Title
Spatially Dependent Multiple Testing Under Model Misspecification, With Application to Detection of Anthropogenic Influence on Extreme Climate Events

Journal
Journal of the American Statistical Association, 114(525)

ISSN
0162-1459

Authors
Risser, MD
Paciorek, CJ
Stone, DA

Publication Date
2019-01-02

DOI
10.1080/01621459.2018.1451335

Peer reviewed

Spatially Dependent Multiple Testing Under Model Misspecification, With Application to Detection of Anthropogenic Influence on Extreme Climate Events

Mark D. Risser,Christopher J. Paciorek &Dáithí A. Stone

Abstract

The Weather Risk Attribution Forecast (WRAF) is a forecasting tool that uses output from global climate models to make simultaneous attribution statements about whether and how greenhouse gas emissions have contributed to extreme weather across the globe. However, in conducting a large number of simultaneous hypothesis tests, the WRAF is prone to identifying false “discoveries.” A common technique for addressing this multiple testing problem is to adjust the procedure in a way that controls the proportion of true null hypotheses that are incorrectly rejected, or the false discovery rate (FDR). Unfortunately, generic FDR procedures suffer from low power when the hypotheses are dependent, and techniques designed to account for dependence are sensitive to misspecification of the underlying statistical model. In this article, we develop a Bayesian decision-theoretical approach for dependent multiple testing and a nonparametric hierarchical statistical model that flexibly controls false discovery and is robust to model misspecification. We illustrate the robustness of our procedure to model error with a simulation study, using a framework that accounts for generic spatial dependence and allows the practitioner to flexibly specify the decision criteria. Finally, we apply our procedure to several seasonal forecasts and discuss implementation for the WRAF workflow. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

Keywords: Bayesian nonparametrics, Climate models, Decision theory, Empirical orthogonal functions, Event attribution, False discovery rate, Generalized double Pareto

1. Introduction

Event attribution (EA) is a field of study that seeks to understand and describe the influence of greenhouse gas emissions and other human activities on extreme weather (Stott et al. 2013; National Academies of Sciences and Medicine 2016). The increasing interest in this field arises from the realization that a major fraction of past, current, and future climate impacts and climate change-related impacts result from the occurrence of extreme weather (Arent et al. 2014; Smith et al. 2014). Risk-based EA studies quantify the effect of greenhouse gas (GHG) emissions and other anthropogenic factors on weather by comparing two climate scenarios: a factual real-world scenario (the “world as it is”) and a counterfactual, nonanthropogenic world (the “world as it might have been”). Then, using a probabilistic framework (Allen 2003; Stone and Allen 2005; Hansen et al.

2014), a risk-based EA study compares the probabilities of predefined unusual weather in the two scenarios and estimates how much more or less likely extreme events are in the anthropogenically influenced world than they would have been otherwise (note: here and throughout we mean “risk” in the sense of epidemiological or relative risk, not statistical risk). Typically, the probabilities for each of these scenarios are estimated from simulations of climate models.

Risk-based EA studies can either be targeted or systematic in their approach. Targeted studies examine one event (or a small number of events), studying in detail the meteorological mechanisms involved in the event and how the anthropogenic influence is transmitted through them, and are generally reactive in the sense that they are only conducted for an event that has actually occurred (e.g., Stott, Stone, and Allen 2004; Pall et al. 2011). Systematic studies cover a much larger number of events using an identical method for all events, but the rigidity of a single experimental design means that some events are not amenable to investigation (Angélil et al. 2017). An advantage of the systematic approach is that it does not necessarily depend on the occurrence of events, with it being possible to instead perform the analyses on a predefined list of events. This is the approach taken by the Weather Risk Attribution Forecast (WRAF, *http://climate.web.runbox.net/wraf*). To have EA information available in “real-time,” the WRAF performs analyses 1 month in advance using a predefined list of 232 extreme weather events, comprising an unusually hot, cold, wet, and/or dry month over each of 58 regions (Angélil et al. 2014). In the upcoming new version, the number of regions will be increased by a factor of about four (see, e.g., Figure 4; D. Stone, “A hierarchical collection of political/economic regions for analysis of climate extremes,” submitted). Data for both the factual and counterfactual scenarios come from climate model simulations.

Formally, the forecast involves estimating the probability of a predefined extreme event for both climate scenarios in each of the regions. For region $i = 1, \ldots, M$ (in the upcoming version of the WRAF, $M = 237$), the forecast uses the ratio of scenario-specific probabilities p_{Fi} (for the factual scenario) and p_{Ci} (for the counterfactual scenario) or “risk ratio” $RR_i = p_{Fi}/p_{Ci}$ to formally test for changes in the probability of an extreme month. In other words, a collection of statistical tests are conducted that have null hypotheses of the form

$$H_i : RR_i \leq c, \quad i = 1, \ldots, M, \qquad (1)$$

where, for example, $c = 1$ if we are interested in determining whether anthropogenic influence has resulted in an increase in the event probability. Ultimately, we wish to separately test collections of hypotheses like (1) for extreme temperature (both hot and cold) and precipitation (both wet and dry).

Of course, when the number of tests M is large, a classical testing procedure is prone to identifying false "discoveries," or incorrectly rejecting null hypotheses (commonly referred to as Type I errors). As such, the testing procedure is often adjusted, attempting to control the false discovery rate (FDR), which is the proportion of true null hypotheses that are incorrectly rejected. Since the data arise from physical climate models, it is anticipated that the hypotheses might be dependent: in other words, there is likely strong dependence within each spatial field of probabilities. This dependence might arise from the spatial proximity of the regions (i.e., strong dependence between p_{Fi} and p_{Fj} for adjacent regions i and j) but also from potentially unspecified long-range teleconnections (in which two probabilities p_{Fi} and p_{Fj} might be highly correlated even if regions i and j are far apart) that are common for atmospheric climate variables considered over the globe (see, e.g., Cressie and Wikle 2011). Unfortunately, while classical FDR procedures (Benjamini and Hochberg 1995; see Section 2) are theoretically valid for positively correlated hypotheses (Benjamini and Yekutieli 2001), they are also known to suffer from low power when the test statistics from each test are not independent (e.g., see Sun and Cai 2009). And, while the literature contains a number of methods for applying FDR procedures under dependence, the methods are outlined for specific underlying probability models and are sensitive to improper specification of this model (Sun et al. 2015).

In this article, we develop an approach to the multiple testing problem for spatially dependent hypotheses in a systematic and decision-theoretical framework. Focusing on procedures that account for dependence among tests, we provide an overview of the diverse literature on false discovery control, including traditional methods and both Frequentist and Bayesian decision-theoretical approaches. The framework we use, originally introduced by Müller et al. (2004), allows the practitioner to flexibly specify the decision criteria for false discovery control, and we explore practical comparison of various FDR procedures and decision criteria that can be used when an empirical estimate of the correlation among tests is available. Furthermore, we introduce a robust yet practical modeling framework for addressing spatial dependence among hypotheses, address sensitivity of the decision rule's performance to statistical model misspecification, and demonstrate the robustness of our modeling framework for FDR control. While the methodology is designed specifically for the hypothesis testing setting of the WRAF, our framework is useful for a broader set of problems involving multiple testing over a spatial domain, particularly in the case where an empirical correlation estimate is available, which is often the case for climate science scenarios. In this context, the methods outlined in this article could be used for general hierarchical Bayesian models beyond just considering the probability of extremes or the risk ratio.

A reader familiar with the climate science literature will be aware of the concept of statistical field significance (Livezey and Chen 1983), which is an alternative multiple testing approach that seeks to evaluate the collective significance of a set of statistics. While field significance techniques are well-established in climate science, we instead seek to control FDR following the arguments outlined in Ventura, Paciorek, and Risbey (2004), the most important of which is that field significance provides no specific information about which individual tests are significant. Interestingly, the idea of FDR-control is growing in popularity among climate scientists, as evidenced by a recent article by Wilks (2016). Finally, note that if one is only concerned with a real-time attribution statement for a single region in advance, then the multiple testing framework presented here is not required.

The article proceeds as follows. In Section 2, we introduce a decision-theoretical framework for FDR control and present a systematic and flexible Bayesian approach to the problem. In Section 3, we introduce our nonparametric Bayesian framework for modeling the factual and counterfactual probabilities and extreme ratios, while in Section 4 we conduct a simulation study to assess the sensitivity of the Bayesian FDR procedure to misspecification of the statistical model and identify a data-driven approach that robustly controls the FDR. In Section 5, we apply the method to a real dataset to be used for the WRAF; Section 6 concludes the article.

2. Decision-Theoretical Approaches for False Discovery Control

The WRAF is generated based on monthly simulations of the Community Atmospheric Model version 5.1 (CAM5.1; see Section 5.2 for more details). Temperature and precipitation from the CAM5.1 climate model ensembles are aggregated monthly for each region, for both the factual (F) and counterfactual (C) scenarios. A climate model ensemble is a set of climate model runs such that each ensemble member has the same boundary conditions (e.g., atmospheric chemistry or sea ice concentrations) but stochastically perturbed initial conditions. Denote the resulting collection of random variables $\{Y_{kil}$: $i = 1, \ldots, M; k \in \{F, C\}; l = 1, \ldots, n_{\text{ens}}\}$, where Y generically represents either average monthly temperature or total monthly precipitation and nensnens is the ensemble size (or number of replicates). Formally, for an extreme event type (e.g., cold months, wet months) in region $i = 1, \ldots, M$, define random variables

$$Z_{Fi} = \sum_{l=1}^{n_{\text{ens}}} I(Y_{Fil} > y_i), \quad Z_{Ci} = \sum_{l=1}^{n_{\text{ens}}} I(Y_{Cil} > y_i) \qquad (2)$$

(for hot and wet extremes; replace “>” with “<” for cold and dry extremes), where the extreme event is defined in terms of a region-specific threshold y_i (e.g., exceeding a monthly average temperature threshold of 290 K; note that the event definition is common across scenarios). Define scenario-specific data $\mathbf{Z} = \{(Z_{Fi}, Z_{Ci}) : i = 1, \ldots, M\}$; these binomial random variables can be used to estimate the event probabilities in each scenario $\{(p_{Fi}, p_{Ci})\}$ and establish evidence regarding the null hypotheses $\{H_i: i = 1, \ldots, M\}$ from (1).

For each null hypothesis, define a corresponding collection of unknown parameters that represent the true state for each hypothesis:

$$\theta_i = \begin{cases} 0 & \text{if the true state of hypothesis } i \text{ is null} \\ 1 & \text{if the true state of hypothesis } i \text{ is nonnull} \end{cases} \qquad i = 1, \ldots, M.$$

The testing problem involves generating a decision rule $\delta \equiv \delta(Z) = \{\delta_i : i = 1, \cdots, M\}$, such that

$$\delta_i = \begin{cases} 0 & \text{if hypothesis } i \text{ is classified as null} \\ 1 & \text{if hypothesis } i \text{ is classified as nonnull} \end{cases} \qquad i = 1, \ldots, M.$$

In addition to specifying the form of the decision rule (often based on a test statistic, *p*-value, etc.), an underlying probability model must be specified to estimate the decision rule. The false discovery proportion (FDP) is defined as $\mathrm{FDP} = [\sum_{i=1}^{M}(1 - \theta_i)\delta_i]/[1 \vee \sum_{i=1}^{M}\delta_i]$. Note that the FDP is simply a function of unknown parameters (θ_i) and random variables (δ_i), and is hence fundamentally neither frequentist nor Bayesian.

For a full summary of classical, model-specific approaches to the multiple testing problem, we refer the interested reader to Appendix A in the supplemental materials. The original FDR procedure given by Benjamini and Hochberg (1995, henceforth BH) controls the (frequentist) FDR, defined as the expected FDP, that is, $\mathrm{FDR} \equiv \mathrm{E(FDP)}$, where the expectation is taken over repeated experiments. Their remarkably simple procedure ensures that $\mathrm{FDR} \leqslant \alpha$; the proof in Benjamini and Hochberg (1995) is established for independent test statistics and any configuration of false null hypotheses. One alternative to BH is the adaptive FDR procedure (Benjamini and Hochberg 2000; Genovese and Wasserman 2002). Other alternatives to BH are based on a random mixture model formulation of the multiple testing problem, where the θ_i are Bernoulli random variables and $(Z_{Fi}, Z_{Ci})|\theta_i \sim \theta_i F_0 + (1 - \theta_i)F_1$, where F_0 and F_1 are the null and alternative distributions, respectively. Using this framework, procedures were developed to control either the positive FDR $pFDR = \mathrm{E}(\mathrm{FDP}|\sum_{i=1}^{M}\delta_i > 0)$ (Storey 2003), the marginal FDR $mFDR = \mathrm{E}(\sum_{i=1}^{M}(1 - \theta_i)\delta_i)/\mathrm{E}(\sum_{i=1}^{M}\delta_i)$ (Storey 2003), and Bayesian approaches to the problem using local FDR (Efron et al. 2001; Efron 2007) and the *q*-value (Storey 2003). Yet another alternative approach uses a weighted classification approach, wherein the decision rule δ is constructed by minimizing the classification risk $\mathrm{E}[L_\lambda(\theta, \delta)]$, where the loss function is

$$L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{M}\sum_{i=1}^{M}\Big\{\lambda(1-\theta_i)\delta_i + \theta_i(1-\delta_i)\Big\}; \qquad (3)$$

here, $\lambda > 0$ is the loss attached to a false positive error (relative to a false negative error).

