9.3

"The stippling shows statistically significant gridpoints": How Research Results are Routinely Overstated and Over-interpreted, and What to Do About It

Daniel S. Wilks
Cornell University, Ithaca NY
dsw5@cornell.edu

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 . . . from a lack of understanding of the combined effects of number and interdependence of set numbers." (Livezey and Chen, 1983)

More than thirty years have passed since the seminal paper by Livezey and Chen (1983) pointed out that collections of multiple statistical tests, often in the setting of individual tests at many spatial gridpoints, are very often interpreted incorrectly and in a way that leads to research results being routinely overstated. That paper also proposed an approach to dealing with and protecting against that problem, which they called assessment of "field significance". The idea was to construct a "meta-test" using as input the results of the many tests, to address the "global" null hypothesis that all individual "local" (e.g., gridpoint) null hypotheses are true. If a global null hypothesis cannot be rejected, one cannot conclude with adequate confidence that any of the individual local tests show meaningful violations of their respective null hypotheses. Thus, failure to achieve field significance protects the analyst to a degree from being mislead into believing results from the many erroneous rejections of true local gridpoint null hypotheses that will invariably occur.

Unfortunately, very little has changed over the intervening decades with respect to the over-interpretation of multiple hypothesis tests in the atmospheric sciences literature. For example, of the 281 papers published in the *Journal of Climate* during the first half of 2014, 97 (34.5%) included maps described in part by some variant of the quotation in the title of this paper. These papers implicitly but wrongly represented that *any* individual gridpoint test exhibiting nominal statistical significance was indicative of a physically meaningful result. By contrast, only 3 of the 281 papers (1.1%) considered the effects of multiple testing on their scientific conclusions. (The remaining 64.4% of these papers either had no maps, or did not attempt statistical inference on any of the mapped quantities.)

These are disturbing but unfortunately quite representative statistics. A frustrated *Journal of Climate* editor (who wishes to remain anonymous) wrote in this context, "The use and misuse of statistical tests is an important topic to me – I see a lot of abuses as *J. of Climate* editor, and sometimes I feel that our field is not even science." Although this may seem to be an extreme viewpoint, it is undeniable that the consequences of the widespread and continued failure to address the issue of multiple hypothesis testing are overstatement and over-interpretation of the scientific results, to the detriment of the discipline.

The purposes of this paper are to highlight problems relating to interpretation of multiple statistical tests, to provide some of the history related to this issue, and to describe and illustrate a straightforward and statistically principled approach – control of the False Discovery Rate (FDR) – to protecting against overstatement and over-interpretation of multiple testing results.

2. Exposition of the multiple-testing problem

Computation of a single hypothesis test involves defining a null hypothesis (H_0), which will be rejected in favor of an alternative hypothesis (H_A) if a sufficiently extreme value of the test statistic is observed (e.g., Wilks 2011). Rejection of H_0 at a test level α occurs if the test statistic is sufficiently extreme that the probability (called the *p*-value) of observing it or any other outcome even less favorable to H_0, *if that null hypothesis is true*, is no larger than α. If H_0 is rejected with $\alpha = 0.05$ (the most common, although an arbitrary, choice), the result is said to be significant at the 5% level[1].

Although perhaps intuitively attractive, it is quite incorrect to interpret a *p*-value as the probability that the null hypothesis is true, given the evidence expressed in the observed test statistic (e.g., Ambaum 2010). The correct interpretation is opposite: a *p*-value is a probability related to the magnitude of a test statistic, assuming the truth of H_0. The implication is that any true null hypothesis will be rejected with probability α (if the test has been formulated correctly), so that a collection of N_0 hypothesis tests whose null hypotheses are all true will exhibit, on average, αN_0 erroneous rejections. However, any particular collection of N_0 hypothesis tests whose null hypotheses are all true will rarely exhibit exactly αN_0 erroneous rejections, but rather the number of erroneous rejections will be a random variable. That is, the number of erroneous rejections will be different for different (possibly hypothetical) batches of the same kind of data, and for any particular batch this number will behave as if it had

[1] In the atmospheric sciences literature this conclusion is often expressed as significance "at the 95% level", but that convention is inconsistent with mainstream terminology (e.g., Jolliffe 2004).

been drawn from a probability distribution whose mean is $\alpha\, N_0$.

If the results of these N_0 hypothesis tests are statistically independent, then the probability distribution for the number of erroneously rejected null hypotheses will be binomial, yielding the probabilities for the possible numbers of erroneously rejected tests, x,

$$\Pr\{x\} = \frac{N_0\,!}{x\,!(N_0 - x)!}\,\alpha^{x}\,(1-\alpha)^{N_0 - x}\;,\; x = 0, 1, \ldots, N_0 \qquad (1)$$

One implication of this equation is that, unless N_0 is relatively small, erroneously rejecting at least one of the true null hypotheses is nearly certain: for example if $\alpha = 0.05$ and $N_0 = 100$ this probability is 0.994. Thus some true null hypotheses will almost always be erroneously rejected in any realistic multiple testing situation involving gridded data. Even though this number will be $\alpha\, N_0$ on average, Equation (1) specifies nonnegligible probabilities for numbers of erroneous rejections that may be substantially larger than $\alpha\, N_0$. When the members of the collection of hypothesis tests are not independent, which is the usual situation for gridded data, Equation (1) is no longer valid and the probabilities for numbers of erroneous rejections larger than $\alpha\, N_0$ are even higher.

The problem of interpreting the results of N multiple simultaneous hypothesis tests is further complicated by the fact that the fraction of true null hypotheses N_0/N is unknown, and also that some of the $N_A = N - N_0$ false null hypotheses may not be rejected. How, then, can a spatial field of hypothesis test results be interpreted in a statistically principled and meaningful way?

3. Historical development of multiple testing in the atmospheric sciences

3a. The Walker test

The question just posed has been confronted in the atmospheric sciences for more than a century, apparently having been addressed first by Walker (1914). Katz and Brown (1991) and Katz (2002) provide a modern perspective on Walker's thinking on this subject.

