"THE STIPPLING SHOWS STATISTICALLY SIGNIFICANT GRID POINTS"

How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It

BY D. S. WILKS

Controlling the false discovery rate provides a computationally straightforward approach to interpretation of multiple hypothesis tests.

> "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 30 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 grid points, are very often interpreted incorrectly and in a way that leads to research results being 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 "metatest" using as input the results of the many individual tests to address the "global" null hypothesis that all individual "local" (e.g., grid point) null hypotheses are true. If the 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 misled into believing results from the many erroneous rejections of true local gridpoint null hypotheses that will invariably occur.

Unfortunately, very little has changed during the intervening decades with respect to the overinterpretation 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 studies implicitly but wrongly represented that any individual gridpoint test exhibiting

AFFILIATIONS: WILKS—Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York
CORRESPONDING AUTHOR: Dan Wilks, Department of Earth and Atmospheric Sciences, 1104 Bradfield Hall, Cornell University, Ithaca, NY 14853
E-mail: dsw5@cornell.edu

The abstract for this article can be found in this issue, following the table of contents.
DOI:10.1175/BAMS-D-15-00267.1

In final form 3 March 2016

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 hypothesis 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. Consequences of the widespread and continued failure to address the issue of multiple hypothesis testing are overstatement and overinterpretation 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 overinterpretation of multiple-testing results.

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 collections 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 quantity. 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 been drawn from a probability distribution whose mean is αN_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,

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

$$x = 0, 1, \ldots, N_0.$$

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 αN_0 on average, Eq. (1) specifies nonnegligible probabilities for numbers of erroneous rejections that may be substantially larger than αN_0. When the members of the collection of hypothesis tests are not independent, which is the usual situation for gridded data, Eq. (1) is no longer valid and the probabilities for numbers of erroneous rejections much larger than αN_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 scientifically meaningful way?

HISTORICAL DEVELOPMENT OF MULTIPLE TESTING IN THE ATMOSPHERIC SCIENCES.

Walker's 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

[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).

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. 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 (e.g., Wilks 2006)

$$\alpha_{\text{Walker}} = 1 - \left(1 - \alpha_0\right)^{1/N_0}. \qquad (2)$$

Of course $\alpha_{\text{Walker}} = \alpha_0$ for a single ($N_0 = 1$) test. To limit the probability of erroneously rejecting any of $N_0 = 100$ true null hypotheses to the level $\alpha_0 = 0.05$, only those tests having p values smaller than $\alpha_{\text{Walker}} = 0.000513$ would be regarded as significant according to this criterion. In contrast, as noted above, naïvely evaluating each of $N_0 = 100$ independent 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 Eq. (2) will yield relatively few rejections of true null hypotheses, the Walker criterion is quite strict since $\alpha_{\text{Walker}} \approx \alpha_0/N$, which compromises the sensitivity of the procedure for detecting false null hypotheses.

The field significance approach. Von Storch (1982) and Livezey and Chen (1983) cast the problem of evaluating multiple hypothesis tests as a metatest, 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 [Eq. (1)] allows calculation of the minimum number of locally significant tests required to reject a global null hypothesis—that is, 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) 5 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 10 or more such rejections are required in order to have smaller than $\alpha_{\text{global}} = \alpha_0 = 0.05$ 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. The usual condition of spatial correlation among the local gridpoint tests implies that even more local test rejections than implied by Eq. (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" $N_{\text{eff}} < N$ and to use this value in Eq. (1), although often N_{eff} cannot be estimated rigorously (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).

The Livezey–Chen procedure has other drawbacks beyond its sensitivity to spatial correlation. The most important of these are as follows:

i) 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). This problem is particularly acute when the fraction of false null hypotheses is small.

ii) 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 overinterpret the data in an attempt to explain them.

