A Remark on Chervin-Schneider's Algorithm to Test Significance of Climate Experiments with GCM's

HANS V. STORCH

Meteorologisches Institut der Universität Hamburg, Hamburg, F.R. Germany

11 March 1981

ABSTRACT

It is shown by two examples that the algorithm proposed by Chervin and Schneider (1976) is of little use in deciding, with a given risk, whether a GCM result differing from others (or nature) is caused by chance or by prescribed changes of some boundary values. What remains is that one can *believe* in a "significant" change if the rate of rejections of local null hypotheses is quite large (e.g., three times expectation or more). Additionally, the Chervin-Schneider algorithm can be used to gain a first guess.

1. Introduction

For about 10 years, a large number of experiments with general circulations models (GCM's) were performed to predict consequences of sea surface temperature anomalies, ice- and snow-cover anomalies or man-produced thermal pollution on climate. Here climate is defined as ensemble averages of time averages. Therefore, an impact on climate can be decided only by statistical testing. The main advantage of such testing is that one is able to accept an alternative against a null hypothesis with a given probability to err (risk).

A commonly used algorithm, recently proposed by Chervin and Schneider (1976), estimates an ensemble average of time averages (climate mean) and standard deviations (noise levels) by means of some (i) randomly differing experiments for each grid point. The time averages obtained by a prescribed change experiment are subtracted from the climate mean and divided by the noise level. This number (r) multiplied by a constant dependent only on i, is t distributed with i-1 degrees of freedom (e.g. Chervin and Schneider, 1976, p. 410). For reasons of completeness, it should be mentioned that Warshaw and Rapp (1973) previously applied an F-test to zonal statistics.)

Hasselmann (1979) criticized Chervin and Schneider's method, because it made a lot of univariate (local) decisions instead of one single multivariate (global) decision. That a sum of local decisions generally cannot substitute for a global decision with a given risk has been known since the early 1950's [see, e.g., the summary given by Chung and Fraser (1958)].

Unfortunately, both the results of medical statistics and the example of Hasselmann did not get the attention they deserved (Chervin, 1980). Therefore, in what follows, we show by means of two simple examples that the algorithm is of little statistical value (Sections 2 and 3). But this does not mean that it is useless, because it can be applied very well to gain a first guess (Section 4). In what follows, a significance level of 5% was selected, though this value is arbitrary; any other from the interval (0, 1) could have been specified.

2. Theoretical criticism for the case of independent local decisions

For the idealized situation that all local decisions are statistically independent, the distribution of the number of false rejections of local null hypotheses is known, provided the global null hypothesis is valid (no climate change). This is shown by the following example:

Assume there are m boxes filled with 20 balls each. The local null hypothesis is 19 balls are black, one is white, and its alternative is that at least two balls are white. A reasonable test for the local (i.e., for one box) decision is simply to draw randomly one ball from this box, to reject the local null hypothesis if the ball is white, and to accept it if otherwise. The risk for an erroneous acceptance of the alternative is 5%. (Whether this test is an intelligent one is of less importance in this context, as we are interested in risk not in power.) To work with the algorithm proposed by Chervin and Schneider would mean to draw from each of the m boxes one sample and to accept the local alternative for those boxes from which a white ball was drawn.

The global null hypothesis is "no box contains more than one white ball" and its alternative "at least one box contains more than one white ball". If the global null hypothesis is valid, the probability to draw maximally n white balls is given by the cu-

0022-4928/82/010187-03$04.75

TABLE 1. Frequency of erroneous rejections of local null hypotheses ($m = 768$).

	Experiment				
i	3	4	5	6	7
2	4.2%	4.9%	6.3%	5.3%	4.0%
3		3.0%	9.0%	7.2%	2.9%
4			10.0%	9.7%	2.3%
5				8.7%	0.7%
6					0.5%

mulative binomial-distribution $P(n;\ m,\ 5\%)$, which is centered at $0.05m$, i.e., on an average $n = 0.05m$ local alternatives will falsely be accepted. But by chance n can be greater. If, for example, $m = 768$ (the number used below) the event $48 < n$ (expectation plus 11) will occur with a probability of 5% only. For this special case a reasonable global test could be: reject the null hypothesis at a significance level of 5%, if at least 49 local alternatives are accepted. Unfortunately, in GCM applications the assumption of independence is not valid.