Unfortunately, proofs for the optimality of all of these procedures rely on the notion of independent hypotheses, and the optimality is called into question when the hypotheses are instead dependent. The decision rules of Benjamini and Hochberg (1995), Benjamini and Hochberg (2000), Efron et al. (2001), and Sun and Cai (2007) are "simple," meaning that δ_i is a function only of the data corresponding to hypothesis *i*. It is easy to imagine that in the case of correlated hypotheses, compound decision rules (i.e., decision rules δ such that δ_i depends on data corresponding to the other hypotheses) are preferred in that they might be able to identify nonnulls with a smaller signal by pooling information across tests. As a result, Sun and Cai (2009) extended the compound decision framework for multiple testing in the presence of dependence, specifically when the unknown θ_i arise from a hidden Markov model (HMM). Two recent articles by Sun et al. (2015) and Shu, Nan, and Koeppe (2015) extend this work further to provide similar results for spatial random fields and multi-dimensional Markov random fields (MRFs), respectively. However, proofs for the optimality of these procedures are model-specific; furthermore, Sun et al. (2015) also found that "the precision of [their] testing procedure shows some sensitivity to model misspecification."

To move away from the classical model-specific procedures, we are motivated to consider fully Bayesian approaches to the multiple testing problem, first presented by Newton et al. (2004), Müller et al. (2004), and Müller, Parmigiani, and Rice (2006). Whereas the Frequentist FDR is defined as an expectation over repeated experiments, Müller et al. (2004) defined a Bayesian FDR $\overline{\mathrm{FDR}} \equiv \mathrm{E}\left(\mathrm{FDP}\,|\boldsymbol{Z}\right) = \int \mathrm{FDP}\, dp\left(\boldsymbol{\theta}|\boldsymbol{Z}\right)$ (i.e., the posterior expected FDP), where the expectation is with respect to the posterior distribution of the unknown states conditional on the data. Conditioning on the data and marginalizing with respect to θ, Müller et al. (2004) showed that $\overline{\mathrm{FDR}} = \left[\sum_{i=1}^{M}\delta_i\pi_i\right]/\left[1 \vee \sum_{i=1}^{M}\delta_i\right]$, where $\pi_i = P\left(\theta_i = 0|\boldsymbol{Z}\right)$ is the posterior probability that the *i*th hypothesis is null. A similar expression can be obtained for the Bayesian false nondiscovery rate

$\overline{\mathrm{FNR}} = \left[\sum_{i=1}^{M} (1-\delta_i)(1-\pi_i)\right] / \left[1 \vee \left(M - \sum_{i=1}^{M} \delta_i\right)\right]$, as well as count versions $\overline{\mathrm{FD}} = \sum_{i=1}^{M} \delta_i \pi_i$ and $\overline{FN} = \sum_{i=1}^{M} (1-\delta_i)(1-\pi_i)$.

A Bayesian decision criteria that is similar in nature to the frequentist approaches (e.g., Sun and Cai 2007; Sun et al. 2015) is to minimize the $\overline{\mathrm{FNR}}$ subject to the constraint that $\overline{\mathrm{FDR}} \leq \alpha$. Müller et al. (2004) showed that the optimal decision rule for this criteria is $\delta^*_i = I(\pi_i < t^*_\alpha)$, where the threshold depends on the desired α. Interestingly, this Figure 1 decision rule can be written like the decision rule in Sun et al. (2015): after ranking the π_i such that $\pi_{(1)} < \pi_{(2)} < \cdots < \pi_{(M)}$, find

$$r_1 = \max\left\{ j : \frac{1}{j} \sum_{i=1}^{j} \pi_{(i)} \leq \alpha \right\}; \qquad (4)$$

then $t^*_\alpha = \pi_{(r_1+1)}$, so that we reject $H_{(1)}, \cdots, H_{(r_1)}$. The difference between the decision rule in Sun et al. (2015) and (4) is that the former involves a probability conditional on the hyperparameters while the latter involves a probability that marginalizes over the hyperparameters. In other words, the fully Bayesian posterior probability π_i is almost the same as the oracle statistic in Sun et al. (2015), but accounts for uncertainty in the hyperparameters. (Note, however, that while the oracle statistic in Sun et al. 2015 is derived using a frequentist criteria, it is calculated using a Bayesian framework and coincides exactly with (4). Their simulation study verifies that this strategy controls the frequentist FDR.)

The optimality of (4) for controlling $\overline{\mathrm{FDR}} \leq \alpha$ is true for “any probability model with nonzero prior probability for both the null and alternative hypotheses” (Müller, Parmigiani, and Rice 2006), which is quite powerful in light of the extensive work to develop model-specific oracle procedures in the Frequentist setting (e.g., Sun and Cai 2007, 2009; Sun et al. 2015; Shu, Nan, and Koeppe 2015). Of course, the Bayesian FDR (FDR ¯≡E(FDP |Z)) is not the same as the frequentist FDR (FDR ≡ E(FDP)), but Müller et al. (2004) and Müller, Parmigiani, and Rice (2006) showed that controlling the Bayesian FDR implies frequentist FDR control when tests are independent. Unfortunately, this is not necessarily true for dependent hypotheses (Pacifico et al. 2004; Guindani et al. 2009), although the relationship between the decision rule in Sun et al. (2015) and (4) suggests a similarity between the two approaches.

A benefit of the decision-theoretical framework is that classification errors can be controlled in a variety of ways, beyond just the rate of false discoveries. In addition to the decision criteria that controls the Bayesian FDR introduced in the previous paragraph by minimizing a posterior expected loss (henceforth R_1), Müller et al. (2004) defined two other decision criteria. The first (denoted R_2) is similar to the classification risk for (3):

$$R_2(\boldsymbol{\delta}, \boldsymbol{z}) = \lambda_1 \overline{\mathrm{FD}} + \overline{\mathrm{FN}} = \sum_{i=1}^{M} \left\{ \lambda_2 \delta_i \pi_i + (1 - \delta_i)(1 - \pi_i) \right\}.$$

This criteria minimizes the number of false negatives and false discoveries, where λ_2 represents the cost for a false discovery relative to a false negative. Like R_1, Müller et al. (2004) showed that the optimal decision rule for R_2 is a threshold rule, that is, $\delta^*_i = I(\pi_i < t^*_\lambda)$, where the optimal threshold is $t^*_\lambda = 1/(\lambda_2 + 1)$. The second (denoted R_3) is similar in nature to R_1, although instead of controlling the *rate* of false discoveries we control the *number* of false discoveries, that is, R_3minimizes the $\overline{\mathrm{FN}}$, subject to $\overline{\mathrm{FD}} \leq \gamma$. The optimal rule is again a threshold rule, now $\delta^*_i = I(\pi_i < t^*_\gamma)$, and we can write the optimal threshold as a step-up procedure: findr

$$r_3 = \max \left\{ j : \sum_{i=1}^{j} \pi_{(i)} \leq \gamma \right\} \qquad (5)$$

and set $t^*_\gamma = \pi_{(r_3+1)}$ so that we reject $H_{(1)}, \cdots, H_{(r_3)}$. Note that by definition, R_2 and R_3 do not specifically control the false discovery rate. However, given that the optimal decision rule for both criteria is a threshold rule (like R_1), they do imply FDR control at some level determined in an indirect way via λ_2 and γ.

With all of these tools at our disposal, which should we use? On one hand, the three different decision criteria R_1, R_2, and R_3 allow the decision maker to choose a criteria based on their application of interest and what feels most natural. On the other hand, the criteria do not yield equivalent inference, even if the thresholds are "equivalent." To illustrate this point, consider Figure 1, which simultaneously visualizes the three criteria by plotting artificial posterior probabilities $\pi_i = P(\theta_i = 0 | Z)$ for $M = 100$ tests along with the threshold statistics corresponding to R_1, R_2, and R_3. The x-axis corresponds to the posterior probabilities π_i, the light gray histogram in the background shows the distribution of the π_i, and each y-axis corresponds to one of the decision criteria. First, the y-axis on the left side of the plot displays the threshold quantity for R_2 (where $\pi = 1/(\lambda_2 + 1) \leftrightarrow \lambda_2 = 1/\pi - 1$, in blue). This axis can be thought of as the minimum $\lambda_2$value for which a given π_i would lead to rejection. The y-axes on the right show the threshold quantities for R_1 (the cumulative average of the $\pi_{(i)}$, in green) and R_3 (the cumulative sum of the $\pi_{(i)}$, in red).

Figure 1. A comparison of the various decision criteria, for a bimodal distribution of $M = 100$ artificial posterior probabilities. The triangular points are plotted on the scale of R_1; the square points are

plotted on the scale of R_2; the circular points are plotted on the scale of R_3. The horizontal threshold line illustrates the cutoff for all three decision criteria: R_1, where we want to make sure that fewer than 20% of our discoveries are false; R_2 (which thresholds the raw probabilities), when we have specified a false discovery to be four times more costly than a false negative; and R_3, where we want to make sure that we have fewer than 20 total false discoveries.

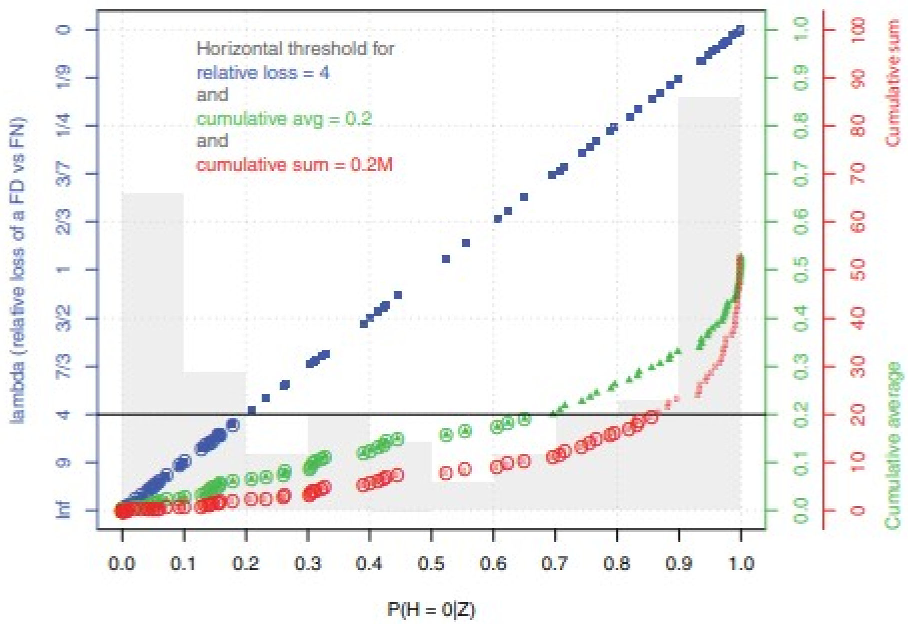


Figure 1 shows “equivalent” horizontal thresholds at $\alpha = 0.2$ for R_1 (meaning that we want to control $\overline{\mathrm{FDR}}$ at 20%), $\lambda_2 = 4$ for R_2 (meaning that we specify a false discovery to be four times as costly as a false negative), and $\gamma = 0.2*100 = 20$ threshold for R_3 (meaning that we want to ensure we have fewer than 20 false discoveries). The π_i values for rejected null hypotheses are circled. The main point of Figure 1 is to show how threshold values from the decision criteria relate to each other. First, if we are willing to think of the R_1 and R_2 cutoffs as equivalent (i.e., that controlling $\overline{\mathrm{FDR}}$ at 20% is equivalent to a false discovery being four times as costly as a false negative), then we can see that R_2 is more conservative than R_1. This is true in general: R_2 thresholds the raw posterior probabilities π_i, while R_1 thresholds the cumulative average. Also, note that while a statistical model can use information from all regions to estimate the individual posterior probabilities (see Section 3), if one uses the R_2 criteria then the distribution of the π_i is unimportant: all posterior probabilities less than the threshold are classified as rejections, regardless of how they are distributed over (0, 1). Alternatively, R_1 (and R_3) considers the *cumulative* posterior probabilities when deciding the classification rule: for example, if there are many posterior probabilities near zero, then tests with posterior probabilities much larger than α can still be rejected (in Figure 1, note that a test with $P(H = 0|\boldsymbol{Z}) \approx 0.65$ is rejected).

Similarly, if we are willing to think of the R_1 and R_3 cutoffs as equivalent (i.e., $\alpha = \gamma/M$), then we can see that R_1 is more conservative than R_3 (again, this is true in general). However, while equating the thresholds for R_1 and R_2 is reasonable, it is much more difficult to equate the thresholds for R_1 and R_3; therefore, it might not make sense to compare R_1 and R_3. The reason for this difference is that R_1considers a *rate* of false discoveries, while R_3 considers a count: as such, the total number of discoveries or rejections is very important. For example, out of 100 tests, setting out to control the FDR at 20% (using R_1) means that if 10 tests are rejected, then having 2 of those 10 *rejections* be incorrect is acceptable. This is quite different than being happy with 20 false discoveries out of 100 *tests* (which is the corresponding statement for R_3).

