

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

Supplemental Materials

Mark D. Risser

Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory
and

Christopher J. Paciorek

Department of Statistics, University of California, Berkeley
and

Dáithí A. Stone

Computational Research Division, Lawrence Berkeley National Laboratory

November 13, 2017

A Classical model-specific decision theory approaches for false discovery control

Using a Frequentist perspective, [Sun and Cai \(2007\)](#) frame the multiple testing problem in a compound decision theory framework. This thread of research considers controlling the marginal FDR, using the fact that, under weak conditions, $\text{mFDR} = \text{E}(\text{FDR}) + \mathcal{O}(M^{-1/2})$ ([Genovese and Wasserman, 2002](#)). [Sun and Cai \(2007\)](#) note that two approaches can be taken to address the multiple testing problem. First, one can set out with the goal of separating the non-null hypotheses from the nulls, using a weighted classification approach. In other words, the decision rule δ is constructed by minimizing the classification risk $\text{E}[L_\lambda(\theta, \delta)]$, where the loss function is

$$L_\lambda(\theta, \delta) = \frac{1}{M} \sum_{i=1}^M \left\{ \lambda(1 - \theta_i)\delta_i + \theta_i(1 - \delta_i) \right\}; \quad (\text{A.1})$$

here, $\lambda > 0$ is the loss attached to a false positive error (relative to a false negative error). Alternatively, one can set out with the goal of discovering as many true findings as possible while incurring a low proportion of false positive findings: in other words, find δ with the smallest false non-discovery rate (FNR) among all rules with the FDR bounded by $\alpha \in (0, 1)$. [Sun and Cai \(2007\)](#) go on to show that these two approaches are equivalent as long as a monotone likelihood ratio condition is satisfied; that is, the optimal solution to the classification problem (where λ depends on the desired α) is also optimal for the multiple testing approach, in the sense that the classification rule yields the smallest marginal false negative rate (mFNR) among all procedures that bound $\text{mFDR} \leq \alpha$.

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. On one hand, [Benjamini and Yekutieli \(2001\)](#) show that FDR is controlled at the stated level for dependent hypotheses using either the original approach in [Benjamini and Hochberg \(1995\)](#) or the adaptive procedure in [Benjamini and Hochberg \(2000\)](#). However, on the other hand, [Efron \(2007\)](#) found that non-zero correlation between tests can result in testing procedures that are either too conservative or too anti-conservative; [Schwartzman and Lin \(2011\)](#) show that the procedure can fail to be consistent as the number of tests grows under certain types of dependence. [Sun and Cai \(2009\)](#) also note that in dealing with the effects of correlation on an FDR procedure,

the efficiency of the procedure should be the focus (not just the validity), and that failing to model any known dependence structure can impact the optimality of the procedure. 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 Z_i ; i.e., $\delta_i(\mathbf{Z}) = \delta_i(Z_i)$, and therefore symmetric, meaning that $\delta(\tau(\mathbf{Z})) = \tau(\delta(\mathbf{Z}))$ for all permutation operators τ ([Sun and Cai, 2007](#)). 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 the other Z_j , $j \neq i$) are preferred in that they might be able to identify non-nulls with a smaller signal by pooling information across tests. For example, when hypotheses are positively correlated within a temporal or spatial domain, one would expect that the non-null θ_i would appear in groups or clusters ([Sun and Cai, 2009](#)).

As a result, [Sun and Cai \(2009\)](#) extend the compound decision framework for multiple testing in the presence of dependence. Specifically, modeling the unknown θ_i as random effects arising from a hidden Markov model (HMM), [Sun and Cai \(2009\)](#) prove that the optimal classification rule for the loss function (A.1) is of the form $\delta_i = I(T_i < t_\lambda)$, where

$$T_i = P_\xi(\theta_i = 0|\mathbf{Z}) \quad (\text{A.2})$$

is the so-called “oracle statistic” and ξ is a vector of all hyperparameters in the HMM. It is important to note that the derivation of (A.2) in [Sun and Cai \(2009\)](#) as the oracle statistic is specific to the HMM framework. Furthermore, because the HMM satisfies a monotone likelihood ratio condition, T_i is also the optimal statistic for the multiple testing problem, in that $\delta_i = I(T_i < t_\lambda)$ yields the smallest mFNR subject to $\text{mFDR} \leq \alpha$. The relationship between λ and α can be seen by writing the decision rule as a step-up procedure (like [Benjamini and Hochberg, 1995](#)): first, rank the oracle statistics $T_{(1)} \leq \dots \leq T_{(M)}$, and find