A PRINCIPLED AND STRAIGHTFORWARD SOLUTION—CONTROLLING THE FALSE DISCOVERY RATE. The problems just noted can be addressed by controlling 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. It has become the dominant, mainstream approach to evaluation of multiple hypothesis test results, both in the statistics literature and in the scientific literature more broadly, with Google Scholar listing more than 34,000 citations of the original (Benjamini and Hochberg 1995) paper. 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). In particular, FDR control addresses (and, in a sense, puts a ceiling on) the probability that a rejected local null hypothesis is in fact true, whereas a p value quantifies the probability of results at least as inconsistent with the null hypothesis as the observed test statistic, under the assumption that the null hypothesis is true. The former notion is more closely aligned with common intuition, and indeed p values are commonly misinterpreted in this way (e.g., Storey and Tibshirani 2003; Jolliffe 2004; Ambaum 2010), presumably because investigators often prefer this framing of scientific answers.

The FDR procedure is similar in spirit to Walker's approach in that it requires a higher standard (i.e., smaller p values) in order to reject local null hypotheses. The algorithm operates on the collection of p values from N local hypothesis tests p_i, with $i = 1, \ldots, N$, which are first sorted in ascending order. Using a standard statistical notation, these sorted p values are denoted using parenthetical subscripts, so that $p_{(1)} \le p_{(2)} \le \ldots \le p_{(N)}$. 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:

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

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 [Eq. (2)] is very nearly the same as Eq. (3) if $i = 1$, so that the FDR procedure will be more sensitive to detecting false null hypotheses to the extent that Eq. (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 Eq. (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 $\alpha_{\text{global}} = \alpha_{\text{FDR}}$ (Wilks 2006).

Even though Eq. (3) assumes statistical independence among the local test results, the FDR procedure is (as will be illustrated in the following section) approximately valid (while being somewhat conservative) even when those results are strongly correlated, unlike the use of Eq. (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

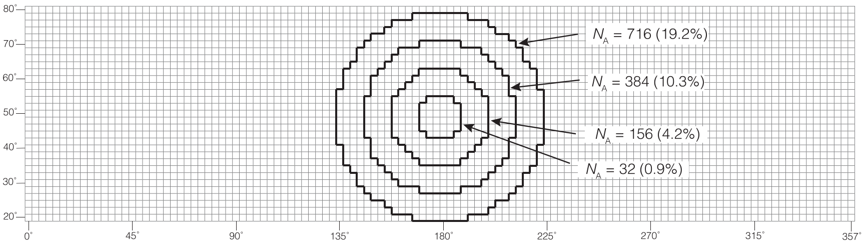

Fig. 1. Hypothetical 3720-gridpoint domain, representing the Northern Hemisphere from 20° to 80°N. Concentric thick outlines indicate regions where local null hypotheses are not true.

local p values, the most complicated computation required is merely their sorting into ascending order so that Eq. (3) can be evaluated.

SYNTHETIC-DATA EXAMPLES.

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 Fig. 1. The ordinate represents the 31 latitudes from 20° 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 thick outlines indicate regions, ranging in extent from 0.9% to 19.2% of the total number of grid points, where the local null hypotheses are not 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

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

where d is the great-circle distance between two grid points. These eight spatial autocorrelation functions range in e-folding distance (i.e., average distance at which the data correlations drop below $1/e = 0.3679$) from 0.1×10^3 km (nearly spatially independent) to 10×10^3 km (very strongly dependent). The star symbols in Fig. 2 indicate data for spatial autocorrelation of the Northern Hemisphere 500-hPa 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.

One of the strengths of the FDR method is that it is applicable to collections of p values from hypothesis tests of any form. In this section the FDR method is illustrated using p values from one-sample t tests, with one local t test being computed for each of the 3720 grid points shown in Fig. 1. The underlying synthetic data are random Gaussian fields with spatial correlations governed by Eq. (4), generated using methods described in Wilks (2011, p. 499). That is, the statistical distribution of the synthetic values at each grid point is standard Gaussian—that is, having zero mean and unit variance. For each realization of 3720 local hypothesis tests, 25 of these fields were generated, yielding 3720 sample means and 3720 sample standard deviations (which are not assumed to be equal across the domain). From these quantities, the test statistics for 3720 local one-sample t tests with H_0: $\{\mu = 0\}$ having 24 degrees of freedom at each grid point were computed. The alternative hypothesis in each case is two sided: that is, H_A: $\{\mu \neq 0\}$. In experiments where some of the local null hypotheses are false, all gridpoint sample means within one of the outlines shown in Fig. 1 were increased above zero by amounts $\Delta\mu$ ranging from 0.05 to 1.00. Using $c = 0.42$ yields spatial

