JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 00, NO. 0, 1–16: Theory and Methods
https://doi.org/10.1080/01621459.2020.1775613

Detection of Local Differences in Spatial Characteristics Between Two Spatiotemporal Random Fields

Sooin Yun[a], Xianyang Zhang[b], and Bo Li[a]

[a]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL; [b]Department of Statistics, Texas A&M University, College Station, TX

ABSTRACT

Comparing the spatial characteristics of spatiotemporal random fields is often at demand. However, the comparison can be challenging due to the high-dimensional feature and dependency in the data. We develop a new multiple testing approach to detect local differences in the spatial characteristics of two spatiotemporal random fields by taking the spatial information into account. Our method adopts a two-component mixture model for location wise p-values and then derives a new false discovery rate (FDR) control, called mirror procedure, to determine the optimal rejection region. This procedure is robust to model misspecification and allows for weak dependency among hypotheses. To integrate the spatial heterogeneity, we model the mixture probability as well as study the benefit if any of allowing the alternative distribution to be spatially varying. An EM-algorithm is developed to estimate the mixture model and implement the FDR procedure. We study the FDR control and the power of our new approach both theoretically and numerically, and apply the approach to compare the mean and teleconnection pattern between two synthetic climate fields. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2018
Accepted May 2020

KEYWORDS
FDR; Mirror procedure; Multiple testing; Spatial dependency; Spatiotemporal random field

1. Introduction

Comparing spatiotemporal random fields and identifying the locations where the difference occurs are often at demand. For example, climate scientists are interested in learning the difference between synthetic climate fields generated by different general circulation models, for the purpose of studying the mechanism of climate models. In the study of paleoclimate, a key step is to compare the reconstructed climate field to its target to evaluate the reconstruction method. In those examples, the characteristics of climate fields to be compared typically include the spatial mean and the spatial covariance structure. Furthermore, it will be more appealing if we can identify where the difference is if any. These questions can be answered by comparing the two random fields at the location level. However, hypothesis testing at each location will lead to multiplicity issue due to the large number of spatial locations. In addition to the high-dimensional feature, the dependency in the climate fields makes this multiple testing problem more challenging.

Various approaches have been proposed to compare two spatiotemporal random fields. Enlightened by Lund and Li (2009), Li and Smerdon (2012) proposed to compare climate field reconstructions by integrating the difference in both the mean and covariance structure based on the two-sample Kolmogorov–Smirnov test of whitened random fields. By treating the spatiotemporal random fields as replicates of two-dimensional functional data, Zhang and Shao (2015) and Li, Zhang, and Smerdon (2016) developed test statistics for evaluating the difference of several properties between two random fields. Another related work is the comparison of two spatial random fields in terms of user-chosen loss functions (Hering and Genton 2011). Note all these papers provided only global tests for the entire random field but not the comparison at local level.

False discovery rate (FDR) has become a popular criterion for large-scale hypothesis testing, for its ability to attain high power while controlling the FDR or the number of incorrect rejections at a prespecified level. Classical multiple testing procedure was first introduced by Benjamini and Hochberg (1995) (hereafter BH) which has been widely used in many disciplines. Later, Benjamini and Hochberg (2000) and Benjamini and Yekutieli (2001) presented an adaptive p-value procedure as an alternative to the BH method and Genovese and Wasserman (2002) presented a test procedure that controls FDR and false nondiscovery rate (FNR) simultaneously. Storey (2002) (hereafter JH) developed a method to improve the power of BH by estimating the proportion of null hypotheses. A number of other alternatives to BH method adopted the Bayesian viewpoint by formulating the multiple testing through a mixture model, $x_i|\theta_i \sim \theta_i f_0 + (1-\theta_i)f_1$, where x_i's are the observations, θ_i is a Bernoulli random variable indicating the state, and f_0 and f_1 are the null and alternative distributions, respectively. Based on the mixture model, Storey (2003) controlled the positive FDR (pFDR) and Efron et al. (2001) and Efron (2004) introduced the Bayesian framework for parameter estimation and multiple testing.

CONTACT Bo Li libo@illinois.edu Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, IL 61820.
Yun and Zhang are joint first authors.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

The majority of the multiple testing methods assume independence among the hypotheses. Efforts have been made to apply the FDR to correlated data. Shen, Huang, and Cressie (2002) developed a nonparametric enhanced FDR (EFDR) procedure to test the difference between the mean function of two spatial random fields. Their method first performed a wavelet transformation to the gridded data, and then applied the EFDR procedure to test whether the independent wavelet coefficients are simultaneously zero. Later, Pavlicová, Santner, and Cressie (2008) developed an FDR procedure also in wavelet space but enhanced by a p-value adaptive thresholding approach, motivated by a brain image data. Benjamini and Heller (2007) controlled the FDR of spatial signals by proposing a two-stage hierarchical procedure that first identifies clusters of pixels where the signal might be present and then removes pixels where the signal is absent.

In presence of dependence among hypotheses, Sun and Cai (2009) introduced a new test statistic, the local index of significance (LIS), which takes into account the dependency in adjacent observations through hidden Markov models (HMM). However, since the optimality of LIS requires the estimates of the unknown model parameters to be consistent, LIS in general may not be the optimal multiple test for dependence structures other than HMM. Sun et al. (2015) tackled more general dependency structure in identifying spatial signals from a random field that is contaminated by correlated errors. Their method is developed based on an oracle statistic that represents the posterior probability of whether the signal occurs. This posterior probability is computed using Bayesian models that take the spatial correlation into account. Although Sun et al. (2015) accounted for dependence in hypotheses, it involves the posterior probability that conditions on the hyperparameters. Thus, the optimality of their procedure is model-specific and appears to be sensitive to model misspecification. Very recently, by adopting the decision rules in Müller, Sansó, and De Iorio (2004) and Muller, Parmigiani, and Rice (2006), Risser, Paciorek, and Stone (2019) developed a robust Bayesian decision theoretical approach to control FDR for dependent multiple testing via constructing a nonparametric hierarchical statistical model for estimating the dependence in the mixture probability of hypotheses. For multiple testing problems that exhibit spatial patterns, Tansey et al. (2018) (hereafter FDRS) introduced a FDR smoothing method which automatically finds the spatially localized regions of significant test statistics and then relaxes the thresholds of significance within these regions.

We develop a multiple testing approach to compare characteristics of two spatiotemporal random processes at a local level. Our approach models the mixture probability of hypotheses as a smooth function over space to account for spatial variability, and models the alternative density of p-values either as an invariant nonparametric function or a spatially varying semiparametric function. To control FDR, we propose a new multiple testing procedure (called the mirror procedure) that is not only robust to model misspecification but also allows for weak dependence among hypotheses while boosting power compared to traditional FDR methods. An EM-algorithm is developed to estimate the mixture model.

The article is organized as follows. Section 2 derives the mirror FDR procedure based on a mixture model of p-values and shows the asymptotic property of this new procedure. Section 3 develops an EM-algorithm to estimate the mixture model with either a nonparametric or a semiparametric alternative density function in the mixture model. The Monte Carlo experiments in Section 4 demonstrate our methods in comparing the mean and covariance structure of two spatiotemporal random fields. In Section 5, we apply our method to paleoclimate reconstruction to detect the deviation of the reconstruction from its target in terms of the mean and teleconnection pattern. Section 6 concludes with a brief discussion.

2. FDR Control Procedure

Let $X(\mathbf{s},t)$ and $Y(\mathbf{s},t)$ be two spatiotemporal random fields observed over spatial locations, $\mathbf{s} \in D$, and time points, $t \in \mathbb{Z}$. At each location $\mathbf{s}$, we assume that the sequences $\{X(\mathbf{s},t)\}_t$ and $\{Y(\mathbf{s},t)\}_t$ are both stationary over time. Our goal is to compare the spatial characteristics of the two spatiotemporal random fields at each location while adjusting the multiplicity due to multiple comparison. Denote by $\theta_X(\mathbf{s})$ ($\theta_Y(\mathbf{s})$) certain spatial characteristics of the distribution of $\{X(\mathbf{s},t)\}_t$ ($\{Y(\mathbf{s},t)\}_t$), that is, $\theta_X(\mathbf{s}) = T(F_{X(\mathbf{s},t)})$, where T denotes a functional on the distribution of $X(\mathbf{s},t)$. For instance, $\theta_X(\mathbf{s})$ can be the mean of $X(\mathbf{s},t)$. More generally, $\theta_X(\mathbf{s})$ ($\theta_Y(\mathbf{s})$) can denote some characteristics of the joint distribution of $\{X(\mathbf{s},t), X(\mathbf{s}_0,t)\}_t$ ($\{Y(\mathbf{s},t), Y(\mathbf{s}_0,t)\}_t$) for a fixed location $\mathbf{s}_0$, for example, the covariance between $X(\mathbf{s},t)$ and $X(\mathbf{s}_0,t)$. Denote by $H_{0,\mathbf{s}}$ and $H_{a,\mathbf{s}}$ the null and alternative hypotheses associated with location $\mathbf{s}$. We are interested in testing:

$$H_{0,\mathbf{s}} : \theta_X(\mathbf{s}) = \theta_Y(\mathbf{s}) \quad \text{versus} \quad H_{a,\mathbf{s}} : \theta_X(\mathbf{s}) \neq \theta_Y(\mathbf{s})$$

simultaneously for all $\mathbf{s} \in D$.

We will develop a FDR control method for this multiple testing problem that is defined over spatial locations. To motivate the FDR control procedure, we assume a two-component mixture model for location wise p-values. The rest of this section first presents the mixture model for spatially varying p-values, and then derives the optimal rejection region followed by the theoretical derivation of the cutoff value.

2.1. Model for p-Values

Suppose we are given the samples $\{X(\mathbf{s},t) : \mathbf{s} \in D, t = 1,2\ldots,n_1\}$ and $\{Y(\mathbf{s},t) : \mathbf{s} \in D, t = 1,2\ldots,n_2\}$. Let $p_\mathbf{s}$ be the p-value for testing $H_{0,\mathbf{s}}$ against $H_{a,\mathbf{s}}$ obtained based on a suitable testing procedure. We assume that $p_\mathbf{s}$ follows a two-component mixture model:

$$f(p_\mathbf{s};\mathbf{s}) = \pi(\mathbf{s})f_0(p_\mathbf{s}) + (1-\pi(\mathbf{s}))f_1(p_\mathbf{s};\mathbf{s}), \tag{1}$$

where $\pi(\mathbf{s}) \in [0,1]$ is the probability that the p-value is from the null, and f_0 and f_1 are the null and alternative distributions, respectively. Throughout, we shall assume that f_0 is mirror conservative (Lei and Fithian 2018), that is,

$$\int_{a_1}^{a_2} f_0(p)dp \leq \int_{1-a_2}^{1-a_1} f_0(p)dp, \quad 0 \leq a_1 \leq a_2 \leq 0.5. \tag{2}$$

This condition means the density f_0 is at least as large at $1-p$ as at p for $p \leq 0.5$. The varying null probability $\pi(\mathbf{s})$ reflects

the relative importance of each hypothesis, and the varying alternative density $f_1(\cdot;\mathbf{s})$ emphasizes the heterogeneity among signals at different locations.

2.2. Optimal Rejection Region

We derive the optimal rejection region under model (1). Similar results have been obtained in recent literature (see, e.g., Sun et al. 2015; Basu et al. 2018; Lei and Fithian 2018). Following Sun et al. (2015), we define the local FDR (LFDR) at location $\mathbf{s}$ and p-value $p_\mathbf{s}$ as

$$\mathrm{LFDR}_\mathbf{s}(p_\mathbf{s}) = \frac{\pi(\mathbf{s})f_0(p_\mathbf{s})}{\pi(\mathbf{s})f_0(p_\mathbf{s}) + (1-\pi(\mathbf{s}))f_1(p_\mathbf{s};\mathbf{s})}.$$

Suppose we reject the hypothesis at $\mathbf{s} \in D$ when $p_\mathbf{s} \le q_\mathbf{s}$ for some cutoff value $q_\mathbf{s}$.

Let F_0 and $F_1(\cdot;\mathbf{s})$ be the null and alternative distribution functions associated with f_0 and $f_1(\cdot,\mathbf{s})$, respectively. Write $\mathbf{q} = \{q_\mathbf{s} : \mathbf{s} \in D\}$. Our goal is to control the average FDR (aFDR) over m hypotheses defined as

$$\mathrm{aFDR}_m(\mathbf{q}) := \frac{\sum_{\mathbf{s}\in D}\pi(\mathbf{s})F_0(q_\mathbf{s})}{\sum_{\mathbf{s}\in D}\{\pi(\mathbf{s})F_0(q_\mathbf{s}) + (1-\pi(\mathbf{s}))F_1(q_\mathbf{s};\mathbf{s})\}}, \quad (3)$$

while maximizing the number of rejections. This can be formulated into the following constraint optimization problem:

$$\max_{\mathbf{q}} \frac{1}{m}\sum_{\mathbf{s}\in D}(1-\pi(\mathbf{s}))F_1(q_\mathbf{s};\mathbf{s}) \quad \text{subject to} \quad \mathrm{aFDR}(\mathbf{q}) \le \alpha. \quad (4)$$

We show that the optimal thresholds in this case are level surfaces of the $\mathrm{LFDR}_\mathbf{s}$, that is, $\mathrm{LFDR}_\mathbf{s}(\cdot)$ evaluated at the optimal threshold is independent of the location $\mathbf{s}$. To see this, note that the Lagrangian function of (4) is given by

$$\mathcal{L}(\mathbf{q},\lambda) = \frac{1}{m}\sum_{\mathbf{s}\in D}\big\{-\lambda(1-\alpha)\pi(\mathbf{s})F_0(q_\mathbf{s}) + (1+\lambda\alpha)(1-\pi(\mathbf{s}))F_1(q_\mathbf{s};\mathbf{s})\big\}.$$

Thus under similar conditions as those in Theorem 2 of Lei and Fithian (2018), the Karush–Kuhn–Tucker (KKT) condition indicates that

$$-\lambda(1-\alpha)\pi(\mathbf{s})f_0(q_\mathbf{s}) + (1+\lambda\alpha)(1-\pi(\mathbf{s}))f_1(q_\mathbf{s};\mathbf{s}) = 0,$$

which after some simple algebra shows that $\mathrm{LFDR}_\mathbf{s}(q_\mathbf{s})$ is a constant independent of $\mathbf{s}$. This insight can be formally stated as the following theorem, which motivates our choice of rejection regions.

