the chance result for every experiment, and anything in excess of this could not have been an accident. This would be the case for an infinite collection of unrelated significance tests. Here neither qualifier is true; the grid has 936 points, and the data is highly correlated in space. Consequently, the probability distribution has finite width. Thus to find the required percent of area for a given confidence level, the same properties, finiteness and dependence, that were taken into account in the time series must also be quantitatively accounted for in the spatial network. Almost without exception researchers in the past have failed to do so. For example, frequently the assumption has been made that any excess of the significant area over 5%, required to ensure less than one in twenty odds of obtaining the field by chance, is negligible. This is equivalent to the assumption that the distribution approximates a spike so that the 95% acceptance level is practically indistinguishable from the 50% level.

We intend to demonstrate, both theoretically in Section 2 and by example in Section 3, that for many applications in our principal area of interest, intermonthly to interannual climate fluctuations, this presumption is unjustified. The effects of finiteness and interdependence will be treated separately in both sections, leading naturally to a two-step field significance testing procedure. Through our example in Section 3 we will try to demonstrate the power of Monte Carlo techniques for the second step. Within the framework developed in Section 2, we will argue

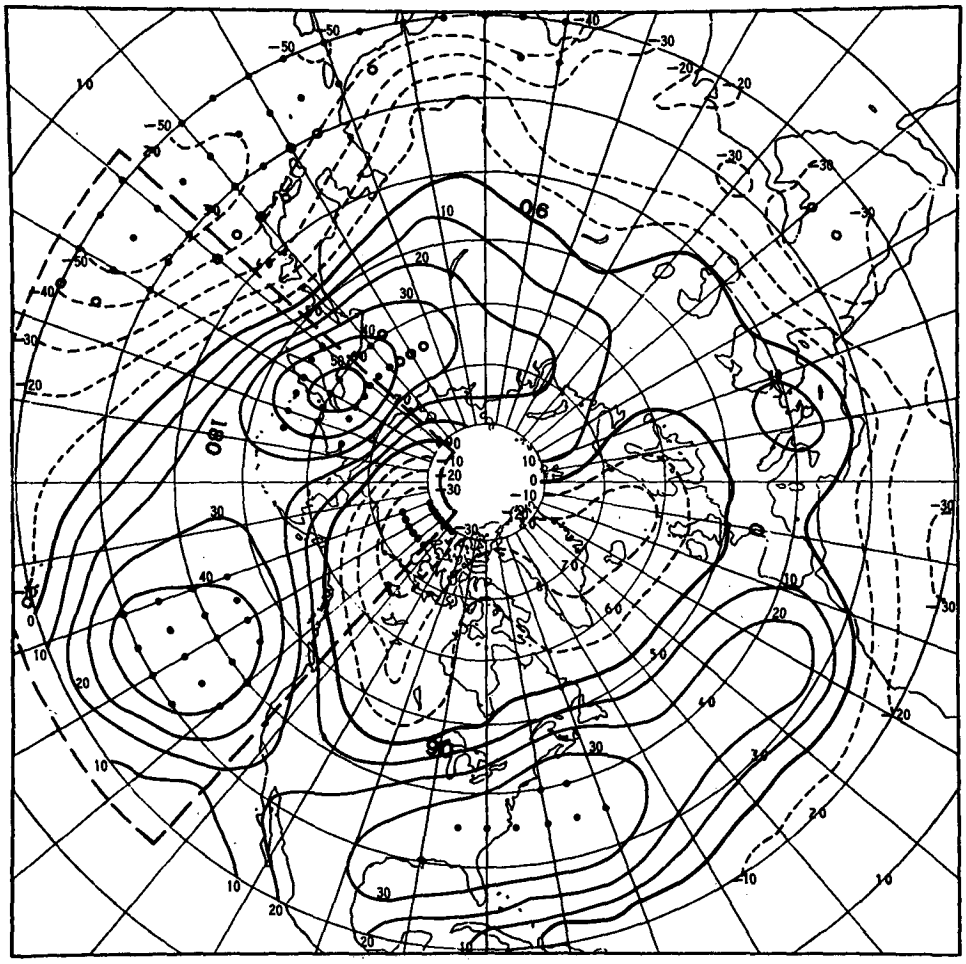

FIG. 1. Correlation of winter season averages of a Southern Oscillation Index (SOI) and 700 mb heights in hundreds. Negative and positive isopleths are shown as thin dashed lines and solid lines, respectively, with the two separated by heavy solid zero lines. A subarea approximating the Pacific basin, used for the experiment summarized in Fig. 5c, is enclosed by a heavy dashed line. Points on a 5 × 5 degree latitude–longitude grid at which correlations are 95% significant are indicated by solid dots; open dots represent additional points significant with a more liberal test.

in Section 4 that the statistical significance of results presented in at least three recent papers is considerably overstated. At the same time, however, specific Monte Carlo tests capable of resolving any residual uncertainties will be suggested. Finally, in our concluding remarks (Section 5) we will discuss the potential for extension of these techniques to significance testing of general circulation model (GCM) sensitivity tests.

Before we continue, some discussion of the two kinds of significance levels employed below is necessary. These consist of the acceptance level used to test for individual or *local* significance, and the level used to test for field or *global* significance. The values assigned to these levels are arbitrary choices of the researcher and need not be equal. However, if a goal is to establish statistical significance, whether local or global, these choices should be guided by accepted testing standards and the degree to which an *a priori* hypothesis has been formulated from different data or theoretical arguments. Generally, the level should be high for experiments with nonexistent or nonspecific *a priori* hypotheses, and lower for experiments backed up by specific independent predictions. When only a global test is needed, the choice of local testing level need not be so rigidly constrained. This is because now the local level's function is mainly to define a *pattern* of significance and to determine the value of the parameter (percent of area significant in our example above) tested against the second level. Thus, it can be set high or low by the researcher depending on the strength of signal or relationship he

wishes to highlight. In all but one of the studies that will be examined in Sections 3 and 4, only local significance was originally checked. Therefore, for the purpose of discussing collective significance in these experiments, we will adopt the respective author's choices of local acceptance levels as also appropriate for global criteria. Because all three studies are fundamentally *a posteriori* in character, this procedure seems reasonable.

2. Theoretical and empirical considerations

Even if no interdependence exists in a finite set of time series, it is still possible to conclude erroneously that a set of statistics derived from them is statistically significant (at the 95% level) if over 5% of the members are individually significant. Consequently, the property of finiteness will be examined first.

a. Finiteness and the binomial distribution

In terms of probability theory, a collection of N independent significance tests of random number statistics is perfectly analogous to N tosses of a loaded coin. In both instances there are only two outcomes with fixed probabilities for each test or toss. Instead of heads or tails with odds based on the coin load, the outcomes for a 95% significance test are test passed (probability equal to 0.05) or test failed (probability equal to 0.95).

All such two-outcome situations are probabilistically described by the binomial distribution. Therefore, we can apply this distribution directly in the design of overall significance tests for N independent tests. Suppose 95% confidence is required, then the following question must be answered first in order to conduct the test. What number of passed tests M_0 must be equalled or exceeded experimentally in N tests such that the probability of this result occurring by accident is less than 0.05? Once M_0 is known, the overall test is performed by simple comparison of this number to the experimental result M; if $M \geq M_0$ the collection of statistics is statistically significant at the 95% level, otherwise it is not.