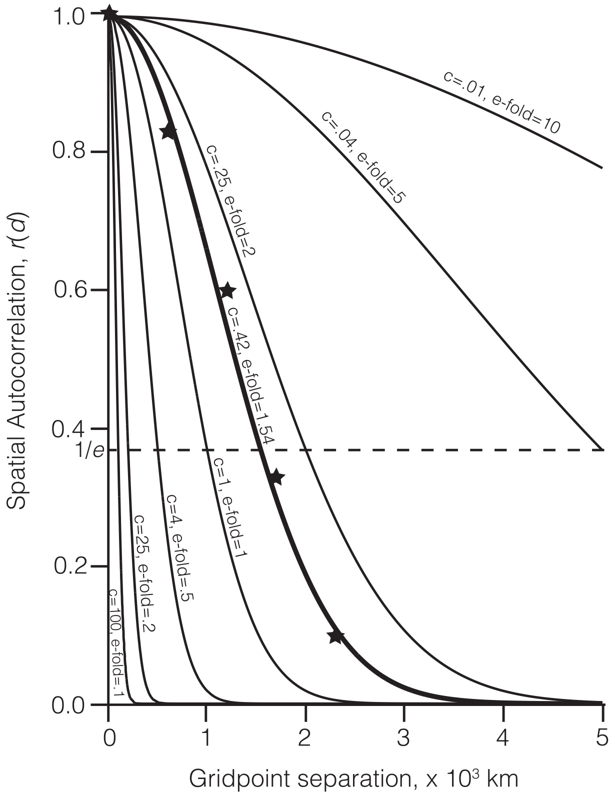

Fig. 2. Eight spatial autocorrelation functions of the form in Eq. (4). Stars indicate correlations for Northern Hemisphere 500-hPa heights from Polyak (1996).

correlation corresponding to the magnitude typically exhibited by Northern Hemisphere 500-hPa height fields. Although the correlation function in Eq. (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.

Global test properties. Figure 3 illustrates the operation of the FDR procedure (diagonal lines), in contrast to the "naïve stippling" 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 grid points 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 Eq. (3) using $\alpha_{\text{FDR}} = 0.10$, according to which $p^*_{\text{FDR}} = 0.003998 = p_{(150)}$. 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 dots below the dashed diagonal line. The twelve circles above the dashed diagonal line represent false null hypotheses that were not rejected. The six crosses 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 Eq. (3) when $\alpha_{\text{FDR}} = 0.20$, in which case $p^*_{\text{FDR}} = 0.009502 = p_{(183)}$. 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 naïve stippling approach of rejecting any local null hypothesis for which the p value is less than $\alpha_0 = 0.05$ (dashed horizontal line) detects all 156 false null hypotheses, but at the expense of erroneously rejecting 189 true null hypotheses (crosses and dots 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 Eq. (3)], which ideally will equal $\alpha_{\text{global}} = \alpha_{\text{FDR}}$. These probabilities are approximated in Fig. 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. That is, when the spatial correlation

