Controlling the Proportion of Falsely Rejected Hypotheses when Conducting Multiple Tests with Climatological Data

VALÉRIE VENTURA

Department of Statistics, and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania

CHRISTOPHER J. PACIOREK

Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts

JAMES S. RISBEY

Centre for Dynamical Meteorology and Oceanography, Monash University, Clayton, Victoria, Australia

(Manuscript received 19 August 2003, in final form 23 February 2004)

ABSTRACT

The analysis of climatological data often involves statistical significance testing at many locations. While the field significance approach determines if a field as a whole is significant, a multiple testing procedure determines which particular tests are significant. Many such procedures are available, most of which control, for every test, the probability of detecting significance that does not really exist. The aim of this paper is to introduce the novel ‘‘false discovery rate’’ approach, which controls the false rejections in a more meaningful way. Specifically, it controls a priori the expected proportion of falsely rejected tests out of all rejected tests; additionally, the test results are more easily interpretable. The paper also investigates the best way to apply a false discovery rate (FDR) approach to spatially correlated data, which are common in climatology. The most straightforward method for controlling the FDR makes an assumption of independence between tests, while other FDR-controlling methods make less stringent assumptions. In a simulation study involving data with correlation structure similar to that of a real climatological dataset, the simple FDR method does control the proportion of falsely rejected hypotheses despite the violation of assumptions, while a more complicated method involves more computation with little gain in detecting alternative hypotheses. A very general method that makes no assumptions controls the proportion of falsely rejected hypotheses but at the cost of detecting few alternative hypotheses. Despite its unrealistic assumption, based on the simulation results, the authors suggest the use of the straightforward FDR-controlling method and provide a simple modification that increases the power to detect alternative hypotheses.

1. Introduction

Climate research often involves an assessment of the statistical significance of a quantity, such as an observed correlation or trend. If the quantity is measured at multiple locations, this requires testing many hypotheses simultaneously. Such a situation can arise when correlating an atmospheric field with a nonspatial quantity, such as a teleconnection pattern, or when correlating two atmospheric fields. It can also arise when evaluating time trends or model performance at many locations. The usual setting for such multiple testing in climatological studies involves quantities measured over time, with time providing the replication necessary for calculating the chosen test statistic, such as correlation, trend, or model fit, at each of the locations. This paper addresses the problem of evaluating statistical significance when many hypothesis tests are performed simultaneously.

A single test performed at significance level α has probability α of rejecting the null hypothesis when it is in fact true. Hence if n such tests are performed when all n null hypotheses are true (the collective null hypothesis), then the average number of tests for which the null is falsely rejected is $n\alpha$. For example, with $\alpha = 5\%$, testing for a trend at 1000 locations at which no change really occurred would yield 50 significant locations on average; this is unacceptably high.

Climatologists have long recognized the problem of accounting for multiple tests; as early as 1914, when conducting multiple tests, Walker adjusted the significance level used for rejection (Katz 2002). More recently, climatologists have taken the alternative approach of testing for field significance. In a seminal paper, Livezey and Chen (1983) proposed a method that

Corresponding author address: Christopher Paciorek, Department of Biostatistics, 655 Huntington Avenue, Harvard School of Public Health, Boston, MA 02115.
E-mail: paciorek@alumni.cmu.edu

determines if a field of individual, or local, hypothesis tests are collectively significant. This approach is popular in the climatological literature, with over 300 citations of Livezey and Chen (1983) since its publication. Statistical climatology texts cover the method as the primary way to deal with multiple hypothesis tests (Wilks 1995; von Storch and Zwiers 1999). The basic approach relies on the fact that if the collective null hypothesis holds, and if the *p*-values are independent, they can be viewed as a sample from a binomial distribution with sample size *n* and probability of ''success'' (correctly accepting the null), $1 - \alpha$. Using properties of the binomial distribution, one can calculate the probability *p* of observing as many or more significant *p*-values as were actually observed. If *p* is less than α, the observed field is said to be field, or globally, significant at level *p*. If the *p*-values are not independent, *p* can be obtained by Monte Carlo simulation. Wilks (1997) discusses modifying the approach to make use of the bootstrapped null distribution of quantities other than the *p*-values for determining field significance. Another variant on the Livezey and Chen (1983) approach is to find the number of effectively independent locations, also termed degrees of freedom, n^*, to use in place of the actual number, *n*, of locations in the binomial approach.

Livezey and Chen (1983) illustrate the field significance method on the correlation between the Southern Oscillation index and the 700-mb geopotential height at many locations over a 29-yr period. Subsequent researchers have used the method or variants on a wide variety of problems. The method screens out situations in which apparent spatial effects, such as a coherent area of positive correlations in one region, could plausibly be accounted for by chance as a result of spatial correlation in the observations used to calculate the field. Researchers who find field significance then proceed to interpret the field.

We now present an example of a multiple testing problem in climatology and give the results of several multiple testing approaches to highlight the issues at hand. Paciorek et al. (2002) analyze time trends of multiple indices of winter storm activity during 1949–99 to assess if activity has changed over 51 yr. For illustrative purpose here, we focus on one such index, the variance of the bandpass-filtered daily mean temperature at 500 hPa, which reflects the variability in temperature at scales of 2.5–10 days, and acts as a measure of large-scale atmospheric variability brought about in part by the passage of extratropical cyclones. To test for a change, Paciorek et al. (2002) use the fitted slopes from linear regressions of the index against time (51 yr) at each of $n = 3024$ locations on a 2.5° by 2.5° latitude–longitude grid in the Northern Hemisphere, 20°–70°N.

The method of Livezey and Chen (1983) detects field significance, with $p < 0.001$, but gives no indication about which particular locations are significant, even though this is of crucial interest. For that, we must test for a change at each location. Traditional methods typically declare the change to be significant if it exceeds some multiple of its standard deviation σ; the most commonly used threshold is 2σ, which corresponds to a probability of about $\alpha = 5\%$ of erroneously rejecting a null hypothesis. An alternative is the recent false discovery rate (FDR) procedure of Benjamini and Hochberg (1995), which controls the proportion *q* of falsely rejected null hypotheses relative to the total number of rejected hypotheses.

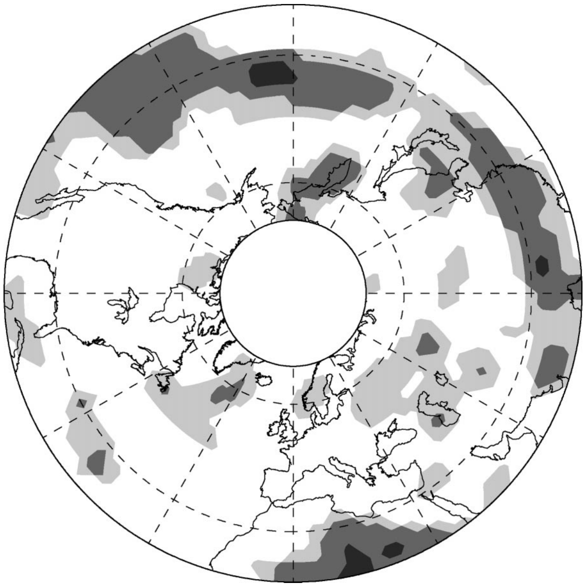

FIG. 1. Map of locations with significant trends in the temperature variance index, 1949–99, in the area 20°–70°N based on three multiple testing procedures: $\alpha = 5\%$ (light + medium + dark gray), FDR (medium + dark gray), and Bonferroni (dark gray).

Figure 1 shows the significant locations as determined by testing at significance level $\alpha = 5\%$, testing at a more stringent Bonferroni-corrected $\alpha = n^{-1} \times 5\%$ with $n = 3024$ (section 2b), and testing by the FDR procedure with the false discovery rate set at $q = 5\%$. These three procedures pick vastly different numbers of significant locations—941, 19, and 338, respectively—and perhaps the only reassuring (and not surprising, as we shall see later) fact is that there appears to be an order: the Bonferroni-significant locations are a subset of the FDR-significant locations, which in turn are a subset of the σ-significant locations. Which result should we report?

The purpose of this paper is to introduce FDR procedures to members of the climatology community, who are often faced with problems of multiple testing, and to show the advantages of FDR compared to traditional procedures.

