



# Accounting for multiple testing in the analysis of spatio-temporal environmental data

José Cortés<sup>1</sup> · Miguel Mahecha<sup>2</sup> · Markus Reichstein<sup>2</sup> · Alexander Brenning<sup>1</sup>

Received: 30 August 2019 / Revised: 5 April 2020 / Published online: 12 May 2020  
© The Author(s) 2020

## Abstract

The statistical analysis of environmental data from remote sensing and Earth system simulations often entails the analysis of gridded spatio-temporal data, with a hypothesis test being performed for each grid cell. When the whole image or a set of grid cells are analyzed for a global effect, the problem of multiple testing arises. When no global effect is present, we expect  $\alpha\%$  of all grid cells to be false positives, and spatially autocorrelated data can give rise to clustered spurious rejections that can be misleading in an analysis of spatial patterns. In this work, we review standard solutions for the multiple testing problem and apply them to spatio-temporal environmental data. These solutions are independent of the test statistic, and any test statistic can be used (e.g., tests for trends or change points in time series). Additionally, we introduce permutation methods and show that they have more statistical power. Real-world data are used to provide examples of the analysis, and the performance of each method is assessed in a simulation study. Unlike other simulation studies, our study compares the statistical power of the presented methods in a comprehensive simulation study. In conclusion, we present several statistically rigorous methods for analyzing spatio-temporal environmental data and controlling the false positives. These methods allow the use of any test statistic in a wide range of applications in environmental sciences and remote sensing.

**Keywords** Gridded data · Multiple testing · Nonparametric statistics · Permutation methods · Spatial patterns · Statistical inference

---

Handling Editor: Pierre Dutilleul.

---

✉ José Cortés  
jose.cortes@uni-jena.de

<sup>1</sup> Department of Geography, Friedrich Schiller University, Jena, Germany

<sup>2</sup> Max Planck Institute for Biogeochemistry, Jena, Germany

## 1 Introduction

A common strategy in analyzing gridded spatio-temporal data derived from remote sensing or Earth system models is to fit a statistical model at each grid cell (Julien and Sobrino 2009; Fensholt and Proud 2012; Beck and Goetz 2012; Eckert et al. 2015; Zhang et al. 2017). The statistical model or test employed depends on the researcher's study target. For example, a correlation test, a two-sample  $t$ -test, a trend test, or a linear or even nonlinear model could be the most appropriate. Each of these tests produces a  $p$ -value for each grid cell, and the  $p$ -values can be plotted over the entire study area to create a statistical image. When attempting to analyze this image to assess its collective significance (e.g., to identify significant patterns or an overall effect), we incur the multiple testing problem.

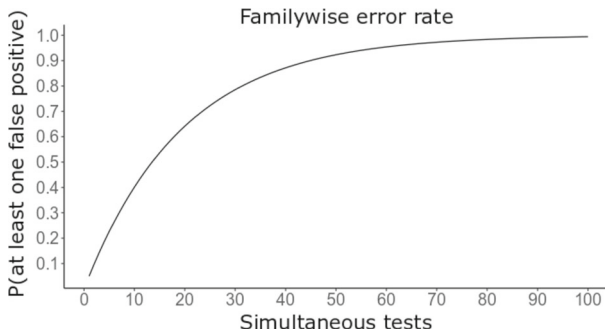
This problem, which results in uncontrolled false-positive test results and consequent false scientific “discoveries”, has received relatively little attention in the environmental sciences and remote sensing, except for a small but growing number of climate science reports (Ventura et al. 2004; Wilks 2006a, b). Therefore, our objective in this study is to raise awareness of this issue. We outline state-of-the-art solutions, including permutation methods, and we demonstrate their potential in real-world applications. We additionally conduct a simulation study.

More specifically, we explore two permutation methods that address the multiple testing problem in the context of trend detection in a comprehensive simulation study. Previous simulation studies have focused on evaluating the Familywise error rate of Bonferroni and related methods, random field theory methods, and permutation methods (Nichols and Hayasaka 2003). Such studies have also compared the Familywise error rate (FWER) and the statistical power of the false discovery rate (FDR) with the case of no correction (Wilks 2016) or with Bonferroni and related methods (Ventura et al. 2004; Wilks 2006a). A permutation method based on clustering has been introduced in neuroimaging (Nichols and Holmes 2002), but it has not yet been evaluated in a simulation study. A necessary step is to compare—in a single study—permutation methods with Bonferroni and related methods, both in terms of their Familywise error rate and their statistical power. As an additional point, we know of no other study that evaluates the performance of the Mann–Kendall trend test in the context of multiple testing.

The paper is structured as follows. In Sect. 2 we present the general conceptual background, and we continue to explain the methods and notation in detail in Sect. 3. In Sect. 4, a simulation study is conducted to assess the validity of the methods and to compare their performance. The methods are then applied to two real-world datasets. Finally, in Sect. 5, we conclude with some remarks and discussion about the methods and the interpretations that can be obtained from each of them, and we suggest possible additional applications.

## 2 Background

In a single statistical hypothesis test, a result is declared to be significant if the test indicates that the observed data are unlikely given that the null hypothesis is true.



**Fig. 1** Probability of at least one false positive test (FWER) as a function of the number of independent simultaneous tests performed ( $\alpha = 0.05$ ), i.e.  $P = 1 - (1 - \alpha)^n$

When multiple tests are performed (e.g., at the grid cell level), as is common in many environmental science studies, the probability of obtaining a significant result by chance (false positive) greatly increases. This probability of at least one false positive among a “family” of tests is called the Familywise error rate (FWER). As with any other statistical test, we wish to constrain the FWER to a desired  $\alpha$  level, which is usually 0.05, although suitable  $\alpha$  levels are problem specific. Figure 1 illustrates the FWER as a function of the number of multiple tests being performed—when performing as few as 100 tests, we are almost guaranteed to have at least one false positive (>99%), when the individual tests are independent. Indeed, the possibility exists that the majority of discoveries or significant results are false positives (Wilks 2016). Further, due to spatial autocorrelation, these false positives can also cluster together, giving the analyst a false impression of a coherent pattern. Both of these drawbacks are further illustrated in the simulation study (Sect. 4.2).

Two common strategies are used to control the FWER. One strategy is to set a new threshold of significance that takes into account the number of tests performed (e.g., Bonferroni, Hochberg); we refer to these methods as Bonferroni-related methods. The other strategy is to set the threshold of significance using the sampling distribution of the maximum statistic (i.e., the  $100 \cdot (1 - \alpha)$ th percentile). Only methods that have strong control of the FWER allow making inferences on specific hypothesis tests (T. Nichols and Hayasaka 2003); for this reason, we focus mainly on such methods.

This work introduces methods that use the distribution of the maximum statistic and compares them to Bonferroni-related methods in the context of geospatial environmental data. Performance is assessed with a simulation study, and two real-world datasets are used for illustration. Methods that use the distribution of the maximum statistic have been applied in other disciplines, most notably in neuroimaging (Nichols and Holmes 2002; Nichols and Hayasaka 2003) and genetics (Dudoit et al. 2003).

**Table 1** Classification of grid cells

	Retain null hypothesis	Reject null hypothesis
Null hypothesis true	$G_{0 0}$	$G_{1 0}$
Null hypothesis false	$G_{0 1}$	$G_{1 1}$

### 3 Methods

#### 3.1 Notation

Let  $H = \{H_i, i = 1, 2, \dots, M\}$  be the hypotheses at each grid cell, where  $H_i = 0$  indicates that the null hypothesis is true, and  $H_i = 1$  indicates that the alternative hypothesis is true. We retain and reject  $H_i$  with  $\hat{H}_i = 0$  and  $\hat{H}_i = 1$ , respectively, depending on the outcome of the corresponding test.  $H_{\text{global}}$  denotes the global null hypothesis; that is, all null hypotheses are true. The test statistic corresponding to each hypothesis is denoted by  $T_i$ , and the image of test statistics is  $T$ . Let  $P = p_i$  be the corresponding p-values. Note that the number of tests,  $M$ , equals the number of grid cells, since we test one hypothesis per grid cell. We indicate the ordered p-values using the standard notation  $p_{(i)}$ , where  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(M)}$ . Finally, we denote the significance level for the test statistic at each grid cell by  $\alpha_{\text{local}}$ , and the significance level for the global null hypothesis by  $\alpha_{\text{global}}$ ; rejecting the global null hypothesis allows the researcher to declare what is referred to as global significance or field significance.

#### 3.2 Types of error

For a statistical test, we use  $\alpha$  to denote the type I error, that is, the probability of rejecting the null hypothesis (no effect) when it is true (false positive). The type II error,  $\beta$ , is the probability of not rejecting the null when we should have (false negative). The statistical power of a test is  $1 - \beta$ . In a single statistical test, we control the probability of a false-positive decision at the specified  $\alpha$  level. The classification of grid cells in multiple testing is summarized in Table 1. In this context,  $G$  represents the number of grid cells, and the subscript is the decision (fail to reject:  $\hat{H}_i = 0$ , reject:  $\hat{H}_i = 1$ ) given the truth ( $H_i = 0$ ,  $H_i = 1$ ); for example,  $G_{0|1}$  is the number of grid cells for which we retained a false null hypothesis.