Two other distributions are shown in Appendix B (in the supplementary materials), comparing the decision criteria for $\{\pi_i\}$ clustered near zero (Figure B.1) and clustered near one (Figure B.2). These figures reiterate the fact that the distribution of the $\{\pi_i\}$ is important for R_1 and R_3. When the π_i are clustered near zero (as in Figure 1), both R_1 and R_3 are quite aggressive and yield qualitatively similar results, rejecting tests for which the posterior probability of the null is large (i.e., tests where $\pi_i \approx 0.65$). Alternatively, when the π_i are clustered near one, R_1 is quite conservative and rejects only a few hypotheses, while R_3 is still quite liberal and rejects many hypotheses. As will be seen later, in Section 4, R_3 is always nonconservative: using this decision criteria will always result in rejecting *at least*$\lfloor\gamma\rfloor$ tests, even when all $\pi_i = 1$.

In conclusion, we reiterate that the choice of decision criteria for a specific application depends on the criteria that feel most natural for the decision maker: indeed, this is one reason that the decision-theoretical approach is so helpful. In light of the differences in R_1, R_2, and R_3, our simulation study (see Section 4) will apply each of these decision rules and summarize the performance of each in terms of their target criteria (i.e., the realized false discovery rate, loss, and false discovery count, respectively).

3. A Robust Nonparametric Bayesian Model with Sparsity for Extreme Ratios

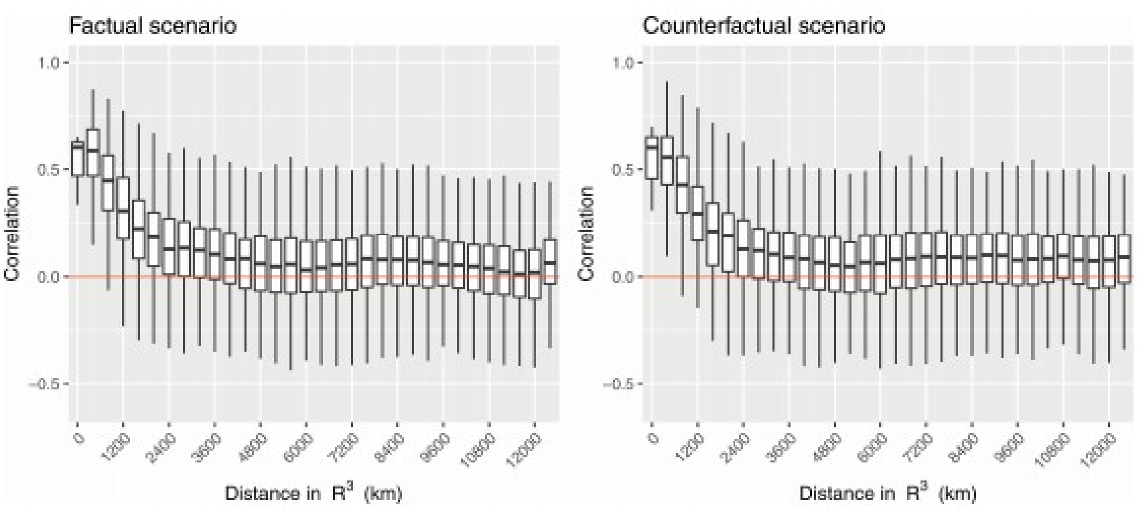
A natural statistical model for the random variables $\boldsymbol{Z} = \{(Z_{Fi}, Z_{Ci}) : i = 1, \cdots, M\}$ from (2) is a binomial likelihood $Z_{ki} \overset{\text{ind}}{\sim} \text{binomial}\,(n_{\text{ens}}, p_{ki})$ which represents a “nonparametric” approach to estimating the event probabilities, as no assumptions need to be made regarding the behavior of the underlying climate variable (as opposed to an extreme value distribution approach). A useful Bayesian framework for this likelihood involves a scenario-specific hierarchical model for the probabilities

$$p_{ki} = \mathrm{logit}^{-1}(\mu_k + \beta_{ki})$$
$$\boldsymbol{\beta}_k \sim G_k \quad (6)$$

for $k \in \{F, C\}$. Here, μ_k are scenario-specific (logit) means, $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kM})$, and $G_k$is a scenario-specific, mean-zero prior distribution for the region-specific effects that flexibly captures dependence among the regions. In principle, G_C and G_F need not be related; however, we model them as arising from the same class but allow for different hyperparameters (and hence the subscript).

While the literature contain a variety of options for how to model the G_k, existing approaches can neither flexibly model (potentially) non-Gaussian behavior nor directly account for irregular or long-range dependence due to teleconnections. As an illustration of the type of correlation, we might expect to see in the scenario-specific probabilities, consider Figure 2, which shows boxplots of the empirical correlations in the logit probability of a hot January over 1959–2014 (binned by distance; see below for details on how this is calculated). Note that the correlations for each scenario tend to be positive, even at long distances. Standard stationary spatial models will likely not be able to account for these irregular dependence relationships.

Figure 2. Empirical correlation between the *logit* probability of a seasonally adjusted hot January (1959–2014; on the anomaly scale) versus distance, for both the factual (left) and the counterfactual (right) scenarios.



Therefore, we seek a model that more robustly uses available data to estimate the covariance between the hypotheses. One way to use prior knowledge to estimate the covariances between regions is to use long time series of climate model simulations: while the WRAF is generated based on monthly simulations of CAM5.1 (again see Section 5.2 for more details), there are also historical simulations of CAM5.1 available for both climate scenarios dating back to 1959. As such, we can use the empirical relationships between the historical simulations of both the factual and the counterfactual to inform the dependence relationships among the hypotheses. Formally, we can estimate monthly probabilities $\left\{\hat{p}_{ki}^{(t,j)} : k \in \{F, C\}; i = 1, \cdots, M; t = 1, \cdots, T; j = 1, \cdots, 12\right\}$ (t represents the year, j represents the month) using a simple beta-binomial Bayesian model (maximum likelihood estimates are not used because zeros are possible), where the corresponding random variables $\{z^{(t)}_{ki}\}$ are calculated using a threshold specific to each month (note: the $z^{(t)}_{ki}$ are different than the random variables introduced in (2)). Both the threshold for what is considered "extreme" and the count variables are calculated based on anomaly data (i.e., the atmospheric variables for each year are mean zero). Then, for a forecast in month j, we have an $M \times T$ matrix of probabilities $\widehat{\boldsymbol{p}}_k^{(j)}$ that can be used to calculate an empirical covariance on the logit scale: $\widehat{\boldsymbol{S}}_k^{(j)} = \text{cov}\left[\text{logit}\,\widehat{\boldsymbol{p}}_k^{(j)}\right]$, where $\widehat{\boldsymbol{p}}_k^{(j)} = \left\{\hat{p}_{ki}^{(t,j)} : i = 1, \cdots, M; t = 1, \cdots, T\right\}$. (Note: the correlation matrices used to create Figure 2 are from the $\text{logit}\,\widehat{\boldsymbol{p}}_k^{(1)}$ for hot months.)

Unfortunately, since for our application we have $T < M$ (the historical simulations only cover $T = 56$ years and there are $M = 237$ regions), the resulting empirical estimate will not be a positive definite matrix; furthermore, it is well-known that the empirical covariance is a poor estimator for the true covariance (see, e.g., Daniels and Kass 2001; Bickel and Levina 2008). Instead, we can use a basis function approach where the basis functions are the eigenvectors of the estimated covariance $\widehat{\boldsymbol{S}}_k^{(j)}$, also known as empirical orthogonal functions (EOFs; see Wikle 2010 or Cressie and Wikle 2011). EOFs are a popular strategy in modeling global climate variables, as the eigenvectors summarize the major modes of variability in a multivariate dataset. Furthermore, it can be shown that the modes of variability (i.e., eigenvectors) are the same for the true covariance and a noisy estimate of the covariance (e.g., Cressie and Wikle 2011). The main idea here is to base the current forecast on past data. While the "past" (here, 1959–2014) is not necessarily a stationary climate (especially for the factual scenario), it can be argued that the modes of variability should be approximately consistent. As an example of the spatial patterns that we are able to capture using the EOF approach, consider the leading empirical EOFs for the logit probability of a hot January, shown in Figures B.3 and B.4 of Appendix B (in the supplemental materials).

Suppressing the j notation, suppose for each month we have a set of p EOF basis functions $\boldsymbol{h}_{kl} = (h_{kl}(\boldsymbol{s}_1), \cdots, h_{kl}(\boldsymbol{s}_M))^\top, l = 1, \cdots, p$, for each scenario, collected into an $M \times p$ matrix $\boldsymbol{H}_k = \left(\boldsymbol{h}_{k1}^\top, \cdots, \boldsymbol{h}_{kp}^\top\right)^\top$ (note that the EOFs are calculated separately for each scenario and event type). Then, following Wikle (2010), we can specify the following model for $\beta_k = (\beta_{k1}, \ldots, \beta_{kM})$:

$$\boldsymbol{\beta}_k = \boldsymbol{H}_k \boldsymbol{\alpha}_k + \boldsymbol{\xi}_k, \quad (7)$$

where $\boldsymbol{\alpha}_k = (\alpha_{k1}, \cdots, \alpha_{kp}) \in \mathbb{R}^p$ is a random vector of basis function coefficients and ξ_k is a residual vector that captures discrepancies from the EOF basis function structure. Because the basis functions are orthogonal, the elements of α_k can be considered independent a priori.

Using (7), we must specify three components: a prior distribution for the residual vector ξ_k, a prior distribution for the basis function coefficients α_k, and the number of EOFs to include in the model (i.e., p).

3.1. Accounting for Non-Gaussian Discrepancy from the EOF Structure

If the number of EOFs is large enough to account for both large-scale and small-scale spatial variability, we can model the residual vector as independent and identically distributed (iid) random variables. A standard approach in nearly all statistical modeling is to assume that error is mean-zero and Gaussian, that is, $\xi_{ki} \overset{\text{iid}}{\sim} N\left(0, \tau_k^2\right)$. However, in modeling extreme probabilities, such an assumption might be tenuous, even on the logit scale. Furthermore, in basing the dependence structure of the current forecast on past data, there is a risk of misspecifying the large-scale structure in the probabilities (i.e., the EOFs). Therefore, it behooves us to use a more flexible approach in accounting for discrepancies from the fixed EOF structure.

One approach to more flexibly model the residual vector is via the skew-t family of distributions (Fernández and Steel 1998; Azzalini and Capitanio 2003; Frühwirth-Schnatter and Pyne 2010), which is a generalization of the Gaussian distribution that allows for skewness and heavy (nonexponential) tails. For a mean-zero random effect, this family involves three parameters: a scale parameter, a skewness parameter, and the degrees of freedom, which controls the heaviness of the tails. Actually, we use what Arellano-Valle and Azzalini (2008) call the "centered" parameterization for the skew-t distribution, where $\sigma_k > 0$ is the scale parameter, $\delta_k \in (-1, 1)$ controls the skewness, and $\nu_k > 0$ is the degrees of freedom (see Appendix C in the supplemental materials for details); we write $\xi_{ki} \overset{\text{iid}}{\sim} ST\left(0, \sigma_k, \delta_k, \nu_k\right)$. In any case, as the degrees of freedom ν approaches zero, the skew-*t* distribution becomes very heavy-tailed and allows quite large deviations from zero. Also, note that the standard Gaussian error approach is a special case of the skew-*t*: for $\delta = 0$ and in the limit as $\nu \to \infty$, the residuals ξ_{ki} are iid Gaussian.

3.2. Sparsity-Imposing Prior for the EOF Coefficients

The EOF framework in (7) is a form of principal component regression (PCR), where the principal components (PCs) of a multivariate dataset are used as regressors or covariates. In general, selecting an appropriate subset of PCs for PCR is extremely important for the sake of interpretation and parsimony of the resulting model; furthermore, we want to avoid overfitting the signal with too many PCs. Existing approaches for selecting the number of PCs (see Jolliffe 2002) generally fall into one of three categories: graphical methods like the scree plot (Cattell 1966); computational methods such as cross-validation (Wold 1978; Josse and Husson 2012); and model-based criteria like reversible jump MCMC (Zhang et al. 2004), marginal likelihood estimation (Minka 2001), and model averaging (Katzfuss et al. 2017). Each of these approaches only consider “nested” PC models, in that a particular PC is included only if all lower-order PCs are included. Other approaches consider data-driven component selection for Bayesian PCR (Wang 2012; Lee and Oh 2013; Junttila et al. 2015), which use various prior distributions to regularize the PC coefficients. For example, Junttila et al. (2015) specified $\alpha_{kl} \overset{\text{iid}}{\sim} N\left(0, v\right)$ where the prior variance *v* is estimated from the data, where the Gaussian prior corresponds to an L_2 penalty in penalized regression. Other recent articles by Hughes and Haran (2013) and Guan and Haran (2018) combine a Gaussian prior for the PC coefficients with either model selection or cross-validation for choosing an appropriate number of PCs.