$$r = \max \left\{ j : \frac{1}{j} \sum_{i=1}^j T_{(i)} \leq \alpha \right\}; \quad (\text{A.3})$$

then, reject $H_{(1)}, \dots, H_{(r)}$. In practice, of course, the T_i (and hence the $\{\delta_i\}$ and r) are unknown: [Sun and Cai \(2009\)](#) outline a data-driven procedure that uses a plug-in estimate $\hat{\xi}$ to estimate $\hat{T}_i = P_{\hat{\xi}}(\theta_i = 0|\mathbf{Z})$ and therefore determine r by replacing $T_{(i)}$ with $\hat{T}_{(i)}$ in (A.3). Since the estimated oracle test statistic for the i th hypothesis depends on the entire vector of data, [Sun and](#)

Cai (2009) note that the decision rule is neither simple nor symmetric.

Two recent papers by Sun et al. (2015) and Shu et al. (2015) extend the work of Sun and Cai (2009) to provide similar results for spatial random fields and multi-dimensional Markov random fields (MRFs), respectively. In spite of the different statistical models, in both cases the oracle statistic is the same as (A.2) and the decision rule can be written as (A.3). However, model-specific proofs are required to verify that (1) the classification risk is indeed minimized by $\delta_i = I(T_i < t_\lambda)$, and (2) the optimal classification (oracle) statistic satisfies a monotone likelihood ratio condition and hence yields the smallest mFNR among all procedures with $\text{mFDR} \leq \alpha$ (here, both mFNR and mFDR are defined in a Frequentist sense). Furthermore, *estimation* of the oracle statistic T_i is, of course, model-specific. Sun and Cai (2009) use random effect prediction conditional on hyperparameter estimates: in the HMM, conditional on $\hat{\xi}$, the oracle statistic can be expressed in terms of forward and backward density variables, which can be calculated recursively. Sun et al. (2015) also conduct random effect prediction (albeit marginalizing over hyperparameters), but, since there is no longer an iterative formula for calculating the \hat{T}_i for a Gaussian random field, they instead utilize the Bayesian computational framework (i.e., Markov chain Monte Carlo) as a way to “extract information effectively from large spatial data sets” and implement their data-driven procedure.

Both Sun and Cai (2009) and Sun et al. (2015) conduct simulation studies to verify that their approach outperforms traditional FDR procedures (e.g., BH and AP) when simulated data arise from the true statistical model (i.e., HMM or Gaussian random field). However, Sun et al. (2015) also find that “the precision of [their] testing procedure shows some sensitivity to model misspecification.”

B Supplemental figures

Supplemental figures for the main text are shown in Figures B.1, B.2, B.3, B.4, B.5, and B.6. Results for the simulation study with the larger WRAF regions (WRAF2 with $M = 68$) are shown in Figures B.7, B.8, and B.9.