Walker realized that an extreme value of a sample statistic (e.g., a small *p*-value) is progressively more likely to be observed as more realizations of the statistic (e.g., more hypothesis tests) are examined, so that a progressively stricter standard for statistical significance must be imposed as the number of tests increases. In order to limit the probability of erroneously rejecting one or more of N_0 true null hypotheses to an overall level α_0, Walker's criterion is that only individual tests with *p*-values no larger than α_{Walker} should be regarded as significant, where

$$\alpha_{\mathrm{Walker}} = 1 - (1 - \alpha_0)^{1/N_0} \quad . \qquad (2)$$

This formula can be derived from the fact that a *p*-value for a test involving a true null hypothesis is equally likely to be any real number between zero and one (e.g., Wilks 2006). Of course $\alpha_{\mathrm{Walker}} = \alpha_0$ for a single ($N_0 = 1$) test. In order to limit the probability of erroneously rejecting any of $N_0 = 100$ true null hypothesis tests to the level $\alpha_0 = 0.05$, only those tests having *p*-values smaller than $\alpha_{\mathrm{Walker}} = .000513$ would be regarded as significant according to this criterion. In contrast, as noted above, naively evaluating each of $N_0 = 100$ tests having true null hypotheses at the $\alpha_0 = 0.05$ level (i.e., ignoring the multiple-testing problem) results in a 0.994 probability that at least one true null hypothesis is erroneously rejected.

Equation (2) was derived under the (often unrealistic) assumption that the results of the individual tests are statistically independent, but in practice it is robust to (only modestly affected by) deviations from this assumption (Katz and Brown 1991, Wilks 2006). On the other hand, although Equation (2) will yield relatively few rejections of true null hypotheses, the Walker criterion is quite strict since $\alpha_{\mathrm{Walker}} \approx \alpha_0/N$, which compromises the sensitivity of the procedure for detecting false null hypotheses.

3b. The "field significance" approach

Von Storch (1982) and Livezey and Chen (1983) cast the problem of evaluating multiple hypothesis tests as a "meta-test", or a "global" hypothesis test whose input data are the results of N "local" hypothesis tests. Because the individual local tests often pertain to a grid or other geographic array, they can be thought of as composing a "field" of test results. Accordingly this approach to multiple testing is generally referred to as assessment of "field significance" (Livezey and Chen 1983). It has become the dominant paradigm for multiple testing in the atmospheric sciences, especially when the individual hypothesis tests pertain to a network of geographic locations.

The global null hypothesis is that all of the local null hypotheses are true, so that failure to reject the global null hypothesis implies that significant results have not been detected anywhere in the field of individual local tests. In the idealized case that the local null hypotheses are statistically independent, the binomial distribution (Equation 1) allows calculation of the minimum number of locally significant tests required to reject a global null hypothesis – i.e., to achieve field significance. For example, again if $N = 100$ independent tests and $\alpha_0 = 0.05$, the global null hypothesis implies $N_0 = N = 100$ so that on average (over many hypothetical realizations of the single testing situation for which we have data) five of the 100 local null hypotheses are expected to be rejected. But in order to reject the global null hypothesis, an unusually large number of local test rejections must be observed. Equation (1) specifies that ten or more such rejections are required in order to have smaller than $\alpha_{\mathrm{global}} = \alpha_0 = 0.05$ probability of observing this or a more extreme result if the global null hypothesis is true. If fewer of these independent local tests have *p*-values smaller than $\alpha_0 = 0.05$, then none of them are regarded as significant according to this criterion.

Assuming statistical independence among the local test results is a best-case situation, so that the usual condition of spatial correlation among the local gridpoint tests implies that even more local test rejections than implied by Equation (1) are required in order to achieve field significance. However, exactly how many local test rejections are required depends on the nature of the underlying spatial correlation, and this threshold may be difficult to determine in a particular multiple-testing setting. One approach is to try to estimate an "effective number of independent tests" $N_{\mathrm{eff}} < N$, and to use this value in Equation (1), although often N_{eff} cannot be rigorously estimated (von Storch and Zwiers 1999). Livezey and Chen (1983) also suggest estimating the frequency distribution for numbers of locally significant tests using Monte Carlo methods (i.e., randomly resampling the available data in a manner consistent with the global null hypothesis, e.g., Mielke et al. 1981, Zwiers 1987). This approach can require elaborate and computationally expensive calculations, especially if the data exhibit both temporal and spatial correlations (Wilks 1997), and in some test settings an appropriate Monte Carlo algorithm may not be available. Ignoring the effect of spatial correlation leads to highly inaccurate test results when using this method, with global null hypotheses being rejected much more frequently than specified by the nominal α_{global} (von Storch 1982, Livezey and Chen 1983, Wilks 2006).

3c. False Discovery Rate (FDR)

The field significance procedure described in the previous section is much better than the common but naive approach characterized by the quotation in the title of this paper, according to which any nominally rejected local null hypothesis is concluded to be false. However it suffers from several drawbacks, some of which have already been mentioned:

(i) Because of the discreteness of counts of locally significant tests, the overall field significance can be conservative (rejecting true global null hypotheses less frequently than a specified α_{global}). In the example above when $N_0 = 100$ independent tests and $\alpha_{\mathrm{global}} = 0.05$, requiring ten local test rejections will lead to rejection of the global null hypothesis with probability $0.028 < \alpha_{\mathrm{global}}$, but adopting the less strict requirement of at least nine local tests significant will be inappropriate because the corresponding probability would then be 0.063. This effect will also slightly degrade the test sensitivity (ability to detect violations of the global null hypothesis).

(ii) The global test statistic involves only the numbers of locally significant tests but not their *p*-values, so that vanishingly small local *p*-values can provide no more evidence against the global null hypothesis than do local tests for which $p \approx \alpha_0$. Test sensitivity is consequently less than optimal because not all the available information is used (Zwiers 1987, Wilks 2006).

(iii) Correlation among the local tests greatly inflates the probability of erroneously rejecting the global null hypothesis. Accounting for the spatial correlations using Monte Carlo methods may be difficult and expensive, particularly when temporal correlation is also present, and in some cases appropriate Monte Carlo methods may not be available.

(iv) Having declared field significance, many of the local tests exhibiting $p < \alpha_0$ will have resulted from random and irreproducible fluctuations rather than physically real effects (Ventura et al. 2004, Wilks 2006). This problem is compounded in the presence of spatial correlation because these spurious "features" will tend to exhibit geographic coherence, potentially leading the analyst to over-interpret the data in an attempt to explain them.