However, when dealing with empirical PCs, it is often the case that several of the PCs have extremely large variance (corresponding to large coefficients) and many have small variance (corresponding to small coefficients). In a penalized framework, an L_2 penalty corresponds to a linear smoother (see Tansey et al. 2017), which over-penalizes large signals and does not induce sparsity. Laplace priors, which correspond to an L_1 penalty, do encourage sparsity but still overshrink large signals in the presence of many near-zero signals due to their light tails. This is problematic in the PCR setting. A recent thread of research that addresses this problem in a Bayesian framework is the generalized double Pareto prior (GDP; Armagan, Dunson, and Lee 2013), which has a spike at zero (like the Laplace prior) but heavy student’s *t*-like tails. The GDP prior has a simple analytic form, yields a proper posterior distribution, and has a simple characterization as a scale mixture of Gaussian distributions: if $X \sim N(0, V)$, $V \sim \mathrm{Exp}(U^2/2)$, and $U \sim \mathrm{Gamma}(s, r)$, then the marginal distribution of X is the GDP

$$p(x|s, r) = \frac{r}{2s}\left(1 + \frac{|x|}{r}\right)^{-(s+1)}, \qquad (8)$$

written X~GDP(s,r), where *s* and *r* are the Gamma shape and rate, respectively (Armagan, Dunson, and Lee 2013). The GDP prior avoids the overshrinkage problems associated with Gaussian or Laplace priors (Tansey et al. 2017) and encourages sparsity (Taddy 2013). (Note: related work by Polson and Scott (2012) and Carvalho et al. (2010) also introduce heavy-tailed shrinkage priors with similar properties, but unfortunately these priors are not available in closed form when marginalized.)

Thus, while shrinkage priors have been used for PC or EOF selection, in a novel approach we propose to use the GDP prior as a more appropriate framework for incorporating EOF selection into the prior specification. Instead of worrying about how many EOFs to include in (7), we will instead incorporate all $T = 56$ EOFs, so that $p = T$. Formally, the prior for the EOF coefficients is $\alpha_{kl} \overset{\text{iid}}{\sim} \mathrm{GDP}\,(s, r)$ for $l = 1, \ldots, T$. Note that using an exchangeable prior on the coefficients in this way ignores information about smoothness of the signals, an aspect which is not present in all regression settings but is present in PCR. When dealing with PCs or EOFs, we expect the variances of the empirical PCs to decay smoothly (following the eigenvalues), so that an exchangeable prior on the coefficients is not quite right. Lee and Oh (2013) explicitly included a “smooth” prior for the coefficients by specifying a functional form for how their prior variances decay. However, in a general setting, it is not immediately obvious what type of decay is most appropriate. In our application, we use a different set of data to calculate the EOFs (i.e., the historical simulations) than what is used to actually estimate the coefficients (i.e., the year of simulations corresponding to the forecast year). In this case, there may be some mismatch between the historical simulations and the new data with respect to the magnitude and ordering of the signals. Using an exchangeable prior can account for this mismatch and also does not require one to specify a functional form for the decay in the coefficient variances.

3.3. Hyperprior Specification and Computation

The hyperparameters for (7) are the mean μ_k, the skew-*t* parameters $\{\sigma_k, \delta_k, \nu_k\}$, and the GDP parameters $\{s, r\}$. For the mean and skew-*t* parameters, we use proper but noninformative priors, namely, $p(\mu_k) = N(0, 10^2)$, $p(\sigma_k) = U(0, 100)$, $p(\delta_k) = U(-1, 1)$, and $p(1/\nu_k) = U(0, 1)$, where $U(a, b)$ is the uniform distribution over the interval (a, b). Note that the prior for the degrees of freedom is actually on $1/\nu_k$ to improve mixing; furthermore, the upper bound of 1 on $1/\nu_k$ limits the tails to be no heavier than those of a Cauchy distribution. The hyperparameters of the GDP are slightly more complicated. Armagan, Dunson, and Lee (2013) suggested fixing these at $s = 1$ and $r = 1$; Taddy (2013) fixed both hyperparameters at $s = 1$ and $r = 1/2$, but found that his results were robust to other values of s. Tansey et al. (2017) also fixed $s = 1$ but encountered major problems when trying to estimate r and instead fit separate models across a discrete grid of fixed values for r and used DIC to choose the best value. Following these suggestions, we fix $s = 1$, but, given the relative simplicity of (7) (compared to Taddy 2013 and Tansey et al. 2017), we were able to estimate r from the data and used $p(r) = U(0, 100)$.

As is usually the case, the posterior distribution for this model is not available in closed form regardless of prior specification, so we resort to Markov chain Monte Carlo (MCMC) methods to obtain samples from the joint posterior distribution. We fit the model using the nimble software for R (de Valpine et al. 2017), which is a BUGS-like system for building and sharing analysis methods for statistical models, particularly for hierarchical frameworks. The MCMC is relatively straightforward, and code to fit the model is available in the online reproducibility documents.

4. Sensitivity to Model Misspecification

The classical FDR procedures in the vein of Sun and Cai (2007) were developed for specific data models, and unfortunately Sun et al. (2015) found that the optimality of the procedure is quite sensitive to model misspecification. While the Bayesian procedures of Müller et al. (2004) are appropriate for more general classes of models, Newton et al. (2004) noted that the bounds on $\overline{\mathrm{FN}}$, $\overline{\mathrm{FD}}$, and $\overline{\mathrm{FDR}}$ are "approximate … because [they] rest on the accuracy of the fitted model." Furthermore, as noted in Section 2, the performance of the Bayesian decision rules for Frequentist FDR is not guaranteed in the presence of correlation (Pacifico et al. 2004; Guindani et al. 2009). As such, we wish to understand how both model misspecification and dependence impact the performance of these decision rules for the WRAF application and, subsequently, ensure that the nonparametric Bayesian framework developed in Section 3 is robust to both correlation and error in specifying a statistical model for the probabilities $\{p_{ki}\}$, $k \in \{F, C\}$.

To compare the performance of our robust nonparametric Bayesian approach outlined in Section 3 (henceforth labeled RNB) as well as several other related models within the various Bayesian decision rules based on R_1, R_2, and R_3, we perform a simulation study to explore the FDR performance for a variety of "true" states. For our simulation study, the number of regions will match that of the WRAF regions, that is, $M = 237$, and we use ensemble sizes of $n_{\text{ens}} = \{50, 100, 400\}$. A total of N_{rep}=100 datasets will be generated from each of six "true states" (see Table 1) that are designed to represent the full space of all possible "states" for the CAM5.1 simulations. The hyperparameters for the true states will be fixed (see Appendix F of the supplemental materials), and the replicates will be drawn from the random effect distributions (as opposed to repeated binomial draws with the same effects). Complete details on the procedure for obtaining samples from each true state is provided in Appendix F, and code to generate samples from each true state is provided in the online reproducibility documents.

For comparison, we also use a variety of standard models that might traditionally be used for the WRAF. Each of these again use the Z_{ki} introduced in (2), as well as the binomial likelihood

$$Z_{ki} \overset{\text{ind}}{\sim} \text{binomial}\left(n_{\text{ens}}, p_{ki}\right)$$

used in Section 3.

Classical likelihood ratio test. To compare new approaches with classical FDR, we first outline a method for calculating a *p*-value for each null hypothesis using a Frequentist likelihood ratio test. Rewriting the hypotheses in terms of the probabilities H: $p_F/p_C \leqslant c$ (for now ignoring the region-specific subscript), the test statistic for a likelihood ratio test considers the ratio of likelihoods for $Z_C = z_C$ and $Z_F = z_F$:

$$\lambda(z_C, z_F) = \frac{\sup_{\Theta_0} L(p_F, p_C | z_C, z_F)}{\sup_{\Theta} L(p_F, p_C | z_C, z_F)},$$

where Θ_0 is the parameter space defined by the null hypothesis and Θ is the entire (unrestricted) parameter space for p_F and p_C. The likelihood is the product of individual Binomial likelihoods:

$$L(p_F, p_C | z_C, z_F) \propto (p_F)^{z_C}(1 - p_F)^{n_C - z_C}(p_C)^{z_F}(1 - p_C)^{n_F - z_F}.$$

It can be shown that the likelihood in the denominator is maximized for the MLEs p^C=zC/nC and p^F=zF/nF. Alternatively, for the numerator, the restricted MLEs arep^CR,p^FR=(p^C,p^F)ifp^F>(p^C/c)(p˜C,p˜C/c)ifp^F≤(p^C/c),where p˜C=(1/4)(-b-b2-8d): b=-[c(1+p^F)+1+p^C], and d=c(p^C+p^F) (Farrington and Manning 1990). Statistical theory says -2logλ(zC,zF)→dχ12 as n_C, $n_F \to \infty$ (Θ involves two free parameters while $\Theta_0$has just one); thus, an asymptotic *p*-value is $p(\chi^2_1 > -2\log\lambda(z_C, z_F))$. Note that when p^F>p^C/c, the likelihood ratio is 1, $-2\log\lambda(z_C, z_F) = 0$, and the null hypothesis will never be rejected. The resulting collection of *p*-values can be used for a classical FDR (Benjamini and Hochberg 1995) or a Bonferroni-style family-wise error rate (FWER) procedure.

Parametric Bayesian models for the risk ratio. Again using the independent binomial likelihood for the Z_{ki} in (2), the simplest Bayesian approach to modeling these probabilities is to estimate each of the p_{Fi} and p_{Ci} independently of each other and all of the other regions (an "independent across regions" model), henceforth M1. For $k \in \{F, C\}$ and $i = 1, \ldots, M$, simply use a conjugate beta prior $\pi(p_{ki}) = \mathcal{B}(a_p, b_p)$ for the binomial likelihood so that the posterior is

$$\pi(p_{ki}|Z_{ki} = z_{ki}) = \mathcal{B}(z_{ki} + a_p, n_{\text{ens}} - z_{ki} + b_p). \qquad (9)$$

Posterior samples can be obtained by direct sampling from (9). Alternatively, mirroring the hierarchical Bayesian framework in (6), a variety of prior models for G_k can be implemented:

M2 Exchangeable Gaussian prior: a random effects framework is useful for borrowing strength across the regions without the notion of spatial dependence. Here, we use $\beta_{ki} \overset{iid}{\sim} N(0, \tau_k^2)$.

M3 Exchangeable skew-*t* prior: however, the Gaussian assumption may be too restrictive, in that the effects could be nonsymmetric and heavy-tailed. Alternatively, we can use the skew-*t* family of distributions (as in Section 3.1), with $\beta_{ki} \overset{iid}{\sim} ST(0, \sigma_k, \delta_k, \nu_k)$.

M4 Conditionally autoregressive (CAR) prior: a natural approach for areal data like the WRAF regions, a CAR prior models $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kM})$ as a spatial random effect (see, e.g., Pascutto et al. 2000; Banerjee, Carlin, and Gelfand 2004). Using a Gaussian model, the joint distribution for $\boldsymbol{\beta}_k$ can be defined in terms of the conditional distributions

$$p\big(\beta_{ki}|\{\beta_{kj} : j \neq i\}\big) = N\left(\frac{1}{|\partial i|}\sum_{j\in\partial i}\beta_{kj}, \frac{\tau_k^2}{|\partial i|}\right), \quad i = 1, \ldots, M,$$

where ∂i is the set of regions that share a border with region i (the "neighborhood") and $|\partial i| = \#$ regions in the neighborhood. This specification is also called an *intrinsic* CAR model which is an improper prior (Rue and Held 2005 outline various ways to address this issue; see Appendix E).

M5 Hybrid CAR/exchangeable prior: the model outlined in Leroux, Lei, and Breslow (2000) offers a compromise between M2 and M4. As outlined in Leroux, Lei, and Breslow (2000), the τ^2_k parameter in M4 represents both overdispersion and spatial dependence, and these features may be in stark contrast. A variety of strategies are used in the literature to address this problem (see, e.g., Cressie 1991); Leroux, Lei, and Breslow (2000) instead specified an approach based on "additive precisions," in which the precision matrix of the random effects is a convex combination of the exchangeable and CAR precision matrices:

$$\Sigma_k^{-1} \equiv \tau_k^{-2}\left[(1-\lambda_k)I + \lambda_k Q\right], \qquad (10)$$

where Q is the CAR precision correlation matrix; $\lambda_k \in [0, 1]$ is a parameter that controls the degree of spatial dependence. Note that in this framework, $\lambda_k = 0$ corresponds to M2 while $\lambda_k = 1$ corresponds to M4; furthermore, Σk-1 is full rank for $\lambda \in [0, 1)$. Using (10), the full conditional distributions for the individual random effects are

$$p\left(\beta_{ki} \mid \{\beta_{kj} : j \neq i\}\right) = N\left(\frac{\lambda_k}{1-\lambda_k+\lambda_k|\partial i|}\sum_{j\in\partial i}\beta_{kj}, \frac{\tau_k^2}{1-\lambda_k+\lambda_k|\partial i|}\right), \quad i = 1, \ldots, M.$$

M6 Gaussian process prior: another alternative to M4 is to use a Gaussian process prior for β_k, defined for the centroids of each region (e.g., Kelsall and Wakefield 2002). Like M5, independent random effects are a special (limiting) case of the Gaussian process prior, such that M6 can flexibly model both independent and dependent effects (unlike M4). For this approach, βk∼NM(0,Σk): Σkij=τk2Mν(||si-sj||/ϕk), where Mν(·) is the Matérn correlation function with smoothness ν, si,sj are the three-dimensional coordinates for the centroids of regions *i* and *j*, and || · || represents Euclidean distance on R3. In practice, since we are fitting a Gaussian process model to areal data and therefore do not observe data at very short distances, we fix $\nu = 0.5$.