Theorem 1. Assume that $f_1(\cdot;\mathbf{s})$ is continuously non-increasing at each $\mathbf{s}$, and f_0 is continuously non-decreasing and uniformly bounded from above. Further assume that

$$\max_{\mathbf{s}\in D}\frac{(1-\pi(\mathbf{s}))f_1(0;\mathbf{s})}{\pi(\mathbf{s})f(0)} > \frac{1-\alpha}{\alpha}. \quad (5)$$

Then (4) has at least one solution and every solution $\{\widetilde{q}_\mathbf{s} : \mathbf{s} \in D\}$ satisfies that $\mathrm{LFDR}_\mathbf{s}(\widetilde{q}_\mathbf{s})$ is independent of the location $\mathbf{s}$.

Here condition (5) is a modification of (ii) in Theorem 2 of Lei and Fithian (2018), which suffices for their arguments to go through in our case. Throughout the following discussions, we assume a monotone likelihood ratio condition:

$$f_1(p_\mathbf{s};\mathbf{s})/f_0(p_\mathbf{s}) \text{ is a monotonically decreasing function of } p_\mathbf{s},$$

for all $\mathbf{s} \in D$. Such an assumption has been made in Storey (2002), Genovese and Wasserman (2002), and Genovese and Wasserman (2004), among others. Under the monotone likelihood ratio condition, $\mathrm{LFDR}_\mathbf{s}(p_\mathbf{s})$ is a monotonically increasing function of $p_\mathbf{s}$. Thus, the rejection rule $p_\mathbf{s} \le \widetilde{q}_\mathbf{s}$ can be expressed as

$$\mathrm{LFDR}_\mathbf{s}(p_\mathbf{s}) \le \mathrm{LFDR}_\mathbf{s}(\widetilde{q}_\mathbf{s}).$$

According to Theorem 1, $\mathrm{LFDR}_\mathbf{s}(\widetilde{q}_\mathbf{s})$ is independent of $\mathbf{s}$, which leads to the following optimal rejection region,

$$\mathrm{LFDR}_\mathbf{s}(p_\mathbf{s}) = \frac{\pi(\mathbf{s})f_0(p_\mathbf{s})}{\pi(\mathbf{s})f_0(p_\mathbf{s}) + (1-\pi(\mathbf{s}))f_1(p_\mathbf{s};\mathbf{s})} \le q, \quad (6)$$

with q to be determined in the next subsection.

2.3. The Mirror Procedure

The cutoff value q in (6) was previously determined by the step-up procedure (see, e.g., Sun and Cai 2007), which typically requires the estimated LFDR to converge to the truth. Here we introduce a more robust procedure to determine the cutoff value which enjoys both the robustness and flexibility over the step-up procedure: the new procedure controls the asymptotic FDR even when (1) the estimated LFDR converges to a limit that is different from the true LFDR (see Assumption 1); and (2) the p-values are weakly dependent (see Assumption 4). This procedure extends the method in Barber and Candes (2016) (also see Lei and Fithian 2018) to incorporate spatial information through the spatial varying functions $\pi(\mathbf{s})$ and $f_1(\cdot;\mathbf{s})$. As our approach is developed by exploiting the mirror conservativeness of f_0, we shall call it the mirror procedure.

To introduce the proposed method, we note that the false discovery proportion (FDP) based on the rejection rule (6) with threshold q is given by

$$\mathrm{FDP}(q) := \frac{\sum_{\mathbf{s}\in S_0}\mathbf{1}\{\mathrm{LFDR}_\mathbf{s}(p_\mathbf{s}) \le q\}}{1 \vee \sum_{\mathbf{s}\in D}\mathbf{1}\{\mathrm{LFDR}_\mathbf{s}(p_\mathbf{s}) \le q\}},$$

where S_0 denotes the set of locations under which the null hypothesis is true, $\mathbf{1}(A)$ denotes the indicator function associated with a set A, and $a \vee b = \max\{a,b\}$. We observe that

$$\mathrm{FDP}(q) \lesssim \frac{\sum_{\mathbf{s}\in S_0}\mathbf{1}\{\mathrm{LFDR}_\mathbf{s}(1-p_\mathbf{s}) \le q\}}{1 \vee \sum_{\mathbf{s}\in D}\mathbf{1}\{\mathrm{LFDR}_\mathbf{s}(p_\mathbf{s}) \le q\}} \le \frac{\sum_{\mathbf{s}\in D}\mathbf{1}\left\{\mathrm{LFDR}_\mathbf{s}(1-p_\mathbf{s}) \le q\right\}}{1 \vee \sum_{\mathbf{s}\in D}\mathbf{1}\left\{\mathrm{LFDR}_\mathbf{s}(p_\mathbf{s}) \le q\right\}} := \mathrm{FDP}_{\mathrm{up}}(q),$$

where "$\lesssim$" means "approximately smaller than" which is due to the mirror conservativeness of f_0. Set $q^* = \max\{q \in [0,1] : \mathrm{FDP}_{\mathrm{up}}(q) \le \alpha\}$. We then reject $H_{0,\mathbf{s}}$ if

$$\mathrm{LFDR}_\mathbf{s}(p_\mathbf{s}) \le q^*.$$

In practice, $f_1(\cdot, \mathbf{s})$ and $\pi(\mathbf{s})$ are often unknown and are replaced by their sample counterparts. Let $\widehat{\text{LFDR}}_{\mathbf{s}}(\cdot)$ be an estimator of $\text{LFDR}_{\mathbf{s}}(\cdot)$ for $\mathbf{s} \in D$, and define

$$\widehat{q} = \max\left\{ q \in [0,1] : \frac{\sum_{\mathbf{s}\in D} \mathbf{1}\left\{\widehat{\text{LFDR}}_{\mathbf{s}}(1-p_{\mathbf{s}}) \le q\right\}}{1 \vee \sum_{\mathbf{s}\in D} \mathbf{1}\left\{\widehat{\text{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le q\right\}} \le \alpha \right\}.$$

The empirical decision rule is to reject $H_{0,\mathbf{s}}$ if

$$\widehat{\text{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le \widehat{q}.$$

Below we show that this empirical procedure provides asymptotic FDR control.

In the following asymptotic analysis, we assume that there is an array of p-values $\{p_{\mathbf{s}} : \mathbf{s} \in D_m \subseteq \mathcal{D} \subseteq \mathbb{R}^2, m \in \mathbb{Z}^+\}$ with $|D_m| \to +\infty$ as $m \to +\infty$. Keep in mind that the set of null locations $S_0 = S_{0,m}$ is also changing with respect to m. The limit below is taken by letting $m \to +\infty$. We assume that $|S_{0,m}|/|D_m| \to \gamma \in (0,1]$ throughout the discussions.

Assumption 1. Assume that there exist nondecreasing and right continuous functions $\Psi_{\mathbf{s}}(\cdot) : [0,1] \to [0,1]$ for $\mathbf{s} \in D_m$ such that

$$\frac{1}{|D_m|}\sum_{\mathbf{s}\in D_m} |\widehat{\text{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) - \Psi_{\mathbf{s}}(p_{\mathbf{s}})| \to^p 0.$$

Assumption 2. Let $\Psi_{\mathbf{s}}$ be defined in Assumption 1. Suppose

$$\frac{1}{|D_m|}\sum_{\mathbf{s}\in D_m} P(\Psi_{\mathbf{s}}(p_{\mathbf{s}}) \le q) \to G_0(q),$$
$$\frac{1}{|D_m|}\sum_{\mathbf{s}\in S_0} P(\Psi_{\mathbf{s}}(p_{\mathbf{s}}) \le q) \to \widetilde{G}_1(q),$$
$$\frac{1}{|D_m|}\sum_{\mathbf{s}\in D_m} P(\Psi_{\mathbf{s}}(1-p_{\mathbf{s}}) \le q) \to G_1(q),$$

for $q \in [0,1]$, where $G_0(q)$, $\widetilde{G}_1(q)$, and $G_1(q)$ are all nondecreasing and differentiable functions with the derivatives bounded from above by some constants $C_0, \widetilde{C}_1$, and C_1 on the interval $[q_0/2, 1]$ for some constant $q_0 > 0$, respectively.

Assumption 3. Suppose there exists a $q' > q_0 > 0$ such that $U(q') := G_1(q')/G_0(q') < \alpha$, where G_i's and q_0 are defined in Assumption 2.

Next we introduce weak dependence assumption on the p-values. Suppose $\rho(\mathbf{a},\mathbf{b}) = \sup_{1\le i\le 2}|a_i - b_i|$ for $\mathbf{a} = (a_1, a_2)$ and $\mathbf{b} = (b_1, b_2)$. For $U, V \subseteq \mathcal{D}$, define the distance between U and V as $\rho(U,V) = \inf_{\mathbf{a}\in U,\mathbf{b}\in V}\rho(\mathbf{a},\mathbf{b})$. Let $\sigma(U) = \sigma(p_{\mathbf{s}}, \mathbf{s} \in U)$ be the sigma-field generated by $p_{\mathbf{s}}$ for $\mathbf{s} \in U \subseteq \mathcal{D}$. The α-mixing and ψ-mixing coefficients for $\{p_{\mathbf{s}}, \mathbf{s} \in \mathcal{D}\}$ are defined, respectively, as

$$\alpha(u,v,r) = \sup_m \sup_{U,V\subseteq D_m} \{\alpha(U,V) : |U| \le u, |V| \le v, \rho(U,V) \ge r\}, \quad (7)$$

$$\psi(u,v,r) = \sup_m \sup_{U,V\subseteq D_m} \{\psi(U,V) : |U| \le u, |V| \le v, \rho(U,V) \ge r\}, \quad (8)$$

where $\alpha(U,V) = \sup\{|P(A\cap B) - P(A)P(B)| : A \in \sigma(U), B \in \sigma(V)\}$ and $\psi(U,V) = \sup\{|P(A|B) - P(A)| : A \in \sigma(U), B \in \sigma(V), P(B) > 0\}$.

Assumption 4. The set $\mathcal{D}$ is infinitely countable. All elements in $\mathcal{D}$ are located at distances of at least $\rho_0 > 0$ from each other (that is $\rho(\mathbf{a},\mathbf{b}) \ge \rho_0$ for all $\mathbf{a},\mathbf{b} \in \mathcal{D}$). The mixing coefficient $\alpha(u,v,r)$ defined in (7) (or $\psi(u,v,r)$ defined in (8)) satisfies that $\sum_{r=1}^{+\infty} r\alpha(1,1,r) < \infty$ (or $\sum_{r=1}^{+\infty} r\psi(1,1,r) < \infty$).

We remark that $\Psi_{\mathbf{s}}$ does not have to be the same as $\text{LFDR}_{\mathbf{s}}$ for our procedure to achieve asymptotic FDR control. Thus, our method is more robust than the classical LFDR-based procedure which requires consistent estimation of $\text{LFDR}_{\mathbf{s}}$. Assumption 1 requires that the empirical L_1 norm of the difference between $\widehat{\text{LFDR}}_{\mathbf{s}}$ and $\Psi_{\mathbf{s}}$ converges in probability to zero. Assumptions 2 and 3 are relatively mild and are similar to those in Theorem 4 of Storey, Taylor, and Siegmund (2004). In Assumption 4, we allow the random field formed by the p-values to be strong mixing. In this sense, our procedure is asymptotically robust to weak dependence.

Define

$$\text{FDR}_m(q) = E\left[\frac{\sum_{\mathbf{s}\in S_0}\mathbf{1}\{\widehat{\text{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le q\}}{1 \vee \sum_{\mathbf{s}\in D_m}\mathbf{1}\{\widehat{\text{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le q\}}\right]. \quad (9)$$

Theorem 2. Under Assumptions 1–4 and further assume that f_0 is mirror conservative, then we have

$$\lim_{m\to+\infty} \text{FDR}_m(\widehat{q}) \le \alpha.$$

The proof for Theorem 2 is deferred to Appendix A.

3. Estimation

We assume $f_0(\cdot)$ to be Uniform(0,1) which satisfies the mirror conservative condition. Alternatively, f_0 can also be estimated from the data (see, e.g., Efron 2004). Then we shall estimate $\pi(\mathbf{s})$ and $f_1(\cdot;\mathbf{s})$ in (1) to implement our FDR control procedures. In the estimation, we assume both functions as nonparametric functions to warrant flexibility, and then additionally consider f_1 also as a spatially varying semiparametric function. Let G be an invertible map from $\mathbb{R} \to [0,1]$, for example, G can be the distribution function of the logistic or normal distribution. We model $\pi(\mathbf{s})$ through G by

$$G^{-1}(\pi(\mathbf{s})) = \sum_{j=1}^{d} \beta_j B_j(\mathbf{s}), \quad (10)$$

where $B_1(\mathbf{s}), \ldots, B_d(\mathbf{s})$ are basis functions defined over D, and d can be determined via the BIC criteria, see Section 3.3 for more details. In this article, we take G to be the logistic distribution function and $B_j(\mathbf{s})$ the B-spline basis for its popularity. Other choices are possible as we find our methods are insensitive to different choices of basis functions through unreported simulations. If we let $\eta_{\mathbf{s}} = G^{-1}(\pi(\mathbf{s}))$, then we can write $\pi(\mathbf{s})$ into

$$\pi(\mathbf{s}) = \frac{e^{\eta_{\mathbf{s}}}}{1+e^{\eta_{\mathbf{s}}}}.$$

For nonparametric form of f_1, we assume that f_1 is spatially invariant and belongs to a class of decreasing densities.

Although we only consider model (10) in our study, note that we can extend the above nonparametric model to include external covariates $Z_{\mathbf{s}} \in \mathbb{R}^{d_0}$ observed at each location $\mathbf{s}$ whenever appropriate. Specifically, we can consider the model

$$\eta_{\mathbf{s}} = \gamma' Z_{\mathbf{s}} + \sum_{j=1}^{d} \beta_j B_j(\mathbf{s}), \quad \gamma \in \mathbb{R}^{d_0}$$

if some useful covariates $Z_{\mathbf{s}}$ are available.