All of these problems can be addressed by controlling the False Discovery Rate (FDR) when analyzing the results of multiple hypothesis tests. The FDR is the statistically expected (i.e., average over analyses of hypothetically many similar testing situations) fraction of local null hypothesis test rejections ("discoveries") for which the respective null hypotheses are actually true. An upper limit for this fraction can be controlled exactly for independent local tests, and approximately for correlated local tests, regardless of the unknown proportion N_0/N of local tests having true null hypotheses. Benjamini and Hochberg (1995) first described this method, with a primary focus on medical statistics (e.g. Storey and Tibshirani 2003). Ventura et al. (2004) introduced its use for multiple hypothesis tests pertaining to gridded atmospheric data, and Wilks (2006) demonstrated its relationship to the traditional field significance framework.

Although it is still not well known within the atmospheric sciences, the FDR method is the best available approach to analysis of multiple hypothesis test results, even when those results are mutually correlated. Its criterion of limiting the fraction of erroneously rejected null hypotheses is more relevant to scientific interpretation than is the traditional approach of limiting the probability that any given local test yields an erroneous rejection (Storey and Tibshirani 2003, Ventura et al. 2004). The remainder of this paper reviews the mechanics of implementing the FDR method, and illustrates its use in an artificial-data setting that highlights its advantages and emphasizes its robustness to the usual strong spatial correlation among the local tests.

4. A principled and straightforward solution – controlling the False Discovery Rate

The FDR procedure is similar in spirit to Walker's approach (Section 3a) in that it requires a higher standard (i.e., *p*-values smaller than a nominal local test level α_0) in order to reject local null hypotheses. It can also be interpreted in the field significance framework described in Section 3b. The algorithm operates on the collection of *p*-values from

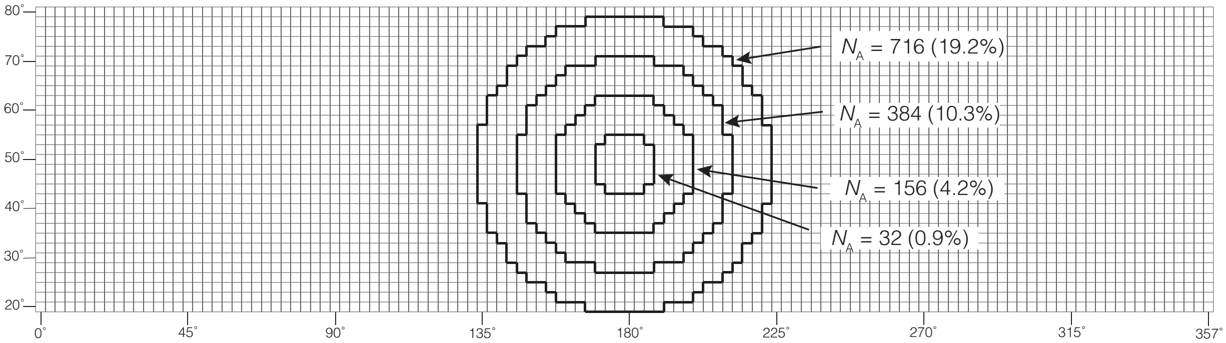

Figure 1. Hypothetical 3720-gridpoint domain, representing the northern hemisphere from 20˚N to 80˚N. Concentric bold outlines indicate regions where local null hypotheses are not true.

N local hypothesis tests, p_i, $i = 1, \ldots, N$, which are first sorted in ascending order. Using a standard notation, these sorted *p*-values are denoted using parenthetical subscripts, so that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(N)}$. Local null hypotheses are rejected if their respective *p*-values are no larger than a threshold level p^*_{FDR} that depends on the distribution of the sorted *p*-values:

$$p^*_{FDR} = \max_{i=1,\ldots,N} \left[p_{(i)} : p_{(i)} \leq \alpha_{FDR}(i/N) \right] \quad , \qquad (3)$$

where α_{FDR} is the chosen control level for the FDR. That is, the threshold p^*_{FDR} for rejecting local null hypotheses is the largest $p_{(i)}$ that is no larger than the fraction of α_{FDR} specified by *i*/*N*.

The Walker criterion (Equation 2) is very nearly the same as Equation (3) if $i = 1$, so that the FDR procedure will be more sensitive for detecting false null hypotheses to the extent that Equation (3) is satisfied by a $p_{(i)}$ with $i > 1$, even as the expected fraction of false detections is maintained below α_{FDR}. In addition the FDR procedure can be interpreted as an approach to field significance. If none of the sorted *p*-values satisfy the inequality in Equation (3), then none of the respective null hypotheses can be rejected, implying also nonrejection of the global null hypothesis that they compose. Furthermore the size of that global hypothesis test (i.e., the probability of rejecting a global null hypothesis if it is true), is $\alpha_{global} = \alpha_{FDR}$ (Wilks 2006).

Even though Equation (3) assumes statistical independence among the local test results, the FDR procedure is approximately valid even when those results are strongly correlated, unlike the use of Equation (1) to evaluate numbers of locally significant tests. This property greatly simplifies statistically principled evaluation of multiple hypothesis test results, since there is no need for elaborate Monte Carlo simulations. Indeed, having obtained the *N* local *p*-values, the most complicated computation required is merely their sorting into ascending order so that Equation (3) can be evaluated.

5. Illustrative Examples

5a. Structure of the synthetic examples

It is instructive to compare the multiple-testing procedures in an artificial yet relatively realistic setting, so that their properties can be evaluated in the context of a completely known data-generating process. In this section, synthetic data will be defined on the $N = 3720$-point grid indicated in Figure 1. The vertical dimension represents the 31 latitudes from 20˚N to 80˚N, at increments of 2˚, and the horizontal dimension represents 360˚ of longitude at 3˚ increments, with a cyclic boundary. The four concentric bold outlines indicate regions, ranging in extent from 0.9% to 19.2% of the total number of gridpoints, where local null hypotheses will not be true.