M7-9 EOF-based structure with a Gaussian prior for a fixed number of coefficients: for comparison, we use (7) with a more traditional exchangeable Gaussian prior on the coefficients, that is, αkl∼iidN(0,σ2α)αkl∼iidN(0,σα2), where σ^2_α is estimated from the data. In this framework, we use three different EOF truncations: $p = 30$, which matches the number of EOFs used in the data generation (henceforth M7); $p = 10$, where we use too few EOFs (henceforth M8); and $p = 50$, where we use too many EOFs (henceforth M9). Models M7–9 will allow us to assess the performance of the GDP prior relative to more traditional priors with a specified truncation.

A summary of all the fitted models and their labels is given in Table 2. Details on the hyperpriors and computation (via MCMC) are given in Appendices D and E (in the supplemental materials).

Table 1. The true states from which the simulated datasets are generated.

Label	True state
G-RE	Gaussian random effects
NG-RE	Gamma (non-Gaussian) random effects
GP-S	(Matérn) Gaussian process, short range of dependence
GP-L	(Matérn) Gaussian process, long range of dependence
EOF-G	Fixed EOF structure, Gaussian discrepancy
EOF-NG	Fixed EOF structure, gamma (non-Gaussian) discrepancy

NOTE: The Matérn correlation function for GP-S and GP-L has smoothness $\nu = 2$. Furthermore, the EOF true states use a fixed number of EOFs ($p = 30$) in the data generation.

Table 2. A summary of the models fit to each simulated dataset. Note that RNB and M1-M9 are implemented in a Bayesian framework.

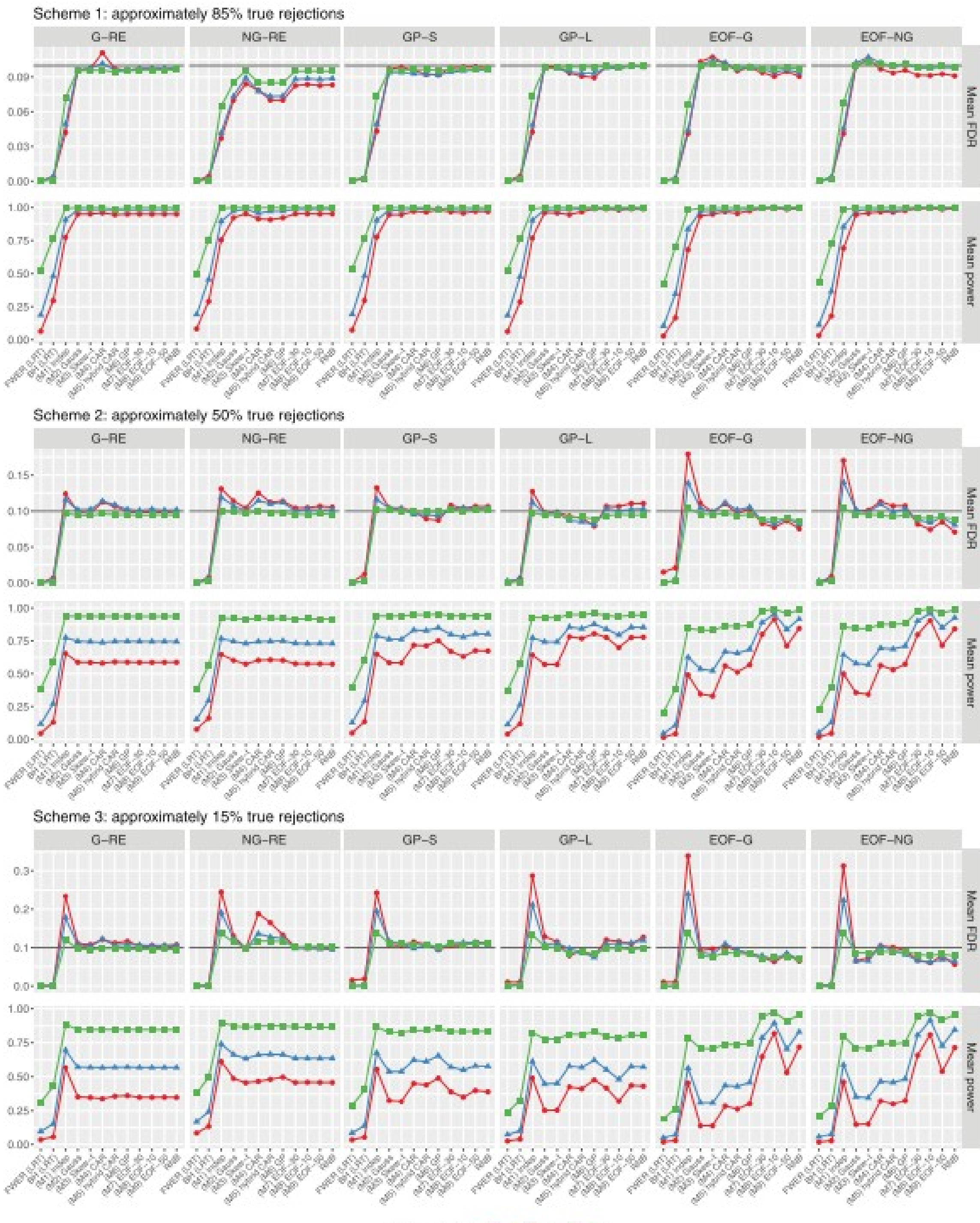
Label	Model description
RNB	Robust nonparametric model with sparsity
LRT	Classical likelihood ratio test (region-specific)
M1	Beta-binomial (independent-across-regions)
M2	Exchangeable Gaussian prior
M3	Exchangeable skew-*t* prior
M4	Conditionally autoregressive (CAR) prior
M5	Hybrid CAR/exchangeable prior
M6	Gaussian process prior with Matérn correlation
M7	EOF-based structure with $p = 30$ EOFs
M8	EOF-based structure with $p = 10$ EOFs
M9	EOF-based structure with $p = 50$ EOFs

For each of the simulated datasets, the performance of the three decision criteria R_1, R_2, and R_3 will be compared by fitting each of the models outlined in Table 2. The null hypothesis for each simulation will use a threshold value of $c = 1$ (testing for an increase in p_F relative to p_C) while attempting to control FDR in ways comparable to the classical 0.1 significance level: for R_1, set $\alpha = 0.10$; for R_2, set $\lambda_2 = 1/0.1 - 1 = 9$; for R_3, set $\gamma = 0.10M = 23.7$. Three different sets of hyperparameters will be used for each model, corresponding to cases in which most tests are true rejections (Scheme 1, ≈ 0.85), around half of tests are true rejections (Scheme 2, ≈ 0.5), and most tests are true nulls (Scheme 3, ≈ 0.15). More details are provided in Appendix F.

Two final notes regarding the model fitting. First, for the Gaussian process model M6, note that the correlation function is set to be exponential, while the true states GP-S and GP-L have a Matérn correlation with smoothness $\nu = 2$. Second, all of the EOF approaches (RNB and M7-M9) require the initial step of estimating the EOF matrix, which is considered fixed when fitting the model. The EOFs used for true states EOF-G and EOF-NG are estimated from the historical simulations of temperature in January as described in Section 3.1 (the same EOFs are used for both generating datasets and model fitting). However, the benefit of the EOF framework is that it can robustly use available data to improve the model; therefore, when fitting RNB and M7-M9 to the other true states (G-RE, NG-RE, GP-S, and GP-L), we first calculate the EOFs using T = 56 replicates drawn from the true state, separately for each scenario. For example, the EOFs used to fit data from GP-L would correspond to the covariance of a stationary Gaussian process with Matérn correlation function.

Results, summarized across simulated replicates. We present results for the R_1criteria here, in the main text of the article, as this decision criteria corresponds most closely with the classical notions of FDR; see Figure 3. The top, middle, and bottom sub-plots show the FDR and power (i.e., the probability of rejecting a false null) for Schemes 1, 2, and 3 (respectively), averaged over the Nrep=100replicated datasets. The sub-panels show the six true states, and the different methods/fitted models are shown along the x-axes.

Figure 3. FDR and power using the R_1 criteria, aggregated over the Nrep=100replicates, for schemes 1, 2, and 3. Note that the x-axis in each subgrid corresponds to the different methods/fitted models. The target of $\alpha = 0.1$ is plotted for FDR.



The first observation to make is that the frequentist *p*-value approaches (FWER and BH) are clearly over-conservative, such that the realized FDR is approximately zero (well below the target $\alpha = 0.1$) across all schemes, ensemble sizes, and true states. This over-conservativeness shows up in the power plots as well, with the FWER procedure in particular suffering from extremely small power, even for the largest ensemble size. Interestingly, the

Bayesian models RNB and M1-M9 each seem to do fairly well at controlling the FDR and maximizing the power (minimizing the FNR) for Scheme 1, across true states (aside from the independence model M1, which has somewhat reduced power particularly for the true states with spatial dependence).

Schemes 2 and 3 tell a different story: for each of these schemes, and across true states, the independence model M1 is anti-conservative and fails to control the FDR (except for the largest ensemble size in Scheme 2). Otherwise, several items are noteworthy: the CAR model M4 performs poorly for the true states that do not include spatial dependence (G-RE and NG-RE); only the models that can accommodate skewness (models M3, M7-M9, and RNB) control the FDR for the NG-RE data. Otherwise, each of the models is mostly able to control the FDR, although major differences show up in the power. While, for example, M2 (a model without spatial dependence) is able to control the FDR for the GP-L simulations in Scheme 2, the power is significantly smaller than for a model that does accommodate dependence, for example, M4.

The EOF models M7-M9 and RNB perform comparably for the independent random effects (G-RE and NG-RE) with respect to both FDR and power, but yield major differences for the true states with spatial dependence. These differences are again most obvious in the power, where we can clearly see how under or overfitting the EOFs plays out. For the GP-S and GP-L effects, M8 (which uses only 10 EOFs) has reduced power relative to M7, M9, and RNB. This is not entirely surprising since 10 EOFs might be insufficient for characterizing a Matérn covariance. For the EOF-G and EOF-NG true states, recall that these data were generated with 30 EOFs, so that M7 (which also uses 30 EOFs) is in a sense the “correct” model. However, as is discussed in Appendix F of the supplemental materials, to mimic the decay present in the corresponding empirical eigenvalues, the last 20 EOF coefficients have very small prior variance (see Table F.5 in the supplement). Therefore, M8 (with 10 EOFs) is also approximately the “correct” model. In fact, M8 performs best for the EOF true states, while overfitting the EOFs as in M9 results in greatly reduced power. Our new approach RNB nearly matches the power of M8 for these true states without having to specify an EOF truncation. Interestingly, the presence of spatial dependence (or lack thereof) in the simulated data has a larger effect on the power than the FDR: when the effects do not include dependence (G-RE and NG-RE), the power is roughly the same for models M2-M9 and RNB. This is true even for the NG-RE effects, for which the Gaussian-based models M2, M4, M5, and M6 struggle to control the FDR.

Therefore, if one were to choose a “best” model for the R_1 decision criteria, the robust nonparametric Bayesian model with a sparsity-imposing prior for the EOF coefficients and skew-*t* discrepancy (RNB) is the clear choice, as it performs well across schemes and true states. RNB is able to control the FDR at approximately the nominal level for every combination of true state/scheme, and yields (almost) the largest power with the exception of the EOF-G/NG true states in Schemes 2 and 3. The only model that performs better for these true states is M8, which is suboptimal for the GP-S/L true states. Thus, RNB performs nearly as well as the best of the other EOF approaches, without requiring the specification of an EOF truncation. In some ways, this is not surprising, since the magnitude of the EOF coefficients together with their GDP shrinkage prior can differentiate between cases both with and without spatial dependence and the flexibility of the skew-*t*residuals can capture both symmetric and nonsymmetric effects. Furthermore, this approach allows us to more robustly use the data at hand (in this case, the historical CAM simulations) to capture irregular (nonstationary) spatial dependence patterns.

The story is largely the same for the R_2 and R_3 criteria (see Figures B.5 and B.6 in Appendix B of the supplemental materials): the RNB model yields the smallest loss (almost always), and controls the number of FDs while minimizing the number of FNs. Therefore, we have good reason to select the RNB model combined with the decision rule of interest as the procedure that best controls the realized loss, FDR, and FD.

5. Applying the Multiple Testing Procedure to the WRAF

Having identified the robust nonparametric Bayesian model as an approach that flexibly controls false discoveries for each of the procedures outlined in Section 2, we now turn to applying the procedure to a real dataset of climate model simulations.