We propose an EM like algorithm to estimate $\pi(\mathbf{s})$ and the alternative distribution f_1. We first consider the case where f_1 is nonparametric and spatially invariant, then we describe in Section 3.4 how to adjust f_1 to allow it to vary with $\mathbf{s}$. Algorithm 1 provides the details of our iterative algorithm. As the EM algorithm is only guaranteed to converge to a local optimum, we shall try multiple initial points in the estimation and select the one that delivers the largest likelihood. Moreover, the optimizations of (11) and (12) in Algorithm 1 are nontrivial and will be elaborated in the following subsections.

Algorithm 1

0. Input the initial values $(\pi^{(0)}, f_1^{(0)})$.

1. **E-step:** Given $(\pi^{(t)}, f_1^{(t)})$, let

$$Q^{(t)}(\mathbf{s}) = \frac{\pi^{(t)}(\mathbf{s}) f_0(p_{\mathbf{s}})}{\pi^{(t)}(\mathbf{s}) f_0(p_{\mathbf{s}}) + (1-\pi^{(t)}(\mathbf{s})) f_1^{(t)}(p_{\mathbf{s}})}.$$

2. **M-step:** Given $Q^{(t)}(\mathbf{s})$, update (π, f_1) through $\pi^{(t+1)}(\mathbf{s}) = G(\sum_{j=1}^{d} \beta_j^{(t+1)} B_j(\mathbf{s}))$ with

$$(\beta_1^{(t+1)}, \ldots, \beta_d^{(t+1)}) = \underset{\beta_1,\ldots,\beta_d}{\operatorname{argmax}} \sum_{\mathbf{s}\in D} \{Q^{(t)}(\mathbf{s}) \log(\pi(\mathbf{s})) + (1 - Q^{(t)}(\mathbf{s})) \log(1-\pi(\mathbf{s}))\}, \tag{11}$$

and

$$f_1^{(t+1)} = \underset{f_1 \in \mathcal{H}}{\operatorname{argmax}} \sum_{\mathbf{s}\in D} (1 - Q^{(t)}(\mathbf{s})) \log f_1(p_{\mathbf{s}}). \tag{12}$$

3. Repeat the above E-step and M-step until the algorithm converges.

3.1. The Optimization Problem in (11)

When G is the logistic distribution function,

$$\log\left(\frac{\pi(\mathbf{s})}{1-\pi(\mathbf{s})}\right) = \sum_{j=1}^{d} \beta_j B_j(\mathbf{s}).$$

Problem (11) thus becomes

$$\max \sum_{\mathbf{s}\in D} \left\{ Q^{(t)}(\mathbf{s}) \log\left(\frac{\pi(\mathbf{s})}{1-\pi(\mathbf{s})}\right) + \log(1-\pi(\mathbf{s})) \right\}$$
$$= \max \sum_{\mathbf{s}\in D} \left\{ Q^{(t)}(\mathbf{s}) \sum_{j=1}^{d} \beta_j B_j(\mathbf{s}) + \log\left(\frac{1}{e^{\sum_{j=1}^{d} \beta_j B_j(\mathbf{s})} + 1}\right) \right\}.$$

To solve this problem and problem (13), we use the Nelder–Mead optimization, a numerical method that has been widely used to find the maximum of an objective function in a multidimensional space. At each optimization step, it reshapes the simplex, one vertex at a time toward an optimal region in the search space. We refer to Singer and Singer (1999) for the details about its computational complexity in different situations.

3.2. The Optimization Problem in (12)

Note that $\mathcal{H}$ in (12) denotes the space of density functions over which we search for the optimal f_1. When $\mathcal{H}$ is the class of decreasing densities, the optimization in (12) could be accomplished by a series of steps.

Step 0: Order the p-values from the smallest to the largest and denote the sorted p-values by $z_1 \le z_2 \le \cdots \le z_{|D|}$, where $|D|$ is the cardinality of the set D. If $p_{\mathbf{s}}$ corresponds to z_i, we denote $Q_i^{(t)} = Q^{(t)}(\mathbf{s})$.

Step 1: The objective function in (12) only looks at the values of f_1 at z_i. The objective function increases if the values $f_1(z_i)$ increases, and the value of f_1 at (z_{i-1}, z_i) has no impact on the objective function (where $z_0 = 0$). Therefore, if f maximizes the objective function, it must be a constant on $(z_{i-1}, z_i]$.

Step 2: Let $y_i = f_1(z_i)$. We only need to find y_i which maximizes

$$\sum_{i=1}^{|D|} (1 - Q_i^{(t)}) \log(y_i),$$

subject to $y_1 \ge y_2 \ge \cdots \ge y_{|D|} \ge 0$ and $\sum_{i=1}^{|D|} y_i (z_i - z_{i-1}) = 1$. It can be formulated as a convex programming problem.

Step 3: Write $Q^{(t)} = \sum_{i=1}^{|D|} (1 - Q_i^{(t)})$. Consider the problem:

$$\min \sum_{i=1}^{|D|} \left\{ -(1 - Q_i^{(t)}) \log(y_i) + Q^{(t)} y_i (z_i - z_{i-1}) \right\}.$$

The solution is given by $\hat{y}_i = \frac{1 - Q_i^{(t)}}{Q^{(t)}(z_i - z_{i-1})}$, which satisfies the constraint $\sum_{i=1}^{|D|} y_i (z_i - z_{i-1}) = 1$ in Step 2.

Step 4: We rewrite the problem in Step 3 as

$$\min \sum_{i=1}^{|D|} (1 - Q_i^{(t)}) \left\{ -\log(y_i) + \frac{Q^{(t)}(z_i - z_{i-1})}{(1 - Q_i^{(t)})} y_i \right\}.$$

Let

$$(\hat{u}_1, \ldots, \hat{u}_{|D|}) = \operatorname{argmin} \sum_{i=1}^{|D|} (1 - Q_i^{(t)}) \left(\frac{Q^{(t)}(z_i - z_{i-1})}{(1 - Q_i^{(t)})} + u_i \right)^2$$

subject to $u_1 \ge u_2 \ge \cdots \ge u_{|D|}$. The solution is given by

$$\min_{a \le i} \max_{b \ge i} \frac{-Q^{(t)} \sum_{j=a}^{b} (z_j - z_{j-1})}{\sum_{j=a}^{b} (1 - Q_j^{(t)})},$$

which can be obtained using the Pool-Adjacent-Violators Algorithm (PAVA). Finally, we obtain the solution to the original problem (12),

$$\tilde{y}_i = -\frac{1}{\hat{u}_i},$$

according to Theorem 3.1 of Barlow and Brunk (1972). It is worth mentioning that PAVA has a computational complexity that is linear in the sample size $|D|$ if skillfully implemented.

3.3. The Choice of d

The number of basis functions d in (10) can be chosen through the BIC criteria. Specifically, the BIC is defined as $-2\log\mathcal{L}+(d+m)\log(|D|)$, where $\mathcal{L}$ denotes the likelihood, m is the degrees of freedom associated with the parameters (other than β_j's) in the model, and $|D|$ is the cardinality of the spatial domain. When f_1 is estimated via the method in Section 3.2, the calculation of m involves the degrees of freedom of the isotonic regression, which turns out to be the number of joined pieces of $\{\hat{u}_i\}_{i=1}^{|D|}$ (Meyer and Woodroofe 2000, Proposition 1).

3.4. Spatially Varying Alternative Distributions

The nonparametric form of f_1 assumes f_1 to be spatially invariant. This assumption might be too restrictive in some situation (e.g., when the strength of signals is varying across locations) and needs to be relaxed.

Generally speaking, if the alternative distribution is allowed to vary with $\mathbf{s}$, model (1) will no longer be estimable unless some extra structural assumption is imposed, because there is only one observation to make inference for $f_1(\cdot;\mathbf{s})$. However, if we assume that $f_1(\cdot;\mathbf{s})$ varies smoothly over $\mathbf{s}$, then one can borrow information from neighboring observations to estimate $f_1(\cdot;\mathbf{s})$. Below we propose a semiparametric model for $f_1(\cdot;\mathbf{s})$ to address this issue.

To capture the heterogeneity of signals, we consider the exponential family of densities on $[0,1]$ given by

$$f_1(p;\mathbf{s}) = \tilde{f}_1(p)\exp\left(\beta_{0,\mathbf{s}} + \mathcal{T}(p)\sum_{j=1}^{d'}\gamma_j B_j(\mathbf{s})\right), \quad p\in[0,1],$$

where $\tilde{f}_1$ is a carrier density function (Efron and Tibshirani 1996) on $[0,1]$ that is independent of $\mathbf{s}$, $\mathcal{T}(p):[0,1]\to\mathbb{R}$ is a prespecified function of p, and

$$\beta_{0,\mathbf{s}} = -\log\left\{\int_0^1 \tilde{f}_1(p)\exp\left(\mathcal{T}(p)\sum_{j=1}^{d'}\gamma_j B_j(\mathbf{s})\right)dp\right\}$$

to ensure that $\int_0^1 f_1(p;\mathbf{s})dp = 1$. Note $f_1(\cdot;\mathbf{s})$ is determined by both $\tilde{f}_1$ and γ_j, and it is spatially varying due to its dependence on the basis functions $\{B_j(\mathbf{s})\}$. We let $\tilde{f}_1$ be the solution from Algorithm 1, that is, the nonparametric form of f_1, and then estimate γ_j by replacing (12) in the M-step of our proposed EM-algorithm by

$$\begin{aligned}\{\gamma_j^{(t+1)}\} = \underset{\gamma_j^{(t+1)}}{\operatorname{argmin}} &-\sum_{\mathbf{s}\in D}(1-Q^{(t)}(\mathbf{s}))\\ &\times\left[-\log\left\{\int_0^1\tilde{f}_1(p)\exp\left(\mathcal{T}(p)\sum_{j=1}^{d'}\gamma_j B_j(\mathbf{s})\right)dp\right\}\right.\\ &\left.+\mathcal{T}(p_{\mathbf{s}})\sum_{j=1}^{d'}\gamma_j B_j(\mathbf{s})\right].\end{aligned} \tag{13}$$

Unreported numerical studies show that our FDR procedure is insensitive to the choice of $\mathcal{T}(p)$. Thus, we choose $\mathcal{T}(p)=p$ for parsimony. Again the number of basis functions d' can be selected using the BIC criteria.

4. Monte Carlo Experiments

We conduct extensive simulations to evaluate the FDR control as well as the power of our method under spatial dependency of various types and strengths. Two examples of hypothesis testing often highly demanded are considered, one for spatial mean comparison and the other for covariance comparison. We also present the results by implementing the classical methods in Benjamini and Hochberg (1995) (BH) and Storey (2002) (JS), and a recent method in Tansey et al. (2018) (FDRS) as references. All tests target FDR at 0.1 and all results are based on 100 simulation runs. Also, all spatial or spatiotemporal processes in the simulation are generated over a 42×42 equidistant grids ($N=1764$) over a spatial domain of $[0,1]\times[0,1]$. Although our proposed method is applicable to randomly generated locations as readily as to grids, we only present the simulation results of spatiotemporal processes over equidistant grids to match with the climate data in Section 5. The simulation results of randomly generated locations show very similar pattern as with the gridded data and are deferred to the supplementary materials.

All location wise p-values in the simulation were derived using the self-normalization test developed by Shao (2010). The self-normalization approach has been shown to provide more accurate null p-values for time series data compared to traditional approaches, and thus generally leads to better FDR control as observed in our unreported study. The details of the implementation of the self-normalization test is described in the supplementary materials.

4.1. Test for Spatial Mean

Consider two spatiotemporal fields $X(\mathbf{s},t)$ and $Y(\mathbf{s},t)$ with spatial mean function $\mu_X(\mathbf{s})$ and $\mu_Y(\mathbf{s})$. Our goal is to identify the locations where the spatial mean of the two spatiotemporal random fields are different after taking the spatial correlation into account. This is equivalent to letting $\theta_X(\mathbf{s})=\mu_X(\mathbf{s})$ and $\theta_Y(\mathbf{s})=\mu_Y(\mathbf{s})$, and testing

$$H_{0,\mathbf{s}}:\mu_X(\mathbf{s})=\mu_Y(\mathbf{s}) \quad \text{versus} \quad H_{a,\mathbf{s}}:\mu_X(\mathbf{s})\neq\mu_Y(\mathbf{s}). \tag{14}$$

We consider three different types of spatial covariance structure in our simulation experiments:

(M1) Spatially independent model

$$\begin{aligned}X(\mathbf{s},t) &= D_X(\mathbf{s},t)+e_X(\mathbf{s},t),\\ Y(\mathbf{s},t) &= \mu(\mathbf{s})\delta(\mathbf{s})+D_Y(\mathbf{s},t)+e_Y(\mathbf{s},t),\end{aligned}$$

where $D_X(\mathbf{s},t)=\rho D_X(\mathbf{s},t-1)+\sqrt{1-\rho^2}\varepsilon(\mathbf{s},t)$ are temporally correlated errors with $\rho=0.3$ and $\varepsilon(\mathbf{s},t)\sim$ iid $N(0,1)$, $e_X(\mathbf{s},t)\sim$ iid $N(0,1)$, and $D_Y(\mathbf{s},t)$ and $e_Y(\mathbf{s},t)$ are defined likewise. The mean of $X(\mathbf{s},t)$ is 0 while the mean of $Y(\mathbf{s},t)$ is $\mu(\mathbf{s})\delta(\mathbf{s})$ where $\mu(\mathbf{s})=\sum_{j=1}^{d}\xi_j B_j(\mathbf{s})$ and $\xi_j=\xi$ for some constant ξ, and $\delta(\mathbf{s})$ is an indicator function taking the value 0 under the null and 1 under the alternative hypothesis.

The indicator function $\delta(\mathbf{s})$ is generated from Bernoulli$(1-\pi(\mathbf{s}))$ where $\pi(\mathbf{s})$ is generated following Equation (10). All notations defined for this setting will be extended to the next two sets of models.

(M2) Spatially correlated model with exponential covariance function

$$\begin{aligned}X(\mathbf{s},t) &= U_X(\mathbf{s},t)+D_X(\mathbf{s},t),\\ Y(\mathbf{s},t) &= \mu(\mathbf{s})\delta(\mathbf{s})+U_Y(\mathbf{s},t)+D_Y(\mathbf{s},t),\end{aligned}$$

where $U_X(\mathbf{s}, t)$ at each time t is a Gaussian random process with mean 0 and covariance matrix Σ_U governed by an exponential covariance function $\sigma_u^2 \exp(-||\mathbf{h}||/\phi)$ for a spatial lag $\mathbf{h}$, and $U_Y(\mathbf{s}, t)$ is defined likewise. Here we fix $\sigma_u^2 = 0.2$ and its range parameter $\phi = 0.1$.