In Fig. 2, the binomial distribution for thirty ($N = 30$) independent 95% ($p = 0.05$) tests is presented in two different ways. In this and subsequent figures and discussion, lower case and upper case "p" will be used respectively to denote the probability of passing a single test (p) and the probability of simultaneously passing M_0 or $M \geq M_0$ out of N tests (P). Note that the probability is 0.045 for *exactly* four out of thirty passed tests (or heads), 0.016 for five, and negligible for more than five. Thus, the cumulative probability is 0.061 for four *or more* passed tests and 0.016 for five or more. Clearly, if thirty significance

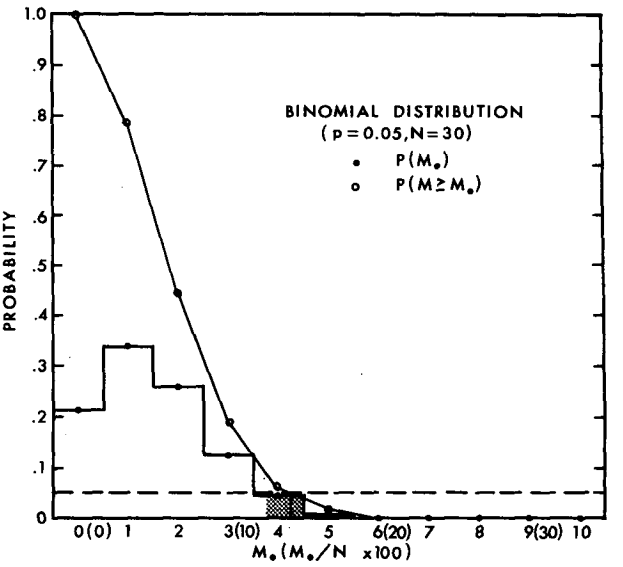

FIG. 2. The binomial probability distribution for $N = 30$ and $p = 0.05$. Closed dots denote probabilities of exact numbers of events, while open dots represent cumulative probabilities of greater than and equal to plotted numbers of events.

tests are performed, five (16.7%) or more must pass to guarantee at least 95% (in this case 98.4%) significance.

Linear interpolation to $P = 0.05$ gives 4.24 (14.1%) or more passed tests, and this threshold criterion for overall significance is plotted as percent of tests in Fig. 3. This reference point is nonrealizable as long as we are dealing with thirty independent tests, but in situations where a much larger number of interdependent tests behave statistically like thirty tests (discussed in the next subsection), a non-integer number of passed tests out of thirty has real meaning. The long curve in Fig. 3 is the result of performing the interpolation described above on cumulative binomial distributions for N's up to 1000. Note that even with 1000 tests the criterion for overall significance is more than 6% and that for less than 300 tests the criterion can be quite large.

A graph like Fig. 3 (other examples can be found in Section 4) or a good hand calculator is the appropriate tool for the first step in evaluating field significance. There is little point in concerning oneself with spatial dependence if an experiment is not significant assuming independence. This possibility can easily be checked by plotting the experimental result (assuming all tests are independent) on Fig. 3. If the plotted point falls below the line, the experiment is not statistically significant and no further consideration need be given to this particular question. If it falls above the line, then it may be significant and the procedures recommended in the next subsection should be followed.

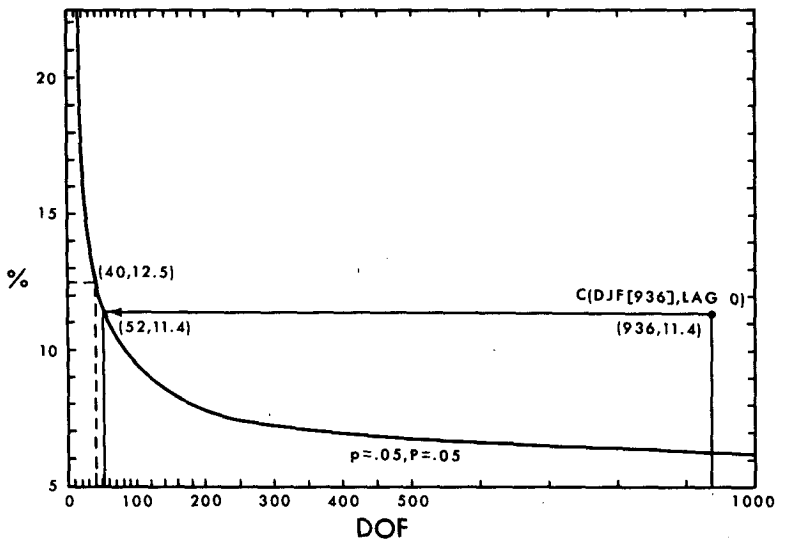

FIG. 3. Estimated percent of independent 95% ($p = 0.05$) significance tests passed that will be equalled or exceeded by accident 5% ($P = 0.05$) of the time versus the number of independent tests N (labeled "DOF" for "degrees of freedom"). The curve is based on the binomial distribution. The plotted point and coordinate lines and points refer to the significance test of Chen's experiment described in the text.

b. Interdependence and Monte Carlo simulation

Although interdependence cannot be treated as conveniently as finiteness, Fig. 3 can still be employed to visualize its impact. Large cross-correlations in a field reduce the so-called "degrees of freedom" of the field in the same manner that large lag autocorrelations reduce the number of effective samples in a time series. Instead of a set of N individual realizations, such a field should be viewed as $n < N$ independent clusters of closely related realizations.

If n could be estimated accurately then Fig. 3 would again be applicable—the effective increase in threshold percent for statistical significance could be read directly, and the second step of the test would be complete. However, even if only a crude estimate of n is available, it still may be possible to complete this part of the test graphically. If the percent of statistically significant points is known from an experiment, the curve in Fig. 3 defines the minimum number of effective degrees of freedom n_0 the field must contain to be significant overall. This number can be determined by finding the intersection of the long curve with a horizontal line drawn through the point plotted in the first step of the test (see Subsection 2a). In the example of Fig. 3, an experimental result of 11.4% gives n_0 equal to $\sim$52. If the estimate of n is *much* larger than n_0 the experimental result can be presumed significant. For many meteorological and climatological studies it is possible to arrive at reasonable estimates of n and examples will be presented in the analyses of Sections 3 and 4.

Incidentally, for testing to this point, at least one alternative exists to the binomial approach emphasized here. This is the technique suggested by Madden and Julian (1971) and implied by Reynolds (1978) where the local acceptance level is raised to the point where *no* tests will be expected to pass by accident. Because this method is over-conservative when dependence exists between individual tests, a two-step strategy different from that outlined above has to be adopted. That is, initial failure to pass the Madden–Julian test (assuming no effective reduction in degrees of freedom) should not terminate testing, whereas success should. If a second step is necessary, the largest significance level attained on the grid in the experiment will uniquely determine n_0, now an upper bound for degrees of freedom rather than a lower bound. If a reasonable estimate of n falls well *below* this n_0, then the test is complete and the results significant.