The effects on the multiple-testing results of eight levels of spatial correlation of the underlying synthetic data will be investigated. Figure 2 shows the spatial autocorrelation functions for these eight levels, of the form

$$r(d) = \exp(-c\, d^2) \quad , \qquad (4)$$

where *d* is the great-circle distance between two gridpoints,

$$d = R_e \cos^{-1}[\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(\lambda_1 - \lambda_2)] \quad .(5)$$

Here ϕ denotes latitude, λ denotes longitude, and $R_e = 6.371 \times 10^3$ km is the earth radius. These eight spatial autocorrelation functions range in *e*-folding distance from 0.1×10^3 km (nearly spatially independent) to 10×10^3 km (very strongly dependent). The star symbols in Figure 2 indicate data for spatial autocorrelation of the northern hemisphere 500-mb height field taken from Polyak (1996), which are closely approximated by the heavy $c = 0.42$ (*e*-folding distance = 1.54×10^3 km) curve.

The underlying synthetic data are random Gaussian fields with spatial correlations governed by Equation (4), generated using methods described in Wilks (2011, p. 499). The statistical distribution of the generated values at each gridpoint is standard Gaussian, i.e., having zero mean and unit variance. Simulations using $c = 0.42$ represent the statistical properties of northern hemisphere 500-mb height

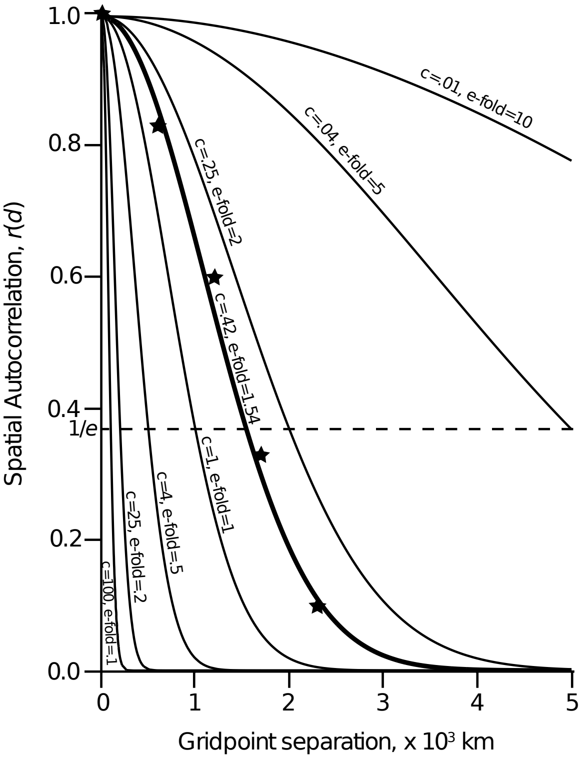

Figure 2. Eight spatial autocorrelation functions of the form in Equation (4). Star symbols indicate correlations for northern hemisphere 500-mb heights from Polyak (1996).

fields. Although the correlation function in Equation (4) does not represent the characteristic wave structures in these fields, these are not important for the purpose of illustrating the effect of spatial correlation on the multiple testing. For each realization of 3720 local hypothesis tests, 25 of these fields were generated and averaged, producing the test statistics for one-sample *t* tests having 24 degrees of freedom at each gridpoint. In experiments where some of the local null hypotheses are false, gridpoint sample means within one of the outlines shown in Figure 1 were increased uniformly by amounts $\Delta\mu$ ranging from 0.05 to 1.00.

5b. Global test properties

Figure 3 illustrates the operation of the FDR procedure (diagonal lines), in contrast to the naive approach of accepting alternative hypotheses at any gridpoint for which a locally significant result occurs (dashed horizontal line). This figure corresponds to a particular realization that will be examined later in more detail. The simulated data were generated with c = 0.42 (realistic spatial autocorrelation), N_A = 156 (4.2% of total gridpoints with false null hypotheses), and using the relatively large alternative-hypothesis mean $\Delta\mu$ = 0.7. The figure shows the smallest 350 of the 3720 sorted *p*-values $p_{(i)}$ as a function of their rank, *i*. The dashed diagonal line indicates the threshold criterion defined by Equation (3) using α_{FDR} = 0.10, according to which p^*_{FDR} = 0.003998 = $p_{(150)}$. That is, in this particular realization the local tests having the 150 smallest *p*-values are declared to exhibit statistically significant results. Of these, 144 are correct rejections, indicated by the small symbols below the dashed diagonal line. The twelve circles above the dashed diagonal line represent false null hypotheses that were not rejected. The six X's below the dashed diagonal represent true null hypotheses that were erroneously rejected, yielding an achieved FDR = 6/150 = 0.04. The inset shows a closer view of the points within the red box.

The dotted diagonal line shows the threshold from Equation (3) when α_{FDR} = 0.20, in which case p^*_{FDR} = 0.009502 = $p_{(183)}$. In this case all N_A = 156 false null hypotheses are detected, but at the expense of erroneously rejecting 27 true null hypotheses, yielding an achieved FDR = 27/183 = 0.15. In contrast, the naive approach of rejecting any local null hypothesis for which the *p*-value is less than α_0 = 0.05 (dashed horizontal line) detects all 156 false null hypotheses, but at the expense of erroneously rejecting 189 true null hypotheses (X's and small symbols above the dashed diagonal), yielding an unacceptably large achieved FDR = 189/345 = 0.55: a majority of the nominally significant results are spurious!

Figure 4 illustrates the performance of the FDR procedure in terms of achieved global test levels, as a function of the degree of spatial correlation. That is, in the situation of all local null hypotheses being true, the achieved level is the probability that the global null hypothesis will be rejected (i.e., that at least one of the sorted *p*-values will satisfy the condition in Equation 3), which ideally will equal $\alpha_{global} = \alpha_{FDR}$. These probabilities are approximated in Figure 4 as the corresponding relative frequencies over 10^5 simulated global tests. As expected, these achieved levels are approximately correct for small spatial correlations, but then decline fairly quickly and stabilize at about half the nominal levels. Thus the FDR procedure is robust to the effects of spatial correlation, yielding a somewhat conservative global test when the spatial correlation is moderate or strong, which is consistent with prior results (Wilks 2006). This result suggests that, for data grids exhibiting moderate to strong spatial correlation, approximately correct global test levels can be produced using the FDR procedure by choosing $\alpha_{FDR} = 2\alpha_{global}$.