(M3) Spatial low rank model

$$X(\mathbf{s}, t) = \sum_{j=1}^{d} R_j^X(t) B_j(\mathbf{s}) + D_X(\mathbf{s}, t),$$

$$Y(\mathbf{s}, t) = \mu(\mathbf{s})\delta(\mathbf{s}) + \sum_{j=1}^{d} R_j^Y(t) B_j(\mathbf{s}) + D_Y(\mathbf{s}, t),$$

where $R_j^X(t)$ and $R_j^Y(t)$ both follow $N(0, \sigma_r^2)$. Since B-spline will be used in the estimation, here we choose a different set of basis functions $B(\mathbf{s}) = \{B_1(\mathbf{s}), \ldots B_d(\mathbf{s})\}$ in the data generation:

$$B(\mathbf{s}) = \{1, s_1, s_2, s_1 s_2, s_1^2, s_2^2, s_1^3, s_2^3, s_1^2 s_2, s_1 s_2^2, s_1^2 s_2^2\},$$
$$\mathbf{s} = (s_1, s_2) \in D.$$

Model (M3) defines a much slower decaying spatial covariance structure compared to model (M2).

Figure 1 shows the spatial mean of $Y(\mathbf{s}, t)$ based on all three models with $\xi = 0.25$ and the null proportion being 70%. The black dots indicate the locations under the alternative hypothesis ($\delta(\mathbf{s}) = 1$). Figures 2 and 3 report the FDR and power of our method together with BH, JS, and FDRS under models (M1), (M2) , (M3-1), and (M3-2) where (M3-1) and (M3-2) are (M3) with $\sigma_r = 0, 3$ and $\sigma_r = 0.5$, respectively. To study the effect of null proportion, the size of mean difference and the spatial correlation strength on the FDR control and power, we repeat simulations with a sequence of null proportions (70%, 80%, and 90%), ξ (0.15, 0.2, 0.25, 0.3) and σ_r values (0.3, 0.5). The null proportion is controlled through varying β_j in (10). A larger ξ generally represents more significant mean difference, and a larger σ_r in model (M3) indicates stronger spatial correlation.

Figure 2 shows that for both models (M1) and (M2) and for two (M3) models with different σ_r values, our mirror procedures control the FDR well at each ξ value, if the null proportion is low. However, when the null proportion is high, our methods tend to inflate FDR if meantime the ξ value is small. The inflation is exacerbated at a larger σ_r. This implies that when there is only a little data to estimate the alternative distribution of p-values and meantime the signal is weak, it is challenging for the mirror procedure to control the FDR. Furthermore, the strong dependency in data makes the FDR control even harder. The semiparametric mirror seems more vulnerable to the above mentioned challenges because the spatially varying semiparametric model requires more information to estimate the model. Overall, the JS method controls the FDR well. BH appears to be conservative by being below the nominal level in most cases. In contrast, FDRS shows inflated FDR for low null proportion combined with weak signal while conservative FDR for high null proportion. The power comparison in Figure 3 shows that in general, all powers increase as ξ increases or the null proportion decreases. This is not surprising as stronger signals or more data available for estimating $f_1(\cdot, \mathbf{s})$ will make the identification of alternatives easier. However, an important

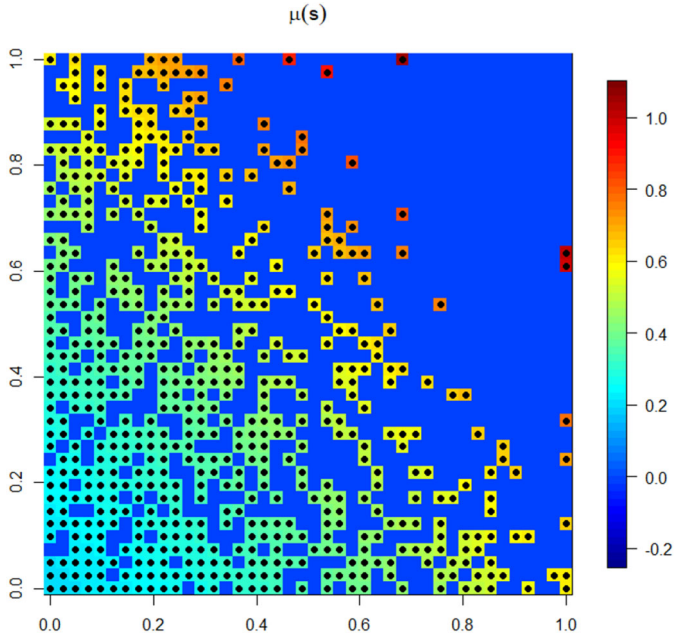

Figure 1. Spatial mean of $Y(\mathbf{s}, t)$. The black dots indicate the locations under the alternative hypothesis.

message we obtain here is that across all values of ξ and null proportions, our methods are uniformly more powerful than the BH, JS, and FDRS methods. The improvement becomes more significant when the signal is weaker (smaller ξ) or the null proportion is higher.

To investigate how the assumption of f_1 being spatially invariant in the nonparametric mirror procedure possibly affects the performance of the multiple testing as opposed to the flexible semiparametric procedure with spatially varying f_1, we consider a simulation experiment where the mean function $\mu(\mathbf{s})\delta(\mathbf{s})$ in Y exhibits more spatial heterogeneity. Unreported results show that even in the scenario that obviously favors the semiparametric method, modeling heterogeneity, that is, allowing f_1 to vary spatially, however, does not lead to more discoveries. In most cases, the nonparametric method appears to be as powerful as the semiparametric method if not more powerful. The nonparametric method also shows a great advantage over the semiparametric one in terms of computational complexity.

The estimation algorithm for the semiparametric procedure involves the Nelder–Mead optimization to estimate $(\gamma_1, \ldots, \gamma_{d'})$ in addition to $(\beta_1, \ldots, \beta_d)$. The Nelder–Mead method optimizes the objective function in multidimensional space by reshaping the simplex toward the optimal region in a search space. Both the number of searches for the simplex and the time it takes for each single simplex search for $\hat{\gamma}$ turn to be much greater than those for $\hat{\beta}$, due to the intricate objective function for estimating γ. It takes approximately 1 sec for the nonparametric procedure versus 1.5 hr for the semiparametric procedure to run one simulation based on an i7 core processor with 8 Gb of RAM.

For the above mentioned two reasons, we recommend to perform the nonparametric mirror procedure for general purposes, especially when comparing large datasets such as climate data.

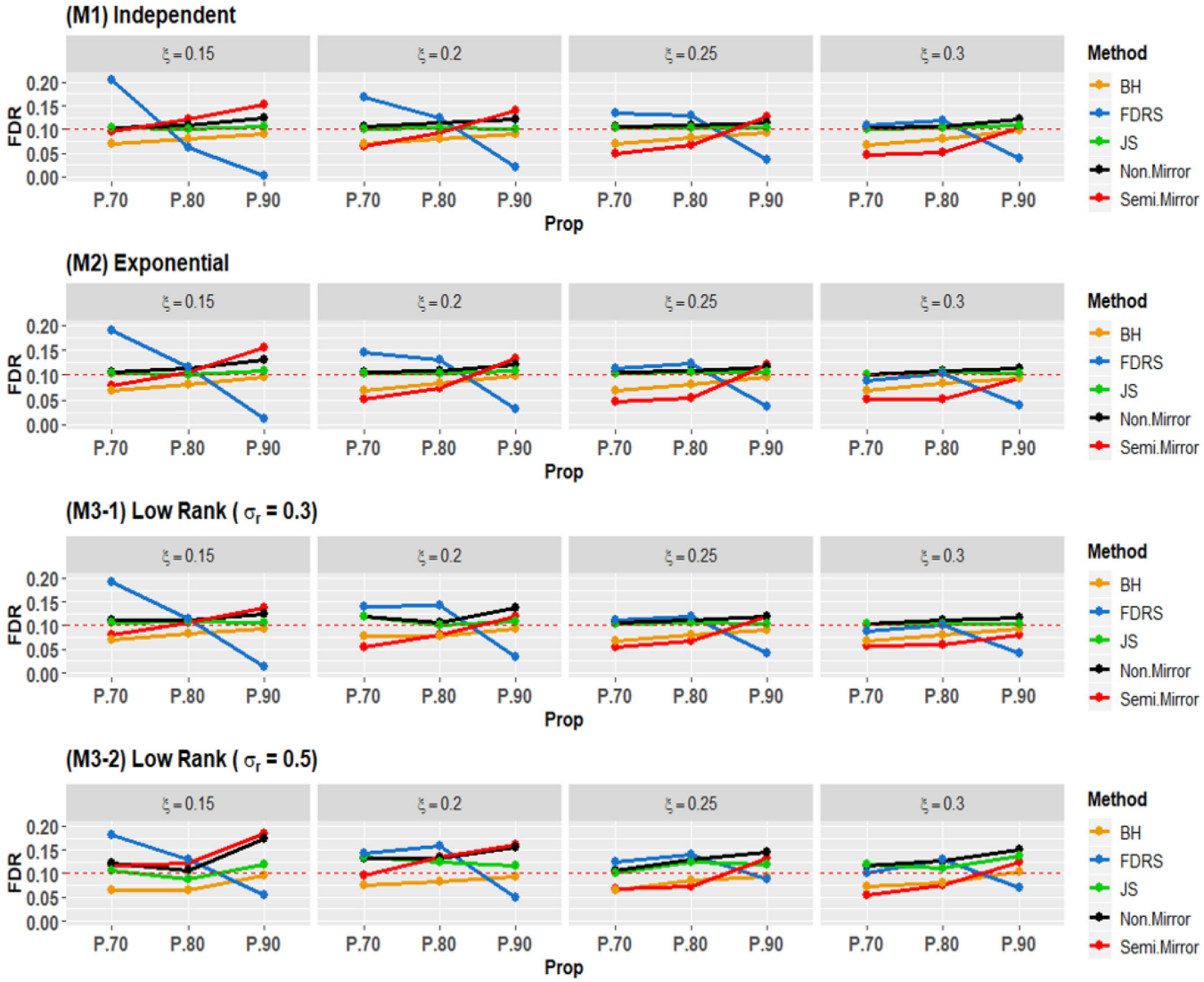

Figure 2. FDR for mean comparison between two spatiotemporal fields. *P*.70, *P*.80, and *P*.90 indicate the null proportions being 70%, 80%, and 90%, respectively.

4.2. Test for Spatial Covariance

Motivated by comparing the teleconnection pattern between different climate fields in Section 5, we evaluate our methods in comparing the spatial covariance structure between two spatiotemporal random fields. The two random fields $X(\mathbf{s}, t)$ and $Y(\mathbf{s}, t)$ are defined again with either an exponential covariance function or a low rank model to reflect different dependence structures. Since $X(\mathbf{s}, t)$ in this section are defined exactly the same as their corresponding models in (M2) and (M3) in Section 4.1, only the models for $Y(\mathbf{s}, t)$ are specified below:

(C1) Spatially correlated model with exponential covariance function

$$Y(\mathbf{s}, t) = K_1(\mathbf{s})U_Y(\mathbf{s}, t) + D_Y(\mathbf{s}, t).$$

(C2) Spatial low rank model

$$Y(\mathbf{s}, t) = K_2(\mathbf{s}) \sum_{j=1}^{d} R_j^Y(t)B_j(\mathbf{s}) + D_Y(\mathbf{s}, t).$$

In both models, the spatially varying process $K_i(\mathbf{s}) = \nu_i(\mathbf{s})\delta(\mathbf{s}) + (1 - \delta(\mathbf{s}))$ where $\nu_i(\mathbf{s}) = \xi_i \sum_{j=1}^{d} B_j(\mathbf{s})$ for constants ξ_i, $i = 1, 2$. We set the parameters for $U_Y(\mathbf{s}, t)$ in (C1) as $\sigma_u^2 = 1.3$ and $\phi = 0.25$. Other notations in (C1) and (C2) are directly adopted from Section 4.1.