Whichever procedure is followed, when estimates of n are unavailable or when experimental results fall within their range, the only recourse to evaluate field signicance is a Monte Carlo simulation. This is because multivariate techniques cannot be applied to large numbers of relatively short geophysical time series. In the Monte Carlo procedure, an experiment modeled after the one to be tested, but in which a null hypothesis is true (that results are by accident) is repeated many times with different random inputs, and consequently, different output statistical fields. A histogram of percent of statistically significant points can then be constructed from all the experi-

ments and used to estimate the five percent tail of the distribution and the threshold fraction for field significance. This threshold can in turn be used to estimate the effective degrees of freedom of the field from curves like the one in Fig. 3.

There are a number of considerations in the design of Monte Carlo simulations. First, the random component must be introduced in such a way as to retain the interdependence whose effects need to be evaluated. Second, probabilities of chance outcomes under the null hypothesis in the actual experiment must be matched in the Monte Carlo simulation. Third, enough simulations need to be performed to estimate the probability distribution accurately. A check on the last two points, as well as the randomness of the input, is the closeness of the mean outcome to its expected value.

The use of Monte Carlo techniques for significance testing in meteorological studies has been quite limited. Neumann *et al.* (1977), in an elegant extension of earlier work by Lund (1970), applied these methods to the testing of multiple regression equations. In turn, Zurndorfer and Glahn (1977), by modeling interrelationships, extended this work to account for correlation between predictors. D. Gilman (personal communication, 1980) took a similar approach in developing statistics for temperature forecasts, using modeled spatial correlations in repeated random simulations. Finally, Barnett and Preisendorfer (1978) successfully employed the Monte Carlo philosophy to assess the significance of higher-order eigenvalues of empirical orthogonal functions (EOF's). Perhaps the main obstacle to their exploitation has been the large number of calculations required, but with recent computer hardware advances, this obstacle has largely been removed. In the next section, a hemispheric field of correlations between a single tropical Pacific index and seasonally averaged hemispheric 700-mb heights will be tested for significance through a Monte Carlo simulation. The example should serve to illustrate both design considerations and the need for such tests.

3. Example of Monte Carlo simulation and testing

In a study by Chen (1981), seasonal mean 700 mb height anomalies on a 5 × 5 degree latitude–longitude grid from 20 to 80°N inclusive from 1951 to 1979 were correlated at various seasonal lags with a Southern Oscillation Index (SOI). This index consists of the seasonal mean of the standardized difference in sea level pressure between the island of Tahiti, and Darwin, Australia. At each point on his grid, Chen tested the statistical significance of correlations by making use of a modified form of a technique used by Davis (1976).

In this approach, a measure of the effective time between independent samples can be estimated from the autoregressive properties of both time series,

$$\tau = [1 + 2 \sum_{i=1}^{N} C_{HH}(i\Delta t)C_{SS}(i\Delta t)]\Delta t. \quad (1)$$

Here Δt is the sampling time, N the number of samples, and the C's the autocorrelations at lags $i\Delta t$ for the height anomaly (H) and SOI (S) respectively. From τ, the effective number of independent samples (or degrees of freedom) in the time series

$$n = N\Delta t/\tau \quad (2)$$

can be determined. To reduce sampling error to acceptable levels at large lags (and small sample sizes), Chen used biased estimates for the C's. This is equivalent to quadratically damping unbiased estimates in inverse proportion to sample size.

With degrees of freedom specified by (2), correlations at each grid point were tested for significance at the 95% level (i.e., $p = 0.05$). The results for wintertime contemporaneous (lag 0) correlations are shown in Fig. 1. Solid dots denote points where the test was passed and represent 11.4% of the total area from 17.5 to 82.5°N.[1] The open dots represent correlations that would have also been statistically significant at the 95% level if a full 29 degrees of temporal freedom had been assumed everywhere, illustrating the conservatism of the alternative approach used.

a. Test step 1

Chen's result is plotted on Fig. 3 without regard to spatial dependence and it falls well above the line. Thus, to complete the test of field significance, spatial dependence must be taken into account.

b. Test step 2

From Fig. 3 it is apparent that the height data set must contain at least 50 spatial degrees of freedom for the positive result of Step 1 to hold up. It would be useful at this point to have an estimate of the actual degrees of freedom n to compare to this.

Meteorological data from standard surface and upper-air reporting networks are unquestionably correlated in space. Generally, surface temperature data is more spatially coherent than precipitation data, while upper-air information typically has larger integral space scales than surface data. Any further spatial smoothing of this data (including gridding by objective analysis) increases length scales, but the effects of time averaging can be considerable as well. Some appreciation for the dominant space scales in monthly means can be obtained from the recent work

[1] If a grid does not approximate an equal area mesh it is more appropriate to consider percent of area.

of Wallace and Gutzler (1981) for hemispheric geopotential heights, and Walsh and Mostek (1980) for United States surface pressure, temperature and precipitation. For the moment, we will focus on the properties of gridded, time-averaged upper-air heights only, but in Section 4 we will frequently cite Walsh and Mostek along with Gilman (1957).

Wallace and Gutzler's teleconnection and teleconnectivity charts as well as those shown by Namias (1980) and Harnack (1980) illustrate the dominance of the largest planetary waves in monthly and seasonal means. Two important features of these upper-air charts are their simplicity and the existence of large remote and local areas of strong absolute spatial correlation. In fact, Wallace and Gutzler were able to account for 48% of the variance of their 500-mb data with only four EOF's.

It is quite likely then that these hemispheric fields contain relatively few degrees of freedom. Estimates ranging from 30 to 60 are possible from consideration of the number of independent modes of either EOF or spherical harmonic decompositions necessary to account for a high percent of the variance. Consequently, $n \sim n_0$, and a conclusive statement of significance is not possible at this point.

To evaluate the field significance of Fig. 1 and Chen's other correlation maps, probability density functions of percent of area statistically significant at the 95% level were approximated with Monte Carlo simulations. The SOI time series was replaced with a series of 29 numbers randomly selected from a normal distribution. The technique used to generate these numbers was first tested to ensure that it well approximated the assumed distribution. More sophisticated modeling of the series was unnecessary because the main interest was the effects of spatial correlation in the height data. As long as experiment and simulation probabilities could be matched, the simpler approach was preferable.

Correlations of this Gaussian noise with seasonal height anomalies were then obtained at every grid point and tested for significance at the 95% level. Because now $C_{SS}(i\Delta t) = 0$ for $i \neq 0$ in (1), the appropriate temporal degrees of freedom become 29 everywhere and $C_{HS} \geq 0.367$ to pass an individual test. Fig. 4 is an example of such a calculation with a striking planetary wave-like pattern of high correlation which resembles patterns presented in Wallace and Gutzler (1981). This experiment was completed a total of 200 times and all results are presented in histogram form in Fig. 5a.