The next three sections are fairly distinct. In section 2, we describe how the competing testing procedures are performed and the different ways in which they control the number of false detections. We argue that FDR controls a more useful and tangible criterion than do the classical methods. This is the most important message of this paper. However, the original FDR procedure of Benjamini and Hochberg (1995), which we describe in section 2 and show the results of in Fig. 1, makes independence assumptions that may be unrealistic for climate data. Section 3 introduces other FDR procedures (Yekutieli and Benjamini 1999; Benjamini and Yekutieli 2001) designed to handle dependent tests; such tests arise often in climatology [in fact, Yekutieli and Benjamini (1999) illustrate their procedure with a climatological example], since observations at nearby locations are likely to be spatially correlated. These new FDR procedures control the number of false detections in the same meaningful way as we describe in section 2, but we know of no simulation study that shows how they actually perform when applied to correlated data. We therefore conduct such a study, based on simulated data consistent with climatological data. Our conclusion is that the FDR procedure of Benjamini and Yekutieli (2001) is too conservative to be useful, while the improvements of the procedure of Yekutieli and Benjamini (1999) over the original procedure of Benjamini and Hochberg (1995) are too minor to warrant the additional computation required. Section 4 introduces recent refinements to the FDR procedure of Benjamini and Hochberg (1995), and section 5 summarizes our findings.

2. Multiple testing procedures and their properties

Consider the general problem of detecting significance, for example, of trends at many locations over some area. The null hypothesis, H_0, states that no change occurred and is tested against the loosely specified alternative hypothesis, H_A, that ‘‘some’’ change (increase, decrease, or either) occurred. Whereas the commonly used test of Livezey and Chen (1983) determines if the data contain evidence against H_0 globally over the whole area, here we are concerned with testing H_0 versus H_A at a potentially large number, n, of locations in the area. That is, we wish to make local determinations of significance.

The first step is to choose a test statistic, T, a function of the data whose value helps us decide if H_0 should be rejected or not. In our temperature variance index example, T is the estimated regression slope of index values against time, with evidence against H_0 provided by large values of $|T|$, since we are interested in detecting either an increase or a decrease. The choice of T is in itself an interesting problem, since some particular T may be better able to discriminate between H_0 and H_A; we will not debate this question here and instead will assume that some suitable T has been identified for the purpose of answering the question of scientific interest.

Letting t_i denote the observed value of T at location $i = 1, \ldots, n$, and assuming that large values of $|T|$ provide evidence against H_0, the corresponding p-values are

$$p_i = \Pr(|T| \geq |t_i| \mid H_0 \text{ true}), \tag{1}$$

where ‘‘$| H_0$ true’’ means that (1) is calculated with respect to the distribution of T under the assumption that H_0 is true. In plain words, (1) measures the probability that $|T|$ could have a value greater than $|t_i|$ when H_0 is true. A high probability means that t_i, the value of T we observed, is perfectly ordinary when H_0 is true. A low probability indicates that the observed t_i is very unusual when H_0 is true, which means that either we observed a rare event (a very unusual value of T), or H_0 is not true. Therefore, a small p-value, p_i, provides strong, although not foolproof evidence against H_0; rejecting H_0 could be the wrong decision. Both traditional and FDR procedures reject H_0 when p_i is small, but they differ in the way they control the number of erroneous decisions that can be made, as described in section 2b. We first describe how the various testing procedures are performed.

a. Multiple testing methods

Before we perform any test, we must first choose, for the traditional procedure, the nominal probability, α, of erroneously rejecting any particular null hypothesis, and for the FDR procedure, the nominal FDR, q, which is the rate we are willing to allow of false rejections out of all rejections. The adjective ‘‘nominal’’ conveys that α and q are chosen a priori, before the tests are performed. The probability α is commonly known as the significance level, or the probability of a type I error, although in the context of multiple testing, α could more suitably be referred to as the false positive rate (FPR), as argued in section 2b; henceforth, we refer to traditional procedures as FPR procedures. The choice of α or q is subjective and should reflect our willingness to make mistakes in order to have more power to detect real changes (see section 2b). The most commonly used value for both α and q is 5%.

A traditional FPR procedure rejects H_0 at location i if p_i is smaller than α; the same threshold is applied to all tests/locations. In contrast, the FDR procedure of Benjamini and Hochberg (1995) rejects H_0 at all locations i for which $p_i \leq p_k$, where

$$k = \max_{i=0,\ldots,n} \left\{ i\colon p_{(i)} \leq q\frac{i}{n} \right\}, \tag{2}$$

with $p_{(i)}$, $i = 1, \ldots, n$; the p-values (1) sorted in ascending order; and $p_{(0)} = 0$.

Figure 2 clarifies this seemingly complicated rule. It displays the outcome of the two testing procedures on a stylized example, with the particular choice of $q = \alpha = 20\%$. The ordered p-values, $p_{(i)}$, are plotted against i/n for $i = 1, \ldots, n$. Then the horizontal line at α and the line with intercept zero and slope q are overlaid. All p-values below the horizontal line are rejected by the traditional FPR procedure, while the FDR procedure rejects all p-values, in ascending order, up to the largest

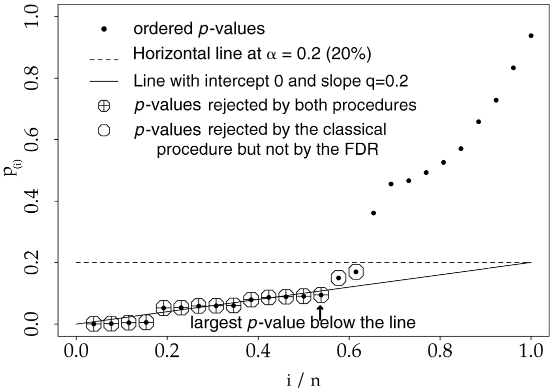

FIG. 2. Illustration of the traditional FPR and FDR procedures on a stylized example, with $q = \alpha = 20\%$. The ordered p-values, $p_{(i)}$, are plotted against i/n, $i = 1, \ldots, n$, and are circled and crossed to indicate that they are rejected by the FPR and FDR procedures, respectively.

p-value that lies below the $(0, q)$ line, indicated on Fig. 2 by an arrow.

Figure 2 also shows that, although the FDR rejection rule is complicated, effectively all p-values below a certain threshold are rejected, since the p-values are plotted in ascending order. This yields two remarks. First, this explains why the three sets of rejected null hypotheses in Fig. 1 were nested subsets: the implicit FDR threshold was between the significance levels of the two FPR procedures, $\alpha = 5\%$ and the Bonferroni-corrected $\alpha = n^{-1} \times 5\%$.

Second, this suggests that the outcome of the FDR procedure could have been obtained with a traditional FPR procedure with some cleverly chosen α. So why bother with an FDR testing procedure? The answer, which we develop further in the next section, is that FDR procedures control false rejections in a meaningful way.

b. Controlling mistakes

When we reject or fail to reject a particular H_0, we may either make the correct decision or make one of two mistakes: reject when H_0 is in fact true or fail to reject when H_0 is in fact false. These mistakes are commonly referred to as false positive and false negative detections and also as type I and type II errors. We denote by n_{FP} and n_{FN} the numbers of such mistakes out of the n tests (see Table 1). Since the truth is unknown, we use testing procedures that control these errors. Both FPR and FDR procedures control the number of false positive detections n_{FP} in different ways, but neither (nor any testing procedure we know) controls the number of false negative detections. It is easy to see why; once α or q is chosen, the decisions about the hypotheses, as carried out in Fig. 2, are determined; there is no room left to control the number of false negatives.

For a traditional FPR procedure, the choice of α determines the properties of the test; α is the probability of rejecting any particular H_0 by mistake, which means that on average, α% of the n_{H_0} locations *for which* H_0 *is true* will be found significant by mistake. We report this in Table 1 as

$$\mathrm{FPP} = n_{\mathrm{FP}}/n_{H_0}, \qquad \mathrm{FPR} = E(\mathrm{FPP}) = \alpha, \qquad (3)$$

where FPP is the observed false positive proportion, and E stands for expectation. The FPP/FPR notation is consistent with standard statistical terminology, where the expectation of an observed ''proportion'' is usually referred to as a ''rate.'' Equation (3) justifies our calling α the FPR.

What (3) means is that the number n_{FP} of false positives that a traditional FPR procedure allows is proportional to the unknown number n_{H_0} of true null hypotheses. So, for example, if most or all locations have n_{H_0} true, this test will yield a large number of false positive detections, as we will later illustrate in Fig. 3.