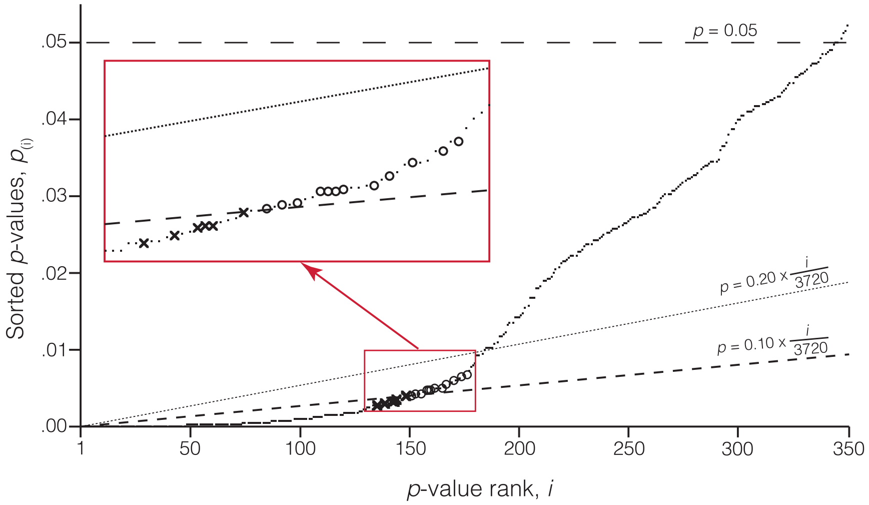

Fig. 3. Illustration of the FDR criterion using $\alpha_{\text{FDR}} = 0.10$ (dashed diagonal line), $\alpha_{\text{FDR}} = 0.20$ (dotted diagonal line), and the naïve stippling approach of rejecting any local test with p value smaller than $\alpha_0 = 0.05$ (dashed horizontal line). Plotted points are 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. Crosses represent six tests with true null hypotheses that were falsely rejected, and circles represent false local null hypotheses that were not rejected, when $\alpha_{\text{FDR}} = 0.10$. Inset shows closer view of points within the red box. The 345 tests with p values smaller than $\alpha_0 = 0.05$ would be declared significant under the naïve stippling procedure, even though a majority of these null hypotheses are true.

is high, the achieved FDR will be smaller (more strict) than the nominal FDR. This is consistent with prior results (Ventura et al. 2004; Wilks 2006). Figure 4 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 $\alpha_{\mathrm{FDR}} = 2\alpha_{\mathrm{global}}$.

In sharp contrast, the achieved test levels for the Livezey–Chen counting procedure, also with no adjustment for spatial correlation, are very strongly permissive. For example, using Eq. (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 $\alpha_0 = \alpha_{\mathrm{global}} = 0.05$. This criterion produces achieved global test levels of 0.0907 and 0.3517 when the *e*-folding distances are 0.2 and 1.54×10^3 km, respectively (results not shown in the figure). The naïve stippling interpretation that any significant local test implies field significance is even worse, as it produces 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 Fig. 4.

Local test interpretations. Often the primary interest will be interpretation of the locations and spatial patterns of the locally significant test results, which might be interpreted as "signal." 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 ("noise"). Figure 5 shows false discovery rates for the FDR method with $\alpha_{\mathrm{FDR}} = 0.10$ (red), the Livezey–Chen counting approach with $\alpha_0 = \alpha_{\mathrm{global}} = 0.05$ (black), and the naïve stippling approach of rejecting any local null hypothesis whose *p* value is no larger than the nominal $\alpha_0 = 0.05$ (brown), as functions of numbers of false local null hypotheses and alternative-hypothesis magnitudes $\Delta\mu$, for the realistic *e*-folding distance 1.54×10^3 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 Fig. 3 are 6/150 = 0.04 for the FDR procedure, 189/345 = 0.55 for the naïve stippling procedure, and zero for the Livezey–Chen counting procedure because fewer than the required 365 local tests[2] 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 $\Delta\mu$. For larger numbers of false local null hypotheses, the Livezey–Chen procedure yields much larger false discovery rates. Worst performance of all is exhibited by the naïve stippling 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.

To help visualize the foregoing more concretely, Fig. 6 shows maps for a particular realization, interpreted according to the FDR procedure with $\alpha_{\mathrm{FDR}} = 0.10$ (Fig. 6a) and the naïve stippling approach using $\alpha_0 = 0.05$ (Fig. 6b). Correct local null hypothesis rejections are indicated by plus signs, failures to reject