Because at least 5% of the trials had greater than

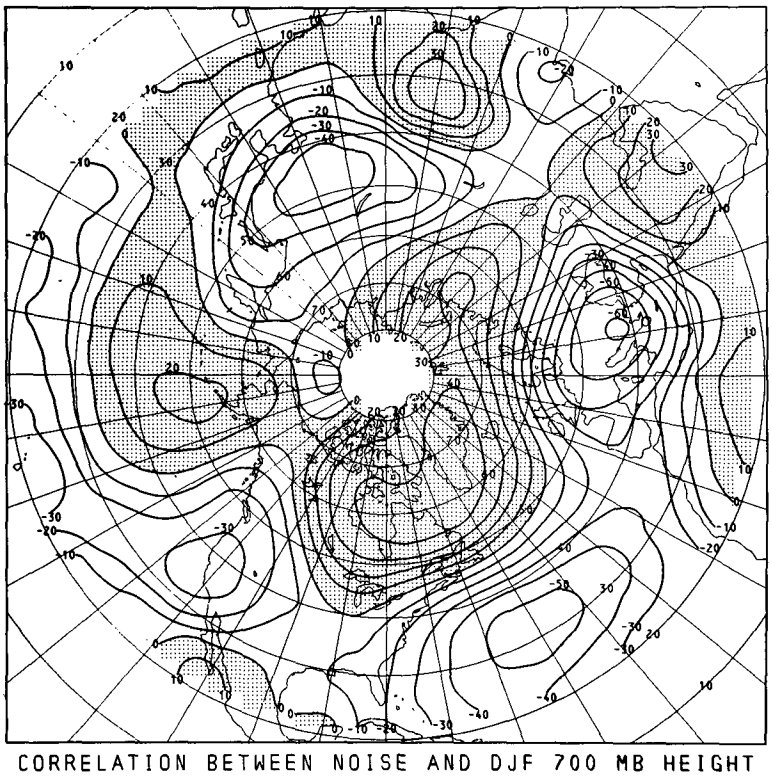

FIG. 4. Correlation in hundreds of winter season averages of 700 mb heights with Gaussian noise. Shading denotes positive values.

12.5% of their area statistically significant at the 95% level, it is not possible to reject the hypothesis (at the 95% level) that the seemingly strong results shown in Fig. 1 were a chance occurrence. Reference again to Fig. 3 suggests that the height anomaly data set actually has fewer than 40 spatial degrees of freedom, considerably less than the more than 50 required. Conversely, lag 1 and 2 correlation patterns (Fig. 5a) are highly significant.

For the quasi-independent summertime experiment, the results were decidedly different in two respects (Fig. 5b). First, there is no evidence for a statistically significant field of correlation at any lag, a conclusion that could have been reached without a Monte Carlo test. Second, the Monte Carlo distribution is somewhat narrower for summer than winter, with the data containing ~55 spatial degrees of freedom compared to 35 for winter. This is simply another manifestation of the dominance of smaller scales during the warmer months.

As a final example, the analogous tests described above were repeated on exactly one-fourth of the winter data set—that area inside the dashed line in Fig. 1—approximating the North Pacific basin. This was done to illustrate the extent that such geographic "selectivity" (cf. Pittock, 1978) and its subsequent reduction in spatial degrees of freedom can influence significance testing. While the percent of significant

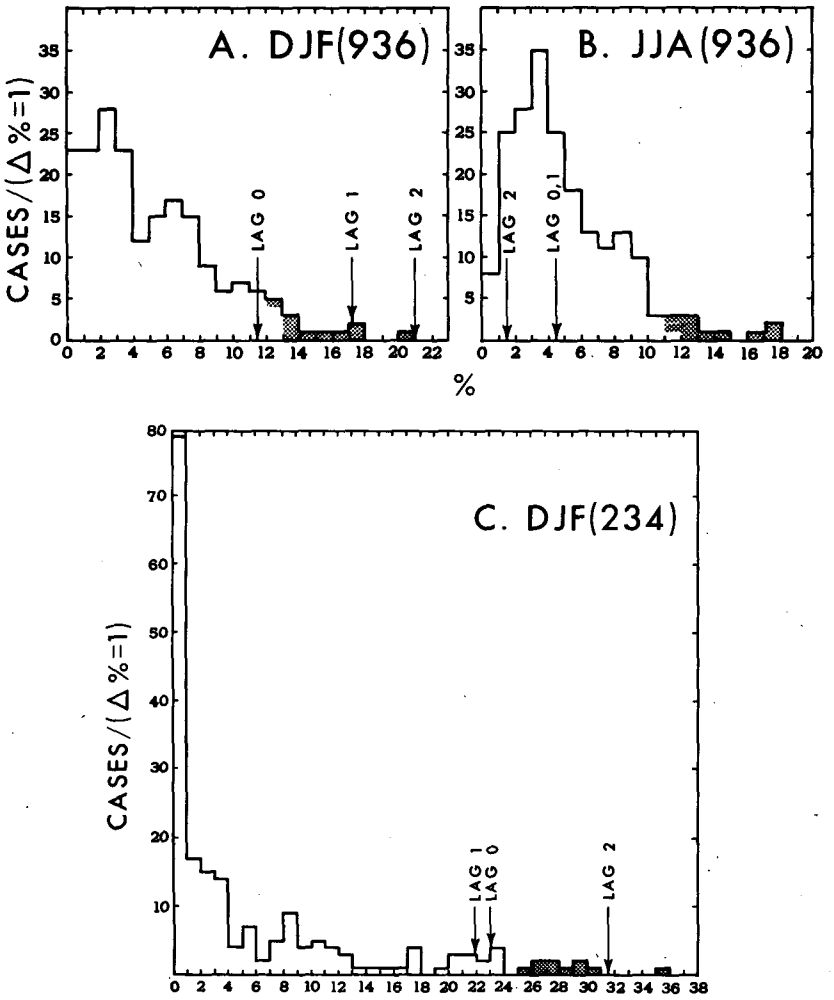

FIG. 5. Histograms of percent of area with correlations of 700 mb heights and Gaussian noise statistically significant at the 95% level (p = 0.05) in 200 Monte Carlo trials for: (A) the winter hemisphere; (B) the summer hemisphere; and (C) the winter north Pacific basin (outlined in heavy dashed line in Fig. 1). The abscissa is percent of area while the ordinate is number of cases for each one percent interval. The 5% tail (P = 0.05) is schematically shown by shading the 10 of 200 largest percents. The results for correlations with seasonally averaged SOI's, with heights lagging by the indicated number of seasons, are shown by vertical arrows.

area doubles for lag 0, as shown in Fig. 5c, conclusions about the field significance remain unaltered from those reached for the full data set. This is because the probability density histogram has spread considerably, or equivalently from Fig. 3, the spatial degrees of freedom have been reduced dramatically (to ~10). The lag 1 pattern is no longer significant for this subarea, perhaps because a greater number of significant points fall outside its boundary than for lag 2.

The three cases embodied in Fig. 5 vividly illustrate the ideas presented in Section 2. In the next section the same considerations will be applied to discussion of three other studies that use time-averaged meteorological data, either at the surface or near the tropopause, to investigate short-term climate fluctuations.