Because the teleconnection is usually calculated as the covariance between a regional averaged time series and local time series of other locations, we conduct our simulation in the following way to mimic the real situation. First, the spatial dependency structure of interest is represented by the covariance between observations at a fixed location $\mathbf{s}_0 \in D$ and every other location $\mathbf{s} \in D, \mathbf{s} \neq \mathbf{s}_0$. Second, we define $X(\mathbf{s}_0, t)$ and $Y(\mathbf{s}_0, t)$ as a denoised version of $X(\mathbf{s}, t)$ and $Y(\mathbf{s}, t)$ by removing their respective $D_X(\mathbf{s}, t)$ and $D_Y(\mathbf{s}, t)$ to mimic the regional average in the teleconnection calculation. Then, we let $\theta_X(\mathbf{s}) = C_X(\mathbf{s}, \mathbf{s}_0)$ ($\theta_Y(\mathbf{s}) = C_Y(\mathbf{s}, \mathbf{s}_0)$) and consider the test:

$$\begin{aligned} H_{0,\mathbf{s}} &: C_X(\mathbf{s}, \mathbf{s}_0) = C_Y(\mathbf{s}, \mathbf{s}_0) \quad \text{versus} \\ H_{a,\mathbf{s}} &: C_X(\mathbf{s}, \mathbf{s}_0) \neq C_Y(\mathbf{s}, \mathbf{s}_0), \end{aligned} \tag{15}$$

where $C_X(\mathbf{s}, \mathbf{s}_0)$ is the covariance between $X(\mathbf{s}, t)$ and $X(\mathbf{s}_0, t)$ and $C_Y(\mathbf{s}, \mathbf{s}_0)$ is defined similarly. The alternative holds wherever $K_i(\mathbf{s}) \neq 1$ given that $K_i(\mathbf{s}_0) = 1$. We set $0 < \nu_i(\mathbf{s}) \leq 1$ so

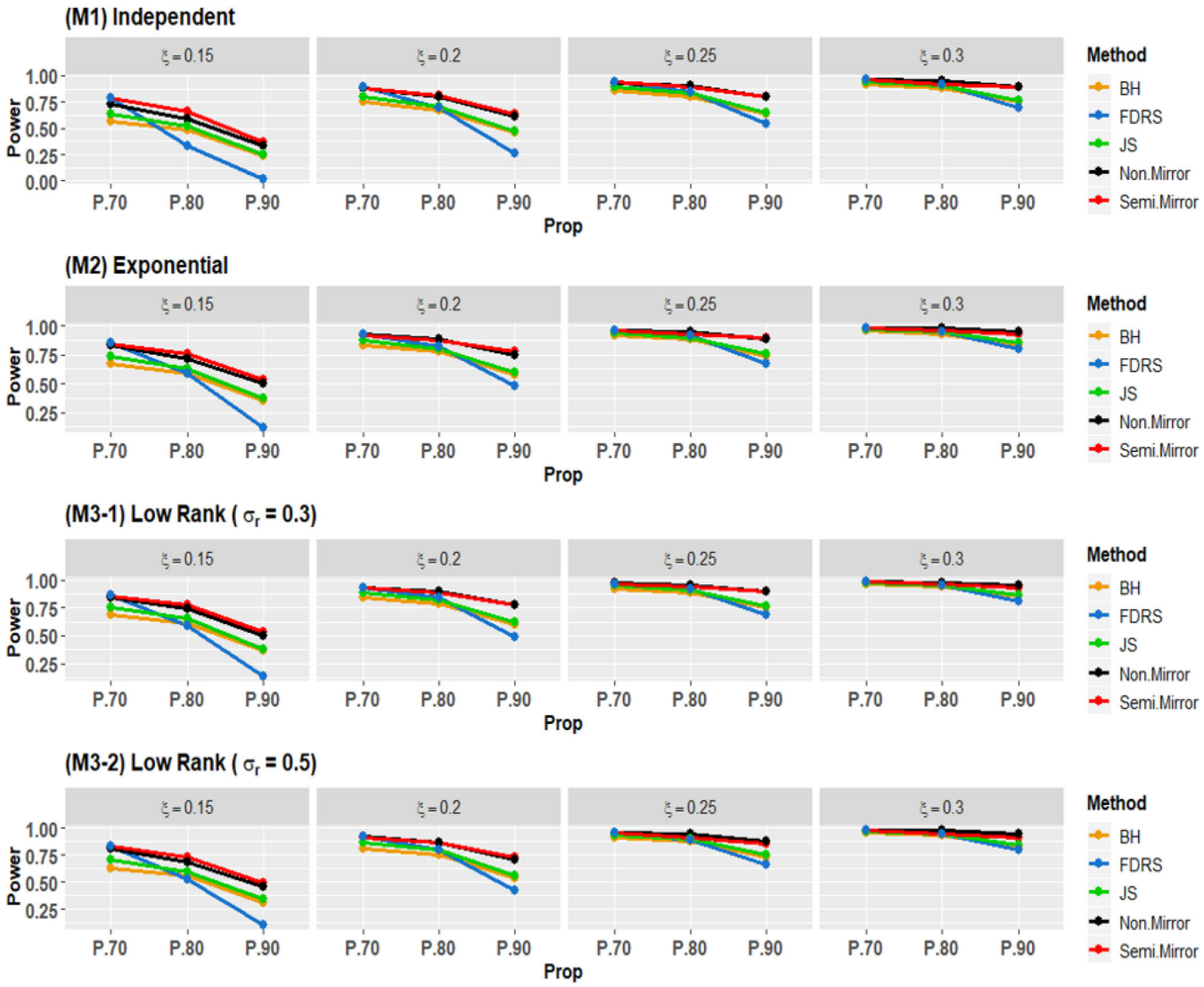

Figure 3. Power for mean comparison between two spatiotemporal fields. *P*.70, *P*.80, and *P*.90 indicate the null proportions being 70%, 80%, and 90%, respectively.

a smaller ξ_i makes $v_i(\mathbf{s})$ further apart from 1 and thus creates stronger alternatives (signal).

Under the exponential covariance (C1) structure, when the distance between an alternative location and the anchored point is beyond a certain value, the correlation and thus the covariance of the observations at those two locations, even in Y, will decay to almost zero, so the H_0 actually holds at that alternative location.

Therefore, we propose to use the practical range for exponential covariance function as the threshold to differentiate whether a selected alternative location should be considered as alternative or null. The practical range is defined as the distance above which the correlation decays to less than 0.05 and is practically nonexistent (Montero, Fernández-Avilés, and Mateu 2015). Figure 4 illustrates this idea. It shows the empirical $C_X(\mathbf{s}, \mathbf{s}_0)$ and $C_Y(\mathbf{s}, \mathbf{s}_0)$ of one simulation run based on the (C1) model with $\rho = 0.3$, $\xi_1 = 0.15$, and the null proportion being 62%. The black dots indicate the actual alternative locations after truncation by the practical range and the red dots are the truncated locations.

Figures 5 and 6 evaluate the performance of covariance comparison between two spatiotemporal random fields governed by either (C1) or (C2), using our nonparametric mirror procedure, BH, and JS methods. Because the covariance comparison essentially transforms to a mean comparison after conducting pairwise product of the random field to empirically estimate the covariance function, we omit the semiparametric method that has shown no obvious advantages and the FDRS which shows unsatisfactory FDR control in the mean comparison for our simulated data. The null proportions are set to be 50%, 56%, and 62% for (C1) model, and 35%, 45%, and 50% for (C2) model. The higher null proportion with (C1) model is due to the truncation of alternative locations which elevates the originally set null proportions. A sequence of $\xi_1 = 0.08, 0.10, 0.12, 0.15$ and $\xi_2 = 0.15, 0.2, 0.25, 0.3$ are tried out to represent strong to weak signals in the simulated data for (C1) and (C2), respectively.

The nonparametric method is superior over other methods although its FDR appears to be slightly inflated at the combination of a weaker signal and a large null proportion. Moreover, the nonparametric method exhibits greater power than BH and

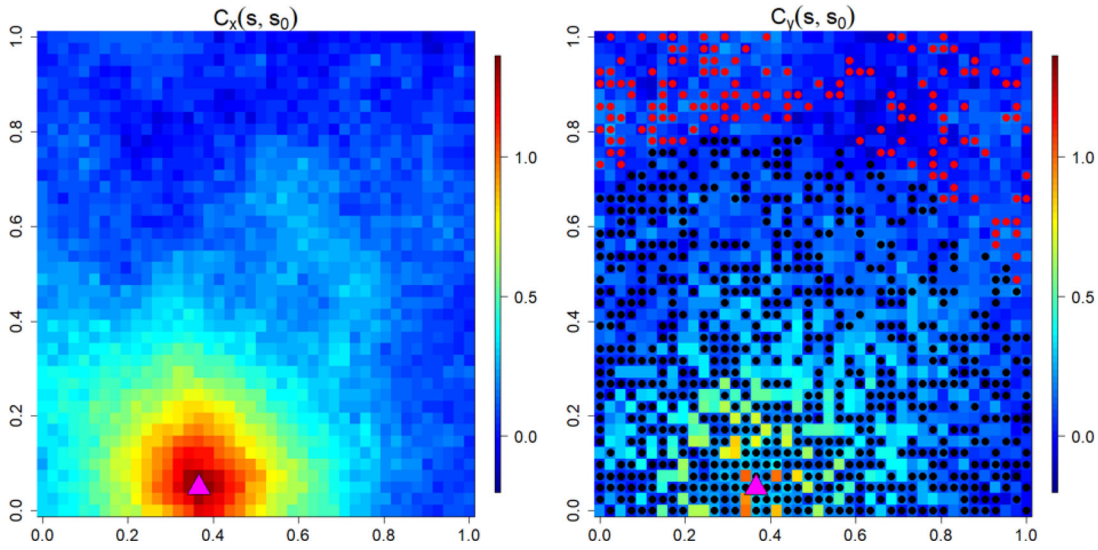

Figure 4. Comparison between $C_X(\mathbf{s}, \mathbf{s}_0)$ and $C_Y(\mathbf{s}, \mathbf{s}_0)$. The black dots and red dots together represent locations with $\delta(\mathbf{s}) = 1$, but only the black dots are within the practical range of the fixed point and considered as alternative locations. The red dots are considered as null locations. The pink triangle marks where $\mathbf{s}_0$ is located.

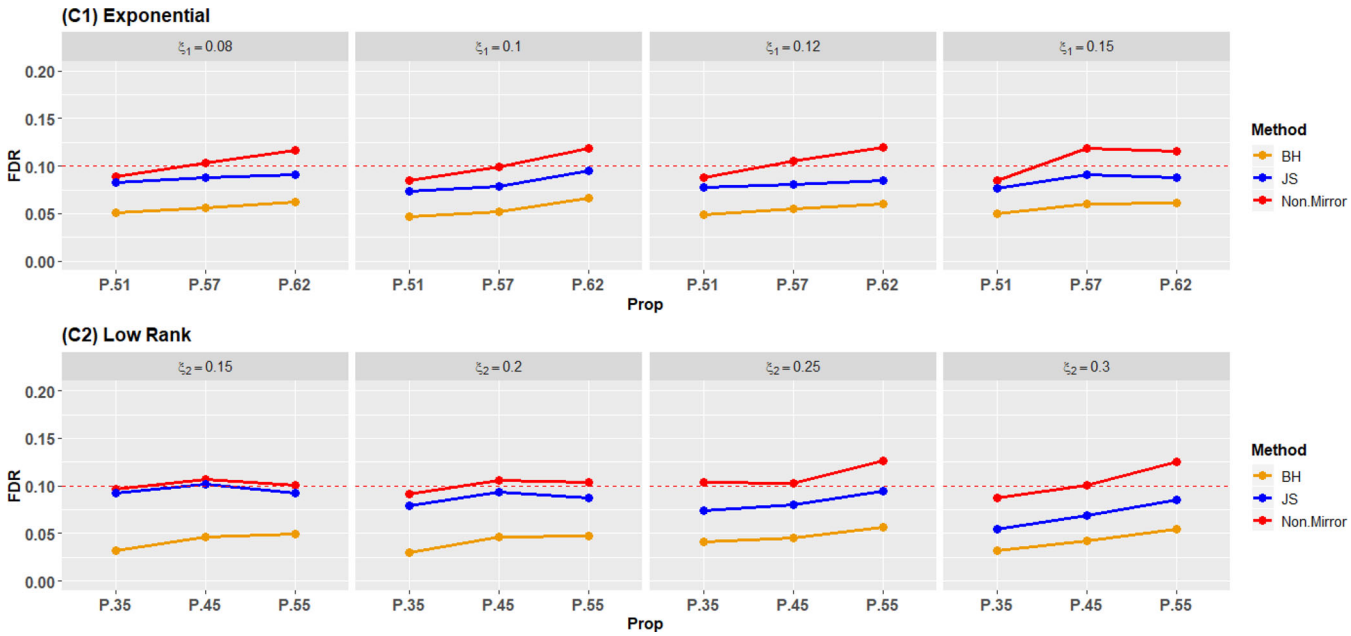

Figure 5. FDR of covariance comparison between two spatiotemporal random fields. *P*.50, *P*.56, and *P*.62 indicate the null proportions being 50%, 56%, and 62%, respectively, for (C1) model; and *P*.35, *P*.45, and *P*.55 indicate the null proportions being 35%, 45%, and 55%, respectively, for (C2) model.

JS methods, especially when the signal is weaker (larger ξ) in both (C1) and (C2) models.

In principal, both the mean and covariance comparisons will be affected by the spatial correlation, because the variability associate with spatial correlation can likely be entangled with the mean surface changing and thus makes it challenging to estimate the spatial mean. This situation could get worse with the low rank model that induces a stronger dependency than the weakly dependent model such as the exponential covariance model. As a consequence, we indeed observe a slight inflation of FDR in the mean comparison when σ_r is large, but the inflation seems very limited. In our simulations, the exponential covariance function has range parameter $\phi = 0.25$, which represents a fairly strongly correlated random field given the spatial domain is a unit length square. The low rank model generates even stronger spatial correlation. The minimum spatial correlation strength induced by the random effects is 0.32 if $\sigma_r = 0.3$ and 0.35 if $\sigma_r = 0.5$. Most real environmental data including the climate data in Section 5 likely exhibit weaker spatial correlation than our simulated data. Despite the more challenging context we set in the simulation, our results seem to indicate that the FDR and power are robust to the existence of spatial correlation using the independent model as a baseline. This suggests that our methods are valid within the realistic spatial correlation strength.

Nevertheless, we still have tried several decorrelation approaches including spectral decomposition of covariance matrix, jointly estimating means and random effect coefficients

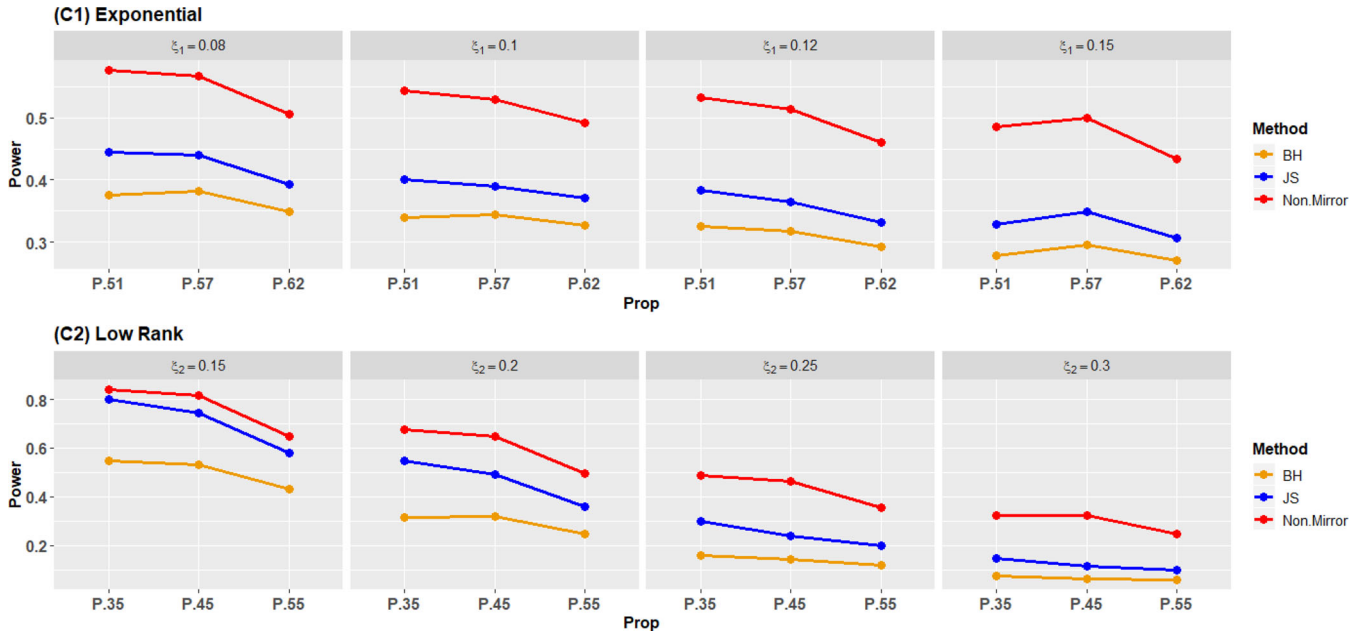

Figure 6. Power of covariance comparison between two spatiotemporal random fields. *P*.50, *P*.56, and *P*.62 indicate the null proportions being 50%, 56%, and 62%, respectively, for (C1) model; and *P*.35, *P*.45, and *P*.55 indicate the null proportions being 35%, 45%, and 55%, respectively, for (C2) model.

with lasso penalty, and Fan, Han, and Gu's (2012) method to filter out the signal in presence of random effect (see details in Appendix A). However, we find all those attempts fail to improve the FDR and power of the test, perhaps due to the fact that those decorrelation approaches can introduce additional uncertainty in the estimation procedures. Therefore, we decide to opt out of the decorrelation.