In sharp contrast, the achieved test levels for the Livezey-Chen counting procedure, also with no adjustment for spatial correlation, is very strongly permissive. For example using Equation (1) and assuming spatial independence yields a requirement for at least 208 locally significant tests (5.6% of local null hypotheses rejected) for field significance with α_0 = α_{global} = 0.05. This criterion produces achieved global test levels of 0.0907 and 0.3517 when the *e*-folding distances are 0.2 and 1.54 x 10^3 km, respectively (results not shown in the figure). The naive interpretation that any significant local test implies field significance is even worse, as it produces

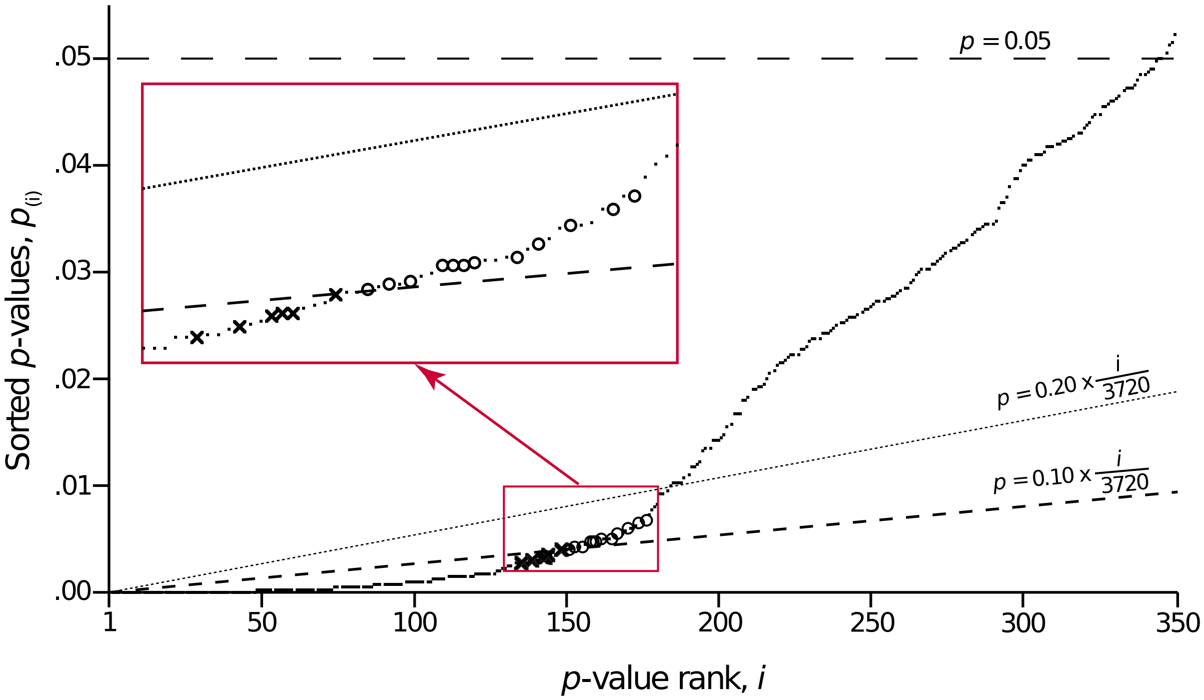

Figure 3. Illustration of the FDR criterion using α_{FDR} = 0.10 (dashed diagonal line) and α_{FDR} = 0.20 (dotted diagonal line), and the naive approach of rejecting any local test with p-value smaller than α_0 = 0.05 (dashed horizontal line). Plotted points are the smallest 350 sorted p-values of 3720 local tests. Points below the diagonal lines represent significant results according to the two FDR control levels. X's represent 6 tests with true null hypotheses that were erroneously rejected, and circles represent false local null hypotheses that were not rejected, when α_{FDR} = 0.10. Inset shows closer view of points within the red box. The 345 tests with p-values smaller than α_0 = 0.05 would be declared significant under the naive procedure, even though a majority of these null hypotheses are true.

an achieved global test level of unity: at least one of the 3720 local tests is virtually certain to exhibit a spurious null hypothesis rejection, regardless of the strength of the spatial correlation within the range considered in Figure 4.

Figure 5 shows global test power (sensitivity for detection of false global null hypotheses) as functions of both the numbers of locally significant tests N_A (Figure 1), and alternative-hypothesis magnitudes $\Delta\mu$ when the local null hypotheses are false, for the realistic case of 1.54 x 10^3 km *e*-folding distance. Red curves show results for the FDR approach when α_{FDR} = 0.10, which Figure 4 indicates should yield global tests near the 5% level. Black curves show results for the Livezey-Chen counting procedure with α_{global} = 0.05, which requires at least 365 (9.8% of 3720) locally significant tests for this degree of spatial correlation[2]. When $\Delta\mu$ = 0 (i.e., all local null hypotheses are true) both methods yield 5% of global tests rejected, which is the correct level. Not surprisingly, probabilities for rejecting the global null hypotheses increase with both the numbers and magnitudes of false local null hypotheses. For large numbers of false local null hypotheses and relatively small $\Delta\mu$ the Livezey-Chen procedure yields somewhat greater power, although as will be seen shortly this comes at the cost of many more erroneous rejections of true local null hypotheses. For the smaller numbers of false local null hypotheses, the power of the Livezey-Chen counting procedure is strongly limited, regardless of $\Delta\mu$. Even if $\Delta\mu$ is large enough for every false null hypothesis to be detected with near certainty, when there are fewer than 365 of these it is not assured that the remainder of the 365-count threshold will be met by false local test rejections. This is a serious deficiency of the counting procedure when the field of tests contains relatively few false null hypotheses, which derives from the fact that the magnitudes of the smallest local p-values do not contribute to the global test.

5c. Local test interpretations

Often the primary interest will be interpretation of the locations and spatial patterns of the locally significant test results. Reliability of these interpretations will of course be enhanced to the extent that they are minimally contaminated with erroneous rejections of true local null hypotheses. Figure 6a shows false discovery rates for the FDR method with α_{FDR} = 0.10 (red), the Livezey-Chen

[2] Note that it is not clear how to design a Monte Carlo procedure to determine this cutoff for field significance in the present setting because it involves a one-sample test, but the 365-count threshold can be computed for this artificial example because the underlying data-generating process is known.