In multiple testing we often wish to control the probability of observing at least one false positive among the whole family of tests,  $P(G_{1|0} > 0)$ ; this is the FWER. A relatively new approach to dealing with multiple testing is the control of the FDR (Benjamini and Hochberg 1995), defined as  $E(G_{1|0}/G_{1\cdot})$ . The FDR is the expected value of the proportion of false positives among all rejected null hypotheses, or “discoveries.” Thus, a small FDR ensures that our discoveries are reliable, without constraining the probability of making at least one false discovery. FDR control is therefore expected to have higher within-image power than FWE methods, which is why it is sometimes preferred.

Importantly, there are two types of control of the FWER: weak and strong. Both allow testing for field significance, but weak control only allows rejecting the global null. Weak control does not enable rejecting individual grid cells’ null hypotheses, which would be required to pinpoint specific significant subregions. In contrast, strong control enables rejecting individual grid cells’ null hypotheses, but at the cost of being a more conservative test and thus having weaker power.

The comparison of weak and strong control is analogous to a comparison of more than two means in an ANOVA. The null hypothesis is that all means are equal, and we use the  $F$  statistic to quantify this global hypothesis. If the statistic is extreme enough (or its p-value lower than the specified  $\alpha$ ), we conclude that the means are different, but we do not know which specific means are different. This situation is equivalent to weak control. To determine which pairs of means are statistically different, several strategies exist, including the Bonferroni correction, Tukey’s HSD, Scheffé’s method, and so forth. After applying any of these post hoc tests, we can determine which means are not equal, which is equivalent to having strong control.

Formally, weak control is defined as

$$P\left(\bigcup_{i \in G} T_i \geq u \mid H_{global} = 0\right) \leq \alpha_{global} \tag{1}$$

where  $T_i = \{1, \dots, M\}$ ; the set  $G$  is the whole study area composed of  $M$  grid cells; and  $u$  is the threshold of significance. To achieve strong control, the false positives must be controlled in any region  $G_0 \subset G$  where the null hypothesis is true:

$$P\left(\bigcup_{i \in G_0} T_i \geq u\right) \leq \alpha_{global} \quad \forall G_0 \subset G \tag{2}$$

In other words, detecting significance in a region should not affect the results of other regions when strong control is achieved (Nichols and Hayasaka 2003). As an example, in a global-scale analysis, detecting significance in a continent will not affect the control of false positives in other continents.

### 3.3 Global tests for significance

Testing for a global effect is also known as testing for “field significance”. In this testing, the objective is to assess if an effect is present anywhere in the region or study area; in other words, we want to reject the global null hypothesis  $H_{global}$ . Testing for a global effect is a consequence of testing a hypothesis in each grid cell and controlling the FWER. If a single grid cell is found to be significant (after accounting for the multiplicity of tests), we can reject  $H_{global}$  and conclude that an overall effect exists in the study area. Depending on the type of FWER control (weak or strong), the null hypotheses of individual grid cells can be rejected. We divide the methods for global significance into two categories: Bonferroni-related methods and methods based on the maximum distribution.

### 3.3.1 Bonferroni-related methods

Perhaps the most popular method for multiple testing corrections is the Bonferroni correction, which establishes the new threshold of significance by dividing  $\alpha$  by  $M$ , the total number of tests (grid cells). All p-values below the threshold are declared significant. This method is very conservative; significance thresholds are often too low to detect any significance, especially with large datasets arising from remote sensing.

An alternative method based on the minimum p-value is the Walker method, whose performance has been evaluated in the context of environmental applications (Wilks 2006a). If the global null hypothesis is true, then the p-values follow a uniform distribution on  $[0,1]$ , and observing a p-value close to 0 would indicate a violation of the global null hypothesis. The Walker test uses the distribution of the minimum p-value, which is known to follow a beta distribution (Wilks 2006b), to establish the significance threshold.

An improvement is to let the threshold  $u$  in Eqs. (1) and (2) vary, replacing it with  $u_i$ , as long as the overall FWER is controlled. This leads to step-up and step-down methods, where the ordered p-values,  $p_{(i)}$ , are sequentially compared with the corresponding threshold,  $v_i$ , using the inequality

$$p_{(i)} < v_i \quad (3)$$

where  $v_i$  is the p-value threshold corresponding to the respective threshold in terms of the test statistic,  $u_i$ . In step-up methods, we start the largest p-value,  $p_{(M)}$ , and compare it with  $v_M$ . We then compare the second largest value,  $p_{(M-1)}$ , to  $v_{(M-1)}$ , and so on. The first  $p_{(i)}$  that satisfies the inequality in Eq. (3) is declared significant, as well as all the smaller p-values. In step-down methods, we start with the minimum p-value,  $p_{(1)}$  and compare it with  $v_1$ . We then compare  $p_{(2)}$  with  $v_2$ , and so on. The first p-value that does not satisfy the inequality in Eq. (3) is declared nonsignificant, and all the p-values below are declared significant. The idea behind this approach is to let the threshold adapt to the signal in the data.

We evaluate the performance of step-up and step-down procedures with two common methods: Holm (step-up) and Hochberg (step-down). In theory, both methods have more statistical power than the Bonferroni method, but with thousands of tests, there is little difference in performance (Nichols and Hayasaka 2003; Dudoit et al. 2003).

The final two Bonferroni-related methods evaluated control the FDR. The method introduced by Benjamini and Hochberg (1995), hereafter referred to as BH, controls the expected number of false discoveries among all discoveries. Benjamini and Yekutieli (2001) modify the BH method, which controls the FDR under positive autocorrelation (or independence) among test statistics, to account for other cases of dependency among the test statistics; we refer to their method as the BY method.

The methods and their respective thresholds are compared in Table 2. Note that, for testing field significance, the threshold  $v_1$  is the same for Bonferroni, Holm, Hochberg, and FDR. Hence they are expected to have similar FWER control. Walker and BY methods also offer similar power; take, for example, the analysis of 10,000 grid cells at a significance level of  $\alpha_{global} = 0.05$ . Bonferroni, Holm, Hochberg, and FDR have

**Table 2** Thresholds for Bonferroni related methods

Method	$v_i$	Control of FWER
Bonferroni	$\alpha/M$	Strong
Walker	$1 - (1 - \alpha)^{1/M}$	Strong
Hochberg (step-down)	$\alpha(1/(M - i + 1))$	Strong
Holm (step-up)	$\alpha(1/(M - i + 1))$	Strong
BH (step-up)	$\alpha(i/M)$	Weak
BY (step-up)	$\alpha(i/M \cdot c), c = \sum_{j=1}^m 1/j$	Weak

$v_1 = \frac{0.05}{10000} = 5 \cdot 10^{-6}$ , while Walker has  $v_1 = 1 - (1 - 0.05)^{\frac{1}{1000}} = 5.13 \cdot 10^{-6}$  and BY has  $v_1 = \frac{0.05}{10000 \cdot \sum 1/i} = 5.11 \cdot 10^{-7}$ . The probability of these methods reaching differing conclusions is extremely low.

### 3.3.2 Maximum distribution

The maximum statistic allows us to control for the FWER (Nichols and Hayasaka 2003). At least one grid cell in an image will be declared significant if and only if the maximum statistic,  $\max T$ , exceeds the threshold  $u$ . If we choose  $u$  to be the  $100 \cdot (1 - \alpha_{global})$ th percentile of the distribution of the maximum statistic, written as  $u = F_{\max T | H_{global}=0}^{-1}(1 - \alpha_{global})$ , we have

$$\begin{aligned}
 P\left(\bigcup_i T_i \geq u | H_{global} = 0\right) &= P(\max T \geq u | H_{global} = 0) \\
 &= 1 - F_{\max T | H_{global}=0}(u) \\
 &= \alpha
 \end{aligned}
 \tag{4}$$

Strong control is achieved if the null distribution of any subset does not depend on the other null hypothesis. This circumstance is called subset pivotality, and it is satisfied if no logical constraints exist between grid cells; that is, any combination of significant/nonsignificant grid cells is possible (Nichols and Hayasaka 2003). It is also referred to as satisfying the free combination condition. An example of this condition not being satisfied is the following comparison of three means,  $\mu_1, \mu_2$ , and  $\mu_3$ . If  $\mu_1 \neq \mu_2$ , then  $\mu_1 = \mu_3$  and  $\mu_2 = \mu_3$  cannot both be true, and so the free combination condition is not satisfied (Bretz et al. 2011). In the case of a test statistic being applied at each grid cell, this condition is satisfied—a test statistic  $T_i$  at the  $i$ th grid cell has no impact on any other test statistic.

Significance is determined by  $u$ , as described above. Any test statistic (grid cell) whose absolute value exceeds the threshold  $u$  is declared significant. Note that  $u$  does not depend on the choice of test statistic; we can use any valid test statistic. In the following section, we describe two test statistics based on the maximum distribution, which is derived via permutation methods.