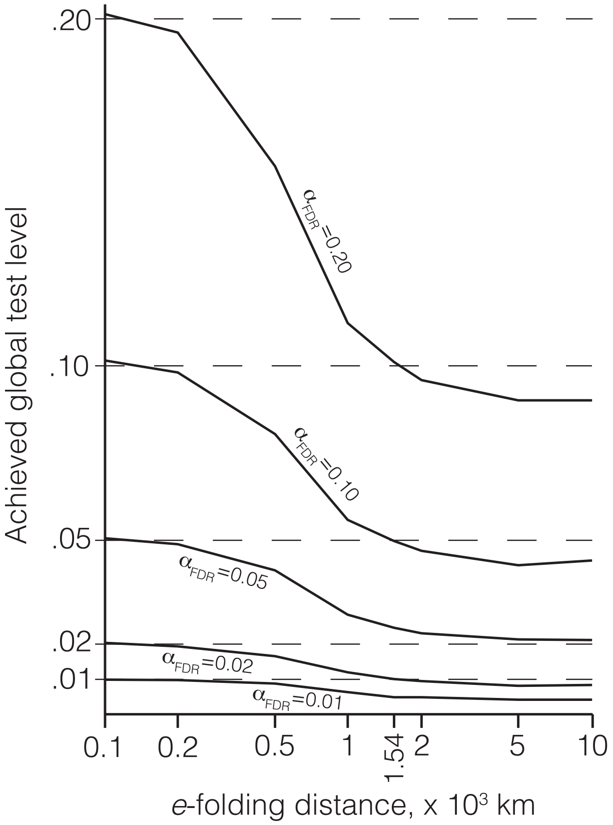

Fig. 4. Achieved global test levels (probabilities of rejecting true global null hypotheses) when using the FDR procedure, as a function of spatial correlation strength. For moderate and strong spatial correlation, approximately correct results can be achieved by choosing $\alpha_{\mathrm{FDR}} = 2\alpha_{\mathrm{global}}$.

[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.

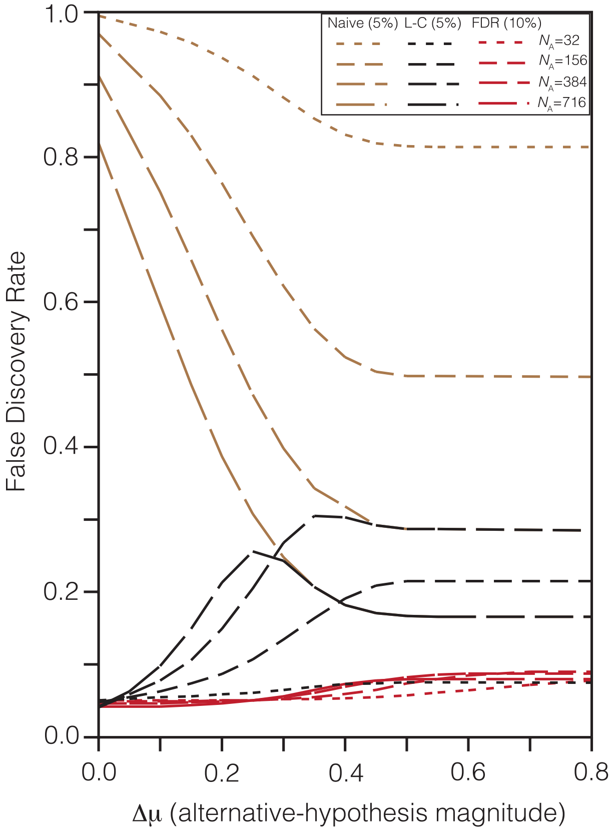

Fig. 5. False discovery rates for the FDR method with $\alpha_{FDR} = 0.10$ (red), the Livezey–Chen counting approach with $\alpha_0 = \alpha_{global} = 0.05$ (black), and the naïve stippling approach with $\alpha_0 = 0.05$ (brown), as functions of numbers of false local null hypotheses and alternative-hypothesis magnitudes $\Delta\mu$, using the e-folding distance 1.54 × 10^3 km.

false local null hypotheses are indicated by circles, and erroneous rejections of true null hypotheses are indicated by crosses. These maps correspond to the ranked p values shown in Fig. 3, with $N_A = 156$, $\Delta\mu = 0.7$, and e-folding distance 1.54 × 10^3 km. In Fig. 6a the FDR procedure fails to reject 12 of the 156 false null hypotheses (circles) but erroneously rejects only 6 true null hypotheses (crosses). The result is that the FDR procedure locates the true signal very effectively while introducing very little noise. By contrast, in Fig. 6b the naïve stippling procedure locates all 156 false null hypotheses but also erroneously indicates another 189 nominally significant grid points. The very large additional noise level in Fig. 6b 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, because the number of false null hypotheses is relatively small, 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 Fig. 6b, even though some of the p values are extremely small.