5. Application With Climate Data

Paleoclimate reconstruction plays a central role in understanding the long term dynamics of climate system because the observed temperatures at global scale are only available since 1850. Several different methods for global temperature reconstruction have been developed yet how to evaluate those space-time reconstructions are still open questions. A widely accepted platform for evaluating reconstructions is through the pseudoproxy experiment (Smerdon 2012; Smerdon, Coats, and Ault 2016) which reconstructs the past temperatures based on the synthetic climate from climate models in the calibration period. Attempts to better understand the spatial reconstructions has been an emerging focus in the literatures (Li and Smerdon 2012; Dannenberg and Wise 2013; Evans et al. 2014; Steiger et al. 2014; Wang et al. 2014; Smerdon, Coats, and Ault 2016). However, most of the comparing methods focus on the visual diagnostics of differences between the reconstruction and modeled data. Li and Smerdon (2012) and Li, Zhang, and Smerdon (2016) pioneered formal statistical comparisons in this application, but their methods only compared the two climate fields globally and could not detect where the differences occur.

We now illustrate how to use our proposed method to identify the locations where the mean and teleconnection are different between the reconstruction and the targeted climate field. We choose a popular reconstruction method developed by Mann et al. (2009) that combines the regularized expectation maximization with truncated total least squares (TTLH hereafter), and take synthetic climate from the National Center for Atmospheric Research Community Climate System version 4 model (CCSM). The TTLH reconstruction based on this modeled climate is available at *http://www.realclimate.org/index.php/data-sources/#GCM_output*. Both the climate model data and the reconstruction are presented on annual scale from 850 to 1850 CE and at each year on a five degree latitude and longitude grid (1732 grids in total). Both modeled and reconstructed climate shows an upward trend in time. Following Li, Zhang, and Smerdon (2016), we remove the temporal trend of both climate fields to attain an approximate temporal stationarity.

In the data application, we again obtain location specific *p*-values using the self-normalization test as in Shao (2010), then apply our estimation procedure and the FDR decision rule to identify the locations where the alternative hypothesis holds at the nominal FDR level 0.1. Because we have demonstrated that the nonparametric mirror procedure is a powerful and robust method, we apply this multiple testing method to our data.

5.1. Mean Comparison

We first evaluate TTLH reconstruction by identifying where its spatial mean is different from that of CCSM climate data. If we let $X(\mathbf{s}, t)$ denote the CCSM climate data and $Y(\mathbf{s}, t)$ denote the TTLH reconstruction, our question can be answered by performing the hypothesis test (14).

Figure 7 displays the mean surface of CCSM (left) and TTLH reconstruction (right). The black dots label the locations where the mean is different under the nonparametric mirror procedure. Smerdon (2012) compared the reconstruction to its targeted CCSM climate and reported the regions with large reconstruction bias. Our results are consistent with

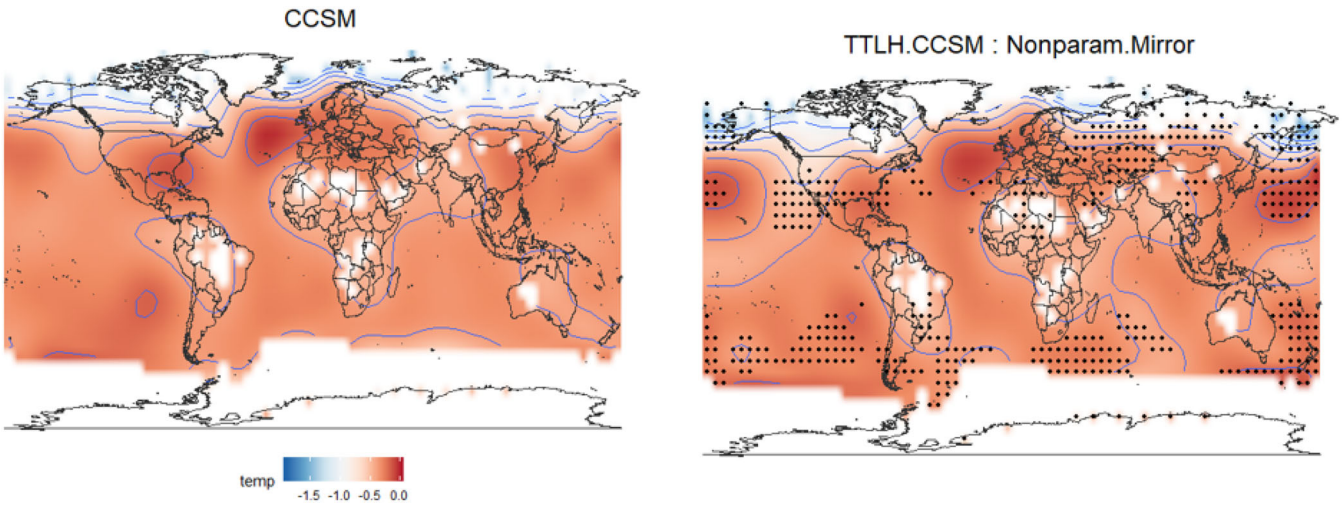

Figure 7. Mean comparison between CCSM and its reconstruction using TTLH. The left plot shows the spatial mean of CCSM climate field and the right plot shows the spatial mean of TTLH reconstruction together with the testing results of nonparametric mirror procedure. The black dots indicate where the mean is different.

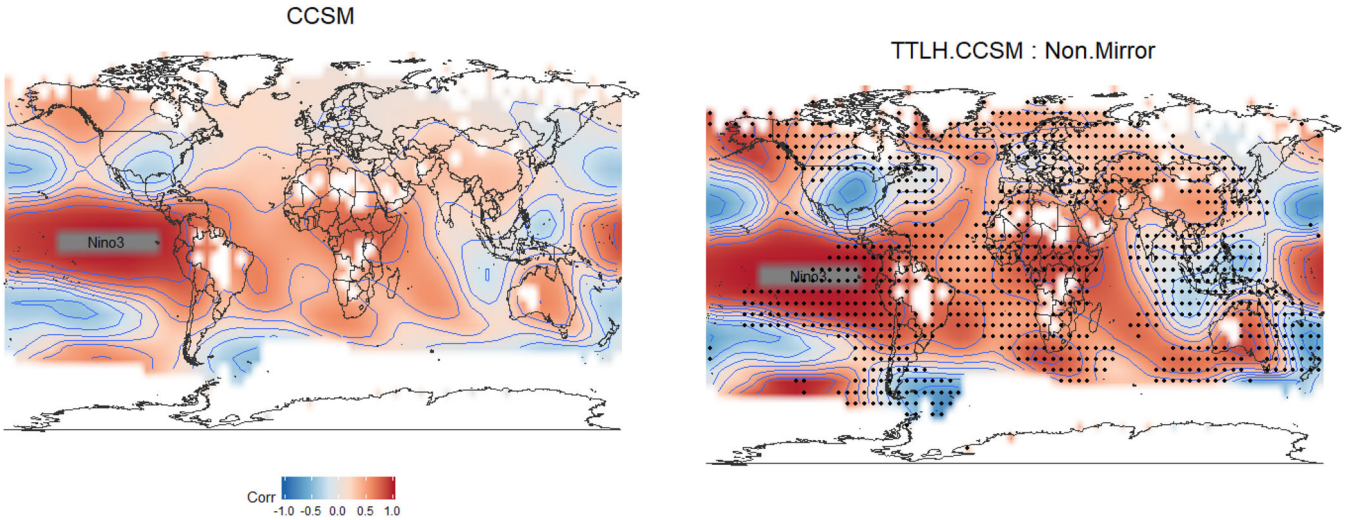

Figure 8. ENSO teleconnection comparison between CCSM and its TTLH reconstruction. The plots show the correlation between the averaged temperature over Nino3 region and local temperatures at each other location with CCSM (left) and TTLH (right). The black dots indicate where the teleconnection strength is different.

Smerdon (2012) in that the locations with substantial bias are largely identified as alternative locations.

5.2. Teleconnection Comparison

El Niño and the Southern Oscillation, also known as ENSO teleconnection, represent a major mode of climate dynamics. It is very important to study the ENSO teleconnection pattern because the presence of El Niño can greatly affect the normal weather conditions in many regions of the world and lead to extreme temperature changes causing socio-economic problems. ENSO occurs when there is a periodic fluctuation in sea surface temperatures and this fluctuation exerts impact on remote regions, affecting extreme weather events worldwide. The ENSO teleconnection is usually measured by the correlation between the averaged sea surface temperature over the tropical region (Nino3, 5N-5S, 150W-90W) and the local temperatures everywhere else. We now evaluate the TTLH by comparing its ENSO teleconnection pattern to that of CCSM. If we define $X(\mathbf{s}_0, t)$ and $Y(\mathbf{s}_0, t)$ as the averaged temperature over the locations in tropical region in their own field, the comparison of teleconnection pattern is equivalent to performing the hypothesis (15).

Figure 8 presents the ENSO covariance structure of CCSM (left) and TTLH (right).

The black dots indicate the locations where the null hypothesis is rejected under our FDR level.

There has been little work to rigorously compare the ENSO teleconnection pattern between the reconstruction and its targeted climate. However, it has been discussed in many papers (Yang and DelSole 2012; Coats et al. 2013) that the teleconnections to North America, Northern Atlantic sea surface, Southern Africa, Indian subcontinent, and western Europe temperature are often preserved well by reconstructions. Our Figure 8 also shows that TTLH reconstruction seems to reproduce the teleconnection to Northern Atlantic ocean and Northern America well, while missing the teleconnection to Southern Africa, Indian subcontinent, and western Europe.

6. Conclusion and Discussion

Motivated by comparing different climate fields, we develop a multiple testing approach to detect the local difference in the characteristics of two spatiotemporal random fields by taking the spatial information into account. A new FDR control for multiple testing over dependent data, mirror procedure, is derived based on a two-component mixture model of location wise p-values with spatially varying mixture probability. Although the mirror procedure is derived based on the mixture model of p-values, it is robust to model misspecification. Moreover, this new procedure allows for weak dependency among the hypotheses. An EM algorithm is developed to estimate the mixture model that allows the alternative density of p-values to be either a nonparametric or a flexible semiparametric function that varies over spatial locations. We study the asymptotic behavior of our proposed multiple testing approach both theoretically and numerically. Our method appears to be more powerful than the traditional FDR control methods.

Examples of using this new approach to compare the mean and covariance between two spatiotemporal random fields are particularly studied and illustrated. The simulation experiments show that both the nonparametric and semiparametric mirror procedures work well for the mean comparison, while the nonparametric mirror seems more robust to different dependency structures and it is computationally much more efficient. Our method can be applied to a wide array of problems that require to evaluate the difference between two correlated data. The paleoclimate reconstruction in our data application is one such example. The comparison between images in medical studies can also benefit from this method.

Our FDR procedure was motivated by detecting local differences between two spatiotemporal random fields. However, our procedure is established based on location wise p-values that are obtained by comparing the characteristics of two time series. Therefore, the procedure does not directly consider the temporal correlation in the data though it takes the spatial correlation of p-values into account. On the other hand, this allows our FDR method to be applicable in a more general situation wherever location wise p-values are available regardless of how they are obtained. The p-values can be derived by comparing two independent samples, two multivariate samples or two stochastic processes. The latter of which is often demanded by studies in industrial engineering.

Finally, our test is established based on location wise but not a continuous process of p-values, hence the decisions cannot be made over unobserved locations. However, observations can first be interpolated to locations of interest and then perform the hypothesis testing using the interpolated values. The effect of the interpolation to the multiple testing needs to be further investigated.

Appendix A

A.1. Proof of Theorem 2

We introduce distances to quantify the discrepancy between two probability measures which will be useful in our proof. For two probability measures μ and ν on $[0,1]$, we define the Kolmogorov–Smirnov distance (over the interval $[t^*,1]$ for $0 \le t^* < 1$) and the Wasserstein distance, respectively, as

$$\mathrm{Kolm}_{t^*}(\mu,\nu) = \sup_{t\in[t^*,1]} |\mu([0,t]) - \nu([0,t])|, \quad 0 \le t^* < 1,$$

$$\mathrm{Wass}(\mu,\nu) = \sup\left\{\left|\int_0^1 f d\mu - \int_0^1 f d\nu\right| : f \text{ is 1-Lipschitz}\right\}.$$

Write $\mathrm{Kolm}_{t^*}(\mu,\nu) = \mathrm{Kolm}(\mu,\nu)$ for $t^* = 0$. An alternative expression for the Wasserstein distance is given by

$$\mathrm{Wass}(\mu,\nu) = \int_0^1 |\mu([0,t]) - \nu([0,t])| dt. \tag{A.1}$$

We first state the following two lemmas that are of independent interest.