TABLE 1. Quantities relevant to traditional FPR and new FDR procedures. The information that is known is indicated in bold. FPP, FNP, FDP, and FNDP indicate, respectively, the observed false positive, negative, discovery and nondiscovery proportions, and FPR, FNR, FDR, FNDR indicate the corresponding expected proportions, which we refer to as rates; for example, $E(\mathrm{FDP}) = \mathrm{FDR}$.

	Decision			
TRUTH	Maintain H_0	Reject H_0	Row totals	Quantities relevant to FPR procedures
H_0	$n_{H_0} - n_{\mathrm{FP}}$ No. correctly maintained	n_{FP} No. of false positives	n_{H_0} No. of true H_0	$\mathrm{FPP} = \frac{n_{\mathrm{FP}}}{n_{H_0}}$ **FPR= $\boldsymbol{\alpha}$**
H_A	n_{FN} No. of false negatives	$n_{H_A} - n_{\mathrm{FN}}$ No. correctly rejected	n_{H_A} No. of false H_0	$\mathrm{FNP} = \frac{n_{\mathrm{FN}}}{n_{H_A}}$ FNR = ??
Column totals	$\boldsymbol{n_{\mathrm{accept}}}$ **No. of maintained $\boldsymbol{H_0}$**	$\boldsymbol{n_{\mathrm{reject}}}$ **No. of rejected $\boldsymbol{H_0}$**	**$\boldsymbol{n}$ (# of tests)**	
Quantities relevant to FDR	$\mathrm{FNDP} = \frac{n_{\mathrm{FN}}}{n_{\mathrm{accept}}}$ **FNDR ≤ $\boldsymbol{a}$** (see section 4)	$\mathrm{FDP} = \frac{n_{\mathrm{FP}}}{n_{\mathrm{reject}}}$ **FDR ≤ $\boldsymbol{q}$**		

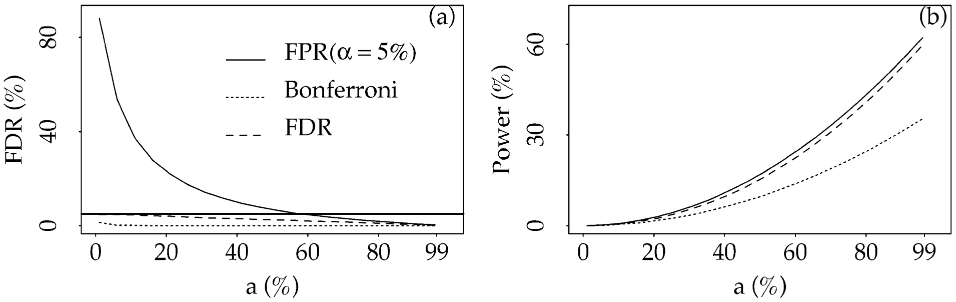

FIG. 3. (a) FDR and (b) power of three testing procedures, as functions of the proportion of true alternative hypotheses, $a = n_{H_A}/n$.

Moreover, because n_{H_0} is unknown, (3) does not help quantify n_{FP}.

But one thing is clear: we probably made at least one mistake. Indeed, when the tests are independent, we can easily show that $P(n_{\mathrm{FP}} \geq 1) = 1 - (1 - \alpha)^{n_{H_0}}$, which, for $\alpha = 5\%$ and $n_{H_0} = 2$, 3, or 10, evaluates to 9.75%, 14.26%, or 40.13%, and increases rapidly with n_{H_0}. That is, the probability of making at least one false positive detection, $P(n_{\mathrm{FP}} \geq 1)$, often referred to as the family-wise type I error, is very high.

Corrections that limit the number of false positives have been suggested. The simplest is to use an arbitrarily smaller α, say 0.5% or 0.01% instead of 5%. The less ad hoc Bonferroni correction uses $\alpha = n^{-1}\alpha'$ for each of the n tests, which guarantees that the family-wise type I error, $P(n_{\mathrm{FP}} \geq 1)$, is smaller than some chosen probability α'; in Fig. 1, we used $\alpha' = 5\%$ and $n = 3024$. Although a smaller α reduces the probability of rejecting null hypotheses by mistake, it also reduces the probability of detecting real changes and thus increases the number, n_{FN}, of false negative errors. Choosing which α to use is a trade-off; increasing the chance of detecting real changes also increases the chance of making false positive detections; the converse is also true.

The false discovery rate procedure of Benjamini and Hochberg (1995) uses a different criterion; it controls the proportion of false rejections out of all rejections, which is referred to as the false discovery proportion (FDP). Choice of a particular q in (2) guarantees that the FDP is on average less than q, which we report in Table 1 as

$$\mathrm{FDP} = n_{\mathrm{FP}}/n_{\mathrm{reject}}, \qquad \mathrm{FDR} = E(\mathrm{FDP}) \leq q, \tag{4}$$

where the FDR is the expectation of the corresponding proportion. For example, setting $q = 5\%$ guarantees that, whatever the number, n_{H_0}, of true null hypotheses and however many tests end up being rejected, at most 5% of the rejected tests are false detections on average. But just as with traditional FPR procedures, choosing q is a trade-off; lowering q to avoid false positive errors decreases our chance of detecting real changes and necessarily increases our chance of making false negative errors. However, the FDR is a more tangible, meaningful, and informative criterion of error control than the FPR. Additionally, since we do know n_{reject} in (4), we can calculate an approximate[1] upper bound for the expected number of false positive detections; that is

$$E(n_{\mathrm{FP}}) \leq q n_{\mathrm{reject}}.$$

In the example of Paciorek et al. (2002; section 1), the FDR procedure with $q = 5\%$ yielded $n_{\mathrm{reject}} = 338$ rejections out of $n = 3024$ tests. If we choose to report this result, we can also report that on average, there are fewer than 338 × 5% = 17 false positive detections, n_{FP}, out of the 338 rejected tests. On the other hand, if we choose to report the results of one of the traditional procedures, we cannot accompany them with estimates of n_{FP} nor of n_{FN}.

Table 1 also reports measures of false negative detections complementary to the FPR and the FDR. The false negative rate (FNR) is the expected proportion of tests we failed to reject out of the n_{H_A} tests that should be rejected. The false nondiscovery rate (FNDR; Genovese and Wasserman 2004) is the expected proportion of tests we should have rejected out of all the n_{accept} tests we did not reject. While the FNR cannot be controlled a priori, it is often of concern to statisticians because it relates to the ‘‘power’’ of a test, where power = 1 − FNR is the probability of rejecting a test for which H_0 is false; the power measures the ability of a test to detect real changes. The FNDR cannot be controlled a priori either, but an upper bound can be estimated. We defer this topic to section 4 to avoid detracting from the main message here, which is that the FDR procedure controls a meaningful measure of false positive detections.

c. Illustration

Consider a simplified version of the example in section 1. Suppose that we are testing for a change at each of $n = 1000$ locations, based on regression slopes. We assume that all locations are independent and defer to section 3 for more realistic scenarios. Assume that the n_{H_0} locations that have experienced no change over time

[1] The upper bound is only approximate because it assumes that n_{reject} is fixed, whereas it is a random variable, since it is determined by the data.

have estimated slopes normally distributed with true mean $\mu_0 = 0$ and standard deviation $\sigma_0 = 1$, while the other $n_{H_A} = n - n_{H_0}$ locations have $\mu_A = 3$ and $\sigma_A = 3$. We used $\sigma_0 > 0$ because, even if true slopes are zero at the n_{H_0} locations with H_0 true, slopes estimated from data will never be exactly zero.

We perform traditional FPR, Bonferroni, and FDR procedures on such samples with $\alpha = 5\%$, the Bonferroni-corrected $\alpha = 1000^{-1} \times 5\%$, and $q = 5\%$, respectively, and for each testing procedure and each sample, we record the false discovery proportion, $\mathrm{FDP} = n_{\mathrm{FP}}/n_{\mathrm{reject}}$. Figure 3a shows the average of the FDP in 1000 samples, plotted against $a = n_{H_A}/n$, the proportion of true alternative hypotheses; note that the y axis reads ‘‘FDR’’ rather than ‘‘average FDP’’ because, since we used a large number of samples, the average FDP is close to its expected value, $\mathrm{FDR} = E(\mathrm{FDP})$, the false discovery rate.[2] The plot clearly shows that the FDR of the traditional procedure with $\alpha = 5\%$ is not controlled, and worse, is uncomfortably high when a is below 25%. In real situations, n_{H_A} is unknown, so we could unknowingly have a large proportion of the rejected hypotheses be false positive detections. On the other hand, the FDR of the FDR procedure is always below the nominal $q = 5\%$, so that, whatever the actual (unknown) numbers of null and alternative hypotheses, n_{H_0} and n_{H_A}, the procedure delivers on its promise that the proportion of erroneous rejections is on average less than q. Note that FDR falls further below q as a increases. We explain why in section 4 and show how to calculate a tighter upper bound for FDR.