A REAL-DATA EXAMPLE.

Figure 7 shows an analysis of linear trends in annual precipitation for the period 1951–2010, modified from an original figure in Hartmann et al. (2013, p. 203). The underlying data are monthly precipitation values interpolated to a 5° × 5° grid from the Global Historical Climatology Network (Vose et al. 1992). The colored patches locate the 408 grid elements having at least 42 (70%) complete calendar years and at least two complete years during 1951–56 and 2005–10 (Hartmann et al. 2013). The 128 grid elements with linear trends exhibiting regression slopes that are large enough in absolute value to achieve local statistical significance at the $\alpha = 0.10$ level, without considering the multiple-testing problem, have been indicated by the plus signs.

The red circles in Fig. 7 locate the 51 grid elements exhibiting linear precipitation trends that are meaningfully different from zero, assessed according to the FDR method with $\alpha_{FDR} = 0.10$. Here $\alpha_{FDR} = \alpha = 0.10$ (the same test level as the original naïve stippling results) has been used because of the relatively weak spatial correlation of the underlying annual precipitation totals. Figure 8 shows these correlations for the pairs of colored grid elements in Fig. 7 separated by no more than 2 × 10^3 km and indicates an approximate e-folding decorrelation distance of 0.62 × 10^3 km, or about 150% of the typical grid element separation of approximately 400 km. Comparing Fig. 3, which was calculated on the basis of a 2° × 3° grid system, 150% of the typical grid separation of 250 km translates to an e-folding correlation distance of approximately 0.38 × 10^3 km, for which choosing $\alpha_{FDR} = \alpha$ produces only very slight test conservatism.

Using $\alpha_{FDR} = 0.10$, $p^*_{FDR} = p_{(51)} = 0.01136$ [Eq. (3)], so that the 51 grid elements whose local tests reject null hypotheses of zero linear trend with p values no larger than 0.01136 can be regarded as meaningful. No more than five of these are expected to be erroneous rejections of true local null hypotheses.

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS.

The problem of simultaneously evaluating results of multiple hypothesis tests, often at a large network of grid points or other geographic locations, is widespread in meteorology and climatology. Unfortunately, the dominant approach

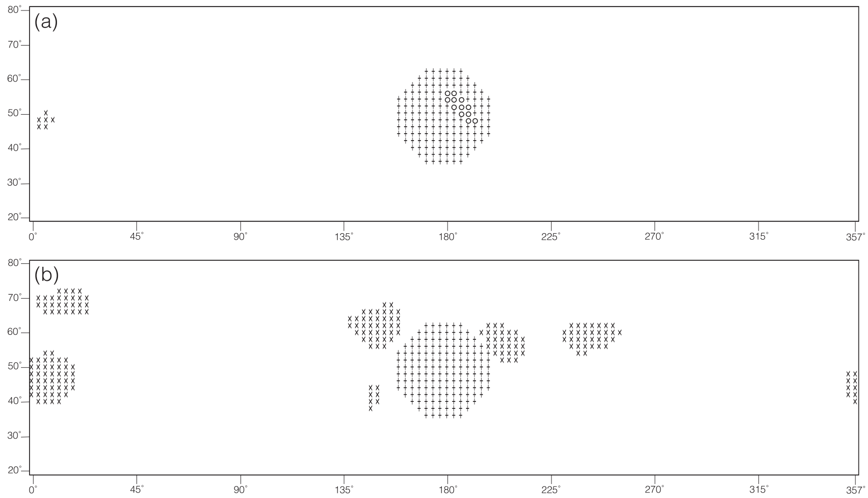

Fig. 6. Maps of local test decisions made by (a) the FDR procedure with $\alpha_{FDR} = 0.10$ and (b) the naïve stippling approach using $\alpha_0 = 0.05$. Correct local null hypothesis rejections are indicated by plus signs, failures to reject false local null hypotheses are indicated by circles, and erroneous rejections of true null hypotheses are indicated by crosses. Results correspond to the ranked p values shown in Fig. 3, with $N_A = 156$, $\Delta\mu = 0.7$, and e-folding distance 1.54×10^3 km.

