Vol. 2: 35–45, 1992 | CLIMATE RESEARCH Clim. Res. | Published July 9

Role of statistics in the validation of general circulation models

Richard W. Katz

Environmental and Societal Impacts Group, National Center for Atmospheric Research*, Boulder, Colorado 80307, USA

ABSTRACT: The field of statistics should play a central role in the validation of general circulation models (GCMs). Both the perspectives of climate modelers and of climate impact researchers are addressed. Confusion over what statistics can or cannot be expected to achieve is described, and an overemphasis of statistical tests of significance is identified. A compelling case is made for assessing the performance of GCMs in terms of their ability to reproduce climate variability, not just average conditions. Unsatisfactory aspects of current statistical practice are pointed out, with the problem of multiple comparisons illustrated by an example. Prospects for the future are considered, emphasizing the need for factoring statistical issues into the design of GCM experiments, as well as the possible application of new statistical techniques. Indicative of ways in which existing techniques could be used more appropriately in the validation of GCM output, an example that involves a multiple confidence interval is presented. If the appropriate role of statistics were recognized, one obstacle currently hindering the improvement of GCMs and limiting their use in climate impact studies would be removed.

INTRODUCTION

General circulation models (GCMs) of the atmosphere/ocean system are becoming increasingly relied upon, both as a research tool for performing climate change experiments and as one source of information for assessments of the impact on society of any anticipated climate change (Houghton et al. 1990). Accompanying this increased reliance is an increased awareness of the need for the validation and intercomparison of the various forms of existing GCMs. Model validation could involve comparing a GCM control run to the corresponding observed climate, or it might even include trying to detect in the observations the 'fingerprint' of climate change (e.g. corresponding to increases in the atmospheric concentration of certain greenhouse gases) as prescribed by a GCM experiment (Barnett & Schlesinger 1987). Model intercomparison would consist of comparing the output produced by 2 or more GCMs, all run in the identical mode (i.e. all control runs or all some particular type of experimental run).

Although the field of statistics has played a relatively minor role in the development of GCMs, it should have a central role in any attempts to validate these models. Collaboration between statisticians and atmospheric scientists has proven valuable in the past on other problems, such as weather modification (e.g. Wegman & DePriest 1980). Moreover, the issues are somewhat analogous to those that have arisen with other types of models, such as for air quality (Fox 1981). This paper focuses on the role of statistics, including both what it can and cannot be expected to achieve. Several recent papers have either reviewed the literature (Livezey 1985) or proposed a general framework for the statistical evaluation of fields of output from GCMs (Preisendorfer & Barnett 1983, Zwiers 1987, Wigley & Santer 1990). Perhaps it is natural that this work has dwelt on refinements of existing statistical techniques to make their application to fields of simulated climate data more justifiable. What these papers could have stated more explicitly is the philosophical issue of why exactly statistics is being applied in the first place.

The perspective to be taken in the present paper does not simply reflect the relatively narrow confines

[*] The National Center for Atmospheric Research is sponsored by the National Science Foundation

0936-577X/92/0002/0035/$ 03.00

of the climate modeler, but attempts to address the needs of the climate impact community as well (Robinson & Finkelstein 1991). After all, it is the members of this community who actually intend to make use of the output produced by GCMs and for whom an assessment of its reliability is crucial if any policy implications are to be taken seriously. One of the reasons for the past neglect of the statistical validation of GCMs certainly stems from the fact that there was originally no intention of producing any information for use in impact studies. GCMs were simply viewed as a research tool for physical scientists. Whether or not GCMs ever provide useful output for climate impact researchers, the idea of delivering model output to a user is still an especially valuable concept when attempting to come to grips with practical goals for model validation.

First, the concept of statistical significance is contrasted with that of practical significance, with an example of hypothesis testing being provided to help clarify the issues. Next, some reasons why information about the variability of climate, not just its average, is required are presented, including the issue of the frequency of extreme events. Then the problem of multiplicity is treated, with the need for multiple comparison techniques being illustrated. Finally, prospects for the future, both design of experiments and new statistical methods, are described, and an example that entails a multiple confidence interval approach to the validation of GCM output is given.

TESTS OF SIGNIFICANCE

What should be the proper role of the field of statistics in the validation of GCMs? Because GCMs are effectively deterministic models of climate, perhaps it could even be argued that it makes no sense to apply statistical methods. At the other extreme, it is plausible to hope that statistical techniques will some day be able to attach a 'seal of approval' to the latest, improved version of GCM, certifying that the output of the model is in perfect statistical agreement with the corresponding real-world observations of climate. The truth is somewhere in between. So long as the standard of comparison for GCMs is the actual climate, it is reasonable to analyze model output as if it were stochastic. On the other hand, no matter how complex GCMs become, they will still remain imperfect models of the atmosphere/ocean system. So no GCM should ever be expected to receive statistical certification.

Recognition of the need for the statistical validation of the performance of GCMs occurred quite a while ago (e.g. Chervin & Schneider 1976, Laurmann & Gates 1977). Typically, the 'all or nothing' approach of hypothesis testing has been followed (e.g Wigley & Barnett 1990). Either a statistically significant difference is present and the GCM is said to have 'failed' the validation test, or no significant difference is found and the model is said to have 'passed' the test. Often the specific hypothesis being tested is not precisely stated. An example is now presented to illustrate these issues in more concrete terms.

Statistical significance example

Suppose that the January mean temperature at a particular grid point is being compared for 2 GCMs, say model number i, $i = 1, 2$. Assume that this variable has a normal distribution with mean μ_i and known common variance σ^2 [written $N(\mu_i,\sigma^2)$], $i = 1, 2$. Under these restrictive assumptions, the only test of significance of interest is

Null Hypothesis: $\mu_1 = \mu_2$,
Alternative Hypothesis: $\mu_1 \neq \mu_2$.