Figure 3b shows the power of each procedure as functions of a; in the notation of Table 1, power $= (1 - \mathrm{FNR})$, with FNR well approximated by the average over the 1000 samples of the false negative proportion, $\mathrm{FNP} = n_{\mathrm{FN}}/n_{H_A}$. We see that the FDR testing procedure is consistently almost as powerful as the traditional FPR procedure, while also successfully limiting the proportion of false positive detections. The Bonferroni method’s consistently low FDR comes at the cost of low power; such a test is often said to be conservative because it fails to reject many alternative hypotheses.

This simple example illustrates that the FDR procedure combines the desirable features of the traditional and Bonferroni procedures without their flaws; it has high power while controlling the FDR to be below q, for whichever q is chosen.

3. FDR procedures for correlated tests

In section 2, we made the argument that controlling the FDR makes more sense than controlling the probability of type I errors (FPR); henceforth we focus only on FDR procedures.

In the example in section 1, we used the original FDR procedure of Benjamini and Hochberg (1995) because it is the simplest to apply. However, it makes the assumption that all n tests are independent, which rarely holds for climate data, for which measurements at nearby locations are often positively correlated. The more recent FDR procedures of Yekutieli and Benjamini (1999) and Benjamini and Yekutieli (2001) are designed to handle this complication. The various FDR procedures differ in their assumptions, as described in section 3a, but under their respective assumptions, all guarantee an FDR below the chosen value of q with probability close to one.

Returning to our example in section 1, the three FDR procedures with $q = 5\%$ detect 338 (Benjamini and Hochberg 1995), 376 (Yekutieli and Benjamini 1999), and 36 (Benjamini and Yekutieli 2001) significant locations, respectively; the results are so different that we felt compelled to investigate how the methods perform when applied to simulated data. We know of no other such study.

Section 3a describes briefly the new FDR procedures. Section 3b explains how datasets consistent with climate data were created, and section 3c reports the results of our simulation study. To summarize in advance, we find that the assumption-free FDR procedure of Benjamini and Yekutieli (2001) is too conservative to be useful. The procedure of Yekutieli and Benjamini (1999) is reasonably successful but requires lengthy computations, while the original straightforward procedure of Benjamini and Hochberg (1995) is fairly resistant to violations of the independence assumption and thus still gives good results for climate data.

a. FDR procedures and assumptions

Let S_{H_0} denote the set of p-values corresponding to locations where H_0 is true, and let S_{H_A} denote the set of all other p-values. Consider the following assumptions:

- Assumption 1: The elements of S_{H_0} are mutually independent.
- Assumption 2: The two sets S_{H_0} and S_{H_A} are independent from one another.

The original FDR procedure of Benjamini and Hochberg (1995) makes both these assumptions, which are clearly unreasonable for data that have spatial dependence. The procedure of Yekutieli and Benjamini (1999) makes assumption 2 only, and while it is likely to be violated by climatological data, the violation is restricted to the spatial boundaries between the two sets S_{H_0} and S_{H_A}. The FDR procedure of Benjamini and Yekutieli (2001) makes no assumptions at all, and therefore applies very generally. For simplicity of terminology, we denote the procedures by FDR-Indep (Benjamini and Hochberg 1995), FDR-Corr (Yekutieli and Benjamini 1999), and FDR-General (Benjamini and Yekutieli 2001).

[2] This result follows from the Weak Law of Large Numbers, which states that as the simulation size goes to infinity, the sample mean approaches the population mean (Casella and Berger 2002, p. 232).

To apply FDR-General is straightforward, we proceed just as with FDR-Indep described in section 2 but replace q with $q/(\Sigma_{i=1}^{n} i^{-1})$ in (2), where n is the total number of locations/tests. For Fig. 2, this translates into replacing the line with slope q with a line with shallower slope, $q/(\Sigma_{i=1}^{n} i^{-1})$. Note that although the only differences between FDR-General and FDR-Indep are the slope of the line used in Fig. 2 and the data assumptions they require, both procedures make the same claim—that FDR $\leq q$. However, for FDR-Indep, the claim is strictly valid only if the tests satisfy assumptions 1 and 2.

FDR-Corr also makes a correction to FDR-Indep, although it is much less immediately obvious. The properties of FDR-Indep rely on the fact that, when the locations are independent, the distribution of the p-values for which H_0 is true is uniform on [0, 1]. However, the uniform distribution no longer holds when locations are correlated, so that FDR control is no longer guaranteed. FDR-Corr involves obtaining p-values from data simulated under the collective null hypothesis (all locations have H_0 true), in much the same way that data type A is simulated in the next section, with the aim of calibrating q in the rejection rule (2), so that the actual FDR is as close as possible to the original nominal FDR q. This FDR method requires substantial computations, based on a collective null hypothesis model estimated from the data; the quality of the test is partly a function of the simulation size (the larger the better) and of how well the estimated collective null hypothesis model represents the hypothetical truth. An additional complication is that the calibration requires an estimate of the number of rejected hypotheses that have H_0 false, which is not easily done. Because FDR-Corr requires simulation and estimation steps, Yekutieli and Benjamini (1999) could only show that FDR control, FDR $\leq q$, is obtained with probability close to but less than one.

b. Simulated datasets

To assess the observed properties of the three FDR procedures, we conduct a simulation study based on three types of simulated datasets that were created specifically so that the strengths of their spatial correlations span most types that arise in climatology; data type A has no spatial dependence, data type B has spatial dependence consistent with the climate data of section 1 (Paciorek et al. 2002), and data type C has stronger spatial dependence. Before we describe how these datasets were simulated, we illustrate the spatial dependence structure in these data.

The correlation structure of a spatial field can be summarized by the spatial autocorrelation function (Cressie 1993), which shows how rapidly the mean correlation between pairs of locations decreases with the distance between the locations. It is obtained by first plotting the sample correlations between all possible pairs of locations against the distances between the locations, then by smoothing the plot; Fig. 4 shows the resulting smooths for the three simulated dataset types, as well as for the two real climate datasets described below.

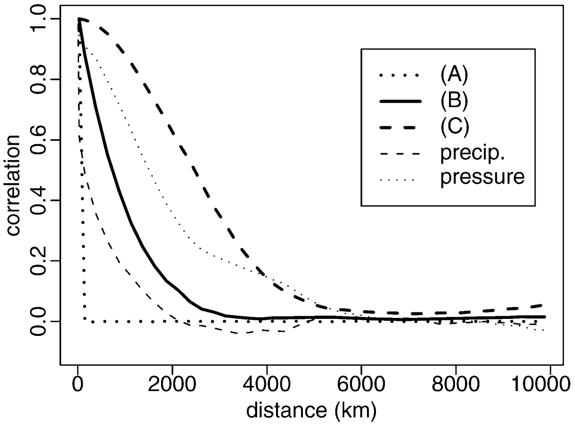

FIG. 4. Spatial autocorrelation for the simulated data types of section 3b. Data type A: no spatial correlation, data type B: actual correlation of data in section 1, and data type C: stronger correlation. Also included are the spatial autocorrelation of station-based monthly mean precipitation and sea level pressure from the GHCN.

For data A, the autocorrelation drops immediately to zero, consistent with the fact that data A have no spatial correlation. It decays more slowly for data C than for data B. For example, the correlation between two locations that are 1000 km apart is 0.5 for data B, whereas it is still about 0.9 for data C; indeed the distance must be greater than 2500 km for the correlation to be below 0.5 for data C. Also, locations that are more than 3500 km apart are almost completely independent for data B, whereas the correlation is still as high as 0.2 for data C.