**Table 3** Example of the permutation approach for a single time series

Data	Mann–Kendall’s S
Original data: 1, 2, 3, 4, 5	10
Permutation 1: 2, 4, 1, 5, 3	2
Permutation 2: 5, 2, 4, 3, 1	− 6
...	...
Permutation k: 3, 5, 1, 2, 4	0

### 3.4 Permutation methods

#### 3.4.1 Background

Permutation tests are nonparametric tests. As opposed to their parametric counterparts, nonparametric tests do not assume a distribution of the test statistic under the null hypothesis; instead, it is derived empirically. This approach is especially useful because deriving the null distribution theoretically requires many assumptions that are hard to meet, as is the case with the maximum statistic distribution. The following is a brief overview of how permutation methods work in this context. Please refer to Nichols and Holmes (2002) for an in-depth treatment of permutation methods and their relation to the maximum distribution of a test statistic.

In a permutation test, the test statistic under all possible rearrangements of the data is calculated. For this calculation, we assume exchangeability; that is, the distribution of the statistic does not change when we change the ordering/labeling of the data under the null hypothesis. For a time series, the ordering/labeling are the time points. When the existence of a trend is being tested, the null hypothesis is that there is no trend. Under this null hypothesis, we assume that the observations are random and could have come from any time point. Thus, our data are exchangeable and we can proceed with a permutation test.

Consider a time series of five observations. If all the values are greater than their previous value, we would be inclined to conclude that the data show an upward trend. Similar to a parametric statistical test, a permutation test allows us to test if the observed data are unlikely given that the null hypothesis is true. Let the values for the time series be 1, 2, 3, 4, and 5, and suppose we want to test for the existence of a trend. We compute Mann–Kendall’s S statistic (as defined in Sect. 4.1) for each permutation, as shown in Table 3.

All the test statistics obtained from the permutations form our distribution of the test statistic under the null hypothesis of no trend. The p-value is given by the proportion of test statistics that are greater than or equal to the observed test statistic. If the total number of possible rearrangements is too large, a subsample is enough (Dwass 1957; Edgington 1969). In this case, there are  $5! = 120$  rearrangements, and the significance for a two-sided test at  $\alpha = 0.05$  is indicated by our test statistic being among the three largest/smallest test statistics (since  $120 \cdot 0.05/2 = 3$ ). In the example, the critical values are  $-8$  and  $8$ , and our observed test statistic is the largest test statistic, and it is significant with a p-value of  $2 \cdot (1/120) = 0.02$ .



### 3.4.2 Controlling the Familywise error rate

The above procedure is a permutation test for a single time series. To control the FWER, we permute the entire image simultaneously and record the maximum statistic (among all the grid cells) for each permutation. Permuting whole images conserves the spatial autocorrelation present in the data. The resulting set of test statistics form the maximum distribution of the test statistic under the null hypothesis, of which we use the  $100 \cdot (1 - \alpha_{global})$ th percentile to establish the significance threshold.

Although the spatial autocorrelation is accounted for by permuting images as a whole, accounting for temporal autocorrelation may still be needed. Temporal autocorrelation violates the exchangeability condition necessary to perform permutations, and ignoring it can lead to false-positive rates (per grid cell) as high as 30% (Yue et al. 2002). To account for temporal autocorrelation, we apply the correction proposed by von Storch (1999): for each grid cell, we calculate the temporal autocorrelation  $\hat{r}$  at lag-1 and replace the original time-series  $x_t$  with the series  $Y_t = x_t - \hat{r}x_{t-1}$ .

Note that while permutation p-values can be obtained as described above, we only use permutation methods to establish the threshold of significance. The p-values used in this paper are derived from the z-statistic obtained from the Mann–Kendall trend test, as described in Sect. 4.1. We use this approach because Bonferroni and related methods have very conservative thresholds of significance, and we would need to perform an enormous number of permutations for p-values to achieve significance. For example, in a dataset of (only) 10,000 grid cells, the Bonferroni threshold at  $\alpha = .05$  is  $v = \frac{0.05}{10,000} = 5 \cdot 10^{-6}$ . With 1000 permutations, the minimum p-value that can be achieved is  $\frac{1}{1000} = 1 \cdot 10^{-3}$ , which is several orders of magnitude larger than required for statistical significance.

The motivation for using the maximum statistic comes from random field theory (RFT), which was first used for statistical analysis in the neuroimaging community (Worsley et al. 1992). An in-depth treatment of RFT can be found in Petersson et al. (1999) and Cao and Worsley (2001) and a more concise one in Nichols and Hayasaka (2003). Essentially, RFT is used to approximate the distribution of the maximum statistic, which is then used to establish a significance threshold that controls the FWER. The drawback of the RFT approach is that it makes assumptions on the image to be analyzed. Among other things, it assumes the data are a realization of a stationary multivariate Gaussian distribution with a known degree of smoothness. In neuroimaging, as well as in environmental sciences, meeting these conditions is difficult, which is why permutation methods are appealing.

We perform two permutation tests in this study: one for the maximum distribution of the test statistic (hereafter called  $\max T$ ), and one for the maximum distribution of the supra-threshold cluster size (STCS) of the test statistic. The STCS is the number of significant grid cells that are adjacent (contiguous first-order queen neighbors). For both methods, we permute whole images simultaneously and recalculate the test statistic at each grid cell. For  $\max T$ , we record the maximum test statistic among all grid cells; for STCS, we record the size of the largest cluster of significant grid cells. We then repeat these steps  $N$  times. These steps are summarized in Table 4. The

**Table 4** Steps for permutation tests

Maximum statistic distribution (max T)	Supra-threshold cluster size (STCS)
Permute images	Permute images
Calculate test statistic at each grid cell	Calculate test statistic at each grid cell
Keep the maximum statistic among all grid cells	Keep the size of the largest cluster of significant grid cells
Repeat N times	Repeat N times

distribution of the maximum statistic is formed from the statistics recorded in each permutation.

Both tests allow us to control the FWER at the desired  $\alpha_{global}$  by setting the threshold of significance,  $u$ , to the  $100 \cdot (1 - \alpha_{global})$ th percentile of their respective maximum distributions. For maxT, any grid cell whose absolute value of the test statistic exceeds  $u$  is declared significant. For STCS,  $u$  is in terms of cluster size, so any cluster larger than  $u$  is declared significant.

The analysis is done entirely in the R programming language (R Core Team 2019). For the STCS method, we identify the clusters with the *osc* package in R (Kriewald et al. 2019). The algorithm selects a random starting point among significant grid cells and checks the neighbors for significance. If significant, it adds them to the cluster and iterates in this manner until no more significant neighbors remain. It then repeats this process until all significant grid cells are assigned a cluster or a maximum of 3 times the number of columns (default setting). We have adapted the algorithm so that it distinguishes between significantly positive and significantly negative grid cells.

## 4 Simulation study and real-world examples

In this section we analyze spatial time series of two widely used environmental datasets that play an important role in the assessment of climate change and its impacts on ecosystems: the GIMMS Normalized Difference Vegetation Index (NDVI) dataset 3rd generation version 1 (Pinzon and Tucker 2014; Tucker et al. 2005) and the NASA GISS Surface Temperature Analysis (GISTEMP) version 5 (GISTEMP Team 2019; Hansen et al. 2010). The NDVI data are available from NASA's Ecological Forecasting Lab repository, <https://ecocast.arc.nasa.gov/data/pub/gimms/>. The GISTEMP data are available from NASA's Goddard Institute for Space Studies repository, <https://data.giss.nasa.gov/gistemp/>.

The simulation study was designed to mimic such real-world situations under controlled conditions with known presence (and magnitude) or absence of trends. It is thus instrumental in evaluating the performance of the Bonferroni-related and permutation methods in terms of control of the FWER, global test power, and percentage of correctly detected trends, considering different rates of change. We use the Mann–Kendall trend test in the real-world and the simulation studies; an overview of the methods analysed is presented in Table 5.

**Table 5** Overall summary of all analyzed methods

Methods	FWER control	Inference on	Observed $\alpha_{\text{global}}$	References
Bonferroni	Strong	Individual grid cells and study area	Conservative	Bonferroni (1936)
Walker	Strong	Individual grid cells and study area	Conservative	Walker (1914)
Hochberg	Strong	Individual grid cells and study area	Conservative	Hochberg (1988)
Holm	Strong	Individual grid cells and study area	Conservative	Holm (1979)
Benjamini-Yekutieli	Weak	Study area	Most conservative	Benjamini and Yekutieli (2001)
Benjamini-Hochberg	Weak	Study area	Conservative	Benjamini and Hochberg (1995)
Maximum statistic	Strong	Individual grid cells and study area	Nominal level	Nichols and Hayasaka (2003) Dudoit et al. (2003)
Supra threshold cluster size	Strong	Clusters and study area	Nominal level	Nichols and Holmes (2002)

#### 4.1 Trend test

Although we could use any test statistic, we focus on the Mann–Kendall’s (MK)  $S$  statistic, which is often used to determine significance of trends obtained with the Theil–Sen estimator. Many reports present a map showing grid cells with a significant trend, but no test for field significance or correction for multiple testing is carried out (Julien and Sobrino 2009; Fensholt and Proud 2012; Beck and Goetz 2012; Eckert et al. 2015; Zhang et al. 2017). Without such correction, the results that are shown may be spurious and there may be no true trend or spatial pattern.