Assume that samples of size n are available for both models, and let $\bar{X}_i$ denote the sample mean for model i. Under the null hypothesis of equal means, the test statistic

$$Z = \frac{\bar{X}_2 - \bar{X}_1}{(2/n)^{1/2}\sigma} \qquad (1)$$

has a standard normal distribution (i.e. zero mean and unit variance).

Suppose that the known common value of the variance is $\sigma^2 = (3\ °\mathrm{C})^2$. Table 1 summarizes the results of 2 tests of significance, one for samples of size $n = 9$ and another for $n = 100$. Both cases have the same observed difference in sample means $\bar{X}_2 - \bar{X}_1 = 1\ °\mathrm{C}$. Table 1 indicates that the null hypothesis would be retained in the $n = 9$ case, but rejected (at the 0.05 level) in the $n = 100$ case.

Table 1 Statistical significance example (difference in January mean temperature between models 1 and 2)

	Sample size	
	$n = 9$	$n = 100$
Sample mean $\bar{X}_1$	0 °C	0 °C
Sample mean $\bar{X}_2$	1 °C	1 °C
Variance σ^2	(3 °C)^2	(3 °C)^2
Test statistic Z	0.707	2.357
p-value	0.480	0.018
95 % confidence interval for $\mu_2 - \mu_1$	(−1.77 °C, 3.77 °C)	(0.17 °C, 1.83 °C)

The point is that any fixed difference in sample means, no matter how small in magnitude, would be declared statistically significant given large enough sample sizes. This effect is perhaps best seen in terms of the confidence interval for the real difference in means, $\mu_2 - \mu_1$, corresponding to the test statistic (1):

$$\overline{X}_2 - \overline{X}_1 \pm 1.96(2/n)^{1/2}\sigma, \quad (2)$$

at the 95 % level. Table 1 shows that the larger common sample size of $n = 100$ has a much shorter confidence interval.

Keeping this example in mind, it is now argued that the typical way in which the performance of GCMs is statistically validated violates the basic tenets underlying the philosophy of hypothesis testing. Ideally, tests of significance should be formulated so that the researcher hopes to reject the null hypothesis. In fact, the alternative hypothesis is sometimes referred to as the 'research hypothesis'. But in the validation of GCMs, especially comparisons between a control run and the corresponding actual climate or between an experimental 'fingerprint' and observed changes in climate, it is expressly desired to retain the null hypothesis of perfect statistical agreement (Wigley & Barnett 1990).

If the null hypothesis is retained, there are 2 possible explanations: (1) the null hypothesis is actually true; or (2) a real difference exists, but the test of significance has insufficient power to detect it. As has already been remarked, the first possibility can be dismissed when dealing with GCMs. For the second possibility, the statistical significance example just presented illustrates that, given large enough sample, the null hypothesis will eventually be rejected. In other words, in the context of the validation of GCMs, the null hypothesis can only be retained for the *wrong reason*.

The confusion over this issue can be attributed to the failure to make the distinction between 'statistical significance' and 'practical significance'. A GCM could provide a very accurate approximation to reality from a physical point of view, yet still produce discrepancies that are statistically significant. At least one group of GCM evaluators stumbled onto this difficulty, but gave the bewildering explanation that a particular test of significance is 'too powerful' (Preisendorfer & Barnett 1983). If a hypothesis test is formulated properly, it can never be too powerful. Whatever the specific needs of climate modelers, they are certainly not being met by the sole reliance on statistical tests of significance. As suggested in the example on statistical significance, one alternative to formal hypothesis testing procedures would be a confidence interval approach.

Confidence intervals

An isomorphism exists between most tests of significance and their confidence interval analogues. Hayashi (1982) advocated the use of confidence intervals to validate GCMs, and Katz (1982, 1983, 1988) included procedures for obtaining them. Nevertheless, confidence intervals have been only rarely employed in the operational validation of GCMs, and apparently their usefulness is still not appreciated.

With a confidence interval, a researcher is presented information about an observed difference in the form of a range of magnitudes, rather than just a 'yes' or 'no' answer concerning statistical significance. Such information can be viewed as providing a measure of the accuracy of the model approximation in terms of one particular metric. Yet the essential information provided by a test of significance is not lost; the null hypothesis is rejected if and only if the hypothesized values of the parameters involved are not contained in the confidence interval. One instance in which confidence intervals are employed in a complex situation is given later in the paper.

It should be noted that Fox (1981) recommended the use of confidence intervals in evaluating air quality models, making an analogous argument. Also, von Storch & Zwiers (1988) and Zwiers & von Storch (1989) introduced the concept of 'recurrence analysis', another way of quantifying any model discrepancies in more physically meaningful terms than ordinary tests of significance. Finally, Solow (1990) proposed a procedure for discriminating among competing models, yet another improvement upon hypothesis testing.

CLIMATE VARIABILITY AND EXTREMES

Climate modelers have usually focused only on the average performance of GCMs, as opposed to their ability to reproduce variability or the relative frequency of extreme events. Climate variability is typically regarded as 'noise', a nuisance that should be removed so that only the 'signal' remains. Indeed, much of the recent impetus for the consideration of how variability and the frequency of extreme events might change with a change in average climate has come from the climate impacts community. This concern arises naturally, since the impacts of climate on society are realized largely through the incidence of variations about 'normal' conditions or of extreme events (such as excursions outside of some range) (Wigley 1985). In the somewhat analogous situation of air quality models, this issue received more attention because standards are specified in terms of extremes (Fox 1981).