B.1 Main text

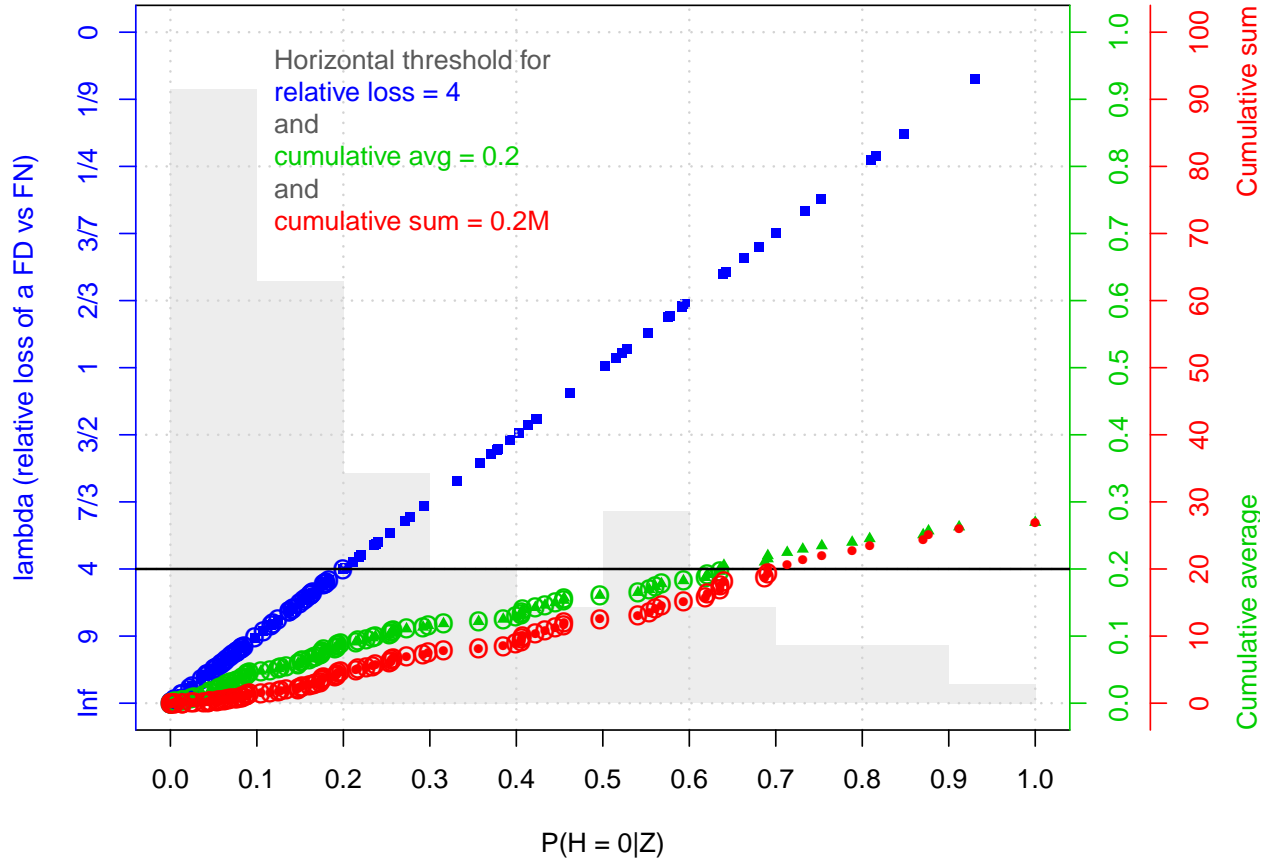


Figure B.1: A comparison of the various decision criteria, for $M = 100$ artificially-generated posterior probabilities clustered around zero. The triangular points are plotted on the scale of R_1 ; the square points are plotted on the scale of R_2 ; the circular points are plotted on the scale of R_3 . 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 4 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.