4. Critical discussion of selected recent results

The papers that have been singled out below are by no means the only ones vulnerable to the line of criticism raised here. A more recent example, for instance, is the work of Egger *et al.* (1981). All three papers, however, are presentations of results of marginal statistical significance in controversial areas, solar–climate relationships or CO_2-related climate change. We hope to underline our view that the tests we advocate or their equivalent should be mandatory parts of such studies, as well as to present a rational challenge to the positive results claimed for these particular experiments.

a. Williams (1980)

Based on 20 stations north of 65°N latitude, Williams (1980) selected the 10 "warmest Arctic winters" and the 10 "warmest Arctic summers" during the period 1900–69. At 121 non-homogeneously distributed North American stations (our count from Williams' figures), mean summer and winter precipitation totals were obtained separately for their respective "warmest Arctic" years and for the remaining years. The differences of these two means were then tested for significance at the 90% level by a two-tailed *t*-test. Of the 121 stations, only one had an individually significant winter difference while 15 had significant summer differences. Williams' interest was in defining potential shifts in precipitation patterns as a result of warming in the Arctic that might result from increases in global CO_2 concentrations. The characterization of the summer results as significant will now be examined.

On Fig. 6, the $p = 0.1$, $P = 0.1$ curve appropriate for Williams' experiment is plotted along with the summer result. Suppose the network of 121 stations are statistically independent, which they are not. Then in order to reject a null hypothesis that the summer result was by accident at the 90% level requires that 13.9% of the differences be individually significant, which is greater than the 12.5% (15 out of 121) that are. The odds are more than 1 in 7 under these conditions that the result was by chance. More realistically, these odds swell to 2 out of 5 and greater if the summer precipitation data contain no more than 75 spatial degrees of freedom. In any case, Williams' results are not statistically significant at the 90% level.

b. Hancock and Yarger (1979)

In a search for possible solar-climate relationships in United States surface weather data, Hancock and

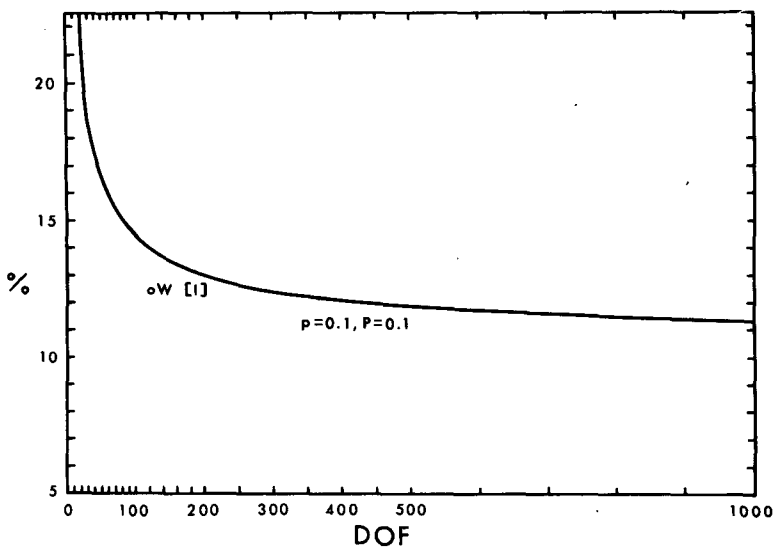

FIG. 6. As in Fig. 3, except for ($p = 0.1$, $P = 0.1$). The open circle and W denote the experimental result of Williams (1980). I denotes insignificant, the conclusion of Section 4a of the text.

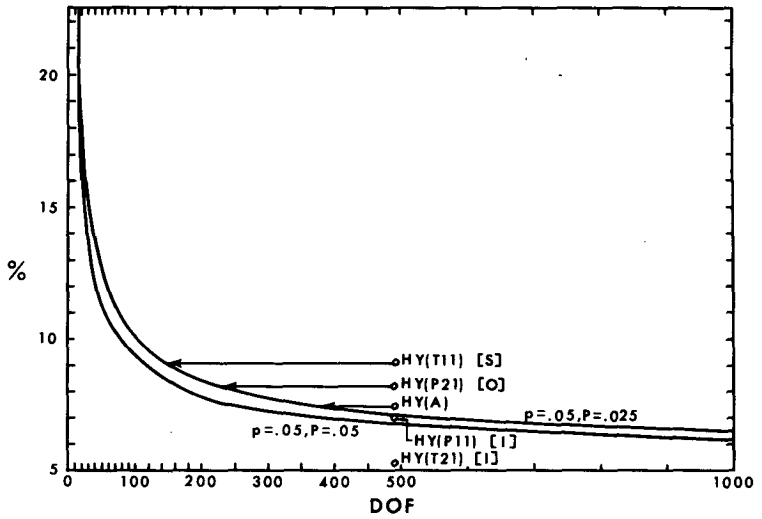

FIG. 7. As in Fig. 3, except for (p = 0.05, P = 0.05) and (p = 0.05, P = 0.025). The open circles and HY denote the experimental results of Hancock and Yarger (1979). Letters and numbers in parentheses denote different experiments: T is temperature, P precipitation, A average or all, and 11 and 21 are cycle lengths in years. Bracketed letters summarize conclusions of Section 4 of the text: I is insignificant, S suspect, and O open. Horizontal arrows locate minimum independent degrees of freedom necessary for an experimental result to be statistically significant for the stated individual probability (p) and level (P). No arrow means that the minimum exceeds the maximum possible for the experiment.

Yarger (1979) (hereafter referred to as HY) estimated 11- and 21-year squared coherences between statewide-averaged monthly temperature and precipitation for 41 states, and the Zurich annual mean sunspot number. They did this separately for each month of the year using data from the period 1891–1974. The statewide monthly averages used were area-weighted means of climate-division monthly means. Spectra were computed by finite Fourier transform techniques and raw spectral estimates smoothed by a five-point uniform filter. For each of the four quasi-independent experiments for temperature and precipitation at 11 and 21 years, the 492 squared coherences (41 states × 12 months) were tested for statistical significance at several levels.

HY correctly expressed results in terms of area rather than number of time series. These are plotted at 492 degrees of freedom in Fig. 7 in terms of percents of total area passing a 95% significance test. Additionally, they took into account the finiteness of the statistics collection by specifying the two-sided 95% interval for a binomial distribution with p = 0.05 and N = 492. This is equivalent to testing for significance at the 97.5% level in terms of the one-tailed tests used elsewhere in the present work. The corresponding curve (p = 0.05, P = 0.025) is also plotted in Fig. 7. Based on this, they claimed significance for three out of four of the experiments. From Fig. 7 and the preceding discussion the required area must in fact exceed 7.1% because of intercorrelations in the experiments. Before a Monte Carlo strategem is recommended to pinpoint the exact minimum requirements, existing studies will be examined to crudely estimate which of HY's results may need further analysis. Attention will be focused on the spatial degrees of freedom in their data.

Gilman (1957), and Walsh and Mostek (1980) (hereafter referred to as WM) conducted parallel analyses of United States monthly mean station data relying heavily on EOF decomposition. In Gilman's study, winter monthly means for the years 1899–1939 at 30 temperature and 64 precipitation stations were decomposed into their respective EOF sets. Only three principal components were needed to explain 80% of the temperature variance, while 20 were required for precipitation.