From a statistical perspective, which test of significance for differences in variability should be employed is not obvious (Katz 1988). For instance, the standard *F*-test for comparing 2 variances is highly sensitive to the assumption of normality (Box 1953). Specifically, this test is affected by the kurtosis (or 4th moment) of the distribution, roughly speaking a measure of the degree of 'peakedness' or 'flatness'. Unlike the case of the mean, this effect does not diminish with larger sample sizes. Since normality is at best only an approximation for climate variables (whether dealing with actual observations or GCM simulations), the *F*-test cannot be unequivocally recommended. Katz (1988) outlined how a simple 'standard error' procedure, that corrects for kurtosis, could be applied to construct tests of significance or confidence intervals for the variance of GCM simulated climate variables.

Tests of significant differences in variability also tend to have relatively low power. The heuristic explanation for this characteristic is that the uncertainty in sample variances involves a 'variance' of variances, not of averages. In other words, the error in estimating the variance arises as a propagation of the error in estimating the mean. Perhaps for this reason, the few attempts to detect changes in variability in GCM climate experiments have been largely inconclusive (e.g. Mearns et al. 1990).

Even the question of how should variability be measured is not routinely answered. Because of the presence of autocorrelation in climate time series, several different possibilities arise, including the so-called 'innovation' variance of the 'prewhitened' series (i.e. obtained by removing the autocorrelation) (Katz 1988). Being based on a filtered, uncorrelated time series, the innovation variance is both more statistically tractable and, in some respects, more physically meaningful than the variance of the original, autocorrelated time series. The variance of a time average of an autocorrelated climate variable, of particular interest in some applications, can be just as succinctly expressed in terms of the innovation variance as in terms of the variance of the original process. Inferences about the innovation variance of GCM output have been made by Wilson & Mitchell (1987) and Mearns et al. (1990).

The neglect of climate variability also stems from a lack of appreciation of just how critical this information is for quantifying the relative frequency of occurrence of extreme events. An example based on the work of Katz & Brown (1992) is now given.

Extreme events example

Consider a climate variable X, with distribution function F, say the daily maximum temperature at a specific location. Assume that F is the $N(\mu, \sigma^2)$ distribution. Suppose that the extreme event that a threshold c is exceeded by X is of interest; that is, the event $E = \{X > c\}$ with probability of occurrence $P(E)$. Here $P(E) = 1 - F(c)$. Of interest is how much this probability changes as either the mean μ or the standard deviation σ change.

Fig. 1 compares $P(E)$ (i.e. the area under the curve to the right of the threshold c) for 3 different normal distributions: (i) the reference case of $N(\mu, \sigma^2)$ (Fig. 1a); (ii) the case in which the mean is changed to a new value μ^*, keeping σ fixed (Fig. 1b); and (iii) the case in which the standard deviation is changed to a new

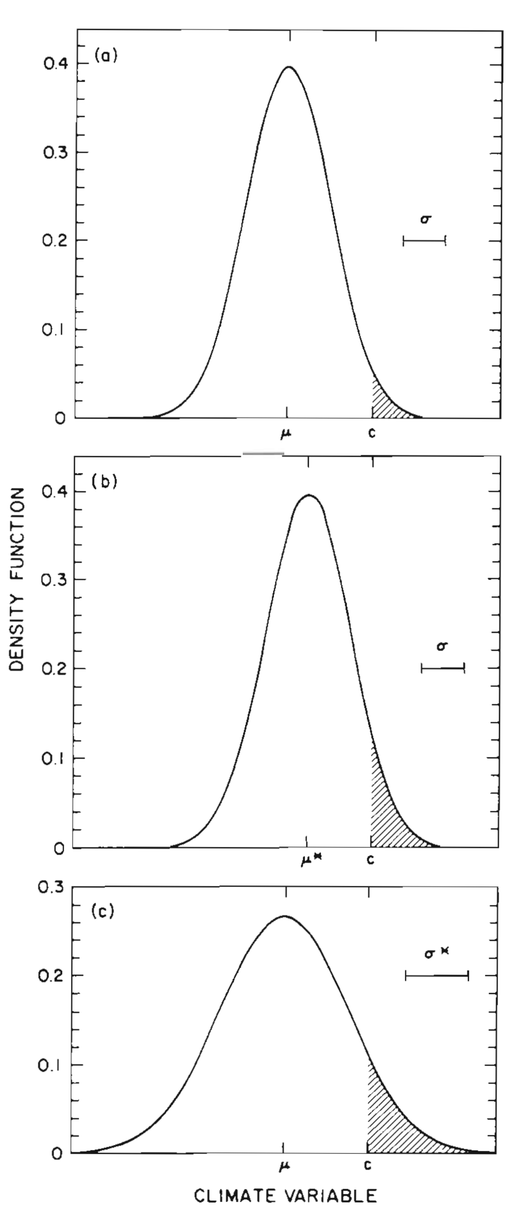

Fig. 1. Probability of extreme event (i.e. exceeding threshold c) for 3 different normal distributions: (a) $N(\mu, \sigma^2)$; (b) $N(\mu^*, \sigma^2)$; and (c) $N(\mu, (\sigma^*)^2)$

value σ^*, keeping μ fixed (Fig. 1c). To make the comparison reasonable, the value of σ^* in case (3) differs from the original standard deviation σ by the same amount as the value of μ^* in case (2) differs from the original mean μ. Despite this constraint, the probability of the extreme event E is considerably higher for the increase in the standard deviation than for the increase in the mean (see Fig. 1). Katz & Brown (1992) established that this result holds much more generally for a wide range of thresholds, for different distributions than the normal, and for more complex forms of extreme event. Mearns et al. (1984) obtained similar results concerning how the relative frequency of extreme high temperature events (e.g. a run of consecutive days on which the maximum temperature exceeds a threshold) changes as the mean and standard deviation change.