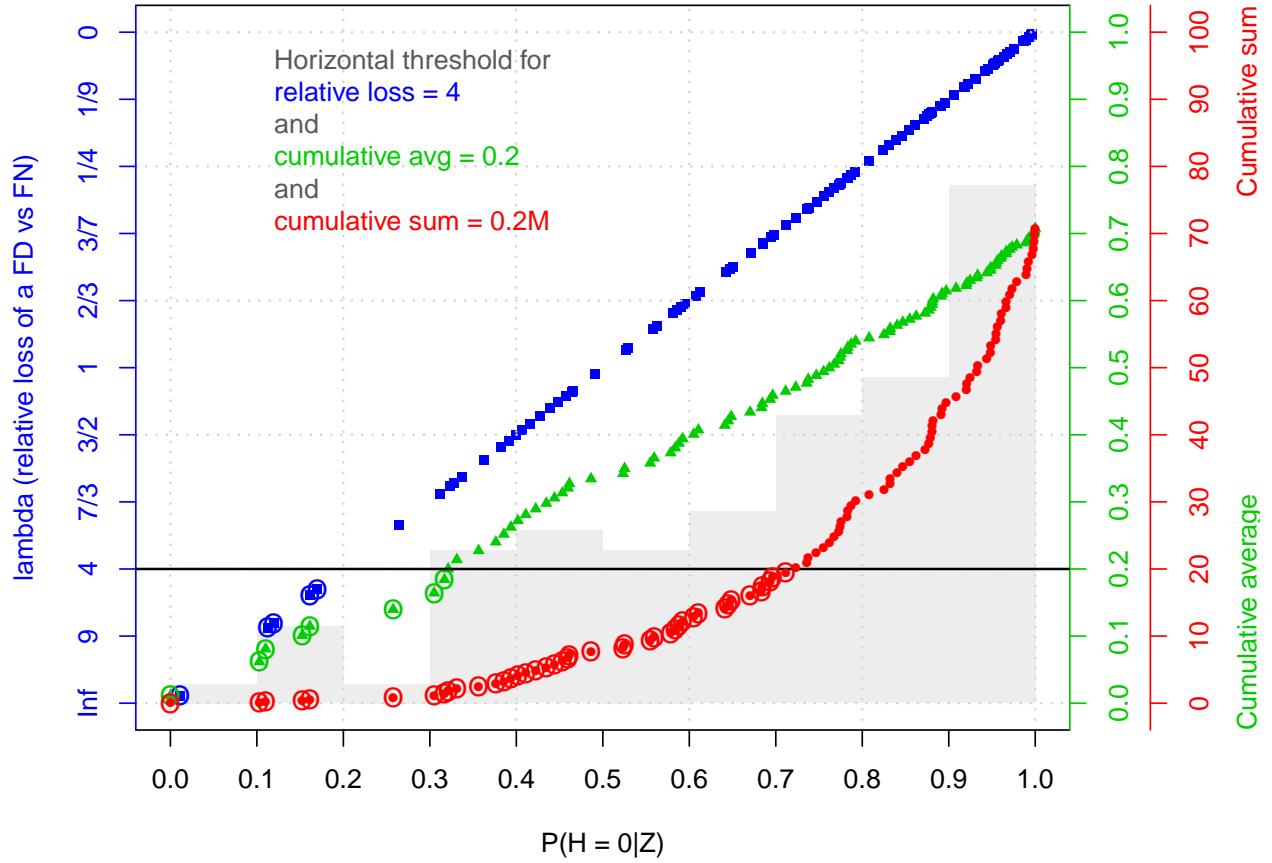


Figure B.2: A comparison of the various decision criteria, for $M = 100$ artificially-generated posterior probabilities clustered around one. The triangular points are plotted on the scale of R_1 ; the square points are plotted on the scale of R_2 ; the circular points are plotted on the scale of R_3 . 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 4 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.

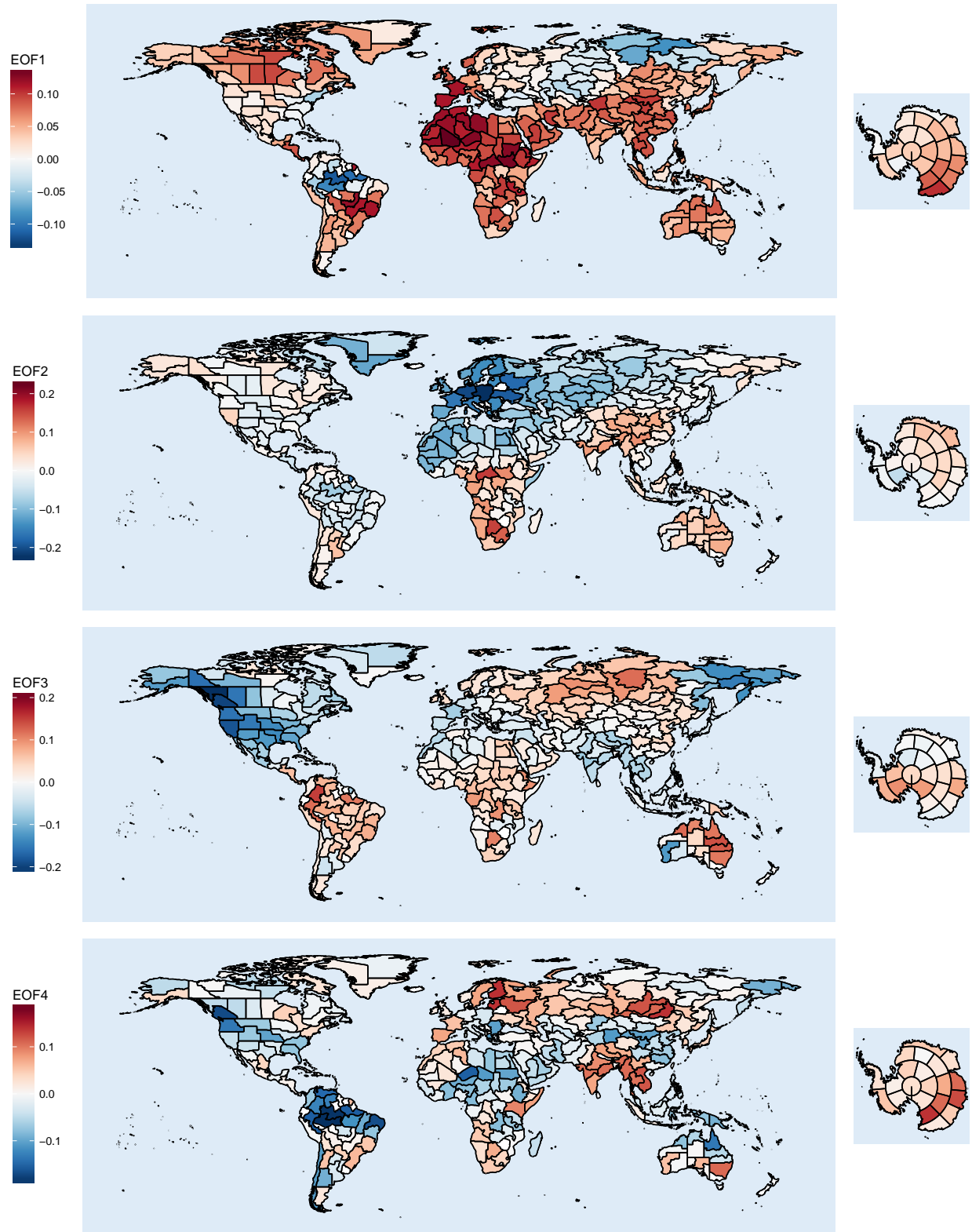


Figure B.3: The first four EOFs for the logit probability of a hot January over 1959-2014, for the factual scenario.