Extending this work, WM performed similar analyses, but for all months taken together, for 61 stations for both temperature and precipitation, and for a longer record, 1900–1977. Additionally, they prefiltered the data by the removal of 30-year running means for each month.[2] Even with the more complex data set, the number of EOF's required to explain 80% of the variance, five for temperature and 31 for precipitation, remained strikingly small. Moreover, when decompositions were performed separately for data

[2] This amounts to removal of the annual cycle and very long-term trends. It will have little or minor damping effect on power at 11 and 21 years, but will reverse the polarity of the 21-year wave (Panofsky and Brier, 1968). Neither of these facts are particularly important to the discussion here.

stratified by month, dominant EOF's changed surprisingly little from month to month. WM also computed the lag 1 autocorrelation of the coefficients for the first several eigenvectors and found important persistences for a number of months and components. Precipitation had less composite persistence than temperature.

These analyses suggest that Gilman's and WM's data sets contain spatially very few (on the order of ten) temperature, and probably less than 50 precipitation, degrees of freedom. Additionally, if months are treated separately as in HY, redundancy will exist between months in terms of both their dominant modes and month to month persistence. Finally, these degree of freedom estimates are for the entire spectrum of interannual frequencies (except those for the long-term trend removed by WM).

Compared to WM's surface data study, HY start with a raw data set two-thirds as large, partitioned by month, and highly smoothed in space. All of these things, particularly the last, reduce the spatial degrees of freedom in any given month's net of time series from the estimates discussed in the last paragraph. Also, the possibility exists that the dominant spatial scales for variance near 11 and 21 years will be larger than for higher frequencies, suggesting an even further downgrading of the estimates. In the context, then, of this more stringent assessment of number of independent tests, HY's results shown in Fig. 7 will now be re-examined.

First, it appears that neither the 21-year temperature nor the 11-year precipitation results are significant at the 97.5% or higher level. Second, even with minimal intermonth dependence, it is unlikely that HY's temperature data contain 160 experimental degrees of freedom, the approximate minimum required for 97.5% significance. Consequently, this result has been labeled "suspect" in Fig. 7. Finally, since it is not clear the total extent to which statewide-averaging, intermonthly dependence, and single-band variance will reduce the total precipitation degrees of freedom, the result considered strongest here, that for 21-year precipitation, has been labeled "open."

The effect of applying more liberal criteria to HY's results, namely for a one-sided 95% (rather than 97.5%) confidence interval, can be seen in Fig. 7 by extending the horizontal lines to the left to the lower curve. With these standards an "S" label is no doubt more appropriate for the 11-year precipitation result, but the other labels should probably remain as they are.

In our opinion much could be gained by answering the questions posed above, particularly those regarding the 21-year precipitation cycle. This can be straightforwardly accomplished by replacing the Zurich mean annual sunspot number by Gaussian noise and repeating the experiments *in all particulars* several hundred times. This procedure will well define threshold percents of area for statistical significance. An obvious check of the Monte Carlo test design will be the closeness of the percent of area means to five percent. An alternative, and maybe better, test distribution is that for the *total* percent of area significant for all four experiments in each Monte Carlo run. If enough experimental degrees of freedom can be gained by lumping results, HY's average result of 7.4% ("A" in Fig. 7) may prove significant at the 95% level. However, their study as well as WM's uncovered important correlations between temperature and precipitation patterns, so gains in degrees of freedom may be partially negated by these additional interdependencies.

c. *Nastrom and Belmont (1980)*

The results examined in this section are those reported by Nastrom and Belmont (1980) (hereafter referred to as NB). Briefly, NB determined the *maximum lag correlation* between seasonally averaged values of 10.7 cm solar flux and seasonal means of temperature and meridional wind components. The former varies predominantly in an 11-year cycle. These correlations were obtained for all seasons and all upper-air levels (standard we assume) at 174 stations where at least 13 years of data from 1949 to 1973 inclusive were available. Correlations were individually tested for significance at the 95% level after temporal degrees of freedom were adjusted by the prescription of Mitchell *et al.* (1966), or by a reduction of two when the lag-1 year autocorrelation of wind was negative. Results of these tests were presented only for the winter season and only at 300 mb for the wind components and partially (in a figure) for the temperature at 500 mb. For the winds, 12 and 19% of the stations for the zonal and meridional components respectively passed NB's significance tests at the 95% level. These results will now be reviewed in light of a number of problems in the testing procedures related to selectivity of results, spatial correlation (the main focus of our paper), and individual temporal correlation significance.

First, the presentation of results only for "winter wind speeds at 300 mb . . . because they generally show the closest relationship with the solar cycle," exemplifies what Pittock (1978) characterizes well as "selectivity" in time and in space. Unlike HY, only the best results have been reported in NB, thereby invalidating application of less stringent *a priori* significance measures. The odds of obtaining a significant result by accident are higher because there is more than one opportunity to obtain it and the best results can be selected. Combination of all levels and seasons inevitably would reduce the overall percent of significant points but increase the total experimental degrees of freedom. The largest impact on both these factors would be from inclusion of the

other seasons, rather than the other levels, because vertical correlations are quite large for seasonal mean winds. Whether or not the significance of results would be lowered cannot be determined here, but either way the partial reporting of tests and the selection of only best results must be taken into account when assessing the significance of NB's 12 and 19% figures. This question will be examined after the discussion of spatial effects.

As shown in Section 3, winter means of hemispheric 700 mb heights contain ~35 spatial degrees of freedom. Regional subsets of this data contain fewer, while heights at 300 mb will at best contain no more and quite likely less (to the extent that the longest waves account for a greater share of the variance at higher levels in the troposphere). Because seasonal mean upper-air winds will be almost exactly related to heights geostrophically outside of the tropics, their spatial length scales and degrees of freedom must be comparable, though teleconnection fields will exhibit different phase relations or preferential directions.

Thus, a reasonable estimate of the maximum spatial degrees of freedom for NB's wintertime 300 mb wind data on the mostly continental net of 174 stations would be 35. Reference to Fig. 8 therefore leads to the conclusion that based on only spatial considerations it is not possible to reject the hypothesis (at the 95% confidence level) that the zonal wind results occurred by accident. Other points that have and will be raised here bring down the confidence level further, thus this result has been labeled "suspect." In contrast, note that only 20 degrees of freedom are required for significance with 19% of the stations individually significant (the meridional component result). As pointed out above, however, NB raised the odds of getting a large percent of significant points (in the context of Fig. 8) by selecting and presenting only the best results. The odds were further inflated in two other ways related to the individual correlation tests.

First, autocorrelations of order greater than one were not taken into account. Chen (1981) found that the lag-1 year contribution to the right side of (1) was generally inconsequential, and that virtually all of the reduction in temporal degrees of freedom and loss of significant points (exemplified in Fig. 1) was the result of higher-order simultaneous autocorrelation. The point of using an approach like that embodied in (1) and (2) is to take into account the fact that sample periods may be too short to include more than a few realizations of important frequencies, a key concern for studies of 11-year cycles in which most station records (as in NB) are less than 22 years long. Careful application of this approach, like Chen's modification of (1), considerably ameliorates this problem.