Because extreme events are inherently rare and because of the limited sample sizes produced by GCMs, it may not be possible to make statistical inferences about the frequency of extreme events directly from GCM output. Indeed, the extreme events produced by GCMs have been examined in even fewer instances than for overall variability. Nevertheless, it is clear that some of the attention now focused on averages should be diverted to model-simulated variability. If GCMs could provide information about both changes in averages and variances, then the techniques of extreme value theory could be employed to translate these changes into terms of the frequency of extreme events (Katz & Brown 1992).

MULTIPLICITY

One serious difficulty that arises when applying tests of significance to the validation of GCMs is the so-called 'multiplicity' problem (e.g. Tukey 1977). This difficulty is related to the multivariate nature of GCM output, which naturally consists of fields of data. Some confusion that currently exists over how best to deal with the multiplicity problem is pointed out, and it is argued that this confusion is really symptomatic of a more fundamental misunderstanding of the proper role of hypothesis testing already discussed.

Early work on the statistical validation of GCMs stressed univariate tests of significance, ordinarily applied separately to each of a multitude of grid points (e.g. Chervin & Schneider 1976). At first, these univariate test results were treated as if literally performed individually in isolation from the others. Then Hasselmann (1979) proposed a fully multivariate approach to the statistical validation of fields of GCM output. Although Hasselmann's approach has never proved to be operationally feasible, soon researchers were castigated for ignoring this multiplicity problem (von Storch 1982).

The first operational procedures for treating fields of GCM output (or of actual climate data), that account for their multivariate nature, are based of the idea of 'resampling' These methods were introduced by Livezey & Chen (1983) and Preisendorfer & Barnett (1983). Later, Zwiers (1987) examined such resampling techniques in more detail from a statistical point of view, whereas Santer & Wigley (1990) and Wigley & Santer (1990) applied them to fields of actual climate data and GCM output in a more systematic fashion.

In general, the problem of multiplicity arises when numerous tests are performed simultaneously. In essence, if each individual test has a fixed nominal level of significance, then the probability of erroneously rejecting at least one null hypothesis rapidly grows (in many cases at a geometric rate) with the number of tests performed. For instance, if $k = 10$ independent tests of significance are performed at the nominal level of 0.05, then the probability of rejecting one or more of the null hypotheses is about 0.40. In this sense, the researcher is really analyzing the data as if an error rate of 40 % were acceptable, presumably much too high in practice.

To combat multiplicity, some forceful arguments can be made in favor of adopting a multivariate approach. Either a statistic can be selected whose form is suggested by the theory of multivariate statistical inference, or one can be obtained through some combination of univariate test statistics applied simultaneously to multiple grid points (Zwiers 1987). Then resampling techniques can be invoked to empirically estimate their distributions under the appropriate null hypothesis. In effect, a single test of significance is performed in one fell swoop, so that the overall level of significance can still be maintained at the desired level.

While the advance from univariate to bona fide multivariate methods certainly is a welcome step, several fundamental issues need to be addressed when these more sophisticated techniques are applied. The sweeping nature of multivariate techniques encourages the validator to allow for the broadest possible null hypothesis. But for the relatively large number of grid points typical of GCM output (and this number is expected to increase in the future with faster supercomputers), the power or sensitivity of the multivariate testing procedure will be extremely low. More importantly, the multiplicity issue has been only temporarily avoided. Once the null hypothesis has been rejected, a formal multiple comparison procedure is needed to identify exactly where the real differences are located

(Miller 1981). Otherwise, the time at which the sin of multiplicity is committed has simply been postponed until a later stage of the data analysis.

Multiple comparison example

This problem is illustrated by way of a relatively simple situation. Suppose that 2 GCMs, denoted by model number i, $i = 1, 2$, are being compared in terms of the mean of some particular climate variable at only $k = 2$ individual grid points. Let $\mu_i(j)$ denote the mean at grid point j, $j = 1, 2$, for model i. Formally, the multivariate approach involves testing for the equality of means simultaneously for both grid points. That is, a composite (or 'global') null hypothesis consisting of 2 individual (or 'local') null hypotheses is tested:

$$\text{Null Hypothesis: } \mu_1(j) = \mu_2(j), \quad j = 1, 2. \qquad (3)$$

Because the alternative hypothesis is, in general, quite complex, it is rarely explicitly stated in the climate literature. Logically, it is just the negation of the null hypothesis; that is, any configuration of the mean values, $\mu_i(j)$, which does not satisfy the conditions (3) specified by the null hypothesis. In this special case of only 2 grid points, 3 different situations are possible:

(i) $\mu_1(1) \neq \mu_2(1), \mu_1(2) = \mu_2(2)$,
(ii) $\mu_1(1) = \mu_2(1), \mu_1(2) \neq \mu_2(2)$,
(iii) $\mu_1(1) \neq \mu_2(1), \mu_1(2) \neq \mu_2(2)$.

Situations (i) and (ii) involve only one of the 2 individual null hypotheses being false, whereas situation (iii) involves both individual null hypotheses being false. The difficulty arises because, simply knowing that the null hypothesis (3) has been rejected, nothing necessarily can be said about which of these 3 specific alternatives is most consistent with the data. Of course, in practice many more than 2 grid points are treated simultaneously, implying that the possible number of specific alternatives (in terms of which of the individual null hypotheses are false) is quite large.

Given a rejection of the global null hypothesis, techniques for identifying which of the individual null hypotheses should be rejected are termed 'multiple comparison' procedures. One popular and conceptually simple procedure is based on the so-called Bonferroni inequality (Miller 1981). Specifically, suppose that k individual tests of significance are being considered. Then conducting each individual test at the local level

$$\alpha_L = \frac{\alpha}{k} \qquad (4)$$

guarantees that the overall global level, α_G say, satisfies $\alpha_G \leq \alpha$, even if the tests were dependent. For instance, if $k = 10$ individual tests are involved, then by (4) the local levels of significance must be reduced to 0.005 to achieve a global level of at most 0.05.