5.1. Selection of Decision Criteria

For this application, we decided not to use the second decision criteria (R_2) because it is not clear for the WRAF what the relative loss for each type of error should be; in other words, there is no obvious way to equate the cost of a false discovery and a false negative. In deciding between R_1 and R_3, we were initially drawn to R_3 because certain choices for the threshold γ allow us to make sure our statements are scientifically significant. In systematically conducting a set of hypotheses regarding the presence of anthropogenic influence on extreme weather, to conclude a significant overall (global) influence we would need to reject a null hypothesis of no anthropogenic influence for some nonzero proportion of the globe, for example, we might want to see rejections for 5% of the globe. In other words, from a practical perspective, we might be willing to make 10 false rejections (about 5% of the 237 regions) because if we find fewer than 10 rejections then there is likely not a scientifically meaningful anthropogenic effect for the entire globe. Furthermore, in making an absolute (instead of a relative) statement about the number of false discoveries, the total number of discoveries relate to overall confidence: rejecting only a few hypotheses indicates low confidence that there is any overall anthropogenic effect, while rejecting many hypotheses indicates high confidence that there is indeed some overall anthropogenic effect.

However, upon further investigation, it became obvious that the R_3 criteria is too liberal. Returning to (5), note that this procedure will always reject *at least* $\lfloor\gamma\rfloor$ tests: in the most extreme case, where $\pi_{(i)} = 1$ for all i (meaning the null hypothesis receives all of the posterior probability), we will still have $\sum_{i=1}^{\lfloor\gamma\rfloor}\pi_{(i)} \leqslant \gamma$, so that in this case $r_3 = \lfloor\gamma\rfloor$. In other words, a set of tests will be rejected, even though the posterior probability that each null hypothesis is true is 1; clearly, it is quite awkward to always reject a set of hypotheses despite the evidence. Therefore, if we use the R_3 criteria, after flagging a set of null hypotheses to reject we must then determine if the results are believable. For example, if $\gamma = 10$ and we only reject 10 hypotheses, then we must conclude that almost all of these are false discoveries; alternatively, if we reject 100 hypotheses, then we can be confident that most of these are true rejections. However, what if we reject 15 hypotheses? Or 20? In these "in-between" cases, we must decide when enough tests have been rejected to conclude that at least some of the rejections are true.

The R_1 criteria, on the other hand, falls more in line with traditional multiple testing procedures, in that the conclusions drawn for a set of hypotheses are more appropriately adjusted for the fact that multiple tests are being conducted. Regardless of how many tests are rejected under this criteria, we can always be sure that (in expectation) only a small proportion of these are being falsely rejected. Furthermore, while the R_3 criteria might flag some hypotheses for rejection in spite of the large posterior probability that the null is true (see the previous paragraph), the R_1 criteria will only begin flagging hypotheses for rejection if the smallest posterior probabilities of the null being true (i.e., $\pi_{(1)}$, $\pi_{(2)}$, etc.) are close to zero. A final benefit of using the R_1 criteria is that the conclusions for nested hypotheses (see Section 5.2.2) will be consistent (e.g., the procedure will only reject $H_i : RR_i^{(\mathrm{wet})} \leq 2$ if $H_i : RR_i^{(\mathrm{wet})} \leq 1$ is also rejected), which is not the case for R_3.

5.2. Case Study Methods

Having opted to use the R_1 criteria, we set $\alpha = 0.1$ (as is done in Section 4). Practitioners often choose an FDR threshold based on common significance levels; here, we do the same, although there is no reason why this should be done (other than the fact that we want the FDR to be small but not too small such that we have no power). We then applied our robust nonparametric Bayesian model with the generalized double Pareto prior for the EOF coefficients with the R_1 criteria to the WRAF for two case studies: (1) hot events in January 2015 and (2) wet events in March 2015. For hot events, we use a more stringent cutoff for the null hypotheses, $c_{\mathrm{hot}}=5$ (i.e., testing for $H_i : RR_i^{(\mathrm{hot})} \leq 5$; this is due to the stronger anthropogenic signal for temperature), while for wet events we use $c_{\mathrm{wet}}=1$.

The data for estimating the p_{ki} ($k \in \{F, C\}$) consist of output from large ensembles of simulations of version 5.1 of the Community Atmospheric Model (CAM5.1) global atmosphere/land climate model, run in its conventional ~ 1° longitude/latitude configuration (Neale et al. 2010; Stone et al. 2018). Simulations have been run under the experiment protocols of the C20C+ Detection and Attribution Project (D. Stone and P. Pall, "A benchmark estimate of the effect of anthropogenic emissions on the ocean surface," submitted), following two historical scenarios (Angélil et al. 2017) and will be regularly updated through time as a contribution to both the C20C+ D&A project and the WRAF. The first set of simulations (for the factual scenario) is driven by observed boundary conditions of atmospheric chemistry (greenhouse gases, tropospheric and stratospheric aerosols, ozone), solar luminosity, land use/cover, and the ocean surface (temperature and ice coverage). The second set of simulations (for the counterfactual scenario) is driven by what observed boundary conditions might have been in the absence of historical anthropogenic emissions: the anthropogenic component of atmospheric chemistry is set to year-1855 values, ocean temperatures are cooled by a seasonally and spatially varying estimate of the warming attributable to anthropogenic emissions, and sea ice concentrations are adjusted for consistency with the ocean temperatures (Stone and Pall, submitted). Simulations within a scenario differ only in the starting conditions. The data and further details on the simulations are available at *http://portal.nersc.gov/c20c*. The simulations for both scenarios cover 01/1959 to 06/2015; the (time-varying) ensemble sizes are given in Table 3[Figure 4].

Table 3. CAM5.1 ensemble sizes from January 1959 to June 2015 used in the case studies.

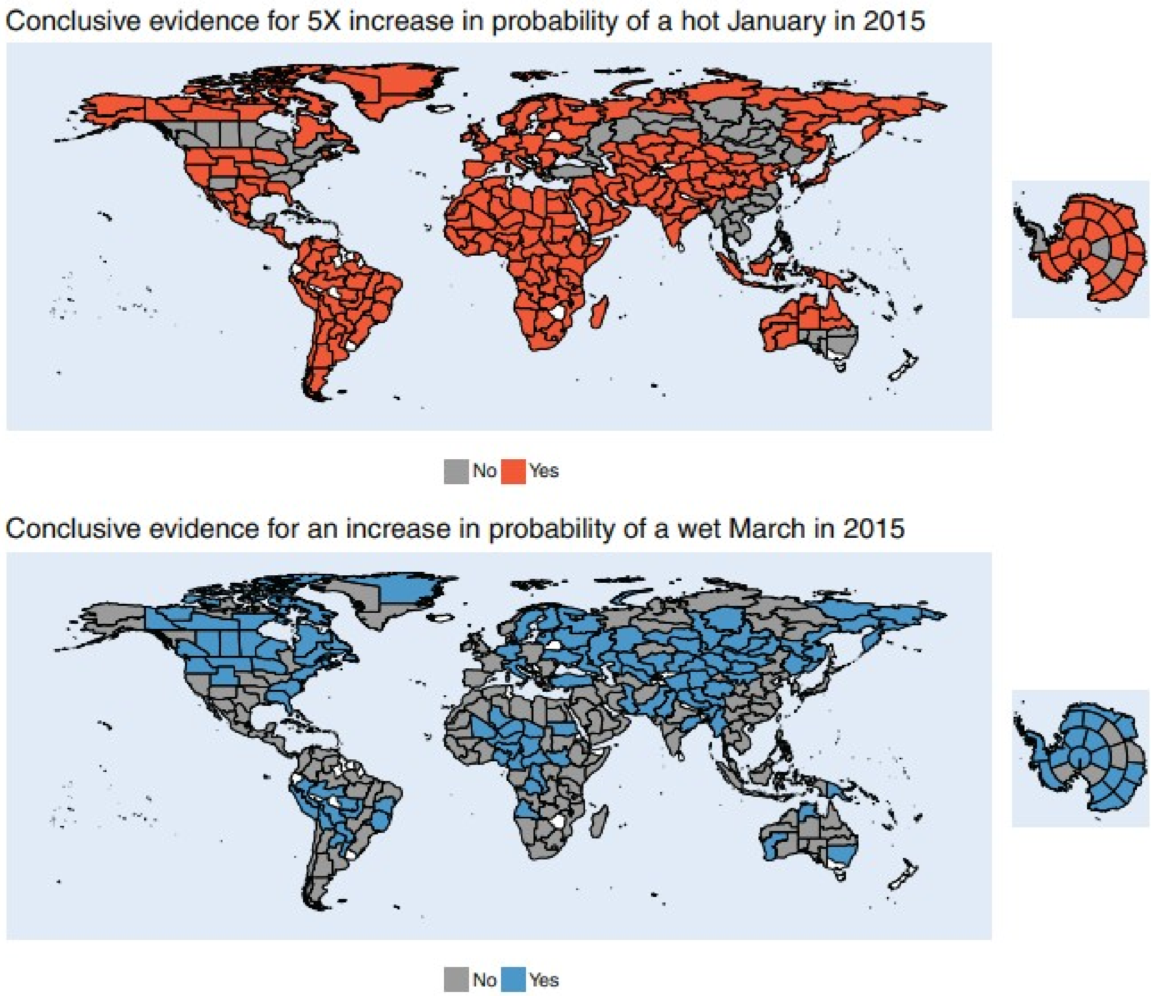
Time range	Ensemble size
01/1959 to 12/1996	50 (both factual and counterfactual)
01/1997 to 12/2009	100 (both factual and counterfactual)
01/2010 to 12/2013	400 (both factual and counterfactual)
01/2014 to 10/2014	100 (both factual and counterfactual)
10/2014 to 06/2015	98 (factual)
10/2014 to 06/2015	99 (counterfactual)

The event of interest for both case studies (used to define the region-specific probabilities p_{ki}) is the occurrence of a month that is more extreme than the third most extreme event expected over the preceding 30 year period. In other words, for a forecast in 2015, the event definition is the 1-in-10 year event which, for hot and wet months, corresponds to the 0.9 quantile of average monthly temperature or precipitation for each region in the factual simulations over 1985–2014 (specifically using the monthly measurements from the 50-member ensemble that covers this entire period). Using a moving time period of fixed length (30 years) ensures that we have accounted for climate change and that the events we consider are extreme in the “current” climate. All of the historical CAM simulations (both the factual and counterfactual) from the entire 1959–2014 period are used to calculate the EOFs following the procedure outlined in Section 3; the simulations from 2015 are used to fit the statistical model and classify the hypotheses. Otherwise, all prior specifications and computation via MCMC are the same as described in the simulation study. Note that an implicit assumption of this application is that the CAM5.1 simulations are suitable for evaluating changes in the probability of extremes (Angélil et al. 2016, 2017).

5.2.1. Results for a Single Set of Hypotheses

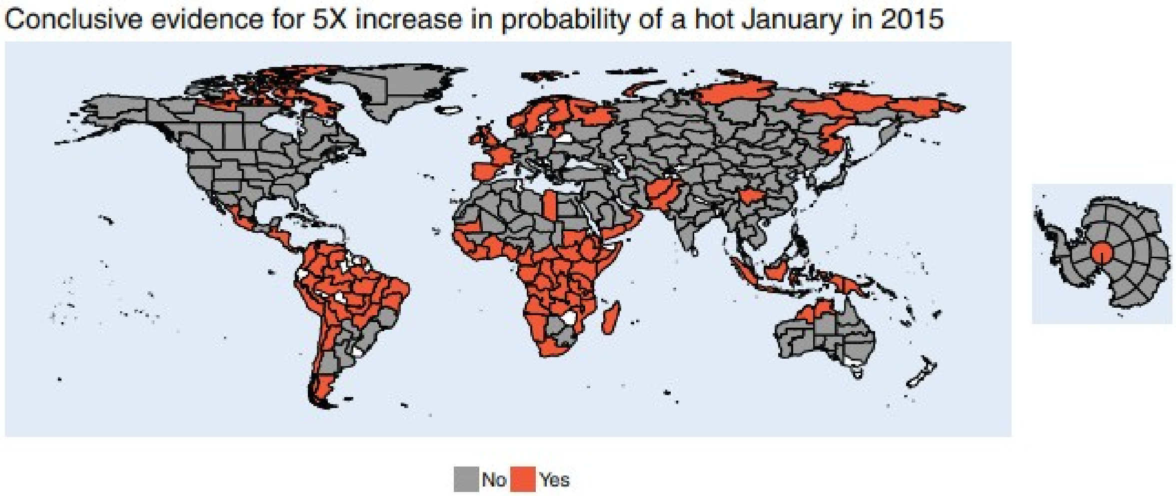
The results for each case study are shown in Figure 4. Even with a larger cutoff for hot Januarys (testing for a five-fold increase as opposed to simply an increase), an overwhelming majority of the regions (194 of 237) have experienced a large degree of anthropogenic warming in 2015, with only a few regions in North America, Southeast Asia, central Russia, and southeast Australia failing to provide conclusive evidence of a five-fold increase in occurrence probability of a hot January 2015 (every region has conclusive evidence against the null hypothesis when the cutoff is relaxed to chot=1). The results are more varied for wet events in March of 2015, as there are many regions with and without conclusive evidence against the null hypothesis. Many regions in the northern extratropics (mid to high latitudes) have an increased probability of a wet event in March 2015 as a result of anthropogenic emissions. An increased probability is the general tendency along an equatorial band as well (although not in Southeast Asia), while the subtropics (arid regions in the tropics) generally lack conclusive evidence; a notable exception is over the northern Sahel, which may indicate an earlier advance of the West African monsoon in this climate model due to anthropogenic emissions (Lawal et al. 2016).