Perhaps more important is the fact that another play of odds was introduced by searching for and testing only the maximum lag correlation at every point. As Panofsky and Brier (1968) point out, without a clear hypothesis of what that lag should be, the chance of obtaining a significant correlation by accident increases. Thus, just as in the selection of the winter 300 mb level, *a priori* criteria for significance cannot be used.

As an illustration of the potential effect of this selectivity of best lag and overall results, and the possible underestimation of temporal degrees of free-

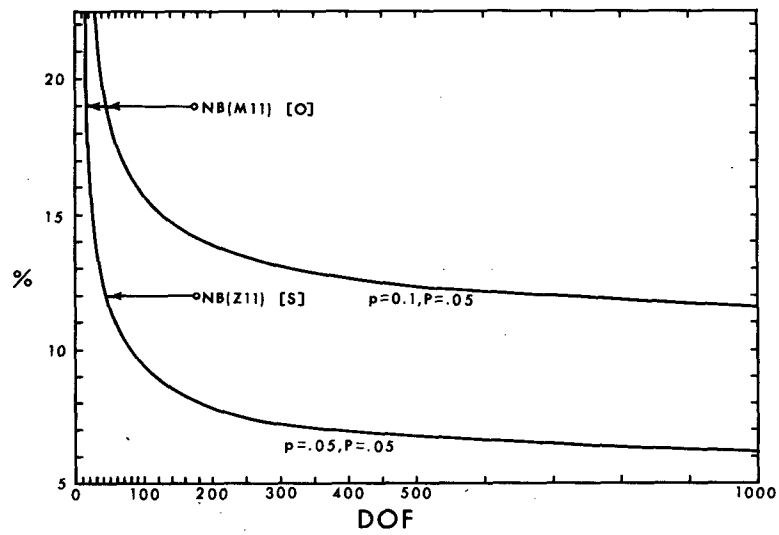

FIG. 8. As in Fig. 3, except for (p = 0.05, P = 0.05) and (p = 0.1, P = 0.05). The open circles and NB denote the experimental results of Nastrom and Belmont, while Z and M denote zonal and meridional, respectively. See Fig. 7 for the remaining notation.

dom, suppose the probability of passing the *a priori* significance test at a given station was effectively doubled to $p = 0.1$. In Fig. 8, the pertinent curve ($p = 0.1, P = 0.05$) for this situation is also plotted. Note now how the significance of NB's meridional results at the 95% level become quite problematical. Despite this, we have labeled them "open" since it is uncertain how much the odds were actually tilted.

To support their results, NB assert that "true significance is perhaps best judged by the organization of the results over the hemisphere and from level to level." They further elaborate, "A stronger case for significance will be made by showing that there are consistent, continuous patterns of atmospheric response among the stations and that these patterns are linked with clearly identified features of the mean circulation."

While these arguments are appealing, they are incorrect. Correlations between NB's data set and random numbers would not only have "well-defined patterns from level to level as well as from station to station" but would also have them from variable to variable. Likewise these patterns would be "linked with clearly identified features of the mean circulation." Indeed, large contiguous centers would be teleconnected to large remote centers of the same or opposite sign. These are all simple consequences of properties of the data set. Statistical significance must be established in spite of these consequences rather than because of them. This should, however, be possible with methods similar to those we described earlier.

Indeed, most of the difficulties associated with NB's testing can be resolved by

1) repeating the point-by-point significance tests using a form of (1) and (2) to take into account higher-order simultaneous autocorrelations,
2) conducting Monte Carlo simulations to take into account spatial correlations and selection of maximum lag correlations; and, optionally,
3) assessing overall significance of correlations at all levels and seasons simultaneously.

The Monte Carlo runs can be performed in two stages. In the first step only contemporaneous correlations with Gaussian noise would be tested in order to calibrate the technique (a mean percent of significant points close to five percent). In the second step, NB's full experiment would be repeatedly simulated: Gaussian noise is correlated at multiple lags with the seasonal mean 300 mb wind data, the maximum correlation is selected at each point and tested for significance, and the percent of points found significant is tabulated. The mean percent of significant points in these experiments will be greater than five, thereby necessitating the preliminary runs.

While tests like these are probably required for a convincing demonstration of statistical significance in studies like NB, demonstration of the converse can often involve considerably less computational effort. In the case of NB, Dr. David E. Venne (personal communication, 1982) of Control Data Corporation recently completed a considerably simplified version of our recommendation 2) above. Venne first estimated the noise level in the 10.7 cm solar flux time series by subtracting out the best fit 11-year period sine wave. The series was then reconstructed by adding a Gaussian random perturbation (with the same standard deviation as the original series residual) to the fitted 11-year wave. With this modeled series, NB's original experiment was repeated with substantially the same results. Similarly, the experiment was conducted with reconstructed time series made up of 5-, 8-, 14-, or 17-year periods and a noise component. For five of the six (the original and five reconstructed) series, the averaged explained variances turned out to be virtually the same, with the five-year period series explaining a little less than the others. Thus, NB's results were not unique and were very likely merely a consequence of the common characteristic of the six series, namely long-period trend.

5. Concluding Remarks

The principal goals of this paper were to demonstrate that, first, number and interdependence of significance tests often have non-trivial impact on the assessment of their collective significance, and that, second, in many cases the combined effect can be straightforwardly taken into account by a two-step test. Number is first taken into account by computing binomial odds of the experimental result. If they are smaller than a preset criterion, estimates of effective degrees of freedom must be compared to the minimum required by the binomial distribution. If these are of the same order of magnitude a Monte Carlo simulation must be performed to finally ascertain significance.

Three out of four of the examples presented in Sections 3 and 4 (Chen, HY and NB) were studies of the relationship of a single time series with the members of a finite set of related time series, while the fourth (Williams) involves a set of differences in means. All lend themselves well to Monte Carlo approaches, including Williams' study in which the random element would be in the selection of the 10 "warmest" years.

This suggests that perhaps an effective Monte Carlo strategy could also be devised to aid the evaluation of the significance of differences between undisturbed and disturbed GCM climates in sensitivity studies (cf. Chervin and Schneider, 1976a, 1976b; Chervin *et al.*, 1976). The need to account for spatial and variable correlation in these models was recognized by Chervin (1981), who also pointed out why multivariate statistical tests cannot be practically applied to this problem. We will describe three possible

Monte Carlo solutions although others are certainly possible.

Suppose a "disturbed" model simulation is to be compared to a six-case mean "undisturbed" field. Point by point tests of significance can be made using Chervin's techniques (thus preserving regional detail), and the percentage of tests passed can be readily determined. To continue, an estimate is needed of the probability distribution of this percentage for random differences of two model simulation means.

One way to proceed depends on whether, in addition to the six undisturbed cases, there are nine more model runs on the shelf. If this is the situation, seven at a time can be randomly selected and six of these compared to the seventh. It is possible to do this more than 6000 different ways, so it would be easy to choose several hundred relatively independent combinations. This approach is conceptually preferable, but may not be practical until the next generation computer is available.

A second tactic requires that the experimenter has some confidence in the model spectral characteristics for the planetary and long baroclinic waves. If this is the case, then more than thirty years of observed atmospheric fields are available to estimate the impact of large scale cross-correlations on the unknown probability distribution. For example, in the case of monthly or seasonal means, thirty realizations permit more than two million different combinations of seven cases. On the other hand, if the researcher does not have confidence in the large-scale statistical properties of the GCM, then from the outset there seems little basis for attaching any significance to a sensitivity test.