The Bonferroni technique works well if a relatively small number of individual hypotheses are being treated. When the total number of tests k is large, the procedure tends to be overly conservative (i.e. producing an overall level of significance that is somewhat lower than the desired value). Katz & Brown (1991) applied the Bonferroni approach to the closely related climate application of searching for teleconnections. This technique can also be employed to produce a multiple confidence interval, as is demonstrated later. Madden & Julian (1971) used the same adjustment (4) to the significance level in the different context of testing for peaks in a spectrum. Walker (1914) was apparently the first to recognize the need to reduce the individual significance levels to correct for multiplicity, using an adjustment based on the assumption of independent tests instead of (4).

Although the simultaneous testing of multiple hypotheses has become quite prevalent in recent years in the validation of GCMs, the need for multiple comparison techniques is apparently still not recognized. Typically, researchers are vague about exactly what can be concluded about local differences at individual grid points upon rejection of the global null hypothesis. The idea of being forced to employ the smaller level of significance specified by (4) is met with resistance and is felt to be unduly conservative (e.g. Livezey & Chen 1983).

Because the implementation of existing multivariate statistical procedures tends to be unfeasible for the relatively large number of grid points characteristic of a GCM, validators have usually resorted to computationally intensive, resampling techniques instead. The operation of 'resampling' is an integral part of several closely related techniques: e.g. 'permutation' or 'randomization' tests (Mielke & Brier 1981), the 'jackknife' and 'bootstrap' (Efron 1982), and 'cross validation'. In essence, the distribution of a given test statistic is evaluated empirically, through some scheme for manufacturing additional samples from the original one.

Are resampling techniques a 'panacea' or a 'Pandora's box'? Such procedures are certainly of potential value in that they circumvent some of the limitations of mathematical statistics. Unfortunately, the conventional wisdom in the application of these techniques to GCM output is that resampling will automatically account for any undesirable features of the statistic being scrutinized. Actually, resampling is no substitute for the principles of statistical inference, on which the original derivation of many optimal procedures whose form is now taken for granted is based.

One instance in which this reliance on resampling has led GCM evaluators astray is now cited. Preisendorfer & Barnett (1983), in one of the most influential papers on resampling in the climate literature, considered the so-called 'Euclidean distance statistic.' As Zwiers (1987) later demonstrated, this statistic has the undesirable property of not being scale invariant, and resampling can do nothing to obviate this feature. In the same vein, Wigley & Santer (1990) treated trial statistics, including the Euclidean distance, without regard to whether they satisfy certain desiderata. Zwiers (1990) provided another instance in which resampling procedures have been misused, this time in the context of solar-climate correlations.

Multivariate techniques are applied to GCM output, apparently without the realization that the null and alternative hypotheses literally involve thousands of different parameter configurations, about which no preferences are specified. Agreement on some relatively small set of more focused hypotheses is needed. Otherwise, there is little hope of ever having a statistical procedure for model validation that is either able to detect local and regional differences with a high enough probability or to quantify these discrepancies with a satisfactory degree of precision. This limitation is not inherently statistical, and multivariate procedures are no cure-all. Too many ill-posed questions are being asked to expect to receive precise answers. In an attempt to achieve a compromise between local and global comparisons, an example involving a multiple confidence interval is presented in the next section.

PROSPECTS FOR THE FUTURE

Design of experiments

Not only have statistical issues been largely ignored in the historical development of GCMs, but these issues have not been an important factor in the design of experiments that make use of GCMs. This neglect is in spite of the fact that some early work on the validation of GCMs specifically called for the statistical design of experiments (Chervin & Schneider 1976, Laurmann & Gates 1977). Such design considerations would involve the question of how long a control run should be performed for purposes of model validation or of how long an experimental run should be performed for the detection of climate change.

Although some calculations of power have been performed (e.g. Zwiers 1987), evidently these results have never been employed to help make decisions on how long a control/experiment run should actually be produced in practice. Of course, computer time is frequently cited as the limiting factor in determining the length of model runs. But a lesson can be gained by recalling the expensive, time-consuming experiments in weather modification in the recent decades that failed to establish any conclusive results (Brillinger et al. 1978). In this somewhat analogous situation, the high degree of temporal and spatial variability characteristic of precipitation resulted in quite low power, a fact only appreciated and lamented with hindsight.

Much would be gained in efficiency by taking into account statistical considerations in the planning of future GCM experiments. Although such calculations might be tedious, they would generate, in effect, 'improved' GCM output, without any improvement in the basic understanding of the atmosphere/ocean system. A need also exists for experiments specifically designed to address certain statistical issues heretofore largely ignored by climate modelers. Recalling the discussion in an earlier section, experiments should be performed to study questions concerning the variability of climate as generated by GCMs, including the attribution of the sources of variation (e.g. air-sea interaction). A calculation of the length of model runs required to achieve a satisfactory level of power when comparing variances might be very sobering (Katz 1988, Mearns et al. 1990).

New techniques

More sophisticated statistical methods for testing hypotheses (e.g. resampling schemes such as the bootstrap) are certainly welcome additions (Willmott et al. 1985, Wigley & Santer 1990). But more fundamental changes in approach are needed as well. To fully address the multivariate nature of climate, a spatial and simultaneously temporal stochastic model for the entire globe should be developed. What is desired is a parsimonious representation of the statistical characteristics of fields of climate data in terms of a few parameters, akin to climate modelers' use of spherical harmonics. Hypothesis testing would be relatively straightforward, since only a few parameters would be involved, instead of the current plethora of variables.