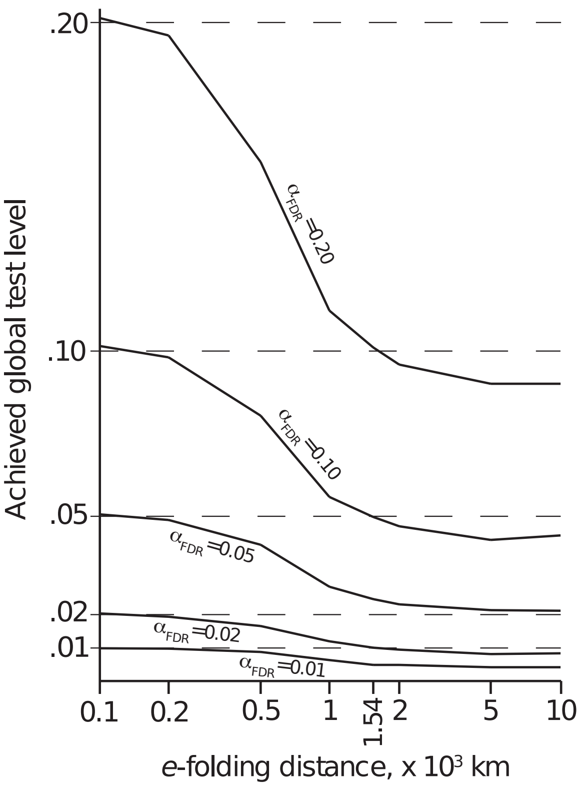

Figure 4. Achieved global test levels (probabilities of rejecting true global null hypotheses) when using the FDR procedure, as a function of spatial correlation strength. For moderate and strong spatial correlation, approximately correct results can be achieved by choosing $\alpha_{\text{FDR}} = 2\alpha_{\text{global}}$.

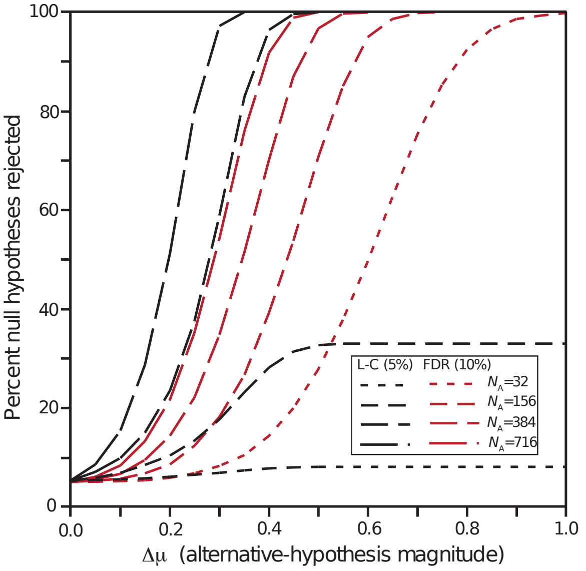

Figure 5. Percent of global null hypotheses rejected (global test power, or sensitivity) for the Livezey-Chen counting procedure with $\alpha_{\text{global}} = 0.05$ (black) and the FDR procedure using $\alpha_{\text{FDR}} = 0.10$ (red), as functions of both numbers of false local null hypotheses and the magnitudes of their nonzero means $\Delta\mu$ for e-folding distance 1.54 x 10^3 km.

counting approach with $\alpha_0 = \alpha_{\text{global}} = 0.05$ (black), and the naive approach of rejecting any local null hypothesis whose p-value is no larger than the nominal $\alpha_0 = 0.05$ (brown); again as functions of numbers of false local null hypotheses and alternative-hypothesis magnitudes $\Delta\mu$, for the realistic e-folding distance 1.54 x 10^3 km. The plotted values are averages over 10^3 realizations, so that for example the quantities contributed to the averages from the particular realization shown in Figure 3 are 6/150 = 0.04 for the FDR procedure, 189/345 = 0.55 for the naive procedure, and zero for the Livezey-Chen counting procedure because fewer than the required 365 local tests were significant at the 5% level (the global null hypothesis could not be rejected). As expected the FDR procedure controls the false discovery rates very tightly. The Livezey-Chen procedure also exhibits small false discovery rates for the smallest number of false local null hypotheses, but primarily because very few global null hypotheses can be rejected regardless of the magnitude of $\Delta\mu$ (compare Figure 5). For larger numbers of false local null hypotheses, the Livezey-Chen yields much larger false discovery rates. Worst performance of all is exhibited by the naive procedure, for which nearly all local test rejections are incorrect when $\Delta\mu$ is small, and which converges to the Livezey-Chen result for large $\Delta\mu$ and N_A since in these cases the Livezey-Chen procedure declares field significance in nearly all realizations.

Figure 6b shows the corresponding result for average proportion of erroneously rejected true local null hypotheses, for the Livezey-Chen (black) and FDR (red) procedures. The proportion of true local null hypotheses rejected is quite small for the FDR procedure. It is also small for the Livezey-Chen procedure for small N_A, but approaches the nominal $\alpha_0 = 0.05$ for sufficiently large $\Delta\mu$ and N_A. Results for the naive procedure are not shown because they yield 5% of true null hypotheses rejected on average for all values of $\Delta\mu$ and N_A.

To help visualize the foregoing more concretely, Figure 7 shows maps for a particular realization, interpreted according to (a) the FDR procedure with $\alpha_{\text{FDR}} = 0.10$, and (b) the naive approach using $\alpha_0 = 0.05$. Correct local null hypothesis rejections are indicated by plus symbols, failures to reject false local null hypotheses are indicated by circles, and erroneous rejections of true null hypotheses are indicated by X's. These maps correspond to the ranked p-values shown in Figure 3, with $N_A = 156$, $\Delta\mu = 0.7$, and e-folding distance 1.54 x 10^3 km. In Figure 7a the FDR procedure fails to reject twelve of the 156 false null hypotheses, but erroneously rejects only six true null hypotheses. The result is that the FDR procedure locates the true signal very effectively while introducing very little noise. By contrast, in Figure 7b the naive procedure locates all 156 false null hypotheses, but also erroneously indicates another 189 nominally significant gridpoints. The very large additional noise