A third approach whose feasibility is being examined by R. Preisendorfer and T. Barnett (personal communication, 1982) uses a greatly reduced sample of undisturbed runs but is more computationally complex. It consists of randomly forming and testing differences in means, without regard to the specific configuration of the actual experiment. In other words, even though the actual experiment might compare a one-case mean to a six-case mean [denoted by (1, 6)], all possible differences of means permitted by the total number of archived undisturbed runs are considered in the Monte Carlo simulations; that is, with eight undisturbed runs (1, 7), (2, 6), (3, 5), (4, 4), (1, 6), (2, 5), (3, 4), . . . , (1, 1). In this example, thousands of differences of means can be formed. If this or one of the other approaches proves feasible and effective, the significance of GCM sensitivity experiments can be described more authoritatively.[3]

One type of empirical problem not dealt with here is the determination of the significance of a grid of point-by-point correlations of one field with another. In this case, Monte Carlo runs could be generated by successive randomizations of the temporal order of one of the data fields. The field interrelationships embodied in this type of data set could perhaps be more efficiently described and their significance evaluated through EOF representation and the analytical methods of either Davis (1976, 1977, 1978) or Barnett and Hasselmann (1979). Nevertheless, for a broad class of empirical studies similar to those we have examined here, Monte Carlo significance testing remains simple, relatively inexpensive and decisive.

Acknowledgments. The authors gratefully acknowledge the support and interest of Dr. Donald L. Gilman. In addition, many of the ideas presented were crystallized and much of the text clarified in discussions with Dr. Richard W. Reynolds. A number of other individuals expertly and critically commented on various aspects of preliminary drafts of the manuscript, and much (but not all) of what they suggested is reflected in the final version. We therefore thank Messrs. R. M. Chervin, C. J. Neumann, R. S. Quiroz and C. F. Ropelewski, and Drs. T. P. Barnett, T. L. Bell, J. B. Blechman, R. N. Hoffman, W. H. Klein, S. K. LeDuc, R. L. Lehman, A. B. Pittock and J. E. Walsh for their advice. Michael C. Gaidurgis drafted the figures while Gail S. Lucas typed the manuscript.

REFERENCES

Barnett, T. P., and R. W. Preisendorfer, 1978: Multifield analog prediction of short-term climate fluctuations using a climate state vector. *J. Atmos. Sci.*, **35**, 1771–1787.

——, and K. Hasselmann, 1979: Techniques of linear prediction, with application to oceanic and atmospheric fields in the tropical Pacific. *Rev. Geophys. Space Phys.*, **17**, 949–968.

Chen, W. Y., 1981: Fluctuations in Northern Hemisphere 700 mb height field associated with the Southern Oscillation. *Mon. Wea. Rev.*, **110**, 808–823.

Chervin, R. M., 1981: On the comparison of observed and GCM-simulated climate ensembles. *J. Atmos. Sci.*, **38**, 885–901.

——, and S. H. Schneider, 1976a: A study of the response of NCAR GCM climatological statistics to random perturbations: Estimating noise levels. *J. Atmos. Sci.*, **33**, 391–404.

——, and ——, 1976b: On determining the statistical significance of climate experiments with general circulation models. *J. Atmos. Sci.*, **33**, 405–412.

——, W. M. Washington and S. H. Schneider, 1976: Testing the statistical significance of the response of the NCAR general circulation model to North Pacific ocean surface temperature anomalies. *J. Atmos. Sci.*, **33**, 413–423.

Davis, R. E., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **6**, 249–466.

——, 1977: Techniques for statistical analysis and prediction of geophysical fluid systems. *Geophys. Astrophys. Fluid Dyn.*, **8**, 245–277.

——, 1978: Predictability of sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **8**, 233–246.

Egger, J., G. Meyers and P. B. Wright, 1981: Pressure, wind and cloudiness in the tropical Pacific related to the Southern Oscillation. *Mon. Wea. Rev.*, **109**, 1139–1149.

[3] Storch (1982) has recently offered a fourth candidate solution that does not require Monte Carlo simulations, but two "similar, but statistically independent" sensitivity experiments. The conditions under which this procedure can be applied were not discussed further and need to be clarified.

Gilman, D. L., 1957: Empirical orthogonal functions applied to thirty day forecasting. Sci. Rep. No. 1, Contract AF19(604)-1283, MIT, 129 pp. [Available from the Micro Reproduction Laboratory, MIT, Rm. 14-0551, Cambridge, MA 02139].

Hancock, D. J., and D. N. Yarger, 1979: Cross-spectral analysis of sunspots and monthly temperature and precipitation for the contiguous United States. *J. Atmos. Sci.*, **36,** 746–753.

Harnack, R. P., 1980: An appraisal of the circulation and temperature pattern for winter 1978–79 and a comparison with the previous two winters. *Mon. Wea. Rev.*, **108,** 37–55.

Lund, I. A., 1970: A Monte Carlo method for testing the statistical significance of a regression equation. *J. Appl. Meteor.*, **9,** 330–332.

Madden, R. A., and P. R. Julian, 1971: Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci.*, **28,** 702–708.

Mitchell, J. M., B. Dzerdzeevskii, H. Flohn, W. Hofmeyer, H. Lamb, K. Rao and C. Wallen, 1966: *Climatic Change.* Tech. Note No. 79. WMO, 79 pp.

Namias, J., 1980: Causes of some extreme Northern Hemisphere climatic anomalies from Summer 1978 through the subsequent winter. *Mon. Wea. Rev.*, **108,** 1333–1346.

Nastrom, G. D., and A. D. Belmont, 1980: Evidence for a solar cycle signal in tropospheric winds. *J. Geophys. Res.*, **85,** 443–452.

Neumann, C. J., M. B. Lawrence and E. L. Caso, 1977: Monte Carlo significance testing as applied to statistical tropical cyclone prediction models. *J. Appl. Meteor.*, **16,** 1165–1174.

Panofsky, H. A., and G. W. Brier, 1968: *Some Applications of Statistics to Meteorology.* The Pennsylvania State University, 224 pp.

Pittock, A. B., 1978: A critical look at long-term sun–weather relationships. *Rev. Geophys. Space Phys.*, **16,** 400–420.

Reynolds, G., 1978: Two statistical heresies. *Weather*, **33,** 74–76.

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

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.*, **109,** 784–812.

Walsh, J. E., and A. Mostek, 1980: A quantitative analysis of meteorological anomaly patterns over the United States, 1900–1977. *Mon. Wea. Rev.*, **108,** 615–630.

Williams, J., 1980: Anomalies in temperature and rainfall during warm Arctic seasons as a guide to the formulation of climate scenarios. *Climatic Change*, **2,** 249–266.

Zurndorfer, E. A., and H. R. Glahn, 1977: Significance testing of regression equations developed by screening regression. *Preprints, Fifth Conf. on Probability and Statistics in Atmospheric Sciences*, Las Vegas, Amer. Meteor. Soc., 95–100.