For comparison with our simulated data, Fig. 4 also shows the spatial autocorrelation for two real climatological datasets, the Global Historical Climatology Network (GHCN) monthly mean station data from 1961–90 for precipitation and sea level pressure. We see that the spatial autocorrelations of the real data lie within the range of the three simulated datasets. For clarity, we show only the average correlation as a function of distance; in all of the datasets, there are large positive and negative individual correlations between specific locations at all distances shown in the plot. As one would expect, precipitation is less spatially correlated than sea level pressure. We expect that few climatological fields will be substantially more correlated than the pressure field seen here, since pressure fields are typically quite smooth spatially. These comparisons suggest that our simulated datasets cover the range of correlation scales likely to be encountered in climatological data.

Our simulated datasets are all based on the data of Paciorek et al. (2002), described in section 1, so that they are consistent with actual climate data. First, we detrend the time series at all $n = 3024$ locations, with fitted trends obtained from linear regressions of temperature variance against time. We are thus left with a spatially correlated set of n time series $Y_i(t)$, $i = 1, \ldots,$

TABLE 2. Summary of assumptions (defined in section 3a). A √ indicates (left side of table) that a particular FDR procedure requires the assumption under consideration or (right side) that a particular dataset satisfies it. An × indicates (left side) that a procedure does not require the assumption or (right side) that a dataset violates it. An ×× indicates a more serious violation of the assumption. BH95: Benjamini and Hochberg (1995), YB99: Yekutieli and Benjamini (1999), and BYO1: Benjamini and Yekutieli (2001).

	FDR procedure			Data type: Correlation is		
	FDR-Indep (BH95)	FDR-Corr (YB99)	FDR-General (BY01)	zero (A)	‘‘normal’’ (B)	strong (C)
Assumption 1	√	×	×	√	×	××
Assumption 2	√	√	×	√	×	××

n that do not contain any signal; here, i denotes location and t denotes time. Then, letting $\mathbf{t}^*$ denote a vector of 51 yr sampled at random and with replacement from $\mathbf{t} = \{1949, 1950, \ldots, 1999\}$, $Y_i(\mathbf{t}^*)$, $i = 1, \ldots, n$ is a simulated dataset that contains no signal but that has spatial correlation consistent with climate data; in statistical jargon, $Y_i(\mathbf{t}^*)$ is a bootstrap sample. Using the same $\mathbf{t}^*$ at all locations preserves the spatial correlation structure of the original data, yielding datasets of type B, while using different $\mathbf{t}^*$s destroys it, yielding datasets of type A. To create a highly spatially correlated dataset of type C, we start from a dataset of type B and induce stronger spatial dependence by replacing $Y_i(t_j^*)$ at each location i and year j by the average of $Y_k(t_j^*)$ for all locations k within 1913 km (0.3 radians) of i.

Note that so far we have shown how to create spatially independent or correlated time series that contain no signal. We then add nonzero linear trends β_i at each of n_{H_A} locations, giving simulated data,

$$X_i(t_j) = Y_i(t_j^*) + \beta_i t_j.$$

Correlated and uncorrelated datasets all use the same signal β_i; they differ only in their error structure $Y_i(\mathbf{t}^*)$. The trends β_i are chosen such that the signal-to-noise ratio is held constant across these n_{H_A} locations; that is, β_i is chosen to explain a prespecified proportion of variability (R^2) in the regression of $X_i(t)$ on t. Because the magnitude of the random noise differs across locations, had we added the same fixed trend $\beta_i = \beta$ at each of the n_{H_A} locations, it would have been more difficult to understand how the relative levels of signal and noise affect the comparison of the multiple testing procedures.

Finally, we let n_{H_A} vary between 0% and 99% (0, 1, 10, 25, 50, 75, 90, 99) of the total number of locations, $n = 3024$, and use three R^2 levels: 10%, 15%, and 20%. Comparisons between the FDR procedures for each data type at each value of n_{H_A} and R^2 are based on results averaged over 10 000 simulated datasets.

c. Properties of the FDR procedures—A simulation study

We study the properties of the three FDR procedures applied to our three types of simulated data. Table 2 summarizes the assumptions (specified in section 3a) required by each FDR procedure and which types of data satisfy them.

Our main results are summarized in Fig. 5, which has the same structure as Fig. 3. Figure 5a plots, for the three FDR procedures applied to the three simulated data types, the average over 10 000 simulated samples of the FDP versus $a = n_{H_A}/n$, the proportion of true alternative hypotheses; note that since the FDP is averaged over so many samples, it is indeed very close to its expected value, the FDR. Figure 5b shows the corresponding power = (1 − FNR) curves, although for the sake of clarity, we plot only the curves for data type B. Figure 5 shows our results only for $q = 5\%$ and signal-to-noise ratio, $R^2 = 10\%$. Other values of q produced qualitatively the same results. Similarly, other values of R^2 produced quantitatively the same results as in Fig. 5a and qualitatively the same results as in Fig. 5b; not surprisingly, the power increases as a function of the signal-to-noise ratio R^2 since true alternative hypotheses become easier to detect.

We comment on the three FDR procedures in sequence. Because FDR-Indep requires assumptions 1 and 2 (see Table 2), it is only safely applicable to data type A; indeed, our simulation results confirm the theoretical claim in (4) that FDR $\le q$ for all a (Fig. 5a, dotted ‘‘I’’ curve). We had observed this already in Fig. 3. Strictly speaking, FDR-Indep is inappropriate for datasets B and C, since they are spatially correlated, but it is interesting to assess its robustness to violations of the assumptions. The dashed and solid ‘‘I’’ curves in Fig. 5a suggest that the effect of applying FDR-Indep to increasingly correlated data is to further widen the discrepancy between the FDR and q. However, FDR remains below q for all a, so that even though FDR-Indep is theoretically inappropriate for correlated data, it still provides FDR control as stated in (4).

One may feel that an FDR as low as possible is desirable, since it reduces the number of false positive detections. However, a low FDR invariably entails low power, which produces larger numbers of false negative detections, as seen in Figs. 5b and 3b. We reiterate at this point that our aim is not to determine which procedure minimizes the FDR, since all testing procedures make a trade-off between the numbers of false negative and false positive detections, but rather which has FDR

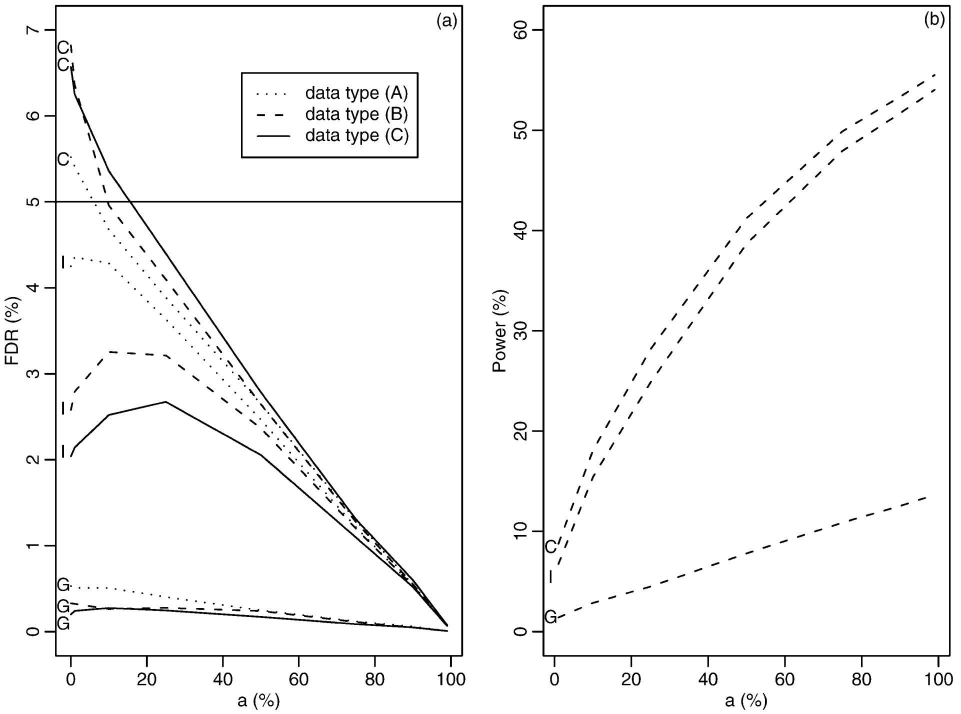

FIG. 5. (a) FDR and (b) power plotted against the proportion of alternative hypotheses, $a = n_{H_A}/n$, for the three FDR testing procedures applied to the three data types described in section 3b. Data types A, B, and C are distinguished with different plotting lines (dotted, dashed, and solid, respectively), and testing procedures are marked by a distinguishing letter at the start of each curve: "C" for FDR-Corr, "I" for FDR-Indep, and "G" for FDR-General. The horizontal line in (a) is at $q = 5\%$. For clarity, in (b) we show only the power for data type B; results for data types A and C are similar.

as close as possible to, and below, q; that is, we want predictable, tight FDR control. Therefore, FDR-Indep provides FDR control as stated by (4) even for positively spatially correlated data, although the tightness of the control degrades as spatial correlation increases, which in turn reduces the power of the test (not shown in Fig. 5 for clarity). In section 4, we explain how to improve FDR-Indep.