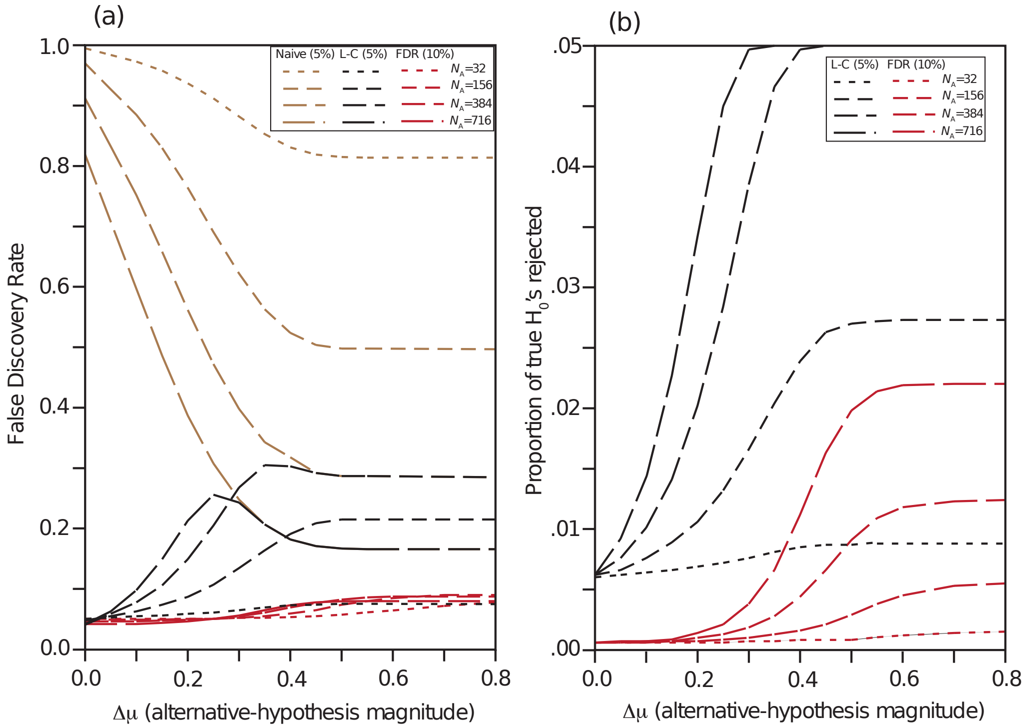

Figure 6. False discovery rates (a) and proportion of true local null hypotheses rejected (b), for the FDR method with α_{FDR} = 0.10 (red), the Livezey-Chen counting approach with $\alpha_0 = \alpha_{global}$ = 0.05 (black), and the naive approach with α_0 = 0.05 (brown), as functions of numbers of false local null hypotheses and alternative-hypothesis magnitudes $\Delta\mu$, using the *e*-folding distance 1.54 x 10^3 km.

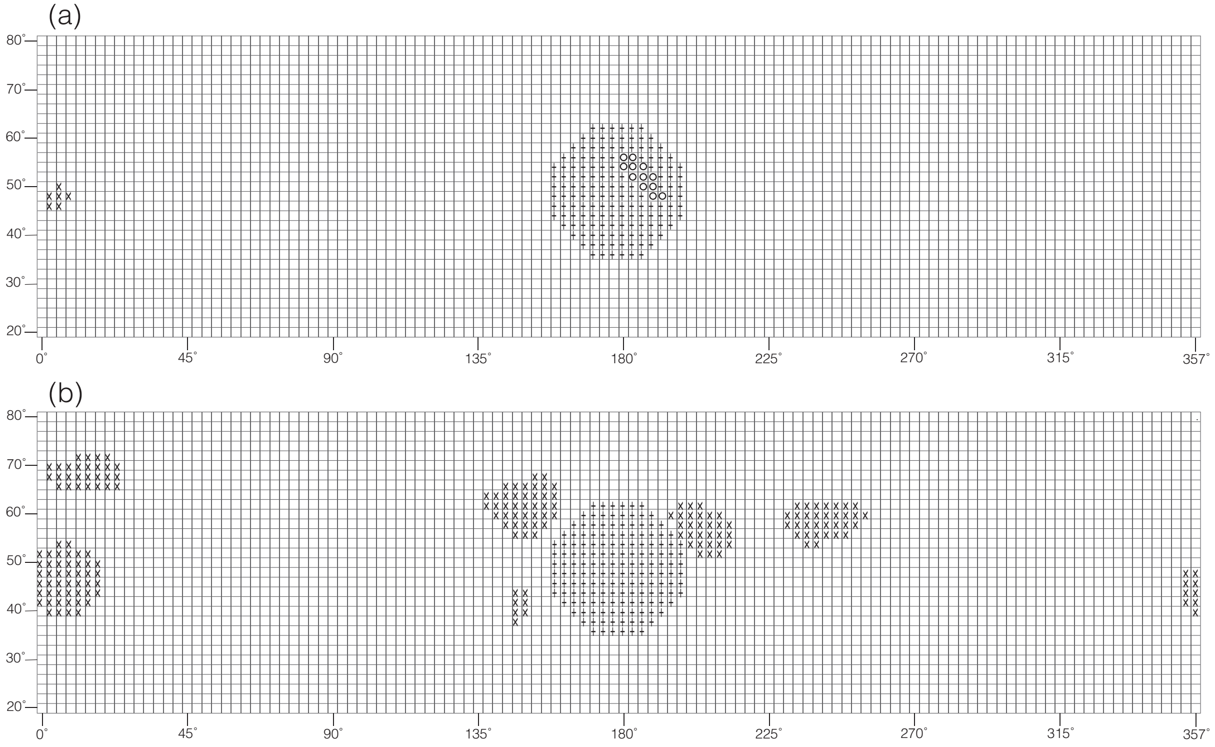

Figure 7. Maps of local test decisions made by (a) the FDR procedure with α_{FDR} = 0.10, and (b) the naive approach using α_0 = 0.05. Correct local null hypothesis rejections are indicated by plus symbols, failures to reject false local null hypotheses are indicated by circles, and erroneous rejections of true null hypotheses are indicated by X's. Results correspond to the ranked *p*-values shown in Figure 3, with N_A = 156, $\Delta\mu$ = 0.7, and *e*-folding distance 1.54 x 10^3 km.

level in Figure 7b would make physical interpretation of this map difficult, possibly leading an analyst to stretch his or her imagination to rationalize the many spurious local test rejections, which may appear to be physically coherent structures because of the strong spatial autocorrelation in the underlying data. Again, in this case the Livezey-Chen procedure would fail to reject the global null hypothesis, leading an analyst to doubt the reality of any of the local test rejections shown in Figure 7b.