to this problem in the literature is to naïvely 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, immediately flagging the results portrayed as almost certainly overstated. This statistically unprincipled practice should be unacceptable 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 metatest 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 was a very substantial advance over the naïve stippling procedure that ignores the effects of multiple testing, it suffers from several drawbacks.

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, intuitive interpretation, and only weak sensitivity to alternative-hypothesis magnitudes and the number of false null hypotheses.

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 resampling procedures addressing both types of autocorrelation simultaneously (e.g., Wilks 1997) are unnecessary. The examples presented here were based on local t tests pertaining to sample means and tests for nonzero regression slopes. However, the method is applicable to collections of multiple hypothesis test results, regardless of the mathematical forms of those

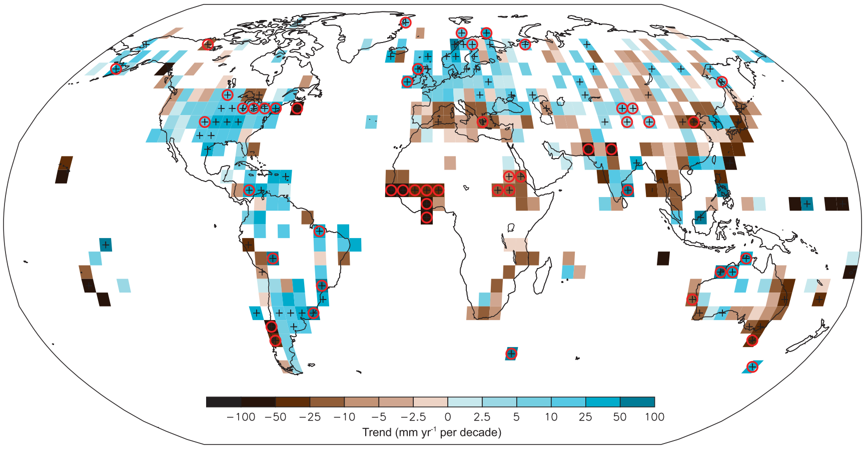

Fig. 7. Linear trends in annual precipitation during 1951–2010, based on data from the Global Historical Climatology Network (Vose et al. 1992). Grid elements with linear trends exhibiting local statistical significance at the $\alpha = 0.10$ level are been indicated by the plus signs, and those with p values small enough to satisfy the FDR criterion with $\alpha_{FDR} = 0.10$ [Eq. (3)] are indicated by the red circles. The figure has been modified from Hartmann et al. (2013, p. 203).

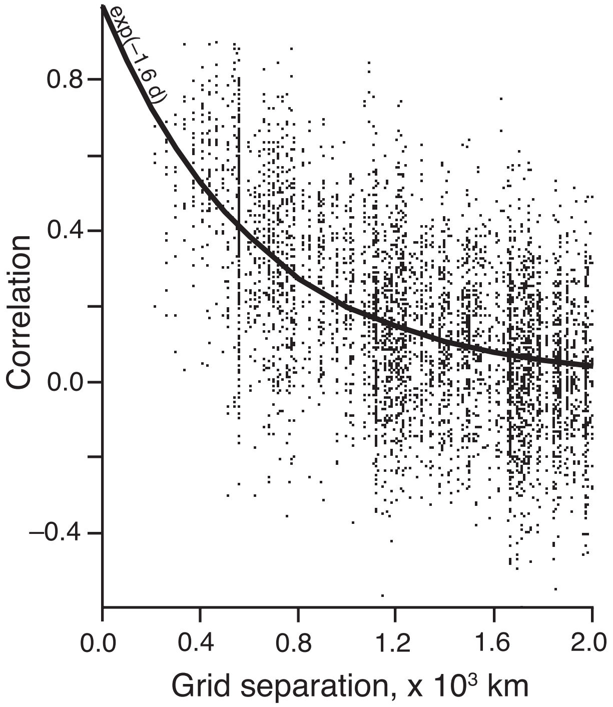

Fig. 8. Correlations for pairs of colored grid elements in Fig. 7 separated by no more than 2 × 10³ km. The fitted exponential decay function indicates an approximate e-folding decorrelation distance of 0.62 × 10³ km, or approximately 150% of the typical grid spacing.