FDR-General (Benjamini and Yekutieli 2001) requires no assumptions, so that (4) should be satisfied for all datasets. And indeed, the corresponding FDR curves ("G" curves in Fig. 5a) are below q for all a, although they are too far below q to make FDR-General practically useful; the excessively low FDR entails very low power, as seen in Fig. 5b. We find that FDR-Indep is practically more useful than FDR-General, even though it requires assumptions that often are not met by the data.

The last FDR procedure considered is FDR-Corr (Yekutieli and Benjamini 1999), with simulation results shown as the "C" curves in Fig. 5. Although it requires fewer assumptions than FDR-Indep (Table 2), FDR-Corr is only strictly applicable to data type A, with the guarantee that (4) holds with probability close to, but less than, one. Note indeed that the FDR curve for data A is below q, except for small a, which is consistent with simulation results in Yekutieli and Benjamini (1999). We also see that the effect of applying FDR-Corr to increasingly correlated data is to increase the FDR, although FDR control as stated in (4) still holds for most a, even for very correlated data. This makes FDR-Corr a competitor to FDR-Indep, especially since it has slightly higher statistical power, as shown in Fig. 5b. However, we feel that the computational effort it requires is excessive. With $n = 3024$ locations and a simulation size of 1000, which we consider a bare minimum, FDR-Corr takes 97 CPU seconds to test the whole field, compared to 7×10^{-5} with FDR-Indep; FDR-Corr also requires programming a complicated algorithm, including generating p-values from the null hypothesis, which may be difficult to replicate and whose simulation differs between datasets.

To summarize, we find that of the two FDR procedures designed to handle dependent data, only FDR-Corr is worth considering. As with the Bonferroni method, FDR-General is too conservative to be useful. We also find that, although it is designed for independent data, FDR-Indep is robust to correlation in simulated data consistent with climate data and therefore can be applied fairly safely, with the considerable advantage that it requires little computational or pro-

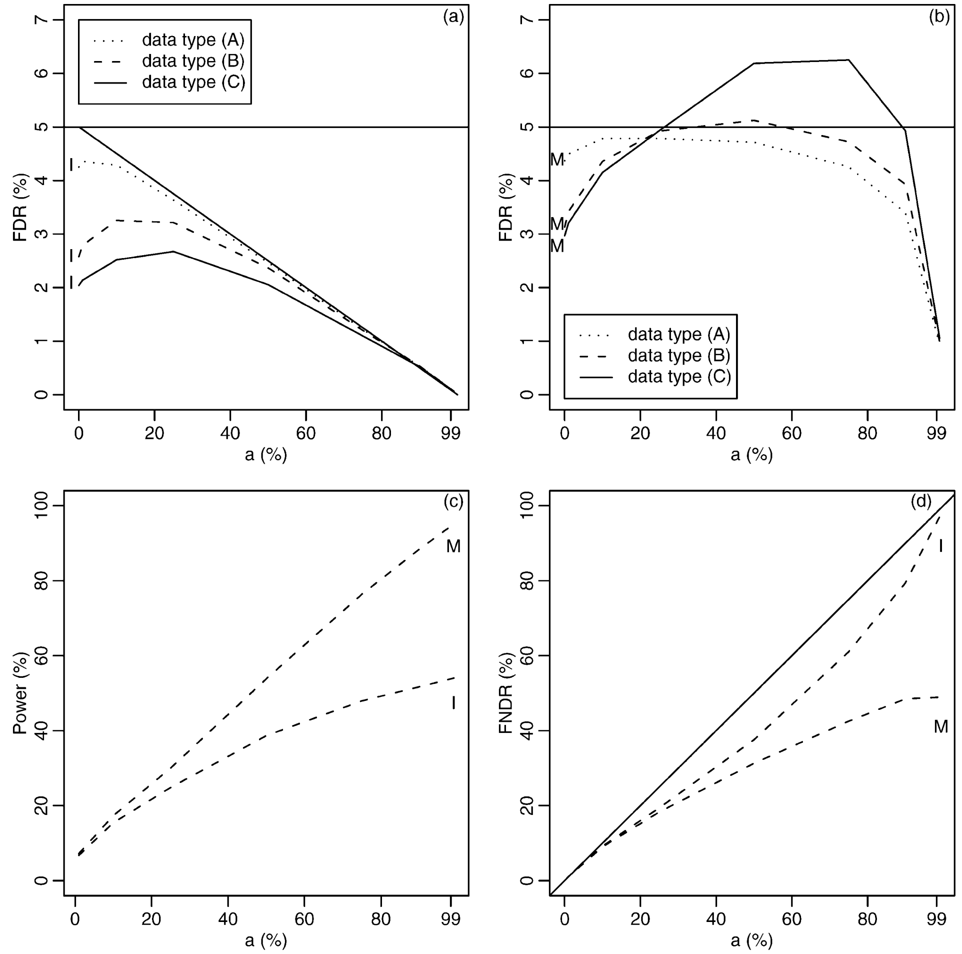

FIG. 6. FDR plotted against $a = n_{H_A}/n$ for (a) the original FDR-Indep procedure and (b) its modified version (section 4), both applied to the simulated data of section 3b. The horizontal lines are at $q = 5\%$; the additional solid line in (a) has slope $(1 - a) \times q$. (c) Power and (d) FNDR for the original and modified FDR procedures are plotted against $a = n_{H_A}/n$ only for data type B for clarity; results for data types A and C are qualitatively similar. The solid line in (d) is $y = x$, the upper bound in (7). The testing procedure is indicated by a distinguishing letter at the start of each curve: "I" for FDR-Indep and "M" for the modified FDR-Indep.

gramming effort. We therefore favor FDR-Indep for its simplicity.

4. A tighter FDR control

We argued in section 2 that, when multiple tests are performed, the FDR is a more relevant quantity to control than the FPR. Then, in section 3, we demonstrated by simulation that the original FDR-Indep procedure of Benjamini and Hochberg (1995) performs sufficiently well when applied to spatially correlated climate data. Here, we show how to tighten the FDR control of FDR-Indep to make it a more powerful procedure.

Letting $a = n_{H_A}/n$ be the unknown proportion of true alternative hypotheses, Genovese and Wasserman (2004) show that for FDR-Indep performed with nominal FDR q,

$$\text{FDR} \leq (1 - a)q. \tag{5}$$

Since $a \in [0, 1]$, then $(1 - a)q \leq q$, so that (5) provides an upper bound for FDR that is tighter than q. This explains why, in Figs. 3a and 5a, the FDR curves fall progressively further below q as a increases; indeed, because the discrepancy between q and $(1 - a)q$ increases with a, so does the discrepancy between q and FDR. Figure 6a is a partial reproduction of Fig. 5a; it shows the FDR curves of FDR-Indep, along with the line $(1 - a)q$, versus a. We see that all FDR curves remain not only below $(1 - a)q$, as expected from (5), but also remain very close to $(1 - a)q$ for all a. We would observe the same in Fig. 3a if we added the line $(1 - \mathrm{a})q$ versus a.

This suggests a way to improve FDR control: to ensure that the FDR is as close as possible to, yet below, a specific rate q', we should perform the FDR-Indep illustrated in Fig. 2 with q such that $q' = (1 - a)q$. We call this the modified FDR-Indep procedure; p-values are rejected according to (2) with

$$q = (1 - a)^{-1}q', \quad \text{which yields} \quad \text{FDR} \leq q'. \quad (6)$$

Tighter FDR control increases the power of the test. This is easily understood from Fig. 2; replacing the line with slope q by the line with steeper slope $(1 - a)^{-1}q$ (steeper since $a \geq 0$) potentially allows more p-values to be under that line and therefore rejected. For example, if half the n locations have H_0 false ($a = 0.5$), then performing FDR-Indep with $q = 10\%$ in Fig. 2, compared to the previous $q = 5\%$, still ensures that the proportion of false discoveries is on average below, yet close to, $q' = 5\%$.