A benefit of using the MK  $S$  statistic is that it can be transformed to a Z-score. When deriving the distribution of the maximum statistic over a set of test statistics, we want them to have a common null distribution so that no single test statistic dominates the maximum distribution (Nichols and Hayasaka 2003). For example, when means are compared, using mean difference as a test statistic (instead of a t-statistic) will result in grid cells with a larger range of values dominating the distribution of the maximum test statistic. Although the FWER would still be controlled, potential significant test statistics could be masked, leading to a loss of power.

Autocorrelation is well known to influence the test statistic; positive autocorrelation inflates the type I errors, while negative autocorrelation makes the test conservative (Yue et al. 2002). Any conclusions drawn from test statistics that fail to control the type I error should be viewed with caution. Before applying the MK trend test, we apply von Storch’s correction, which controls the type I error at the specified  $\alpha$  level

and has only slightly inflated false-positive rates for strong temporal autocorrelation (von Storch 1999).

We stress that the methods here can be applied to any test statistic that produces a p-value, but, for simplicity, we focus only on the MK statistic, calculated by

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sign}(x_j - x_i) \quad (5)$$

where the sign function can be 1, 0, or  $-1$  if the term between the parentheses is positive, zero, or negative, respectively. The variance can be calculated with

$$\text{var}(S) = \frac{n(n-1)(2n+5)}{18}. \quad (6)$$

We can convert this test statistic to a standard normal distribution by

$$Z = \begin{cases} \frac{S-1}{\sqrt{\text{var}(S)}} & \text{for } S > 0 \\ 0 & \text{for } S = 0 \\ \frac{S+1}{\sqrt{\text{var}(S)}} & \text{for } S < 0 \end{cases} \quad (7)$$

All tests performed are two sided, and p-values are calculated by

$$p = 2(1 - \phi(|Z|)) \quad (8)$$

where  $\phi$  is the cumulative distribution function of the standard normal distribution.

## 4.2 Simulation study

### 4.2.1 Setup

Analyses of global gridded datasets commonly include thousands or even millions of grid cells. In this simulation study we generate 1000 realizations of Gaussian random fields on a  $100 \times 100$  grid with different levels of spatial autocorrelation. The spatial autocorrelation is determined by  $r(d) = \exp(-cd^2)$ , where  $d$  represents the Euclidean distance between the grid cell midpoints, and  $c$  is chosen so that  $r(1) = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ . Each realization consists of 34 time steps, where each grid cell has a standard normal distribution ( $\mu = 0, \sigma^2 = 1$ ), and each time step has the specified spatial autocorrelation. This arbitrary square grid is chosen to mimic situations encountered in real-world applications, and it serves as a compromise between limiting computational complexity and avoiding edge effects. Nevertheless, it roughly corresponds to Earth system model results at a  $2^\circ \times 2^\circ$  resolution (assuming global coverage), which contains up to 16,200 grid cells. For comparison, in the GISTEMP dataset we analyze 14,295 grid cells. We do not simulate temporal autocorrelation because it is addressed separately by using an appropriate testing procedure (Wilks 2016). For all scenarios, the permutation methods are based

on 1000 permutations. The proportion of images with at least one significant grid cell is recorded, and this is the FWER, since no trend was added.

Two other correlation functions have been used. An exponential correlation function  $r(d) = \exp(-c|d|)$  with the values for  $c$  chosen as above, and a Matérn correlation function  $r(d; \phi, \kappa) = (2^{k-1} \Gamma(\kappa))^{-1} (d/\phi)^k K_\kappa(d/\phi)$ , where  $\phi$  and  $\kappa$  are nonnegative parameters of the covariance,  $\Gamma(\cdot)$  is the Gamma function, and  $K_\kappa(\cdot)$  is the modified Bessel function of the third kind, of order  $\kappa$ . The range  $\phi$  is fixed to be 1, and the smoothness  $\kappa$  is allowed to vary, with  $\kappa = .5, 1, 2, 3, 4, 5, 6, 7, 8$  representing spatial correlation ranging from low to high. These results can be found in Appendix A.

The magnitude of the trend is unknown in real-world studies. To assess the ability to detect a trend, we simulate correlated Gaussian fields as above, and we induce a varying amount of trend, ranging from 0.001 to 0.1 per time step. This approach covers a wide range of trend magnitudes, from “hardly appreciable” on a visual basis to “impossible to overlook.” The trend is added only to a square in the center; for each of the trend magnitudes, we also vary the size of the trend-affected region: we use a  $5 \times 5$  square, a  $20 \times 20$  square, and a  $50 \times 50$  square, which correspond to 0.25, 4, and 25 percent of all grid cells, respectively. The proportion of images that are identified with field significance (at least one significant grid cell) is recorded, which is the global test power.

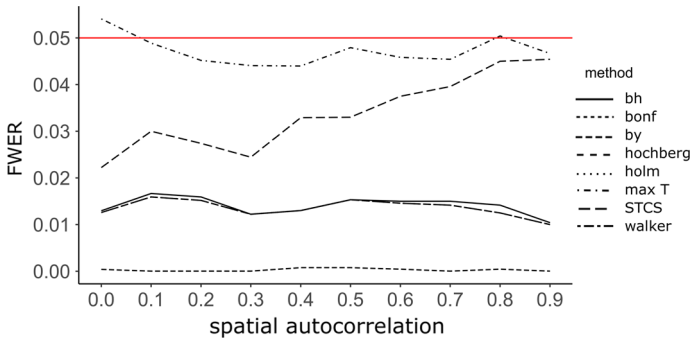
Besides global test power, the researcher may also be interested in how much of the signal is identified. For each simulation in which a trend is induced, we calculate the proportion of correctly identified grid cells in the trend-affected region, which is the within-image test power.

#### 4.2.2 Results

All methods control the FWER, although only the permutation method based on the maximum test statistic effectively controls the FWER at the desired nominal  $\alpha_{global}$  level. The STCS method is slightly conservative at low spatial autocorrelation (FWER between 0.02 and 0.03), but it approaches the nominal level as the autocorrelation becomes stronger. Other methods are very conservative—their FWER is much lower than the nominal level. Bonferroni and related methods achieve a FWER of 0.01, regardless of the strength of spatial autocorrelation. The BY method stands out for being the most conservative, with an observed FWER of 0.0002 across all levels of spatial autocorrelation.

With regard to the global test power, the permutation method and the clustering method consistently have better global test power than the other methods. Among the permutation methods, the STCS is consistently better, with an exception occurring when two conditions are met: a small trend-induced region ( $5 \times 5$  square or 0.25% of all grid cells) and strong spatial autocorrelation ( $\geq 0.8$ ). The BH method offers global test power comparable to the  $\max T$  method, but its power is always slightly lower. Among the Bonferroni-related methods, it offers the highest power.

As a test for field significance, both permutation approaches outperform the others. The clustering procedure only fails in the case mentioned before. For all other cases, the clustering method proves to be the best option to test for field significance, followed



**Fig. 2** Achieved FWER of the multiple testing corrections as a function of spatial autocorrelation

closely by the permutation method. The BH correction follows closely, with the rest of the Bonferroni-type corrections achieving the same type I and type II errors. As expected, the BY correction is last because it is highly conservative.

In terms of within-image test power, the STCS and BH are the best performing methods. The STCS has increased within-image test power for all cases except the one discussed above: small trend-induced region (0.25% of all grid cells) and strong spatial autocorrelation ( $\geq 0.8$ ). The second-best Bonferroni-related method is the BY method, which outperforms the max $T$  method in all cases except those with a small trend-induced area (0.25% of all grid cells). The other methods have consistently low within-image test power—even for the best-case scenario of no correlation and large trend-induced area (Fig. 4g), the methods detect only around 30% of the signal. We can always detect 100% of the signal by declaring all grid cells as significant. Therefore, it is of interest to know the proportion of true negatives that are correctly identified. For all methods, this proportion is  $\sim 100\%$  (not shown).

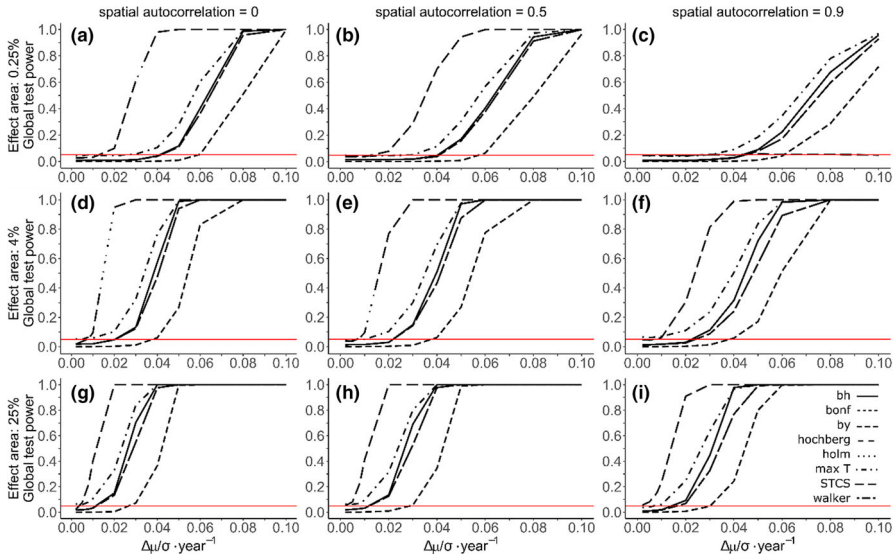
Because Bonferroni-related methods (Bonferroni, Hochberg, Holm, Walker) are unlikely to differ in terms of FWER, global test power, and within-image test power, their individual lines cannot be observed (Figs. 2, 3, 4) because they overlap for most of the simulation scenarios.