Existing techniques for the reduction of fields of climate data are inadequate. For instance, principal component (or empirical orthogonal function) analysis does not actually make any direct use of the spatial location of the data. This information is only taken into account indirectly through the spatial correlations of climate variables. A genuine space-time stochastic model would be in the spirit of stochastic climate models (as proposed by Hasselmann 1976), and might even be a viable replacement for existing GCMs for some purposes. Of course, such a model is nowhere near being developed, both because of its inherent complexity and because so little is currently known about the spatial structure of climate statistics.

Nevertheless, one of the themes of the present paper is that the most pressing need is not new, improved techniques, but more appropriate use of existing techniques. The following example is designed to indicate some directions that this improved use might take. Two different issues are combined, a 'regional' scale compromise between local and global comparisons and a switch from hypothesis testing to confidence intervals.

Multiple confidence interval example

Output from a relatively long control run of the Oregon State University (OSU) GCM is analyzed. It has been suspected that climate 'drift' is present in this model run, rather than stationary, equilibrium conditions. With this apparent effect in mind, the annual mean temperature is compared for 2 different 10 yr time periods, years numbered 11 to 20 and 51 to 60. It is desired to quantify the differences in temperature between the 2 periods, as opposed to just testing whether or not they are statistically significant. Consequently, a confidence interval approach is adopted.

A single region consisting of 5 grid points is selected, to help demonstrate how a compromise between local and global comparisons can be achieved operationally. The latitudes and longitudes of these grid points are listed in Table 2, and their locations are indicated on Fig. 2. This region includes a major portion of the U.S. Corn Belt, an important agricultural area. Although 5 grid points might appear to be a relatively small number, when assessing the societal impacts of climate the individual regions would ordinarily not be of any larger scale than this one.

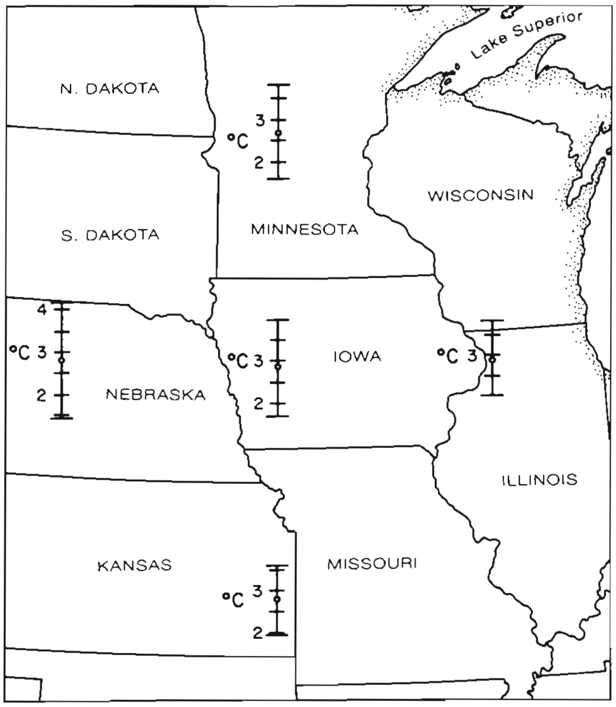

Fig. 2. Conservative 95 % multiple confidence interval (equivalent to 99 % individual level of confidence) for differences in mean annual temperature (Years 51 to 60 minus Years 11 to 20) produced by OSU GCM control run for a region in the USA consisting of 5 grid points. (o) Grid point location and midpoint of confidence interval

The same Bonferroni technique, discussed in the context of tests of significance in the previous section, can be employed to obtain an approximate multiple (or regional) confidence interval from individual (or local) intervals. Suppose that a joint confidence interval for

Table 2. Multiple confidence intervals for annual mean temperature for 5-grid point region as simulated by OSU GCM control run

Grid point	Time period (yr)	Sample mean (°C)	Sample variance (°C)^2	95 % confidence interval for difference in means[a] (°C)	95 % confidence interval for ratio of variances[a]
38° N,	11–20	13.88	0.531	2.79 ± 0.81	(0.062, 2.67)
95° W	51–60	16.67	0.217		
42° N,	11–20	6.04	0.621	2.82 ± 1.36	(0.370, 15.82)
100° W	51–60	8.86	1.502		
42° N,	11–20	9.98	1.017	2.84 ± 1.12	(0.068, 2.93)
95° W	51–60	12.82	0.455		
42° N,	11–20	10.89	0.581	2.93 ± 0.85	(0.071, 3.04)
90° W	51–60	13.82	0.270		
46° N,	11–20	7.27	1.051	2.70 ± 1.11	(0.048, 2.07)
95° W	51–60	9.97	0.333		

[a] Joint level of confidence is at least 95 % (equivalent to 99 % individual level)

k parameters with level $100(1 - \alpha)$ % is desired. The combination of k individual confidence intervals with level $100(1 - \alpha_L)$ %, where α_L is again given by (4), guarantees that the overall regional level, $1 - \alpha_R$ say, satisfies $100(1 - \alpha_R)\ \% \geq 100(1 - \alpha)\ \%$.

Table 2 gives the 10 yr sample means and variances for the 2 time periods at the 5 grid points. It is evident that the second period is consistently warmer on the average than the first period, with the difference in sample means always being slightly less than 3 °C. The pattern in the sample variances is more erratic, in most cases being larger in the first period than in the second.

Let $\mu_i(j)$ denote the true mean temperature for time period i ($i = 1$ indicating years 11 to 20; $i = 2$ indicating years 51 to 60) and grid point j, $j = 1, 2, \ldots, k$ (with $k = 5$). Here the simplifying assumption is made that this mean is constant within the individual 10 yr time periods. A joint, say 95 %, confidence interval is desired for the 5 grid point differences in means: $\mu_2(j) - \mu_1(j)$, $j = 1, 2, \ldots, 5$. So 5 individual 99 % confidence intervals are needed.