6. Summary, conclusions and recommendation

The problem of simultaneously evaluating results of multiple hypothesis tests, often at a large network of gridpoints or other geographic locations, is widespread in meteorology and climatology. Unfortunately, the dominant approach to this problem in the literature is to naively examine each gridpoint test in isolation, and then to report as "significant" any result for which a local null hypothesis is rejected, with no adjustment for the effects of test multiplicity on the overall result. As a consequence, language similar to the hypothetical quotation in the title of this paper is distressingly common, which immediately flags the results portrayed as almost certainty overstated. This statistically unprincipled practice should be unacceptable both to reviewers and editors of scientific papers.

The necessity of correcting for the effects of simultaneous multiple test results has been known in the atmospheric sciences literature for more than a century, dating at least from Walker (1914). More recently, this problem has been cast as a meta-test on the collective results of many individual test results, and known as the assessment of "field significance" (Livezey and Chen 1983). Although the field significance approach is a very substantial advance over the usual naive procedure of ignoring the effects of multiple testing, it suffers from several drawbacks. One of these is that the approach lacks statistical power (sensitivity for detection of global null hypothesis violations) when the features to be detected occupy a small fraction of the domain. Another is that it is very sensitive to the usual strong spatial correlation among the individual gridpoint tests, and elaborate Monte Carlo calculations are generally required to compensate (Livezey and Chen 1983), particularly if the underlying data exhibit temporal autocorrelation as well (Wilks 1997). In some settings, such as that used in Section 5, appropriate Monte Carlo procedures may not be available at all. Furthermore, even when the overall test results are strong enough to achieve field significance, many of the nominally significant gridpoint tests will have resulted from spurious local null hypothesis rejections, which complicates the physical interpretation.

Controlling the FDR (Benjamini and Hochberg 1995, Ventura et al. 2004, Wilks 2006) has many favorable attributes, including only modest sensitivity to spatial autocorrelation in the underlying data. The examples employed here were constructed without temporal autocorrelation in order to simplify the exposition. However, because the FDR method is robust to spatial autocorrelation, effects of temporal autocorrelation can be addressed with appropriate testing procedures (e.g., Katz 1982, Zwiers and Thiébaux 1987, Wilks 2011) in the individual gridpoint calculations, so that complex procedures addressing both types of autocorrelation simultaneously (e.g., Wilks 1997) are unnecessary. Indeed, the method is applicable to collections of multiple hypothesis test results, regardless of the mathematical forms of those tests, so long as the individual tests operate correctly (i.e., with proportion of true null hypotheses rejected close to the nominal test level α_0).

Perhaps the greatest advantage of the FDR approach is that, by design, a control limit is placed on the fraction of significant gridpoint test results that are spurious, which greatly enhances the interpretability of the spatial patterns of significant results. Because the FDR approach is not only effective, but is also easy and computationally fast, it should be adopted whenever the results of simultaneous multiple hypothesis tests are reported or interpreted. Its main computational demand is only that the individual gridpoint *p*-values be sorted and examined in light of Equation 3. The usual strong spatial correlation encountered in gridded atmospheric data can be accommodated by choosing $\alpha_{\text{FDR}} = 2\alpha_{\text{global}}$, as illustrated in Figure 4. The consequence of employing this statistically principled procedure — in stark contrast to the all-too-common naive approach — is that there is much reduced scope for overstatement and over-interpretation of the results. In particular the analyst is not tempted to construct possibly fanciful rationalizations for the many spurious local test rejections that competing methods produce, which may appear to be physically coherent structures because of the strong spatial autocorrelation.

Acknowledgement: This research was supported by the National Science Foundation under Grant AGS-1112200.

References

Ambaum, M.H.P., 2010: Significance tests in climate science. *Journal of Climate*, **23**, 5927-5932.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.,* **57B,** 289–300.

Jolliffe, I.T., 2004: P stands for ...*Weather*, **59**, 77-79.

Katz, R.W., 1982: Statistical evaluation of climate experiments with general circulation models: a parametric time series modeling approach. *Journal of the Atmospheric Sciences*, **39**, 1446–1455.

Katz, R.W., 2002: Sir Gilbert Walker and a connection between El Niño and statistics. *Statistical Science*, **17**, 97-112.

Katz, R.W., and B.G. Brown, 1991: The problem of multiplicity in research on teleconnections. *International Journal of Climatology*, **11**, 505-513.

Livezey, R.E, and W.Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Monthly Weather Review*, **111**, 4659.

Mielke, P.W., K.J. Berry, and G.W. Brier, 1981: Application of multi-response permutation procedures for examining seasonal changes in monthly mean sea level pressure patterns. *Monthly Weather Review*, **109**, 120-126.

Polyak, I., 1996: *Computational Statistics in Climatology*. Oxford University Press, New York, 358 pp.

Storey, J.D., and R. Tibshirani, 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Science*, **100**, 9440-9445.

Ventura, V., C.J. Paciorek, and J.S. Risbey, 2004: Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *Journal of Climate*, **17**,4343-4356.

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

von Storch, H., and F.W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge UK, 484 pp.

Walker, G.T. 1914: Correlation in seasonal variations of weather. III. On the criterion for the reality of relationships or periodicities. *Memoirs of the Indian Meteorological Department* **21**(9), 13–15.

Wilks, D.S., 1997: Resampling hypothesis tests for autocorrelated fields. *Journal of Climate,* **10**, 65-83.

Wilks, D.S., 2006: On "field significance" and the false discovery rate. *Journal of Applied Meteorology and Climatology*, 45, 1181-1189.

Wilks, D.S., 2011: *Statistical Methods in the Atmospheric Sciences*, 3rd Ed. Academic Press, Amsterdam, 676 pp.

Zwiers, F.W., 1987: Statistical considerations for climate experiments: Part II: Multivariate tests. *Journal of Climate and Applied Meteorology*, **26**, 477-487.

Zwiers, F.W., and H.J. Thiébaux, 1987: Statistical considerations for climate experiments. Part I: scalar tests. *Journal of Climate and Applied Meteorology*, **26**, 465–476.