No observable difference appears in the results when we change the correlation function. All methods control the FWER regardless of the correlation structure of the data, with the individual methods performing similarly across all correlation functions. In terms of global test power and within-image test power, we also observe similar performance across all three correlation functions.

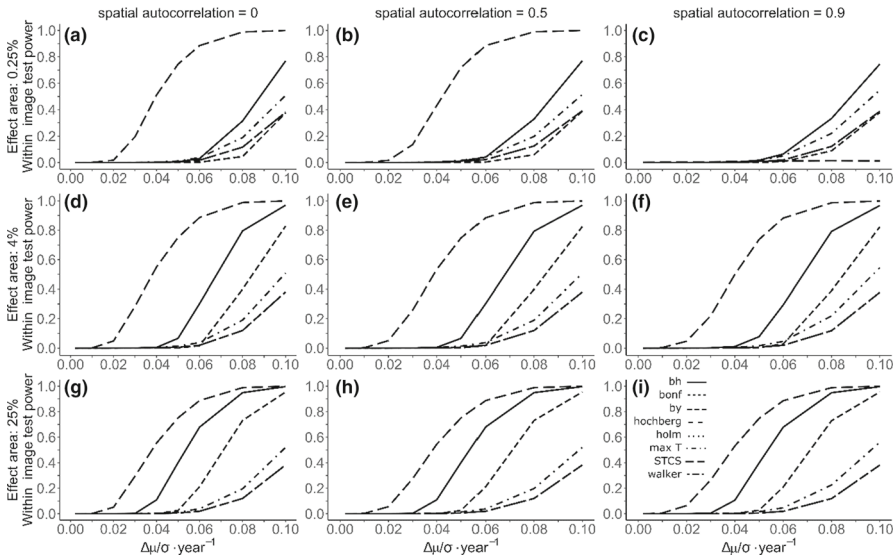
## 4.3 Real-world examples

### 4.3.1 Data

We analyze NDVI trends from the GIMMS dataset for the years 1982 to 2015 globally. The spatial resolution is  $1/12^\circ \times 1/12^\circ$  and the values are aggregated yearly to obtain a time series of 34 years.



**Fig. 3** Global test power for the case of no spatial autocorrelation, moderate spatial autocorrelation (0.5) and strong spatial autocorrelation (0.9) as a function of trend magnitude. The red line indicates the  $\alpha_{\text{global}} = 0.05$  level. The effect area is the percentage of all grid cells where a trend was induced



**Fig. 4** Within-image test power for the case of strong no spatial autocorrelation, moderate spatial autocorrelation (0.5) and strong spatial autocorrelation (0.9) as a function of trend magnitude. The effect area is the percentage of all grid cells where a trend was induced

Long-term temperature trends are analyzed from the NASA GISTEMP product globally for the years 1951 to 2018. The spatial resolution is  $2^\circ \times 2^\circ$  and the values are aggregated yearly to obtain a time series of 68 years.

**Table 6** Results of FWE methods applied to data sets,  $\alpha_{global} = 0.05$ 

Data	# of grid cells	Unadjusted	Bonferroni	Walker	Holm	Hochberg
NDVI	9331,200	217,585	0	0	0	0
GISTEMP	14,295	9152	1	1	1	1
Data	# of grid cells	Unadjusted	BH	BY	max T	STCS
NDVI	9,331,200	217,585	0	0	22	0
GISTEMP	14,295	9152	5962	0	97	8076

Shown are number of significant grid cells identified by each method

### 4.3.2 Results

All methods control the FWER at a significance level of  $\alpha_{global} = 0.05$ , and the permutation distributions are derived from 5000 permutations. In the GIMMS NDVI data, only the permutation method,  $maxT$ , detected a trend in 22 grid cells (not shown), scattered across northeast Africa and southwest Yemen. The detected grid cells showed increasing monotonic trends.

In the GISTEMP temperature data, almost all methods detect significant grid cells with increasing temperature trends, indicating global warming. As expected, Bonferroni, Walker, Holm, and Hochberg all are rather conservative, detecting only a single grid cell (the same grid cell, located in southeast Angola). The STCS, BH, and  $maxT$  methods detect much more significant trends, with 8076, 5962, and 97 grid cells, respectively (Table 6). The BY method does not detect any trend. None of the methods detect significant decreasing trends in any location.

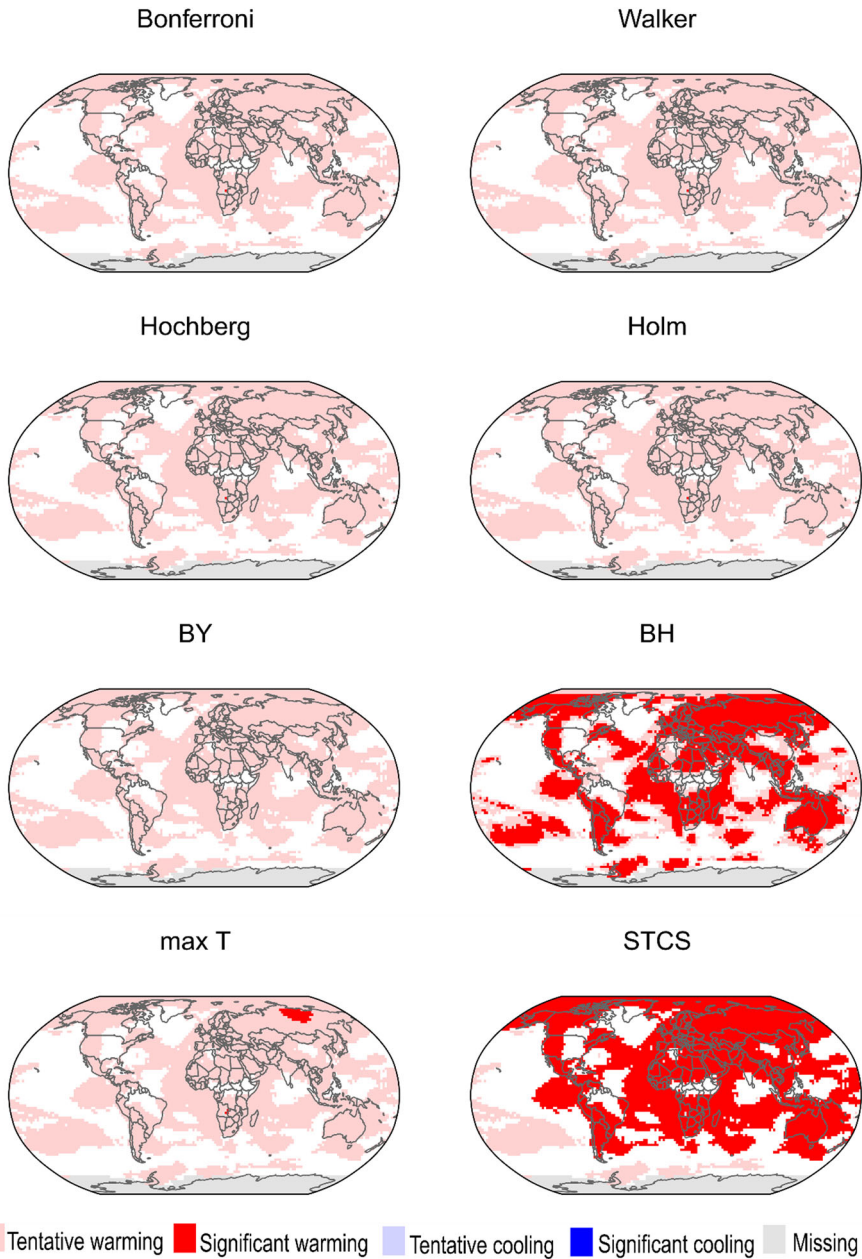
The BH method detects significant grid cells around the globe, identifying significant warming trends around the world. The  $maxT$  method detects fewer grid cells, but it shows a distinguishable spatial pattern of warming in Siberia and a smaller hotspot in southeast Angola (Fig. 5).

Not accounting for the multiplicity can lead to an incorrect conclusion of field significance or misinterpretation of a spatial pattern. Both datasets exhibit large number of grid cells with a significant trend when there is no correction for multiple testing. When multiplicity is not corrected for, the MK test flags only  $\sim 2\%$  of grid cells as significant in the GIMMS NDVI data, while for the GISTEMP temperature data  $\sim 64\%$  of the grid cells are flagged as significant. After correction for multiple testing, the GIMMS NDVI data show no clear spatial pattern and most methods detect no trends. The GISTEMP data retain the overall spatial patterns (STCS and BH methods) and identify smaller regions of interest ( $maxT$  method), while most methods indicate field significance.