In view of the apparent differences in temperature variability given in Table 2, approximate t-confidence intervals are obtained without making the usual assumption of equal variances. An approximate $100(1 - \alpha)$% confidence interval for $\mu_2(j) - \mu_1(j)$ is given by:

$$\overline{X}_2(j) - \overline{X}_1(j) \ \pm\ t_\alpha s[\overline{X}_2(j) - \overline{X}_1(j)], \qquad (5)$$

where

$$s^2[\overline{X}_2(j) - \overline{X}_1(j)] \ =\ \frac{s_1^2(j) + s_2^2(j)}{n}.$$

Here the $\overline{X}_i(j)$ and $s_i^2(j)$ are the sample means and variances, and $n = 10$ is the common sample size. The critical value t_α is obtained from the t-distribution, with the approximate degrees of freedom being estimated from the 2-sample variances by a complex expression, known as the Smith-Satterthwaite procedure and typically used in statistical software programs (e.g. Ryan et al. 1985).

Table 2 lists these 5 individual 99 % confidence intervals [i.e. $\alpha = 0.01$ in (5)], and they are shown in Fig. 2. The midpoints are virtually identical in all 5 instances, whereas the lengths of the intervals vary substantially due to the differences in individual variances. Incidentally, since at least 1 (in fact, all 5) of the intervals does not contain the value zero, the null hypothesis that all 5 differences in means equal zero is rejected at the 0.05 regional level.

Formally, the joint regional confidence interval is a parallelopiped in 5-dimensional Euclidean space $\Re^5$ (i.e. $\Re$ is the real line and $\Re^5 = \Re \times \Re \times \cdots \times \Re$). A conservative 95 % confidence interval for $[\mu_2(1) - \mu_1(1), \mu_2(2) - \mu_1(2), \ldots, \mu_2(5) - \mu_1(5)]$ is (1.98 °C, 3.60 °C) × (1.46, 4.18) × ··· × (1.59, 3.81). Of course, it is not feasible to present a k-dimensional parallelopiped visually. But either 3-dimensional graphical techniques or 2-dimensional ones with color enhancement could be used to improve upon Fig. 2.

Let $\sigma_i^2(j)$ denote the true variance of temperature for time period i, $i = 1, 2$, and grid point j, $j = 1, 2, \ldots, 5$. Again, the assumption is made that this variance is constant within the 10 yr time period. Table 2 includes a joint 95 % confidence interval for the 5 variance ratios $[\sigma_2^2(1)/\sigma_1^2(1), \sigma_2^2(2)/\sigma_1^2(2), \ldots, \sigma_2^2(5)/\sigma_1^2(5)]$, based on the F-statistic (but recall the limitations of this procedure noted previously). All of these individual intervals contain the value 1 (i.e. the case of equal variances). In addition, the intervals are quite wide, indicative of the low power of tests for variances mentioned earlier.

These regional confidence intervals achieve a compromise between local and global comparisons. Nevertheless, the question remains of how to choose the appropriate size of region (in particular, the number of grid points k). Although no simple answer is available, this issue can be investigated without having to construct new displays for every possible value of k. Essentially, the idea is just to employ the relationship (4) between local and regional significance or confidence levels in reverse. If individual $100(1 - \alpha_L)$% levels are shown on the display, then any region consisting of k of these intervals would have a joint level of at least $100(1 - k\alpha_L)$%.

It is possible to conceive of systematically dividing up the globe into disjoint regions, each consisting of k grid points, and then repeatedly making use of this multiple confidence interval technique. It is important to keep in mind, however, that the exact regions considered must be specified prior to any confirmatory statistical analysis. Moreover, the analyst will still be confronted with a sort of multiplicity problem, if any attempt is made to reconcile the outcomes among the regions.

SUMMARY AND CONCLUSIONS

The role of statistics in the validation of GCMs has been clarified. Among other things, a logical fallacy concerning the way in which tests of significance are applied has been identified. As an alternative to hypothesis testing, confidence intervals are recommended as being more in harmony with the goals of climate modelers. The need to examine the variability of GCM output has been motivated by its relationship to the frequency of extreme events, about which information is especially important for climate impact researchers. The lack of appreciation of the multiple comparison problem is pointed out, along with more widespread confusion over the use of multivariate tests of significance based on resampling schemes. Because these issues are not unique to climate modeling, much could be learned from previous discussions of validation for other types of models.

What are the future prospects for the use of statistics in validating GCMs? At least in some respects, they are not so promising. Too often, the inconclusive nature of statistical analysis is lamented after the fact, rather than factoring statistical issues into the design of experiments before the fact. Recalling the experience with multivariate techniques, it is feared that the debate over how to validate GCMs will eventually become bogged down in a quagmire over which particular new, more sophisticated statistical technique ought to be employed.

On a more positive note, one way in which the dilemma over local or global significance could be resolved has been proposed. Multiple confidence intervals, based on the Bonferroni technique, offer an approach to making regional-scale comparisons. They serve both to simplify matters, and to produce results more in accord with the needs of climate modelers as well as impact researchers. If the proper role of statistics were recognized, scenarios for the future development and use of GCMs would be more favorable. An improved understanding of model output could be attained, even without any breakthroughs in physics.

Acknowledgements. The author thanks W. Lawrence Gates for suggesting this topic, Stanley L. Grotch for providing data, and Barbara G. Brown and Mary W. Downton for help in drafting figures. The comments of 3 anonymous reviewers are gratefully acknowledged. This research was performed while the author was at the University of California at Davis (UC-Davis), on sabbatical from the National Center for Atmospheric Research. It was supported in part by the Lawrence Livermore National Laboratory under a cooperative agreement with UC-Davis.