Lemma 1. Let μ and ν be two probability measures. Suppose ν is absolutely continuous with respect to the Lebesgue measure with the density that is bounded from above by a constant C on the interval $[t^*/2,1]$. Then

$$\mathrm{Kolm}_{t^*}(\mu,\nu) \le 2\sqrt{C\mathrm{Wass}(\mu,\nu)} \le 2\sqrt{C\mathrm{Kolm}(\mu,\nu)}.$$

Proof of Lemma 1. Consider a point $t \in [t^*,1)$ and fix an $0 < \epsilon < t^*/2$. Let $g_1(x) = 1$ on $[0,t]$ and $g_1(x) = 0$ on $[t+\epsilon,1]$, and linear interpolation in between. Let $g_2(x) = 1$ on $[0,t-\epsilon]$ and $g_2(x) = 0$ on $[t,1]$, and linear interpolation in between. Then g_1 and g_2 form upper and lower envelopes for $\mathbf{1}\{x \in [0,t]\}$. So

$$\begin{aligned}
&\mu(t) - \nu(t)\\
&= \int_0^t d\mu(x) - \int_0^t d\nu(x)\\
&\le \int_0^1 g_1(x) d\mu(x) - \int_0^1 g_1(x) d\nu(x) + \int_0^1 g_1(x) d\nu(x)\\
&\quad - \int_0^1 \mathbf{1}\{x \in [0,t]\} d\nu(x).
\end{aligned}$$

As g_1 is $1/\epsilon$-Lipschitz,

$$\int_0^1 g_1(x) d\mu(x) - \int_0^1 g_1(x) d\nu(x) \le \frac{1}{\epsilon}\mathrm{Wass}(\mu,\nu).$$

Moreover, let $h(x) = d\nu(x)/dx$ be the Radon–Nikodym derivative of ν with respect to the Lebesgue measure. Then

$$\begin{aligned}
&\int_0^1 g_1(x) d\nu(x) - \int_0^1 \mathbf{1}\{x \in [0,t]\} d\nu(x)\\
&= \int_t^{t+\epsilon} (g_1(x) - \mathbf{1}\{x \in [0,t]\}) d\nu(x)\\
&\le \int_t^{t+\epsilon} (g_1(x) - \mathbf{1}\{x \in [0,t]\}) h(x) dx \le C\epsilon,
\end{aligned}$$

where we have used the fact that $h(x) \le C$ for $x \ge t \ge t^*/2$. The same bound holds for $\nu(t) - \mu(t)$ using similar argument based on g_2. The first inequality follows by optimizing over ϵ. The second inequality follows as $\mathrm{Wass}(\mu,\nu) \le \mathrm{Kolm}(\mu,\nu)$ by (A.1) and the definition of Kolm. □

Lemma 2. Define the empirical distribution functions $F_m(t) = m^{-1}\sum_{i=1}^m \mathbf{1}\{x_i \le t\}$ and $G_m(t) = m^{-1}\sum_{i=1}^m \mathbf{1}\{y_i \le t\}$ based on $\{x_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^m$, respectively, with $x_i, y_i \in [0,1]$.

Let F be a distribution on $[0,1]$ which has a density that is bounded from above by a constant C on the interval $[t^*/2,1]$. Then we have

$$\mathrm{Kolm}_{t^*}(F_m,G_m) \le 2\sqrt{C}\left\{2\sqrt{\mathrm{Wass}(F_m,F)} + \sqrt{m^{-1}\sum_{i=1}^m |x_i - y_i|}\right\}.$$

Proof of Lemma 2. By Lemma 1,

$$\text{Kolm}_{t^*}(F_m, F) \le 2\sqrt{C\text{Wass}(F_m, F)},$$
$$\text{Kolm}_{t^*}(G_m, F) \le 2\sqrt{C\text{Wass}(G_m, F)}.$$

By the triangle inequality and the elementary inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$, we have

$$\begin{aligned}
&\text{Kolm}_{t^*}(F_m, G_m) \\
&\le \text{Kolm}_{t^*}(F_m, F) + \text{Kolm}_{t^*}(F, G_m) \\
&\le 2\sqrt{C}\left\{\sqrt{\text{Wass}(F_m, F)} + \sqrt{\text{Wass}(G_m, F)}\right\} \\
&\le 2\sqrt{C}\left\{\sqrt{\text{Wass}(F_m, F)} + \sqrt{\text{Wass}(G_m, F_m) + \text{Wass}(F_m, F)}\right\} \\
&\le 2\sqrt{C}\left\{2\sqrt{\text{Wass}(F_m, F)} + \sqrt{\text{Wass}(G_m, F_m)}\right\}.
\end{aligned}$$

Note that

$$\begin{aligned}
\text{Wass}(G_m, F_m) &= \sup\left\{\left|m^{-1}\sum_{i=1}^m (f(x_i) - f(y_i))\right| : f \text{ is 1-Lipschitz}\right\} \\
&\le m^{-1}\sum_{i=1}^m |x_i - y_i|,
\end{aligned}$$

as $|f(x_i) - f(y_i)| \le |x_i - y_i|$. The result thus follows. □

We now apply the above result to our problem. Let

$$\widehat{G}_{m,0}(q) = \frac{1}{|D_m|}\sum_{\mathbf{s}\in D_m} \mathbf{1}\{\widehat{\text{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le q\},$$
$$\widehat{\widehat{G}}_{m,1}(q) = \frac{1}{|D_m|}\sum_{\mathbf{s}\in S_0} \mathbf{1}\{\widehat{\text{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le q\},$$
$$\widehat{G}_{m,1}(q) = \frac{1}{|D_m|}\sum_{\mathbf{s}\in D_m} \mathbf{1}\{\widehat{\text{LFDR}}_{\mathbf{s}}(1 - p_{\mathbf{s}}) \le q\}.$$

Similarly, we define $\mathcal{G}_{m,0}(q)$, $\widetilde{\mathcal{G}}_{m,1}(q)$, and $\mathcal{G}_{m,1}(q)$ by replacing $\widehat{\text{LFDR}}_{\mathbf{s}}$ with $\Psi_{\mathbf{s}}$ in the above definitions.

Lemma 3. Under Assumptions 1 and 2, we have

$$\sup_{q\ge q_0}\left|\widehat{G}_{m,0}(q) - G_0(q)\right| \to^p 0, \tag{A.2}$$
$$\sup_{q\ge q_0}\left|\widehat{\widehat{G}}_{m,1}(q) - \widetilde{G}_1(q)\right| \to^p 0, \tag{A.3}$$
$$\sup_{q\ge q_0}\left|\widehat{G}_{m,1}(q) - G_1(q)\right| \to^p 0. \tag{A.4}$$

Proof of Lemma 3. By Lemma 2, we have

$$\begin{aligned}
\text{Kolm}_{t_0}(\mathcal{G}_{m,0}, \widehat{G}_{m,0}) \le& 2\sqrt{C_0}\Bigg\{2\sqrt{\text{Wass}(\mathcal{G}_{m,0}, G_0)} \\
&+ \sqrt{|D_m|^{-1}\sum_{\mathbf{s}\in D_m}|\widehat{\text{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) - \Psi_{\mathbf{s}}(p_{\mathbf{s}})|}\Bigg\}.
\end{aligned} \tag{A.5}$$

Under Assumption 4, $\{\mathbf{1}\{\Psi_{\mathbf{s}}(p_{\mathbf{s}}) \le q\} : \mathbf{s} \in D_m\}$ for any $q \in [0,1]$ is a bounded strong mixing spatial process. By the weak law of large numbers for strong mixing process, see, for example, Theorem 3 of Jenish and Prucha (2009), we have $|\mathcal{G}_{m,0}(q) - E[\mathcal{G}_{m,0}(q)]| = o_p(1)$ for any $q \in [0,1]$. By Assumption 2, we obtain $|\mathcal{G}_{m,0}(q) - G_0(q)| = o_p(1)$ for $q \in [0,1]$. Following the proof of the weak Glivenko-Cantelli Theorem, we have $\text{Kolm}(\mathcal{G}_{m,0}, G_0) \to^p 0$, which implies that $\text{Wass}(\mathcal{G}_{m,0}, G_0) \to^p 0$ in view of (A.1). Together with (A.5) and Assumption 1, we get $\text{Kolm}_{t_0}(\mathcal{G}_{m,0}, \widehat{G}_{m,0}) \to^p 0$. Finally by the triangle inequality,

$$\text{Kolm}_{t_0}(\widehat{G}_{m,0}, G_0) \le \text{Kolm}_{t_0}(\mathcal{G}_{m,0}, \widehat{G}_{m,0}) + \text{Kolm}_{t_0}(\mathcal{G}_{m,0}, G_0) \to^p 0,$$

which implies (A.2). The proofs for (A.3) and (A.4) are similar and we omit the details. □

Lemma 4. Under Assumptions 1 and 2, we have

$$\sup_{q\ge q'}\left|\frac{\widehat{G}_{m,1}(q)}{|D_m|^{-1}\vee\widehat{G}_{m,0}(q)} - \frac{G_1(q)}{G_0(q)}\right| \to^p 0, \tag{A.6}$$
$$\sup_{q\ge q'}\left|\frac{\widehat{\widehat{G}}_{m,1}(q)}{|D_m|^{-1}\vee\widehat{G}_{m,0}(q)} - \frac{\widetilde{G}_1(q)}{G_0(q)}\right| \to^p 0. \tag{A.7}$$

Proof of Lemma 4. We only prove (A.6) as the arguments for the other result is similar. By the monotonicity of G_0, $\min_{q\ge q'} G_0(q) = G_0(q') > 0$ as $U(q') = G_1(q')/G_0(q') < \alpha$. By Lemma 3, we have

$$\begin{aligned}
&\left|\frac{\widehat{G}_{m,1}(q)}{|D_m|^{-1}\vee\widehat{G}_{m,0}(q)} - \frac{G_1(q)}{G_0(q)}\right| \\
=&\left|\frac{(\widehat{G}_{m,1}(q) - G_1(q))G_0(q) - G_1(q)(|D_m|^{-1}\vee\widehat{G}_{m,0}(q) - G_0(q))}{G_0(q)\{|D_m|^{-1}\vee\widehat{G}_{m,0}(q)\}}\right| \\
\le&\frac{G_0(1)|\widehat{G}_{m,1}(q) - G_1(q)| + G_1(1)||D_m|^{-1}\vee\widehat{G}_{m,0}(q) - G_0(q)|}{G_0(q')\{G_0(q) - \sup_{\lambda\ge q'}||D_m|^{-1}\vee\widehat{G}_{m,0}(\lambda) - G_0(\lambda)|\}} \\
\le&\frac{\begin{array}{c}G_0(1)\sup_{\lambda\ge q'}|\widehat{G}_{m,1}(\lambda) - G_1(\lambda)| \\ +G_1(1)\sup_{\lambda\ge q'}||D_m|^{-1}\vee\widehat{G}_{m,0}(\lambda) - G_0(\lambda)|\end{array}}{G_0(q')\{G_0(q') - \sup_{\lambda\ge q'}||D_m|^{-1}\vee\widehat{G}_{m,0}(\lambda) - G_0(\lambda)|\}} \to^p 0
\end{aligned}$$

uniformly for any $q \ge q'$. □

Lemma 5. Suppose f_0 is mirror conservative. Under Assumption 2, we have $\widetilde{G}_1(q) \le G_1(q)$ for any $q \in [0,1]$.

Proof of Lemma 5. As $\Psi_{\mathbf{s}}$ is right continuous, denote by $\Psi_{\mathbf{s}}^{-1}(q) = \inf\{u \in [0,1] : \Psi_{\mathbf{s}}(u) \ge q\}$. Because f_0 is mirror conservative (see (2)), we have for $\mathbf{s} \in S_0$ and $\Psi_{\mathbf{s}}^{-1}(q) \le 0.5$,

$$\begin{aligned}
&P(\Psi_{\mathbf{s}}(p_{\mathbf{s}}) \le q) \\
&= \int_0^1 \mathbf{1}\{p \le \Psi_{\mathbf{s}}^{-1}(q)\}f_0(p)dp \\
&\le \int_0^1 \mathbf{1}\{1 - p \le \Psi_{\mathbf{s}}^{-1}(q)\}f_0(p)dp = P(\Psi_{\mathbf{s}}(1 - p_{\mathbf{s}}) \le q).
\end{aligned}$$

For $\Psi_{\mathbf{s}}^{-1}(q) > 0.5$, we see that

$$\begin{aligned}
&P(\Psi_{\mathbf{s}}(p_{\mathbf{s}}) \le q) \\
&= \int_0^1 \mathbf{1}\{p \le 1 - \Psi_{\mathbf{s}}^{-1}(q)\}f_0(p)dp \\
&\quad + \int_0^1 \mathbf{1}\{1 - \Psi_{\mathbf{s}}^{-1}(q) \le p \le \Psi_{\mathbf{s}}^{-1}(q)\}f_0(p)dp \\
&\le \int_0^1 \mathbf{1}\{1 - p \le 1 - \Psi_{\mathbf{s}}^{-1}(q)\}f_0(p)dp \\
&\quad + \int_0^1 \mathbf{1}\{1 - \Psi_{\mathbf{s}}^{-1}(q) \le p \le \Psi_{\mathbf{s}}^{-1}(q)\}f_0(p)dp \\
&= \int_0^1 \mathbf{1}\{1 - p \le \Psi_{\mathbf{s}}^{-1}(q)\}f_0(p)dp = P(\Psi_{\mathbf{s}}(1 - p_{\mathbf{s}}) \le q).
\end{aligned}$$

It thus implies that

$$\widetilde{G}_1(q) = \lim_{|D_m|\to+\infty} \frac{1}{|D_m|} \sum_{\mathbf{s}\in S_0} P(\Psi_{\mathbf{s}}(p_{\mathbf{s}}) \le q) \le \lim_{|D_m|\to+\infty} \frac{1}{|D_m|} \sum_{\mathbf{s}\in D_m} P(\Psi_{\mathbf{s}}(1-p_{\mathbf{s}}) \le q) = G_1(q). \quad \text{(A.8)}$$