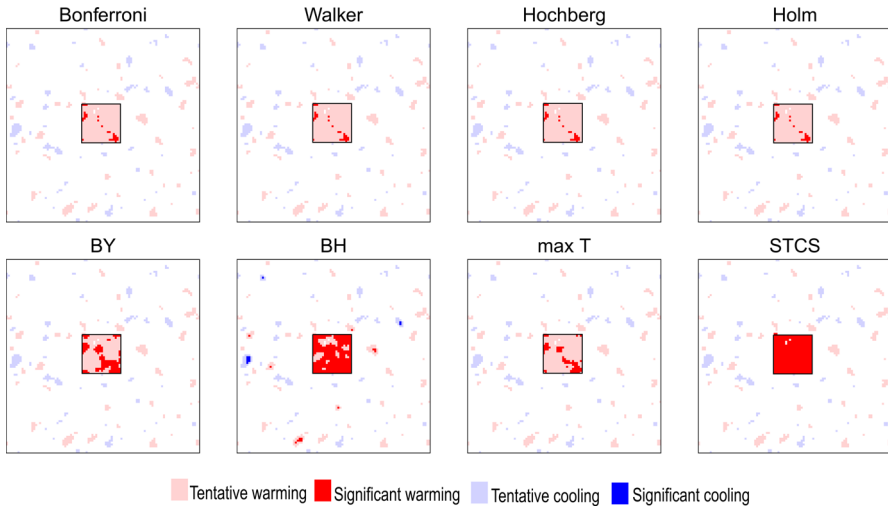
## 5 Discussion

We have compared different strategies to account for the multiple testing issue that arises in the environmental sciences. Previously used strategies, referred to as





**Fig. 5** Results of applying the MK test for the GISTEMP data with different corrections for multiple testing. White grid cells indicate no significance ( $p > 0.05$ ), light red and light blue indicate unadjusted significance ( $p < 0.05$ ) for increasing and decreasing trends, respectively, and bright red and bright blue indicate significance for increasing and decreasing trends, respectively, after correcting for multiple testing with the specified correction at a significance level of  $\alpha_{\text{global}} = 0.05$



**Fig. 6** Results of applying the MK test for the simulated data with different corrections for multiple testing. White grid cells indicate no significance ( $p > 0.05$ ), light red and light blue indicate unadjusted significance ( $p < 0.05$ ) for increasing and decreasing trends, respectively, and bright red and bright blue indicate significance for increasing and decreasing trends, respectively, after correcting for multiple testing with the specified correction at a significance level of  $\alpha_{\text{global}} = 0.05$ . The simulated scenario has spatial autocorrelation of 0.9, an effect area of 4%, and an induced trend ( $\Delta\mu/\sigma$ ) of 0.08 per time step, which corresponds to Fig. 4f. The  $20 \times 20$  black square in the middle highlights the area of the induced trend

Bonferroni-related methods, were compared with two recent permutation alternatives that have been successfully applied in other fields (Nichols and Holmes 2002; Nichols and Hayasaka 2003; Dudoit et al. 2003). All methods were applied to two real-world data sets and were evaluated in a simulation study, in terms of their achieved FWER, global test power and within-image test power.

Permutation methods take into account the spatial autocorrelation in the data because it is captured by the maximum statistic. This is not the case with Bonferroni-related methods. Although Bonferroni-related methods are robust to the spatial autocorrelation (Fig. 2; Wilks 2006a, b), it comes at the price of the tests being conservative, which affects their ability to detect a signal. We observe that Bonferroni-related methods achieve a FWER well below the nominal level (Fig. 2), which affects their global test power and within-image test power (Figs. 3, 4, 6). This outcome can also be seen in the results from real-world datasets, in which Bonferroni-related methods—excluding the Benjamini and Hochberg (1995) method—lead to fewer grid cells being declared as significant relative to permutation methods. Similar results from real-world datasets are observed in Nichols and Holmes (2002), Nichols and Hayasaka (2003), and Dudoit et al. (2003). Although developments have occurred in the FDR methodology in the spatial setting (e.g., Risser et al. 2019; Shen et al. 2002; Sun et al. 2015; Ventura et al. 2004), we focus on the original FDR procedure for two reasons: (1) it is the most commonly used in the environmental sciences, and (2) we choose to focus on methods with strong control of the FWER.

The permutation methods introduced present a favorable alternative for addressing the multiple testing issue in environmental sciences. Specifically, the clustering

method controls the FWER and outperforms all other methods in both global test power and within-image test power in all but one scenario—when there is a small area with an effect (0.25%) and strong spatial correlation ( $\geq 0.8$ ). Other studies in which permutation methods are used to derive the distribution of the maximum statistic have also found increased global test power in simulations (Dudoit et al. 2003) and increased within-image test power in real-world datasets (Nichols and Holmes 2002; Nichols and Hayasaka 2003; Dudoit et al. 2003).

The case in which the clustering method fails is not unexpected. The clustering method only takes into account the size of the cluster, not the individual p-values; therefore, with stronger spatial autocorrelation, the clusters that appear randomly become larger than the size of the cluster with the induced trend. This situation makes the trend-induced cluster invisible to the STCS method. In general, this outcome highlights a possible ambiguity of trend detection and spatial random variability.

In terms of the within-image test power, Benjamini and Hochberg's (1995) method identifies more signal than the  $\max T$  method (Figs. 4, 6), but this benefit comes at a cost of more false positives and no localizing power—individual grid cells cannot be declared significant. The  $\max T$  method identifies fewer grid cells, but it allows individual grid cells to be declared significant. This characteristic can be of critical importance in geospatial data analysis.

Not correcting for multiple testing would almost always lead to the conclusion of field significance and possibly the misinterpretation of spurious spatial patterns. By performing corrections for multiple testing, we ensure that the analysis is done in a statistically rigorous way and enhance the reliability and reproducibility of the results.

The type of Familywise error control—weak or strong—determines the statistical conclusions that may be drawn from the analysis. Only methods with strong control of the FWER (Bonferroni, Walker, Holm, Hochberg, and the permutation method based on the maximum statistic) allow us to make inferences on individual grid cells. The permutation method based on clustering allows us to make inferences on regions (clusters). Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) methods, in contrast, do not allow for inferences on specific regions or grid cells.

Permutation methods and other resampling methods have a wide range of applications. In the environmental sciences, these methods have been used to detect a change in temperature trends (Zang et al. 2019), identify memory effects in time series (Kraft et al. 2019), and assess habitat selection (Fattorini et al. 2014). Many parametric methods have a permutation analog. Since the maximum distribution of a test statistic can be derived for any permutation method, controlling the FWER with the methods presented in this paper is straightforward.

Nevertheless, limitations exist for the use of permutation methods. There are cases in which a permutation procedure is not straightforward or simply not possible. For example, if the test statistic is invariant to permutations (e.g., a one-sample  $t$ -test), we would be unable to derive the distribution of the maximum statistic because permuting the data will not change the test statistic. Another consideration is that permutation methods are computationally expensive, although with modern computing power this issue should not be limiting. When such issues arise, Benjamini and Hochberg (1995) method is a viable alternative for analyzing spatial patterns.

In this paper we focus on methods that apply a correction for multiple testing to the test statistics obtained from a model being fit at each grid cell. An alternative approach to field significance is to fit a spatio-temporal model and test a global null hypothesis of whether all regression coefficients vanish. For more on this approach, the reader is referred to DelSole and Yang (2011).

We introduce permutation methods to account for multiple testing in the context of environmental data, and we show their advantages. Although many methods to account for multiple testing have been developed since Livezey and Chen (1983) introduced their method to the environmental sciences, they do not account for the spatial autocorrelation that often occurs in environmental data. The permutation methods introduced here capture the spatial autocorrelation in the maximum statistics' distribution. By accounting for this autocorrelation, we improve upon previous methods: Permutation tests have higher global test power than all other methods compared here, including controlling the FDR with Benjamini and Hochberg (1995) procedure.

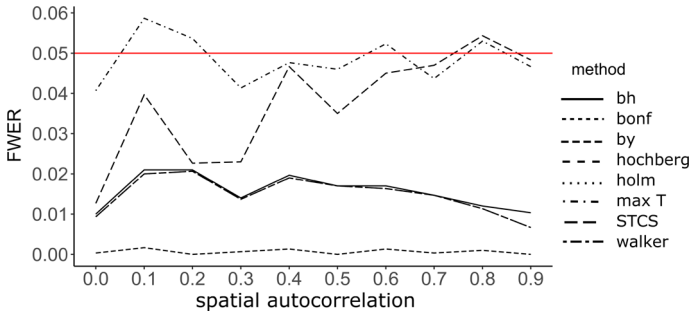
The clustering method introduced here consistently outperforms all other methods in terms of global test power and within-image test power. For analyzing spatial patterns, controlling the FDR remains a powerful tool. However, it comes at the cost of no localizing power—we cannot conclude statistical significance of single grid cells or specific regions. This situation is where permutation methods prove useful because they allow making inferences on specific grid cells or regions, and they can identify more pixels than other commonly used Bonferroni and related methods.

**Acknowledgements** Open Access funding provided by Projekt DEAL. JC is supported by the Free State of Thuringia through a graduate student scholarship and acknowledges support by the International Max Planck Research School for Global Biogeochemical Cycles (IMPRS-gBGC).