3. Empirical criticism for the case of dependent decisions

In the case of dependency, one has to expect, as in the foregoing example, a rate of $0.05m$ false rejections for local null-hypotheses on an average. The distribution of this rate is unknown and can be estimated only by a series of statistically independent experiments. Therefore, the Chervin-Schneider technique is applied to a series of GCM runs, which have different randomly perturbed initial states, i.e., to a series of experiments without any signal to be detected.

Instead of a complete GCM, a much more economic shallow-water equation model, discretized on a 5.6° grid-point space, is used. Northern Hemisphere orography is included as well as some forcing to keep the level of the eddy activity. This model was integrated up to 60 days. The time averaging was done for the height field north of 20°N for the last 30 days, when the model became about stationary (though stationarity is of less importance for this mechanical test). $K = 7$ experiments were performed; $m = 768$ points were taken into account.

For each number $i = 2, \ldots, K - 1$ the $K = 7$ experiments were divided into two groups. The first one, consisting of the first i experiments, was used to estimate the climate mean and the noise level; the second group, with the last $K - i$ experiments, was used to perform the test. (Thus, the number i has the same meaning as the one defined in the Introduction.) It should be emphasized that the K experiments are fully equal, though due to their initial random perturbation statistically independent.

In Table 1 the frequency of rejections of local null hypotheses for the different experiments are listed. Because the global null-hypothesis is true, all rejections are false. On an average, 5.2% of all decisions gave erroneous acceptances of the alternative "local climate change". Table 1 indicates that a rate of rejections three or four times expectation will unlikely occur.

In Fig. 1 a typical hemispheric distribution of the local rejections is sketched.

4. First guess

One correct way to get a global decision with given risk would be to estimate a mean vector and a covariance matrix and, provided the assumption of normality is valid, to apply the χ^2 test (e.g., Hasselmann, 1979). To do so one needs a drastic reduction of components in order to get a capable problem and to increase the power of the test: one has to restrict the degrees of freedom to a relatively small subset of points (or spectral components, if a transformation to some spectral system is done), to a "first guess". Such a first guess can be gained by the Chervin-Schneider algorithm if two similar prescribed change experiments are available. Assume the global alternative to be true, i.e., there are points (or spectral components) where a local climate change is to be correctly stated. Ordinarily, these points will be those with maximum values of the statistic r. Therefore,

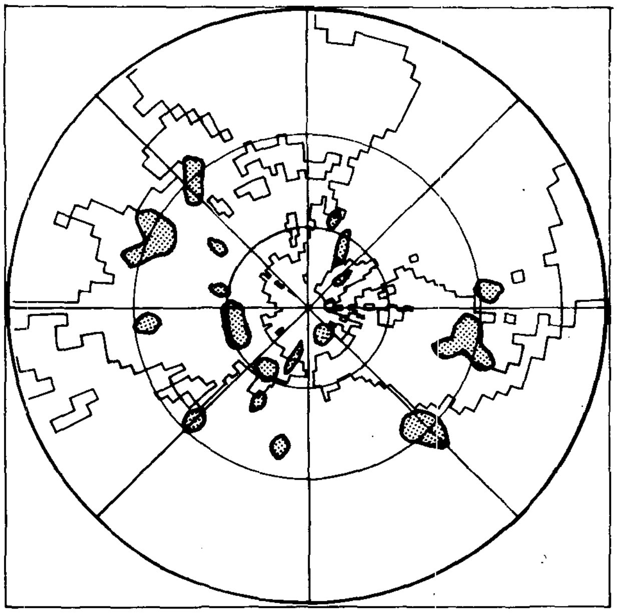

FIG. 1. A hemispheric distribution of erroneous rejections of the local null hypothesis at a significance level of 5%. For the stippled areas a climate change is falsely stated (Experiment 4, $i = 2$).

all points of the first experiment for which the local alternative are accepted with a significance level of, say 10%, could be defined to form the first guess. With a subsequent multivariate test restricted onto this first guess and applied to the results of the second, similar, but statistically independent experiment, the global decision can be carried out.

REFERENCES

Chervin, R. M., 1980: On the simulation of climate and climate change with general circulation models. *J. Atmos. Sci.,* **37,** 1903–1913.

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

Chung, J. H., and D. A. S. Fraser, 1958: Randomization tests for a multivariate two-sample problem. *J. Amer. Statist. Assoc.,* **53,** 529–535.

Hasselmann, K., 1979: On the signal-to-noise problem in atmospheric response studies. *Meteorology of Tropical Oceans.* Roy. Meteor. Soc., 251–259.

Warshaw, M., and R. R. Rapp, 1973: An experiment on the sensitivity of a global circulation model. *J. Appl. Meteor.,* **12,** 43–49.