LITERATURE CITED

Barnett, T. P., Schlesinger, M. E. (1987). Detecting changes in global climate induced by greenhouse gases. J. geophys. Res. 92: 14772–14780

Box, G. E. P. (1953). Non-normality and tests on variances. Biometrika 40: 318–335

Brillinger, D. R., Jones, L. V., Tukey, J. W. (1978). The management of weather resources. Volume II: the role of statistics in weather resources management. Report of the Statistical Task Force to the Weather Modification Advisory Board, U.S. Department of Commerce, Washington, D.C.

Chervin, R. M., Schneider, S. H. (1976). On determining the significance of climate experiments with general circulation models. J. atmos. Sci. 33: 405–412

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia

Fox, D. G. (1981). Judging air quality model performance. Bull. Am. Meteorol. Soc. 62: 599–609

Hasselmann, K. (1976). Stochastic climate models. Part I: theory. Tellus 6: 473–485

Hasselmann, K. (1979). On the signal-to-noise problem in atmospheric response studies. In: Shaw, D. B. (ed.) Meteorology over the tropical oceans. Royal Meteorological Society, London, p. 251–259

Hayashi, Y. (1982). Confidence intervals of a climatic signal. J. atmos. Sci. 39: 1895–1905

Houghton, J. T., Jenkins, G. J., Ephraums, J. J. (eds.) (1990). Climate change: the IPCC scientific assessment. Cambridge University Press, Cambridge

Katz, R. W. (1982). Statistical evaluation of climate experiments with general circulation models: a parametric time series modeling approach. J. atmos. Sci. 39: 1446–1455

Katz, R. W. (1983). Statistical procedures for making inferences about precipitation changes simulated by an atmospheric general circulation model. J. atmos. Sci. 40: 2193–2201

Katz, R. W. (1988). Statistical procedures for making inferences about climate variability. J. Climate 1: 1057–1064

Katz, R. W., Brown, B. G. (1991). The problem of multiplicity in research on teleconnections. Int. J. Climatol. 11. 505–513

Katz, R. W., Brown, B. G. (1992). Extreme events in a changing climate: variability is more important than averages. Climatic Change 21: 289–302

Laurmann, J. A., Gates, W. L. (1977). Statistical considerations in the evaluation of climatic experiments with atmospheric general circulation models. J. atmos. Sci. 34: 1187–1199

Livezey, R. E. (1985). Statistical analysis of general circulation model climate simulation, sensitivity and prediction experiments. J. atmos. Sci. 43: 1139–1149

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

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

Mearns, L. O., Katz, R. W., Schneider, S. H. (1984). Extreme high-temperature events: changes in their probabilities with changes in mean temperature. J Climate appl. Meteorol. 23: 1601–1613

Mearns, L. O., Schneider, S. H., Thompson, S. L., McDaniel, L. R. (1990). Analysis of climate variability in general circulation models: comparison with observations and changes in variability in 2 × CO_2 experiments. J geophys. Res. 95: 20469–20490

Mielke, P. W., Brier, G. W. (1981). Application of multiresponse permutation procedures for examining seasonal changes in monthly mean sea level pressure patterns. Mon. Weather Rev. 109: 120–126

Miller, R. G. Jr (1981). Simultaneous statistical inference (2nd edn). Springer-Verlag, New York

Preisendorfer, R. W., Barnett, T P. (1983). Numerical model-reality intercomparison tests using small-sample statistics. J. atmos. Sci. 40: 1884–1896

Robinson, P. J., Finkelstein, P. L. (1991). The development of impact-oriented climate scenarios. Bull. Am. Meteorol. Soc. 72: 481–490

Ryan, B. F., Joiner, B. L., Ryan, T. A. Jr (1985). MINITAB handbook (2nd edn). Duxbury Press, Boston

Santer, B. D., Wigley, T M. L. (1990). Regional validation of means, variances, and spatial patterns in general circulation model control runs. J. geophys. Res. 95: 829–850

Solow, A. R. (1990). Discriminating between models: an application to relative sea level at Brest. J Climate 3: 792–796

Tukey, J. W. (1977). Some thoughts on clinical trials, especially problems of multiplicity. Science 198: 679–684

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

von Storch, H., Zwiers, F. W. (1988). Recurrence analysis of climate sensitivity experiments. J. Climate 1: 157–171

Walker, G. T (1914). Correlation in seasonal weather, III. On the criteria for the reality of relationships or periodicities. Mem. India Meteorol. Dep. 21: 12–15

Wegman, E. J., DePriest, D. J. (eds.) (1980). Statistical analysis of weather modification experiments. Dekker, New York

Wigley, T M. L. (1985). Impact of extreme events. Nature, Lond. 316: 106–107

Editor: *V. Meentemeyer*

Wigley, T M. L., Barnett, T P. (1990). Detection of the greenhouse effect in the observations. In: Houghton, J. T., Jenkins, G. J., Ephraums, J. J. (eds.) Climate change: the IPCC scientific assessment. Cambridge University Press, Cambridge, p. 239–255

Wigley, T M. L., Santer, B. D. (1990). Statistical comparison of spatial fields in model validation, perturbation and predictability experiments. J. geophys. Res. 95: 851–865

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

Wilson, C. A., Mitchell, J. F. B. (1987). Simulated climate and CO_2-induced climate change over western Europe. Climatic Change 10: 11–42

Zwiers, F. W. (1987). Statistical considerations for climate experiments. Part II: multivariate tests. J. Climate appl. Meteorol. 26: 477–487

Zwiers, F. W. (1990). The effect of serial correlation on statistical inferences made with resampling procedures. J. Climate 3: 1452–1461

Zwiers, F. W., von Storch, H. (1989). Multivariate recurrence analysis. J. Climate 2: 1538–1553

Manuscript first received: November 27, 1991
Revised version accepted: May 4, 1992