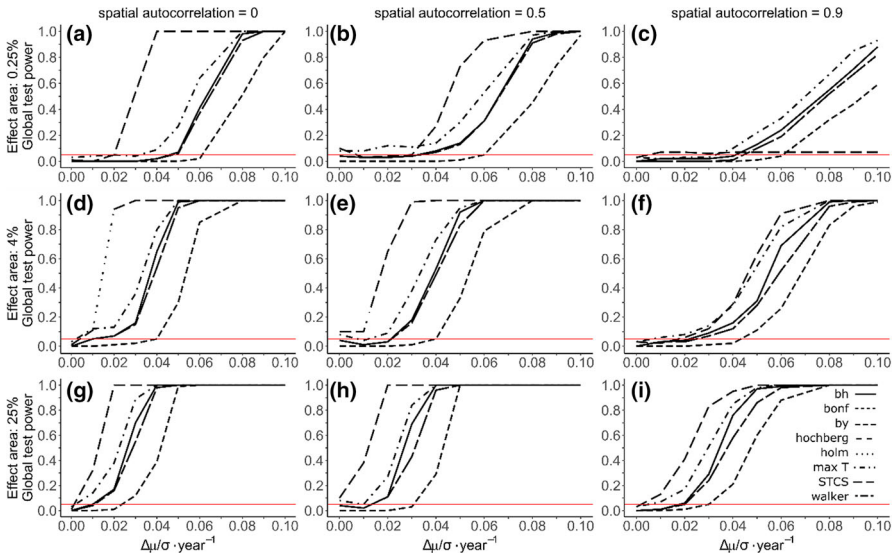
**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Appendix

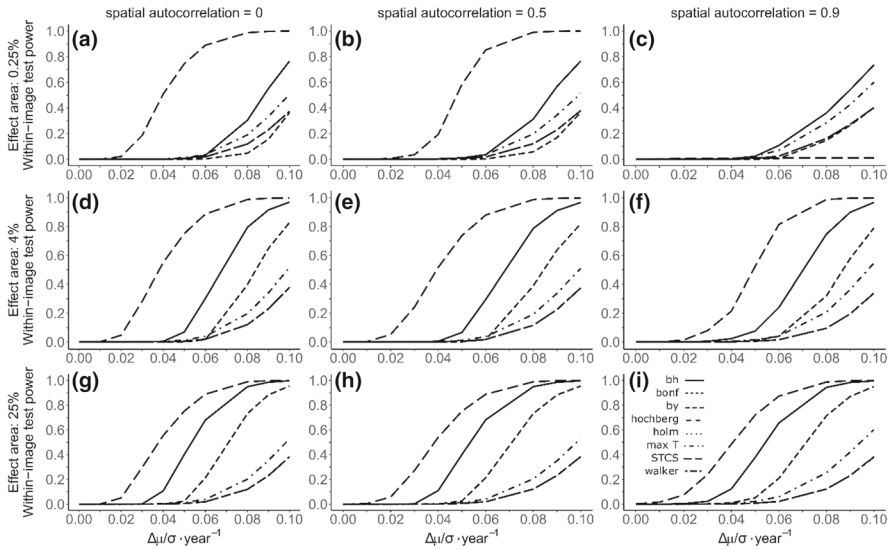
See Figs. 7, 8, 9, 10, 11 and 12.



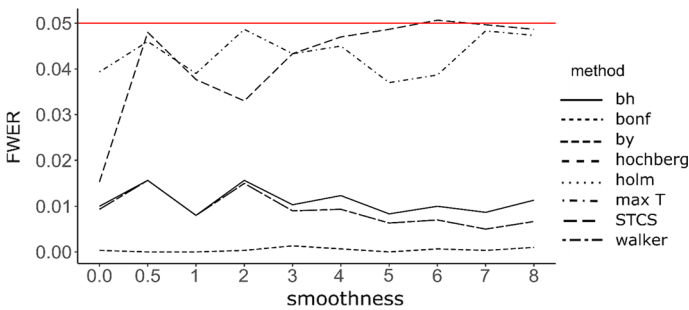
**Fig. 7** Achieved FWER of the multiple testing corrections as a function of spatial autocorrelation. The spatial autocorrelation is determined by  $r(d) = \exp(-c|d|)$ , where  $d$  represents the Euclidean distance between the grid cell midpoints, and  $c$  is chosen so that  $r(1)$  equals the desired spatial autocorrelation



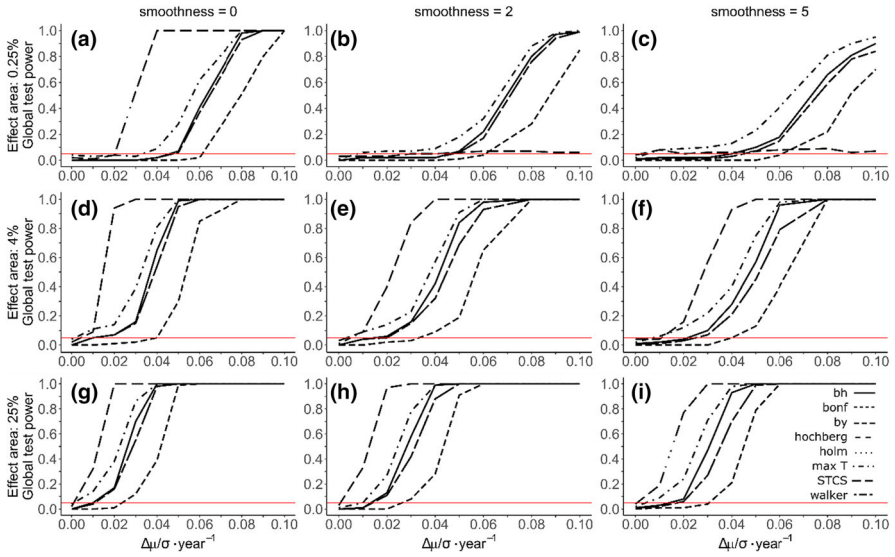
**Fig. 8** Global test power for the case of no spatial autocorrelation, moderate spatial autocorrelation (0.5) and strong spatial autocorrelation (0.9) as a function of trend magnitude. The red line indicates the  $\alpha_{\text{global}} = 0.05$  level. The effect area is the percentage of all grid cells where a trend was induced. The spatial autocorrelation is determined by  $r(d) = \exp(-c|d|)$ , where  $d$  represents the Euclidean distance between the grid cell midpoints, and  $c$  is chosen so that  $r(1)$  equals the desired spatial autocorrelation



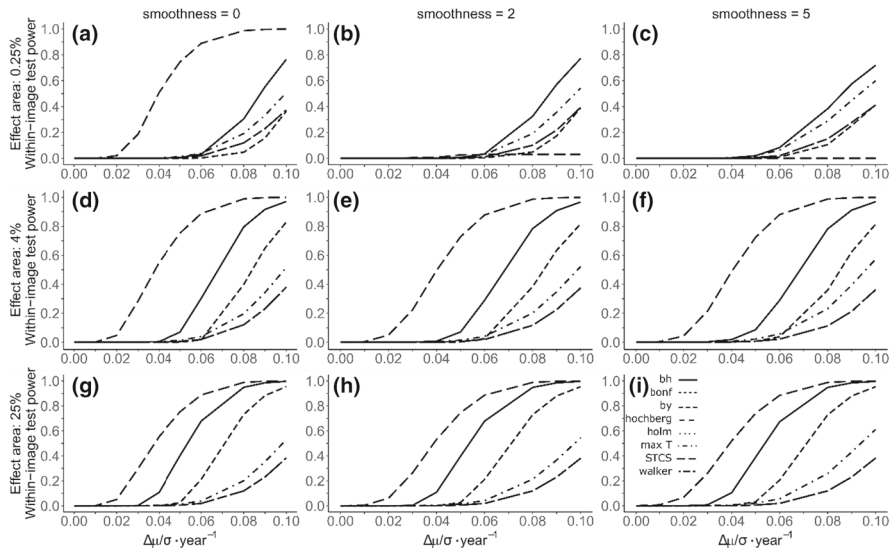
**Fig. 9** Within-image test power for the case of strong no spatial autocorrelation, moderate spatial autocorrelation (0.5) and strong spatial autocorrelation (0.9) as a function of trend magnitude. The effect area is the percentage of all grid cells where a trend was induced. The spatial autocorrelation is determined by  $r(d) = \exp(-c|d|)$ , where  $d$  represents the Euclidean distance between the grid cell midpoints, and  $c$  is chosen so that  $r(1)$  equals the desired spatial autocorrelation



**Fig. 10** Achieved FWER of the multiple testing corrections as a function of spatial autocorrelation. The spatial autocorrelation is determined by the Matérn correlation function with a range parameter of 1 and varying smoothness parameter



**Fig. 11** Global test power for the case of no spatial autocorrelation, moderate spatial autocorrelation ( $\kappa = 2$ ) and strong spatial autocorrelation ( $\kappa = 5$ ) as a function of trend magnitude. The red line indicates the  $\alpha_{\text{global}} = 0.05$  level. The effect area is the percentage of all grid cells where a trend was induced. The spatial autocorrelation is determined by the Matérn correlation function with a range parameter of 1 and varying smoothness parameter



**Fig. 12** Within-image test power for the case of strong no spatial autocorrelation, moderate spatial autocorrelation ( $\kappa = 2$ ) and strong spatial autocorrelation ( $\kappa = 5$ ) as a function of trend magnitude. The effect area is the percentage of all grid cells where a trend was induced. The spatial autocorrelation is determined by the Matérn correlation function with a range parameter of 1 and varying smoothness parameter