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 scientific 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 Eq. (3). The usual strong spatial correlation encountered in gridded atmospheric data can be accommodated by choosing $\alpha_{\mathrm{FDR}} = 2\alpha_{\mathrm{global}}$, as illustrated in Fig. 4. The consequence of employing this statistically principled procedure—in stark contrast to the all-too-common naïve stippling approach—is that there is much reduced scope for overstatement and overinterpretation of the results. In particular, the analyst is not tempted to construct possibly fanciful rationalizations for the many spurious local test rejections, which may appear to be physically coherent structures because of the strong spatial autocorrelation.

ACKNOWLEDGMENTS. I thank the anonymous reviewers, whose comments lead to an improved presentation. This work was supported by the National Science Foundation under Grant AGS-1112200.

REFERENCES

Ambaum, M. H. P., 2010: Significance tests in climate science. *J. Climate*, **23**, 5927–5932, doi:10.1175/2010JCLI3746.1.

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.

Hartmann, D. L., and Coauthors, 2013: Observations: Atmosphere and surface. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 159–254.

Jolliffe, I. T., 2004: P stands for… *Weather*, **59**, 77–79, doi:10.1256/wea.132.03.

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, doi:10.1175/1520-0469(1982)039<1446:SEOCEW>2.0.CO;2.

——, 2002: Sir Gilbert Walker and a connection between El Niño and statistics. *Stat. Sci.*, **17**, 97–112, doi:10.1214/ss/1023799000.

——, and B. G. Brown, 1991: The problem of multiplicity in research on teleconnections. *Int. J. Climatol.*, **11**, 505–513, doi:10.1002/joc.3370110504.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111**, 46–59, doi:10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2.

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. *Mon. Wea. Rev.*, **109**, 120–126, doi:10.1175/1520-0493(1981)109<0120:AOMRPP>2.0.CO;2.

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

Storey, J. D., and R. Tibshirani, 2003: Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, **100**, 9440–9445, doi:10.1073/pnas.1530509100.

Ventura, V., C. J. Paciorek, and J. S. Risbey, 2004: Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *J. Climate*, **17**, 4343–4356, doi:10.1175/3199.1.

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, doi:10.1175/1520-0469(1982)039<0187:AROCSA>2.0.CO;2.

——, and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 484 pp.

Vose, R. S., R. L. Schmoyer, P. M. Steurer, T. C. Peterson, R. Heim, T. R. Karl, and J. Eischeid, 1992: The Global Historical Climatology Network: Long-term monthly temperature, precipitation, sea level pressure, and station pressure data. Environmental Sciences Division Publ. 3912, ORNL/CDIAC-53, NDP-041, 325 pp., doi:10.3334/CDIAC/cli.ndp041.

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

Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate*, **10**, 65–83, doi:10.1175/1520-0442(1997)010<0065:RHTFAF>2.0.CO;2.

——, 2006: On "field significance" and the false discovery rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189, doi:10.1175/JAM2404.1.

——, 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Academic Press, 676 pp.

Zwiers, F. W., 1987: Statistical considerations for climate experiments. Part II: Multivariate tests. *J. Climate Appl. Meteor.*, **26**, 477–487, doi:10.1175/1520-0450(1987)026<0477:SCFCEP>2.0.CO;2.

——, and H. J. Thiébaux, 1987: Statistical considerations for climate experiments. Part I: Scalar tests. *J. Climate Appl. Meteor.*, **26**, 465–476, doi:10.1175/1520-0450(1987)026<0464:SCFCEP>2.0.CO;2.