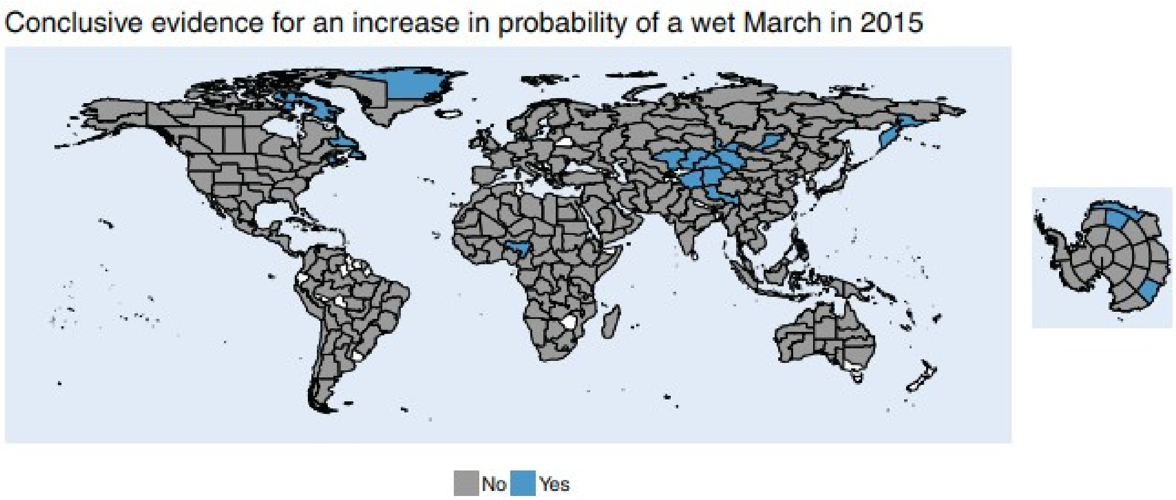
Figure 4. Results of testing a collection of hypotheses Hi:RRi(hot)≤5 (top, i.e., determining if there is conclusive evidence for a five-fold increase in the probability that January 2015 will have an average temperature that exceeds the third hottest expected January over 1985–2014) and Hi:RRi(wet)≤1 (bottom, i.e., determining if there is conclusive evidence for an increase in the probability that the total precipitation in March 2015 will exceed the third wettest expected March over 1985–2014). The white areas (e.g., New Zealand, Zimbabwe) do not satisfy criteria for fitting into political regions of the target 400,000–900,000 km^2 range as described in Stone (2018) and are not analyzed here.



For comparison to standard (frequentist) FDR methods, we show maps for the corresponding forecast using the traditional Benjamini and Hochberg (1995) procedure based on likelihood ratio test *p*-values; see Figure 5. The BH procedure is again much more conservative than the corresponding results using our Bayesian approach (shown in Figure 4), identifying conclusive evidence of changes in extreme probabilities for a greatly reduced subset of the WRAF regions. As an aside, we note that maps like Figure 4 produced using the Bayesian framework with other fitted models (i.e., M2-M9; not shown) yield only mild differences from the map based on our new modeling approach. In this case, where there is no way to assess which fitted model yields the “correct” results, we prefer our new approach based on the results of the simulation study in Section 4.

Figure 5. As in Figure 4, but using the Benjamini and Hochberg (1995) procedure based on likelihood ratio test *p*-values. The white areas (e.g., New Zealand, Zimbabwe) do not satisfy criteria for fitting into political regions of the target 400,000–900,000 km^2 range as described in Stone (2018) and are not analyzed here.





5.2.2. Capturing the Existence and Magnitude of Anthropogenic Influence

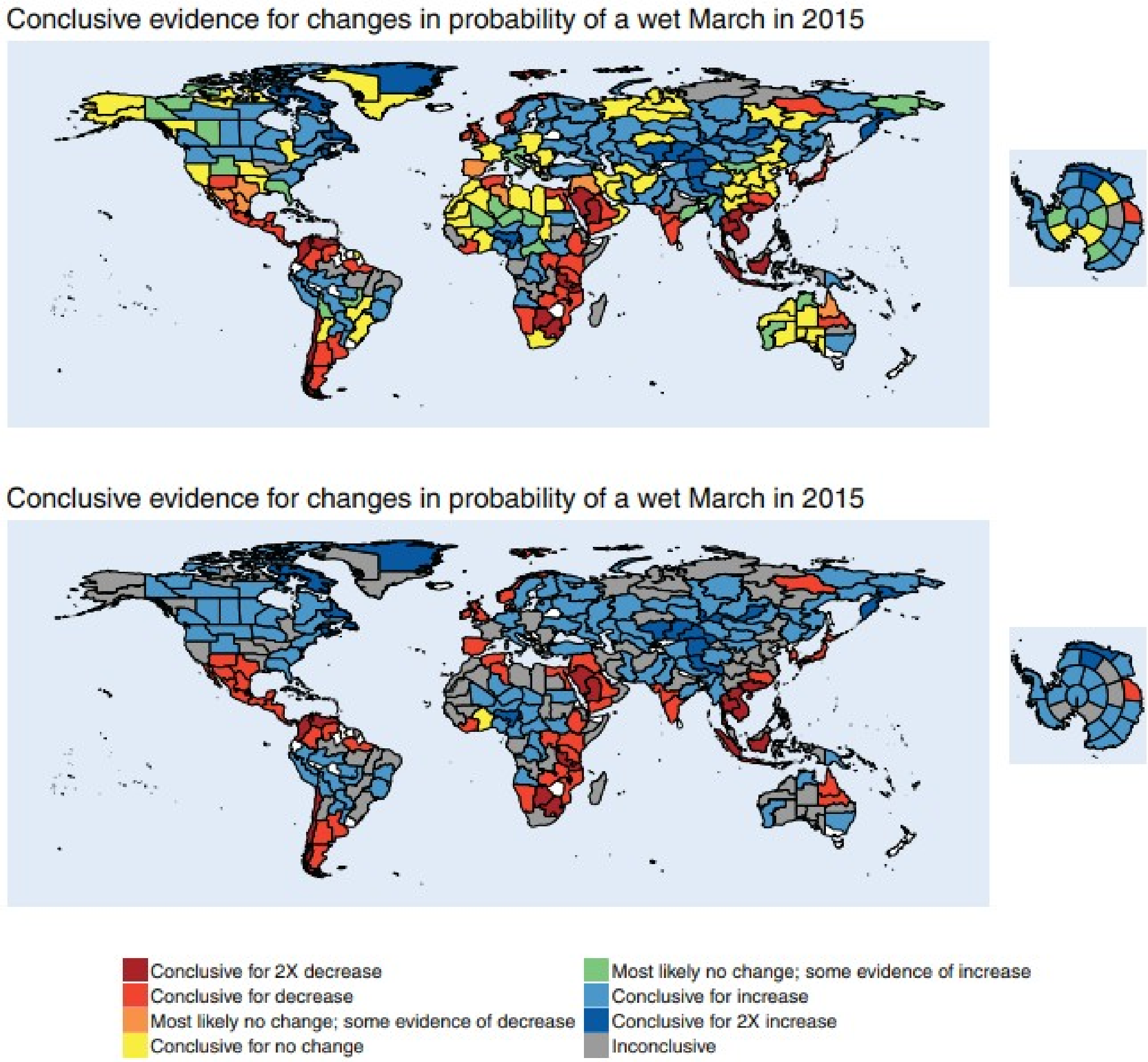
Both the current and planned upcoming versions of the WRAF actually conduct more than one set of hypothesis tests for each forecast: several different thresholds are used (e.g., cwet=1 vs. cwet=2) in conjunction with several different types of null hypotheses (e.g., Hi:RRi(wet)≤cwet vs. Hi:RRi(wet)≥cwet). The purpose of these categories is to make statements that combine confidence in the change in probability as well as the magnitude of this change. As such, the forecast actually involves “multiple–multiple testing,” in that we now have multiple sets of *M* hypotheses to test. This can be accomplished in our framework by simply conducting the classification procedure several times; recall from Section 5.1 that using R_1 yields consistent results for nested hypotheses (unlike R_3). The testing adjustment is done separately for each category, and therefore the existence of any possible false discoveries can be interpreted within each category.

As an example of what the attribution forecast looks like for multiple categories, see Figure 6. A benefit of the Bayesian framework is that we can first test for the *absence* of an anthropogenic effect using a null hypothesis like Hi(1):RRi(wet)≤labsence∪RRi(wet)≥uabsence, where uabsenceand labsence are upper and lower limits, respectively, for an interval including 1 that defines “no anthropogenic influence.” Regions where we can reject $H^{(1)}{}_i$display strong evidence that anthropogenic forcings have not changed the probability of extreme precipitation. Otherwise, the other null hypotheses of interest are

$$H_i^{(2)} : RR_i^{(\text{wet})} \geq 1/2; \quad H_i^{(3)} : RR_i^{(\text{wet})} \geq 1;$$
$$H_i^{(4)} : RR_i^{(\text{wet})} \leq 1; \quad H_i^{(5)} : RR_i^{(\text{wet})} \leq 2;$$

being able to reject these hypotheses indicates conclusive evidence that the probability of extreme precipitation is decreased by a factor of two, decreased, increased, or increased by a factor of two (respectively). There is clearly some overlap between $H^{(1)}{}_i$ and both $H^{(3)}{}_i$ and $H^{(4)}{}_i$; to reflect this, we create two additional categories to indicate regions that reject both $H^{(1)}{}_i$ and $H^{(3)}{}_i$ (orange, indicating that while there is most likely no change in the probability there is some evidence for a decrease) as well as both $H^{(1)}{}_i$ and $H^{(4)}{}_i$ (green, indicating that while there is most likely no change in the probability there is some evidence for an increase). A final category (shown in gray) identifies regions that fail to reject any of the hypotheses and are thus classified as inconclusive.

Figure 6. Results of testing multiple hypotheses per region, to capture the magnitude and direction of the effect of anthropogenic influence. Top: $1/2 \leqslant RR_i \leqslant 2$ defines the “Conclusive for no change” category; bottom: $2/3 \leqslant RR_i \leqslant 3/2$ defines the “Conclusive for no change” category.



Maps of these multi-category results are shown in Figure 6, where we use both a wide interval labsence=1/2 and uabsence=2 as well as narrower limits labsence=2/3 and uabsence=3/2. Angélil et al. (2017Angélil, O., Stone, D., Wehner, M., Paciorek, C. J., Krishnan, H., and Collins, W. (2017), "An Independent Assessment of Anthropogenic Attribution Statements for Recent Extreme Temperature and Rainfall Events," *Journal of Climate*, 30, 5–16. [Crossref], [Web of Science ®], [Google Scholar]) provided justification for using the narrower interval (2/3, 3/2) as the definition of "no anthropogenic influence;" however, the somewhat limited ensemble sizes (≈ 100) in this case study prevent us from conclusively finding no change (except in one region) for the narrower interval (bottom, Figure 6). The wider interval used for the top panel of Figure 6, on the other hand, concludes that a fairly large proportion of the map experiences no change. Clearly, being able to conclude that extreme probabilities are unchanged between the two climate scenarios depends heavily upon both the ensemble size and width of the interval that defines "no change," and this tradeoff can result in significant qualitative differences. Given that the geometric range spanned by (2/3, 3/2) is about half that spanned by (1/2, 2), a quadrupling of the ensemble size (i.e., increased to 400 members) would be expected to result in the identification of a substantial number of regions in the “no change” categories, in contrast to when only 100 members are available (as also noted in analysis of 2013 events when the ensemble sizes are larger; not shown).

6. Conclusions

In this article, we have developed a hierarchical Bayesian modeling framework for estimating the probability of extreme events and the risk ratio over a large collection of land-regions, as well as a decision-theoretical procedure that allows us to flexibly control the number of false discoveries while maximizing the number of true discoveries. The Bayesian hierarchical model robustly uses historical climate model simulations to estimate irregular (nonstationary) dependence patterns among the hypotheses, can account for non-Gaussian behavior in the region-specific effects, and uses an appropriate shrinkage prior that does not require choosing an EOF truncation point. Furthermore, we show that the modeling framework maintains false discovery control even when the true data-generating mechanism arises from a completely different class of statistical models. Finally, we apply our robust statistical model to a real dataset used for making seasonal forecasts for the Weather Risk Attribution Forecast. Moving forward, we plan to operationalize our procedure as described in Section 5.2 to replace the current ad hoc presentation of the forecast.

We have demonstrated the application of our procedure across regions and with multiple hypotheses for each region. However, we have not applied it across event types (e.g., hot, cold, wet, and dry events for a single region) or across multiple months. Application across regions makes sense for several reasons. First, events are presented in global maps of these regions (as in Figure 6) and thus are not only providing information on each region individually but also on the aggregate of all of the regions. Second, even with some correlation across the regions, there remains a large “effective sample size” of tests, whereas testing across event types would yield a small number of tests (such that a multiple testing adjustment has less value). While testing across multiple months (e.g., all months or only the same calendar month from a given period of years) may provide a moderate number of tests, it would be hard to fit into the monthly operational design of the WRAF. Continual updating of past calculations, as further months become available, would pose a presentation and communication challenge. However, in a more retrospective research framework, studying events over a decade for instance, testing across months as well as, or instead of, regions could make sense.