Genovese and Wasserman (2004) also show that

$$\text{FNDR} \leq a, \quad (7)$$

where the FNDR was introduced in section 2 as the rate of tests we should have rejected out of the tests we accepted; FNDR $= E(n_{\text{FN}}/n_{\text{reject}})$ in the notation of Table 1. Unlike the FDR, the FNDR cannot be controlled a priori, so (7) is not useful other than for providing an approximate upper bound for the expected number of false negative errors,

$$E(n_{\text{FN}}) \leq an_{\text{accept}},$$

as illustrated in Fig. 6d.

The implications of (5) and, to some extent, of (7) are very attractive. The difficulty is that a is unknown and therefore must be estimated; this adds variability to the procedure and in turn may invalidate the inequalities in (5) and (7). In the rest of this section, we show how to estimate a, study how (5) and (7) hold on the simulated datasets of section 3, and finally return to our example discussed in section 1.

a. Estimating a

Storey (2002) and Genovese and Wasserman (2004) each propose estimators for a; the method we present here is a hybrid. Let $F_P(x)$ denote the cumulative distribution function (CDF) of the p-values used to perform the tests. For now, we develop the approach using the true CDF; we will later describe how to estimate this using the empirical CDF. Our estimate of a is

$$\hat{a} = I^{-1}\sum_{i=1}^{I}\max\left[0, \frac{F_P(x_i) - x_i}{1 - x_i}\right] \quad \text{with}$$

$$x_i = x_0 + (1 - x_0) \times \frac{(i - 1)}{I}, \quad (8)$$

where x_i takes I regularly spaced values between x_0 and 1; we routinely use $x_0 = 0.8$ and $I = 20$ for reasons justified below.

To make sense of (8), consider first the case $a = 0$; that is, all n locations have H_0 true. It is well known that when the tests are independent, the distribution of the p-values is uniform on [0, 1]; that is,

$$F_P(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x \in [0, 1] \\ 1 & \text{if } x \geq 1. \end{cases} \quad (9)$$

Thus, $F_P(x) - x = 0$ for all $x \in [0, 1]$, and since $x_i \in [0, 1]$ in (8), our estimate, $\hat{a} = 0$, matches the true $a = 0$ perfectly.

Next, consider the general case where $n_{H_A} = an$ locations have H_0 false and the remaining $n_{H_0} = (1 - a)n$ have H_0 true. Then the CDF of the p-values is a mixture distribution:

$$F_P(x) = aF_A(x) + (1 - a)F_0(x), \quad (10)$$

where $F_A(x)$ and $F_0(x)$ are the CDFs of the p-values in the two subpopulations where H_0 is, respectively, false and true. As before, $F_0(x)$ is uniform on [0, 1] when the tests are independent. On the other hand, $F_A(x)$ is completely unknown, other than that it has more mass toward zero than a uniform distribution since p-values that correspond to locations with H_0 false tend to be small. To simplify our argument, assume that all such p-values are smaller than some $x_0 < 1$, so that $F_A(x)$ has all of its mass between zero and x_0. Then $F_A(x) = 1$ for all $x \geq x_0$, so that for $x_i \geq x_0$, (10) reduces to $F_p(x_i) = a + (1 - a)F_0(x_i)$. This implies

$$\frac{F_P(x_i) - x_i}{1 - x_i} = a$$

in (8), yielding $\hat{a} = a$ exactly. More generally, there exists some sufficiently large $x_0 \in [0, 1]$, so that $F_A(x_0) = 1 - \epsilon$, where ϵ is a very small positive number; in this case (8) yields an estimate, $\hat{a}$, such that

$$a\left[1 - \frac{\epsilon}{I}\sum_{i=1}^{I}(1 - x_i)^{-1}\right] \leq \hat{a} \leq a. \quad (11)$$

That is, $\hat{a}$ underestimates a although it is very close to a when ϵ is very small, which happens when x_0 gets close to 1. This explains why we use a fairly large value, $x_0 = 0.8$, in (8).

Note that the potential downward bias of $\hat{a}$ has no adverse effect on the FDR procedure. Indeed, $\hat{a} \leq a$ implies $(1 - a)q \leq (1 - \hat{a})q$, so that FDR control as per (5) is preserved. The only consequence is that the test is slightly more conservative than it would have been had we known a.

The last difficulty is that $F_P(x)$ in (10) is unknown, since neither a nor $F_A(x)$ are known. We estimate $F_P(x)$ with the empirical density function (EDF) of the p-values in (1),

$$\hat{F}_P(x) = n^{-1}\sum_{i=1}^{n} I_{[0,x]}(p_i),$$

where $I_S(p_i) = 1$ if $p_i \in S$, and $I_S(p_i) = 0$ otherwise.

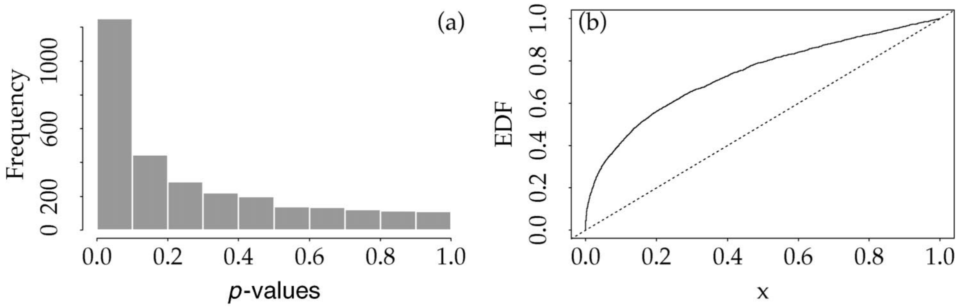

FIG. 7. (a) Histogram of the *p*-values in the example from section 1. (b) Corresponding EDF (solid line) is shown with the CDF of uniformly distributed *p*-values (dashed line) for comparison.

More simply, the EDF is the cumulative histogram of the *p*-values, with histogram bins taken to be the smallest distance between any two *p*-values. Figure 7 shows the histogram and corresponding EDF of the *p*-values in our example in section 1 and Fig. 1, from which, applying (8), we obtain $\hat{a} = 61.7\%$.

The use of $\hat{F}_P(x)$ in place of $F_P(x)$ in (8) explains why we take an average of $I = 20$ values to estimate a in (8); we hope to reduce the variability in $\hat{a}$ associated with estimating $F_P(x)$. Additionally, even though we have $F_P(x) \geq x$ theoretically, $\hat{F}_P(x) \geq x$ does not necessarily hold for all x since $\hat{F}_P(x)$ is variable; to ensure that $\hat{a} \geq 0$, we prevent negative contributions from the x_is for which $\hat{F}_p(x_i) < x_i$ in (8).

b. Properties of the modified FDR-Indep—A simulation study

The results in (5) and (7) are valid for FDR-Indep applied to spatially independent data and when a is known. We established in section 3 that the effect of applying FDR-Indep to spatially correlated data was to decrease the FDR, so that if a is known, then (5) also holds for correlated data. However, a must be estimated, and the method proposed in the last section is strictly valid only for independent tests. Indeed, we have found in the simulations that $\hat{a}$ overestimates a when the correlation is strong. Now, $\hat{a} \geq a$ implies $(1 - \hat{a})q \leq (1 - a)q$, which in turn implies that (5) no longer holds; that is, FDR control is not necessarily guaranteed for spatially correlated data. The purpose of this section is to study how badly FDR control as per (5) degrades as the spatial correlation in the data gets stronger. To do that, we applied the modified FDR-Indep (6) to the simulated datasets from section 3, with results in Fig. 6.

Figure 6a isolates the FDR curves of the original FDR-Indep procedure that were shown in Fig. 5a; Fig. 6b shows the FDR curves of the modified FDR-Indep (6) applied to the same data. Focusing first on A data, which are spatially independent so that $\hat{a}$ does not overestimate a, we see that the FDR is much closer to, yet remains below, $q' = 5\%$ for all a. That is, the modified FDR-Indep provides not only FDR control but tighter FDR control than the original FDR-Indep. Note, however, that the FDR curves drop well below q' for large a; the control degrades. This happens because the larger a is, the less the uniform distribution $F_0(x)$ contributes to (10), so that $\hat{a}$ underestimates a, as described in (11).