□

Proof of Theorem 2. Set $\tau = \alpha - U(q')$. By (A.6), we have

$$P\left(\frac{\widehat{G}_{m,1}(q')}{|D_m|^{-1} \vee \widehat{G}_{m,0}(q')} \le U(q') + \tau/2 < \alpha\right) \to 1.$$

On the above event, $\widehat{q} \ge q'$ and thus $P(\widehat{q} \ge q') \to 1$. Now on the event $\widehat{q} \ge q'$, we have

$$\begin{aligned}
&\alpha - \frac{\sum_{\mathbf{s}\in S_0} \mathbf{1}\{\widehat{\mathrm{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le \widehat{q}\}}{1 \vee \sum_{\mathbf{s}\in D_m} \mathbf{1}\{\widehat{\mathrm{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le \widehat{q}\}} \\
&\ge \frac{\sum_{s\in D_m} \mathbf{1}\left\{\widehat{\mathrm{LFDR}}_{\mathbf{s}}(1-p_{\mathbf{s}}) \le \widehat{q}\right\}}{1 \vee \sum_{s\in D_m} \mathbf{1}\left\{\widehat{\mathrm{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le \widehat{q}\right\}} - \frac{\sum_{\mathbf{s}\in S_0} \mathbf{1}\{\widehat{\mathrm{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le \widehat{q}\}}{1 \vee \sum_{\mathbf{s}\in D_m} \mathbf{1}\{\widehat{\mathrm{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le \widehat{q}\}} \\
&\ge \inf_{q\ge q'} \left\{ \frac{\widehat{G}_{m,1}(q)}{|D_m|^{-1} \vee \widehat{G}_{m,0}(q)} - \frac{G_1(q)}{G_0(q)} + \frac{G_1(q) - \widetilde{G}_1(q)}{G_0(q)} + \frac{\widetilde{G}_1(q)}{G_0(q)} \right. \\
&\quad \left. - \frac{\widehat{\widehat{G}}_{m,1}(q)}{|D_m|^{-1} \vee \widehat{G}_{m,0}(q)} \right\} \\
&\ge \inf_{q\ge q'} \left\{ \frac{\widehat{G}_{m,1}(q)}{|D_m|^{-1} \vee \widehat{G}_{m,0}(q)} - \frac{G_1(q)}{G_0(q)} + \frac{\widetilde{G}_1(q)}{G_0(q)} \right. \\
&\quad \left. - \frac{\widehat{\widehat{G}}_{m,1}(q)}{|D_m|^{-1} \vee \widehat{G}_{m,0}(q)} \right\} = o_p(1),
\end{aligned}$$

where the first inequality follows from the definition of $\widehat{q}$, the third inequality follows from (A.8), and the last equality is due to (A.6) and (A.7). It thus implies that

$$\frac{\sum_{\mathbf{s}\in S_0} \mathbf{1}\{\widehat{\mathrm{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le \widehat{q}\}}{1 \vee \sum_{\mathbf{s}\in D_m} \mathbf{1}\{\widehat{\mathrm{LFDR}}_{\mathbf{s}}(p_{\mathbf{s}}) \le \widehat{q}\}} \le \alpha + o_p(1).$$

Finally applying Lemma 6, we obtain $\lim_{m\to+\infty} \mathrm{FDR}_m(\widehat{q}) \le \alpha$ which completes the proof. □

Lemma 6. Consider the random sequence $\{(X_m, Y_m)\}_m$. Suppose $X_m \le C_0$ and $X_m \le \alpha + Y_m$, where $Y_m = o_p(1)$ and C_0 is some constant. Then we have

$$\limsup_m E[X_m] \le \alpha.$$

Proof of Lemma 6. There exists a subsequence X_{m_k} such that $\limsup_m E[X_m] = \lim_k E[X_{m_k}]$. Along this subsequence, we can pick a further subsequence $Y_{m_{k_j}}$ such that $Y_{m_{k_j}} \to^{a.s.} 0$. Thus with probability one,

$$\limsup_j X_{m_{k_j}} \le \limsup_j Y_{m_{k_j}} + \alpha = \alpha.$$

As $X_{m_{k_j}} \le C_0$, by Fatou's lemma,

$$\limsup_m E[X_m] = \limsup_j E[X_{m_{k_j}}] \le E[\limsup_j X_{m_{k_j}}] \le \alpha.$$

□

A.2. Weakening Spatial Dependence

Our theoretical analysis shows that the mirror procedure is asymptotically robust to weak dependence among the p-values. For low rank dependence, we can use a random effect model to capture and remove the spatial dependence. To illustrate the idea, we consider the case of paired data ($n_1 = n_2 = n$) and let $Z(\mathbf{s}, t) = X(\mathbf{s}, t) - Y(\mathbf{s}, t)$ with $E[Z(\mathbf{s}, t)] = \mu(\mathbf{s})$. We are interested in testing

$$H_{0,\mathbf{s}} : \mu(\mathbf{s}) = 0 \quad \text{versus} \quad H_{a,\mathbf{s}} : \mu(\mathbf{s}) \ne 0.$$

Consider the random effect model,

$$Z(\mathbf{s}, t) = \mu(\mathbf{s}) + W_{\mathbf{s}}^\top \gamma_t + \epsilon(\mathbf{s}, t),$$

where $W_{\mathbf{s}}$'s are the (prespecified) covariates capturing the spatial dependence, γ_t's are the random effects, and $\epsilon(\mathbf{s}, t)$ are the random errors that are independent across $\mathbf{s}$. To estimate γ_t and $\mu(\mathbf{s})$ simultaneously, we use the l_1 penalized regression,

$$\min_{\mu(\mathbf{s}),\gamma_t} \frac{1}{2} \sum_{t=1}^{n} \sum_{\mathbf{s}\in D} (Z(\mathbf{s}, t) - \mu(\mathbf{s}) - W_{\mathbf{s}}^\top \gamma_t)^2 + \lambda \sum_{\mathbf{s}\in D} |\mu(\mathbf{s})|, \quad \text{(A.9)}$$

where $\lambda > 0$ is a tuning parameter (e.g., chosen via cross-validation). Denote by $\hat{\mu}(\mathbf{s})$ and $\hat{\gamma}_t$ the solutions to (A.9). Let $\hat{Z}(\mathbf{s}, t) = Z(\mathbf{s}, t) - W_{\mathbf{s}}^\top \hat{\gamma}_t$. We calculate the p-values for testing $H_{0,\mathbf{s}}$ based on $\hat{Z}(\mathbf{s}, t)$ and apply the FDR control procedure to the resulting p-values.

Supplementary Materials

Supplementary material consists of simulation results for randomly generated locations including plots of FDR and power for mean and covariance comparisons, as well as implementation of the self-normalization test.

Acknowledgments

The authors thank the editor, the associate editor, and the referees for constructive suggestions that have improved the content and presentation of this article.

Funding

Yun and Li's research was partially supported by National Science Foundation grant AGS-1602845 and DMS-1830312. Zhang acknowledges partial support from NSF DMS-1830392 and NSF DMS-1811747.

References

Barber, R. F., and Candes, E. J. (2016), "A Knockoff Filter for High-Dimensional Selective Inference," arXiv no. 1602.03574. [3]

Barlow, R. E., and Brunk, H. D. (1972), "The Isotonic Regression Problem and Its Dual," *Journal of the American Statistical Association*, 67, 140–147. [5]

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018), "Weighted False Discovery Rate Control in Large-Scale Multiple Testing," *Journal of the American Statistical Association*, 113, 1172–1183. [3]

Benjamini, Y., and Heller, R. (2007), "False Discovery Rates for Spatial Signals," *Journal of the American Statistical Association*, 102, 1272–1281. [2]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [1,6]

——— (2000), "On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics," *Journal of Educational and Behavioral Statistics*, 25, 60–83. [1]

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188. [1]

Coats, S., Smerdon, J. E., Cook, B. I., and Seager, R. (2013), "Stationarity of the Tropical Pacific Teleconnection to North America in CMIP5/PMIP3 Model Simulations," *Geophysical Research Letters*, 40, 4927–4932. [12]

Dannenberg, M. P., and Wise, E. K. (2013), "Performance of Climate Field Reconstruction Methods Over Multiple Seasons and Climate Variables," *Journal of Geophysical Research: Atmospheres*, 118, 9595–9610. [11]

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104. [1,4]

Efron, B., and Tibshirani, R. (1996), "Using Specially Designed Exponential Families for Density Estimation," *The Annals of Statistics*, 24(6), 2431–2461. [6]

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160. [1]

Evans, M., Smerdon, J. E., Kaplan, A., Tolwinski-Ward, S., and González-Rouco, J. F. (2014), "Climate Field Reconstruction Uncertainty Arising From Multivariate and Nonlinear Properties of Predictors," *Geophysical Research Letters*, 41, 9127–9134. [11]

Fan, J., Han, X., and Gu, W. (2012), "Estimating False Discovery Proportion Under Arbitrary Covariance Dependence," *Journal of the American Statistical Association*, 107, 1019–1035. [11]

Genovese, C., and Wasserman, L. (2002), "Operating Characteristics and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society*, Series B, 64, 499–517. [1,3]

——— (2004), "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32, 1035–1061. [3]

Hering, A. S., and Genton, M. G. (2011), "Comparing Spatial Predictions," *Technometrics*, 53, 414–425. [1]

Jenish, N., and Prucha, I. R. (2009), "Central Limit Theorems and Uniform Laws of Large Numbers for Arrays of Random Fields," *Journal of Econometrics*, 150, 86–98. [14]

Lei, L., and Fithian, W. (2018), "Adapt: An Interactive Procedure for Multiple Testing With Side Information," *Journal of the Royal Statistical Society*, Series B, 80, 649–679. [2,3]

Li, B., and Smerdon, J. E. (2012), "Defining Spatial Comparison Metrics for Evaluation of Paleoclimatic Field Reconstructions of the Common Era," *Environmetrics*, 23, 394–406. [1,11]

Li, B., Zhang, X., and Smerdon, J. E. (2016), "Comparison Between Spatio-Temporal Random Processes and Application to Climate Model Data," *Environmetrics*, 27, 267–279. [1,11]

Lund, R., and Li, B. (2009), "Revisiting Climate Region Definitions via Clustering," *Journal of Climate*, 22, 1787–1800. [1]

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ammann, C., Faluvegi, G., and Ni, F. (2009), "Global Signatures and Dynamical Origins of the Little Ice Age and Medieval Climate Anomaly," *Science*, 326, 1256–1260. [11]

Meyer, M., and Woodroofe, M. (2000), "On the Degrees of Freedom in Shape-Restricted Regression," *The Annals of Statistics*, 28, 1083–1104. [6]

Montero, J.-M., Fernández-Avilés, G., and Mateu, J. (2015), *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging* (Vol. 998), Hoboken, NJ: Wiley. [9]

Muller, P., Parmigiani, G., and Rice, K. (2006), "FDR and Bayesian Multiple Comparisons Rules," Johns Hopkins University, Department of Biostatistics, Working Paper 115. [2]

Müller, P., Sansó, B., and De Iorio, M. (2004), "Optimal Bayesian Design by Inhomogeneous Markov Chain Simulation," *Journal of the American Statistical Association*, 99, 788–798. [2]

Pavlicová, M., Santner, T. J., and Cressie, N. (2008), "Detecting Signals in FMRI Data Using Powerful FDR Procedures," *Statistics and Its Interface*, 1, 23–32. [2]

Risser, M. D., Paciorek, C. J., and Stone, D. A. (2019), "Spatially-Dependent Multiple Testing Under Model Misspecification, With Application to Detection of Anthropogenic Influence on Extreme Climate Events," *Journal of the American Statistical Association*, 114, 61–78. [2]

Shao, X. (2010), "A Self-Normalized Approach to Confidence Interval Construction in Time Series," *Journal of the Royal Statistical Society*, Series B, 72, 343–366. [6,11]

Shen, X., Huang, H.-C., and Cressie, N. (2002), "Nonparametric Hypothesis Testing for a Spatial Signal," *Journal of the American Statistical Association*, 97, 1122–1140. [2]

Singer, S., and Singer, S. (1999), "Complexity Analysis of Nelder–Mead Search Iterations," in *Proceedings of the 1st Conference on Applied Mathematics and Computation*, pp. 185–196. [5]

Smerdon, J. E. (2012), "Climate Models as a Test Bed for Climate Reconstruction Methods: Pseudoproxy Experiments," *Wiley Interdisciplinary Reviews: Climate Change*, 3, 63–77. [11,12]

Smerdon, J. E., Coats, S., and Ault, T. R. (2016), "Model-Dependent Spatial Skill in Pseudoproxy Experiments Testing Climate Field Reconstruction Methods for the Common Era," *Climate Dynamics*, 46, 1921–1942. [11]

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S., and Roe, G. H. (2014), "Assimilation of Time-Averaged Pseudoproxies for Climate Reconstruction," *Journal of Climate*, 27, 426–441. [11]

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Series B, 64, 479–498. [1,3,6]

——— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-Value," *The Annals of Statistics*, 31, 2013–2035. [1]

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society*, Series B, 66, 187–205. [4]

Sun, W., and Cai, T. T. (2007), "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control," *Journal of the American Statistical Association*, 102, 901–912. [3]

——— (2009), "Large-Scale Multiple Testing Under Dependence," *Journal of the Royal Statistical Society*, Series B, 71, 393–424. [2]

Sun, W., Reich, B. J., Cai, T. T., Guindani, M., and Schwartzman, A. (2015), "False Discovery Control in Large-Scale Spatial Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 77, 59–83. [2,3]

Tansey, W., Koyejo, O., Poldrack, R. A., and Scott, J. G. (2018), "False Discovery Rate Smoothing," *Journal of the American Statistical Association*, 113, 1156–1171. [2,6]

Wang, J., Emile-Geay, J., Guillot, D., Smerdon, J., and Rajaratnam, B. (2014), "Evaluating Climate Field Reconstruction Techniques Using Improved Emulations of Real-World Conditions," *Climate of the Past*, 10, 1–19. [11]

Yang, X., and DelSole, T. (2012), "Systematic Comparison of ENSO Teleconnection Patterns Between Models and Observations," *Journal of Climate*, 25, 425–446. [12]

Zhang, X., and Shao, X. (2015), "Two Sample Inference for the Second-Order Property of Temporally Dependent Functional Data," *Bernoulli*, 21, 909–929. [1]