Finally, while the final version of the forecast will use the M = 237 regions shown in Figure 4 (where each region is approximately 0.5 million km^2; these are the WRAF05 regions), the WRAF will also provide a forecast for aggregates of these regions: 68 regions comprising 2 million km^2 each (WRAF2); 30 regions comprising 5 million km^2 each (WRAF5); and 12 regions comprising 10 million km^2 each (WRAF10). As a demonstration of how our procedure will perform for a smaller number of regions, we conducted a simulation study similar to the one in Section 4using the larger WRAF2 regions (M = 68; these regions are slightly modified from the current version of the WRAF, which has 58 regions); these results are shown in Appendix A.2 (see supplemental materials). Results for the WRAF2 regions are approximately consistent with the simulation study results for the smaller, WRAF05 regions.

References

Allen, M. (2003), "Liability for Climate Change," *Nature*, 421, 891–892.

Angélil, O., Perkins-Kirkpatrick, S., Alexander, L. V., Stone, D., Donat, M. G., Wehner, M.,Shiogama, H., Ciavarella, A., and Christidis, N. (2016), "Comparing Regional Precipitation and Temperature Extremes in Climate Model and Reanalysis Products," *Weather and Climate Extremes*, 13, 35–43.

Angélil, O., Stone, D., Wehner, M., Paciorek, C. J., Krishnan, H., and Collins, W. (2017), "An Independent Assessment of Anthropogenic Attribution Statements for Recent Extreme Temperature and Rainfall Events," *Journal of Climate*, 30, 5–16.

Angélil, O., Stone, D. A., Tadross, M., Tummon, F., Wehner, M., and Knutti, R. (2014), "Attribution of Extreme Weather to Anthropogenic Greenhouse Gas Emissions: Sensitivity to Spatial and Temporal Scales," *Geophysical Research Letters*, 41, 2150–2155.

Arellano-Valle, R. B., and Azzalini, A. (2008), "The Centred Parametrization for the Multivariate Skew-Normal Distribution," *Journal of Multivariate Analysis*, 99, 1362–1382. Special Issue: Multivariate Distributions, Inference and Applications in Memory of Norman L. Johnson.

Arent, D. J., Tol, R. S. J., Faust, E., Hella, J. P., Kumar, S., Strzepek, K. M., Tóth, F. L., Yan, D.,Abdulla, A., and Kheshgi, H. (2014), "Key Economic Sectors and Services," in *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, pp. 659–708.

Armagan, A., Dunson, D. B., and Lee, J. (2013), "Generalized Double Pareto Shrinkage,"*Statistica Sinica*, 23, 119–143.

Azzalini, A., and Capitanio, A. (2003), "Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew t-Distribution," *Journal of the Royal Statistical Society*, Series B, 65, 367–389.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data. Monographs on Statistics and Applied Probability*, London: Chapman & Hall.

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.

——— (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.

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.

Besag, J., York, J., and Mollié, A. (1991), “Bayesian Image Restoration, with Two Applications in Spatial Statistics,” *Annals of the Institute of Statistical Mathematics*, 43, 1–20.

Bickel, P. J., and Levina, E. (2008), “Covariance Regularization by Thresholding,” *The Annals of Statistics*, 36, 2577–2604.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), “The Horseshoe Estimator for Sparse Signals,” *Biometrika*, 97, 465–480.

Cattell, R. B. (1966), “The Scree Test for the Number of Factors,” *Multivariate Behavioral Research*, 1, 245–276.

Cressie, N. A. C. (1991), *Statistics for Spatial Data*, New York: Wiley.

Cressie, N., and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data (Wiley Series in Probability and Statistics)*, New York: Wiley.

Daniels, M. J., and Kass, R. E. (2001), “Shrinkage Estimators for Covariance Matrices,”*Biometrics*, 57, 1173–1184.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R.(2017), “Programming with Models: Writing Statistical Algorithms for General Model Structures with Nimble,” *Journal of Computational and Graphical Statistics*, 26, 403–413.

Efron, B. (2007), “Size, Power and False Discovery Rates,” *The Annals of Statistics*, 1351–1377.

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.

Farrington, C. P., and Manning, G. (1990), “Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk,” *Statistics in Medicine*, 9, 1447–1454.

Fernández, C., and Steel, M. F. (1998), "On Bayesian Modeling of Fat Tails and Skewness,"*Journal of the American Statistical Association*, 93, 359–371.

Frühwirth-Schnatter, S., and Pyne, S. (2010), "Bayesian Inference for Finite Mixtures of Univariate and Multivariate Skew-Normal and Skew-t distributions," *Biostatistics*, 11,317–336.

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.

Guan, Y., and Haran, M. (2018), "A Computationally Efficient Projection-Based Approach for Spatial Generalized Linear Mixed Models," *Journal of Computational and Graphical Statistics*, accepted.

Guindani, M., Müller, P., and Zhang, S. (2009), "A Bayesian Discovery Procedure," *Journal of the Royal Statistical Society*, Series B, 71, 905–925.

Hansen, G., Auffhammer, M., and Solow, A. R. (2014), "On the Attribution of a Single Event to Climate Change," *Journal of Climate*, 27, 8297–8301.

Hughes, J., and Haran, M. (2013), "Dimension Reduction and Alleviation of Confounding for Spatial Generalized Linear Mixed Models," *Journal of the Royal Statistical Society*, Series B, 75, 139–159.

Jolliffe, I. T. (2002), *Principal Component Analysis*, New York: Springer-Verlag.

Josse, J., and Husson, F. (2012), "Selecting the Number of Components in Principal Component Analysis using Cross-Validation Approximations," *Computational Statistics & Data Analysis*, 56, 1869–1879.

Junttila, V., Kauranne, T., Finley, A. O., and Bradford, J. B. (2015), "Linear Models for Airborne-Laser-Scanning-Based Operational Forest Inventory With Small Field Sample Size and Highly Correlated Lidar Data," *IEEE Transactions on Geoscience and Remote Sensing*, 53, 5600–5612.

Katzfuss, M., Hammerling, D., and Smith, R. (2017), "A Bayesian Hierarchical Model for Climate-Change Detection and Attribution," *Geophysical Research Letters*, 44, 5720–5728.

Kelsall, J., and Wakefield, J. (2002), "Modeling Spatial Variation in Disease Risk," *Journal of the American Statistical Association*, 97, 692–701.

Lawal, K. A., Abatan, A. A., Angélil, O., Olaniyan, E., Olusoji, V. H., Oguntunde, P. G.,Lamptey, B., Abiodun, B. J., Shiogama, H., Wehner, M. F., and Stone, D. A. (2016), "The Late Onset of the 2015 Wet Season in Nigeria," *Bulletin of the American Meteorological Society*, 97, S63–S69.

Lee, J., and Oh, H.-S. (2013), "Bayesian Regression Based on Principal Components for High-Dimensional Data," *Journal of Multivariate Analysis*, 117, 175–192.

Leroux, B. G., Lei, X., and Breslow, N. (2000), *Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence*, New York, NY: Springer, pp. 179–191.

Livezey, R. E., and Chen, W. Y. (1983), “Statistical Field Significance and its Determination by Monte Carlo Techniques,” *Monthly Weather Review*, 111, 46–59.

Minka, T. (2001), “Automatic Choice of Dimensionality for PCA,” *Advances in Neural Information Processing Systems*, 13, 598–604.

Müller, P., Parmigiani, G., and Rice, K. (2006), “FDR and Bayesian Multiple Comparisons Rules,” Working Paper 115, Johns Hopkins University, Dept. of Biostatistics.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004), “Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays,” *Journal of the American Statistical Association*, 99, 990–1001.

National Academies of Sciences, E., and Medicine, (2016), *Attribution of Extreme Weather Events in the Context of Climate Change*, Washington, DC: The National Academies Press.

Neale, R. B., Chen, C.-C., Gettelman, A., Lauritzen, P. H., Park, S., Williamson, D. L., Conley,A. J., Garcia, R., Kinnison, D. Lamarque, J.-F., Marsh, D., Mills, M., Smith, A. K., Tilmes, S. Vitt, F., Morrison, H., Cameron-Smith, P., Collins, W. D., Iacono, M. J., Easter, R. C., Ghan, S. J., Liu, X., Rasch, P. J., and Taylor, M. A. (2010), “Description of the NCAR Community Atmosphere Model (CAM 5.0),” *NCAR Tech*, Note NCAR/TN-486+STR.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), “Detecting Differential Gene Expression with a Semiparametric Hierarchical Mixture Method,” *Biostatistics*, 5,155–176.

Pacifico, M. P., Genovese, C., Verdinelli, I., and Wasserman, L. (2004), “False Discovery Control for Random Fields,” *Journal of the American Statistical Association*, 99, 1002–1014.

Pall, P., Aina, T., Stone, D. A., Stott, P. A., Nozawa, T., Hilberts, A. G. J., Lohmann, D., andAllen, M. R. (2011), “Anthropogenic Greenhouse Gas Contribution to Flood Risk in England and Wales in Autumn 2000,” *Nature*, 470, 382–385.

Pascutto, C., Wakefield, J., Best, N., Richardson, S., Bernardinelli, L., Staines, A., and Elliott,P. (2000), “Statistical Issues in the Analysis of Disease Mapping Data,” *Statistics in Medicine*, 19, 2493–2519.

Polson, N. G., and Scott, J. G. (2012), “Good, Great, or Lucky? Screening for Firms with Sustained Superior Performance using Heavy-Tailed Priors,” *The Annals of Applied Statistics*, 6, 161–185.

Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications (Monographs on Statistics and Applied Probability*, Vol. 104), London: Chapman & Hall.

Shu, H., Nan, B., and Koeppe, R. (2015), “Multiple Testing for Neuroimaging via Hidden Markov Random Field,” *Biometrics*, 71, 741–750.

Smith, K. R., Woodward, A., Campbell-Lendrum, D., Chadee, D. D., Honda, Y., Liu, Q.,Olwoch, J. M., Revich, B., and Sauerborn, R. (2014), “Human Health: Impacts, Adaptation, and Co-Benefits,” in *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, pp. 709–754.

Stone, D. A., and Allen, M. R. (2005), “The End-to-End Attribution Problem: From Emissions to Impacts,” *Climatic Change*, 71, 303–318.

Stone, D. A., Risser, M. D., Angélil, O. M., Wehner, M. F., Cholia, S., Keen, N., Krishnan, H.,O'Brien, T. A., and Collins, W. D. (2018), “A Basis Set for Exploration of Sensitivity to Prescribed Ocean Conditions for Estimating Human Contributions to Extreme Weather in CAM5. 1-1Degree,” *Weather and Climate Extremes*, 19, 10–19.

Storey, J. D. (2003), “The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value,” *Annals of Statistics*, 31, 2013–2035.

Stott, P. A., Allen, M., Christidis, N., Dole, R. M., Hoerling, M., Huntingford, C., Pall, P.,Perlwitz, J., and Stone, D. (2013), *Attribution of Weather and Climate-Related Events*,Netherlands, Dordrecht: Springer, pp. 307–337.

Stott, P. A., Stone, D. A., and Allen, M. R. (2004), “Human Contribution to the European Heatwave of 2003,” *Nature*, 432, 610–614.

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.

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

Sun, W., Reich, B. J., Tony Cai, 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.

Taddy, M. (2013), “Multinomial Inverse Regression for Text Analysis,” *Journal of the American Statistical Association*, 108, 755–770.

Tansey, W., Athey, A., Reinhart, A., and Scott, J. G. (2017), “Multiscale Spatial Density Smoothing: An Application to Large-Scale Radiological Survey and Anomaly Detection,”*Journal of the American Statistical Association*, 112, 1047–1063.

Ventura, V., Paciorek, C. J., and Risbey, J. S. (2004), “Controlling the Proportion of Falsely Rejected Hypotheses when Conducting Multiple Tests with Climatological Data,” *Journal of Climate*, 17, 4343–4356.

Wang, L. (2012), “Bayesian Principal Component Regression with Data-Driven Component Selection,” *Journal of Applied Statistics*, 39, 1177–1189.

Wikle, C. K. (2010), “Low-Rank Representations for Spatial Processes,” in *Handbook of Spatial Statistics*, eds. A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, London:Chapman and Hall/CRC, pp. 107–118.

Wilks, D. S. (2016), “The Stippling Shows Statistically Significant Grid points: How Research Results are routinely Overstated and Overinterpreted, and What to do About it,” *Bulletin of the American Meteorological Society*, 97, 2263–2273.

Wold, S. (1978), “Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models,” *Technometrics*, 20, 397–405.

Zhang, Z., Chan, K., Kwok, J., and Yeung, D. (2004), “Bayesian Inference on Principal Component Analysis Using Reversible Jump Markov Chain Monte Carlo,” in *Proceedings of Nineteenth National Conference on Artificial Intelligence (AAAI?04)*, San Jose, CA, p.372.

Funding

This research was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under contract no. DE-AC02-05CH11231 and used resources of the National Energy Research Scientific Computing Center (NERSC), also supported by the Office of Science of the U.S. Department of Energy, under contract no. DE-AC02-05CH11231.