Figure 6b also shows that, for all a, the FDR curves for B and C data are much closer to q than the original FDR curves in Fig. 6a were to q, although they do sometimes exceed q; FDR $\leq q$ does not hold for all a, as we expected since $\hat{a}$ overestimates a for spatially correlated data. However, as argued in section 3, we constructed data type C with spatial correlation more extreme than most real climate data, so the breakdown of FDR control we see here is as extreme as one is likely to experience. For data B, which have correlation consistent with the temperature variance of section 1, FDR control hardly degrades at all. We therefore conclude that the modified FDR-Indep procedure provides tight FDR control for the vast majority of climate data.

Finally, Figs. 6c and 6d show the power (1 − FNR) and FNDR curves of FDR-Indep and its modified counterpart. For clarity, only the curves for data type B are plotted; the other curves were qualitatively similar. It is clear that the power and FNDR of the modified procedure are respectively larger and smaller than the power and FNDR of the original procedure; indeed, allowing more rejections entails a larger number of false positive detections n_{FP} and therefore a smaller number of false negative ones n_{FN}. The FNDR curves are below a, suggesting that (7) does hold, although a is not a tight upper bound; hence the upper bound on the expected number n_{FN} of false negative discoveries is not tight either.

To conclude, both the original and the modified FDR-Indep procedures control the FDR, but the latter has higher power and lower FNDR; it is therefore a better procedure. Additionally, it is fairly resistant to spatial correlation in the data.

c. Temperature variance example revisited

In our motivating example in section 1, we found significant increases in temperature variance over 51 yr at 941, 19, and 338 locations, using, respectively, the FPR ($\alpha = 5\%$), the Bonferroni ($\alpha = 3024^{-1} \times 5\%$), and the original FDR-Indep ($q = 5\%$) procedures; these

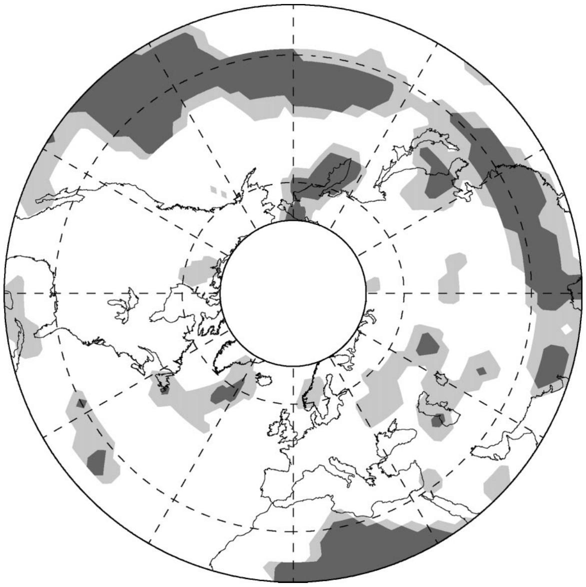

FIG. 8. Map of locations with significant trends in the temperature variance index, 1949–99, in the area 20°–70°N based on FDR-Indep in section 2 (dark gray), and its modified version of section 4 (light + dark gray), both with FDR ≤ 5%.

locations are shown in Fig. 1. We now apply the modified FDR-Indep.

Applying (8) with the estimate $\hat{F}_P(x)$ shown in Fig. 7b yields $\hat{a} = 61.7\%$, which indicates that approximately 61.7% of the locations have H_0 false. Based on Fig. 6, we trust that using $\hat{a}$ in place of the unknown a will yield reasonable results, since the temperature variance data have the same correlation structure as data type B. Such a large value of $\hat{a} = 61.7\%$ first surprised us because FDR-Indep had only rejected $n_{\text{reject}} = 338$ out of $n = 3024$ locations, barely more than 10%. This discrepancy can be explained. First, many locations that have H_0 false might not show enough change to reach statistical significance. Second, we know from (5) and Fig. 5 that when a is large, the FDR-Indep procedure does not provide good FDR control, in the sense that FDR is well below q. The modified procedure of this section will help tighten the FDR control greatly and give more power to detect locations with H_0 false.

Accordingly, we perform FDR-Indep once more, this time with $q = (1 - 0.617)^{-1} \times 5\%$ to ensure that FDR $\leq 5\%$ as per (5). We now find that $n_{\text{reject}} = 807$ locations show a significant change in temperature variance over the last 50 yr, as indicated in Fig. 8. Moreover, (5) and (7) suggest that the number n_{FP} of false positive discoveries is approximately less than $807 \times 5\% = 41$, while the number n_{FN} of locations we failed to detect is less than $(3024 - 807) \times 61.7\% = 1368$.

This last result suggests that we fail to detect a considerable number of locations with H_0 false. This seems more alarming than it really is. There are three reasons; the first two are specific to FDR procedures, and the last concerns detection of significance at large. First, n_{FN} is probably much less than 1368, because the upper bound for FNDR in (7) is not particularly tight for the modified FDR-Indep procedure, as observed in Fig. 6d. We know of no theoretical result that would provide a tighter upper bound for FNDR. Moreover, our data is spatially correlated, so that $\hat{a}$ likely overestimates a, the true proportion of alternative hypotheses, which further inflates our already loose upper bound for n_{FN}. Last, many locations that may have H_0 false do not show a strong enough signal to be declared significant, so that from a statistical (or any other) point of view, these locations are not distinguishable from the locations that have H_0 true.

5. Conclusions

We started this paper by comparing and contrasting the traditional FPR and the FDR procedures, which control the proportion of false positive discoveries in different ways. We then argued in section 2 that the FDR procedure controls a more useful criterion: the rate of falsely rejected hypotheses out of all rejected hypotheses. In section 3, we investigated the robustness of the original FDR procedure of Benjamini and Hochberg (1995; which we denote FDR-Indep) to spatial correlation in the data and found that it is not only robust, but that its performance is just about as good as that of other FDR procedures that were later developed to handle correlated data, and it is much simpler to apply than its direct competitor FDR-Corr (Yekutieli and Benjamini 1999). In section 4, we presented an improvement for FDR-Indep, recently developed by Genovese and Wasserman (2004), which gives FDR-Indep tighter control of the FDR and thus increases its power to detect significant changes. This requires that the unknown proportion of true alternative hypotheses be estimated. The method we propose overestimates this proportion when high spatial correlation is present in the data, but the effect on FDR control is minimal. Supplementary material consisting of sample R, S-Plis, and Matlab code for implementing the original and modified FDR-Indep procedures is available online at http://dx.doi.org/JCLI3199.s1. This material is also available at the first author's home page (available online at http://www.stat.cmu.edu/~vventura/ClimFDR.html), and in the S and Matlab software archives at StatLib (available online at http://lib.stat.cmu.edu).

In conclusion, we have summarized the current state-of-the-art techniques in multiple testing and have demonstrated the properties of the FDR method based on simulated data that are consistent with climate data. Based on these studies, we believe that the FDR-Indep procedure and its modified version are powerful testing procedures that provide effective FDR control.

Acknowledgments. We are grateful to Larry Wasserman and Richard Rosen for helpful advice and comments.

REFERENCES

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.

——, and D. Yekutieli, 2001: The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.,* **29,** 1165–1188.

Casella, G., and R. Berger, 2002: *Statistical Inference.* Duxbury Press, 660 pp.

Cressie, N., 1993: *Statistics for Spatial Data.* Wiley-Interscience, 900 pp.

Genovese, C., and L. Wasserman, 2004: A stochostic process approach to false discovery rates. *Ann. Stat.,* **32,** 1035–1061.

Katz, R. W., 2002: Sir Gilbert Walker and a connection between El Niño and statistics. *Stat. Sci.,* **17,** 97–112.

Livezey, R., and W. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

Paciorek, C., J. Risbey, V. Ventura, and R. Rosen, 2002: Multiple indices of Northern Hemisphere cyclone activity, winters 1949–1999. *J. Climate,* **15,** 1573–1590.

Storey, J., 2002: A direct approach to false discovery rates. *J. Roy. Stat. Soc.,* **64B,** 479–498.

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

Wilks, D., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

——, 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate,* **10,** 65–83.

Yekutieli, D., and Y. Benjamini, 1999: Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan. Inf.,* **82,** 171–196.