## References

- Beck PSA, Goetz SJ (2012) Corrigendum: satellite observations of high northern latitude vegetation productivity changes between 1982 and 2008: ecological variability and regional differences. *Environ Res Lett* 7(2):029501. <https://doi.org/10.1088/1748-9326/7/2/029501>
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc* 57(1):289–300
- Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29(4):1165–1188. <https://doi.org/10.1214/aos/1013699998>
- Bonferroni CE (1936) *Teoria Statistica Delle Classi e Calcolo Delle Probabilità*. Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commerciali Di Firenze 8:3–62
- Bretz F, Hothorn T, Westfall PH (2011) *Multiple comparisons using R*. CRC Press, Boca Raton
- Cao J, Worsley KJ (2001) Applications of random fields in human brain mapping. In: Moore M (ed) *Spatial statistics: methodological aspects and applications*. Springer, New York, pp 169–182. [https://doi.org/10.1007/978-1-4613-0147-9\\_8](https://doi.org/10.1007/978-1-4613-0147-9_8)
- DelSole T, Yang X (2011) Field significance of regression patterns. *J Clim* 24(19):5094–5107. <https://doi.org/10.1175/2011JCLI4105.1>
- Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. *Stat Sci* 18(1):71–103. <https://doi.org/10.1214/ss/1056397487>
- Dwass M (1957) Modified randomization tests for nonparametric hypotheses. *Ann Math Stat* 28(1):181–187. <https://doi.org/10.1214/aoms/1177707045>
- Eckert S, Hüslér F, Liniger H, Hodel E (2015) Trend analysis of MODIS NDVI time series for detecting land degradation and regeneration in Mongolia. *J Arid Environ* 113:16–28. <https://doi.org/10.1016/j.jaridenv.2014.09.001>
- Edgington ES (1969) Approximate randomization tests. *J Psychol* 72(2):143–149. <https://doi.org/10.1080/00223980.1969.10543491>
- Fattorini L, Pisani C, Riga F, Zaccaroni M (2014) A permutation-based combination of sign tests for assessing habitat selection. *Environ Ecol Stat* 21(2):161–187. <https://doi.org/10.1007/s10651-013-0250-7>
- Fensholt R, Proud SR (2012) Evaluation of earth observation based global long term vegetation trends—comparing GIMMS and MODIS global NDVI time series. *Remote Sens Environ* 119:131–147. <https://doi.org/10.1016/j.rse.2011.12.015>
- GISTEMP Team (2019) *GISS Surface Temperature Analysis (GISTEMP)*. <https://data.giss.nasa.gov/gistemp/>
- Hansen J, Ruedy R, Sato M, Lo K (2010) Global surface temperature change. *Rev Geophys* 48:4. <https://doi.org/10.1029/2010RG000345>
- Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75(4):800. <https://doi.org/10.2307/2336325>
- Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6(2):65–70
- Julien Y, Sobrino JA (2009) Global land surface phenology trends from GIMMS database. *Int J Remote Sens* 30(13):3495–3513. <https://doi.org/10.1080/01431160802562255>
- Kraft B, Jung M, Körner M, Requena Mesa C, Cortés J, Reichstein M (2019) Identifying dynamic memory effects on vegetation state using recurrent neural networks. *Front Big Data* 2:31. <https://doi.org/10.3389/fdata.2019.00031>
- Kriewald S, Fluschnik T, Reusser D (2019) *osc: Orthodromic Spatial Clustering* (Version R package version 1.0.4). <https://CRAN.R-project.org/package=osc>
- Livezey RE, Chen WY (1983) Statistical field significance and its determination by Monte Carlo techniques. *Mon Weather Rev* 111(1):46–59. [https://doi.org/10.1175/1520-0493\(1983\)111<0046:SFAID>2.0.CO;2](https://doi.org/10.1175/1520-0493(1983)111<0046:SFAID>2.0.CO;2)
- Nichols T, Hayasaka S (2003) Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res* 12(5):419–446. <https://doi.org/10.1191/0962280203sm341ra>
- Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp* 15(1):1–25
- Petersson KM, Nichols TE, Poline J-B, Holmes AP (1999) Statistical limitations in functional neuroimaging II. Signal detection and statistical inference. *Philos Trans R Soc Lond* 354(1387):1261–1281. <https://doi.org/10.1098/rstb.1999.0478>



- Pinzon J, Tucker C (2014) A non-stationary 1981–2012 AVHRR NDVI3g time series. *Remote Sens* 6(8):6929–6960. <https://doi.org/10.3390/rs6086929>
- R Core Team (2019) R: a language and environment for statistical computing. R foundation for statistical computing. <https://www.R-project.org/>
- Risser MD, Paciorek CJ, Stone DA (2019) Spatially dependent multiple testing under model misspecification, with application to detection of anthropogenic influence on extreme climate events. *J Am Stat Assoc* 114(525):61–78. <https://doi.org/10.1080/01621459.2018.1451335>
- Shen X, Huang H-C, Cressie N (2002) Nonparametric hypothesis testing for a spatial signal. *J Am Stat Assoc* 97(460):1122–1140. <https://doi.org/10.1198/016214502388618933>
- Sun W, Reich BJ, Tony Cai T, Guindani M, Schwartzman A (2015) False discovery control in large-scale spatial multiple testing. *J R Stat Soc* 77(1):59–83. <https://doi.org/10.1111/rssb.12064>
- Tucker CJ, Pinzon JE, Brown ME, Slayback DA, Pak EW, Mahoney R, Vermote EF, El Saleous N (2005) An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data. *Int J Remote Sens* 26(20):4485–4498. <https://doi.org/10.1080/01431160500168686>
- Ventura V, Paciorek CJ, Risbey JS (2004) Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *J Clim* 17(22):4343–4356. <https://doi.org/10.1175/JCLI17224343>
- von Storch H (1999) Misuses of statistical analysis in climate research. In: von Storch H, Navarra A (eds) *Analysis of climate variability*. Springer, Berlin, pp 11–26. [https://doi.org/10.1007/978-3-662-03744-7\\_2](https://doi.org/10.1007/978-3-662-03744-7_2)
- Walker GT (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 DS (2006a) *Statistical methods in the atmospheric sciences*, 2nd edn. Academic Press, Cambridge
- Wilks DS (2016) “The Stippling Shows Statistically Significant Grid Points”: how research results are routinely overstated and overinterpreted, and what to do about it. *Bull Am Meteor Soc* 97(12):2263–2273. <https://doi.org/10.1175/BAMS-D-15-00267.1>
- Wilks DS (2006b) On “Field Significance” and the false discovery rate. *J Appl Meteorol Climatol* 45(9):1181–1189. <https://doi.org/10.1175/JAM2404.1>
- Worsley KJ, Evans AC, Marrett S, Neelin P (1992) A three-dimensional statistical analysis for CBF activation studies in human brain. *J Cereb Blood Flow Metab* 12(6):900–918. <https://doi.org/10.1038/jcbfm.1992.127>
- Yue S, Pilon P, Phinney B, Cavadias G (2002) The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrol Process* 16(9):1807–1829. <https://doi.org/10.1002/hyp.1095>
- Zang CS, Jochner-Oette S, Cortés J, Rammig A, Menzel A (2019) Regional trend changes in recent surface warming. *Clim Dyn* 52(11):6463–6473. <https://doi.org/10.1007/s00382-018-4524-5>
- Zhang Y, Song C, Band LE, Sun G, Li J (2017) Reanalysis of global terrestrial vegetation trends from MODIS products: browning or greening? *Remote Sens Environ* 191:145–155. <https://doi.org/10.1016/j.rse.2016.12.018>

**Jose Cortes** is a Ph.D. student at Friedrich Schiller University, Jena, Germany. He is a member of the International Max Planck Research School on Global Biogeochemical Cycles (IMPRS-gBGC), a joint program with the Max Planck Institute for Biogeochemistry. His work focuses on spatio-temporal trend detection in environmental data.

**Miguel Mahecha** is a scientist and leader of the research group Empirical Inference of the Earth System in the Department of Biogeochemical Integration at the Max Planck Institute for Biogeochemistry. His current research interests are climate extremes and ecosystem functioning, intrinsic functioning of terrestrial ecosystems, biogeography and ecosystem functioning, and methodological aspects of time series analysis (spatiotemporal environmental data), nonlinear dimensionality reduction (e.g. for vegetation data, remote sensing data), and extrapolation (e.g. upscaling plant traits for Europe).

**Markus Reichstein** is the Director of the Department of Biogeochemical Integration at the Max Planck Institute for Biogeochemistry. His research interests are biogeochemical cycle interactions (C-H<sub>2</sub>O-N-P), data assimilation, data mining, dynamic global ecosystem modelling (DGEM), earth observation, eddy flux data interpretation, and soil-plant feedbacks.

**Alexander Brenning** is Professor of Geographic Information Science at Friedrich Schiller University, Jena, Germany. His research focuses on spatial statistical and computational tools and their application in a variety of contexts, in particular mountain geomorphology and environmental remote sensing. He has published free software extensions for the statistical software R that integrate R with GIS software and implement spatial accuracy assessment and variable importance techniques.