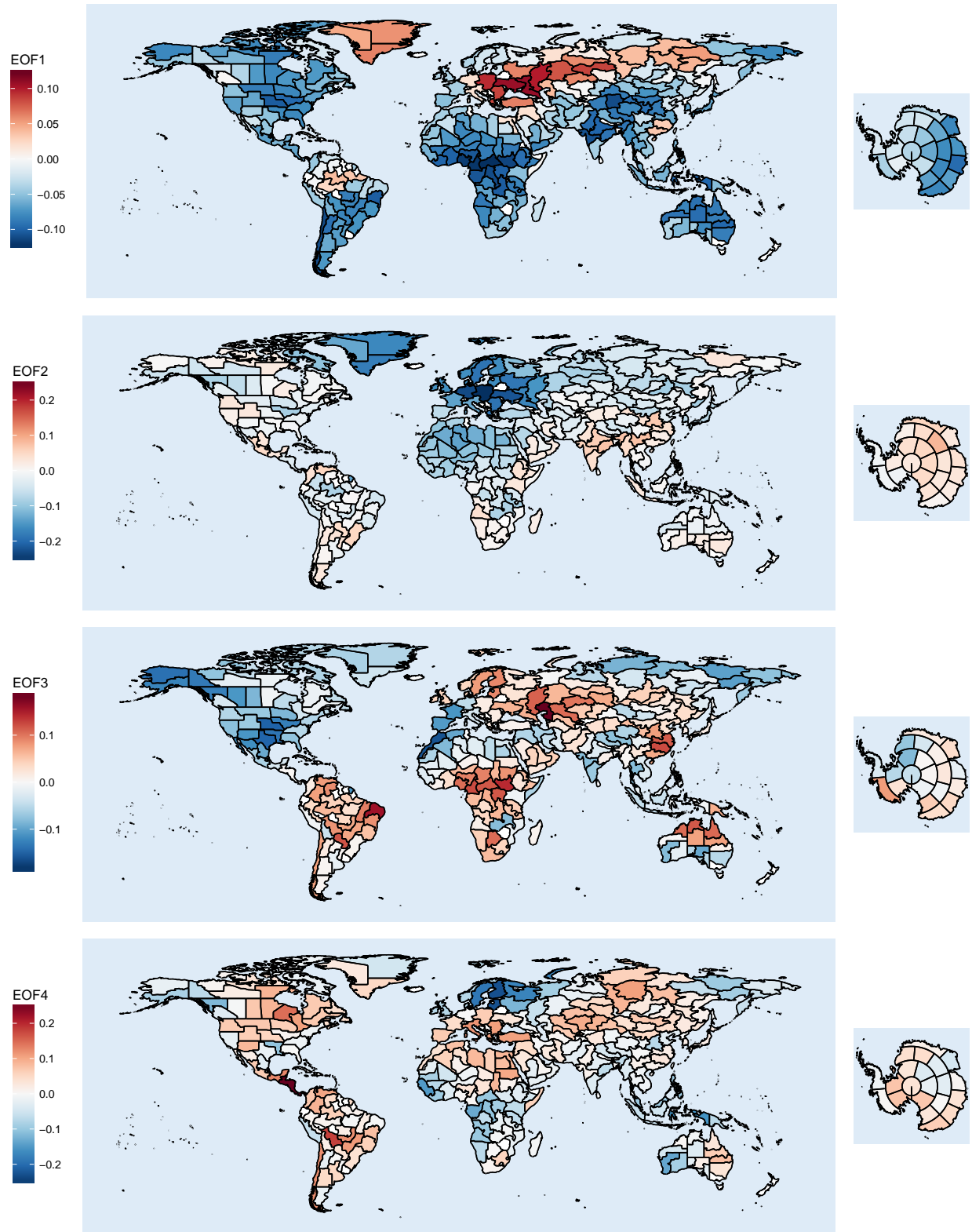


Figure B.4: The first four EOFs for the logit probability of a hot January over 1959-2014, for the counterfactual scenario.

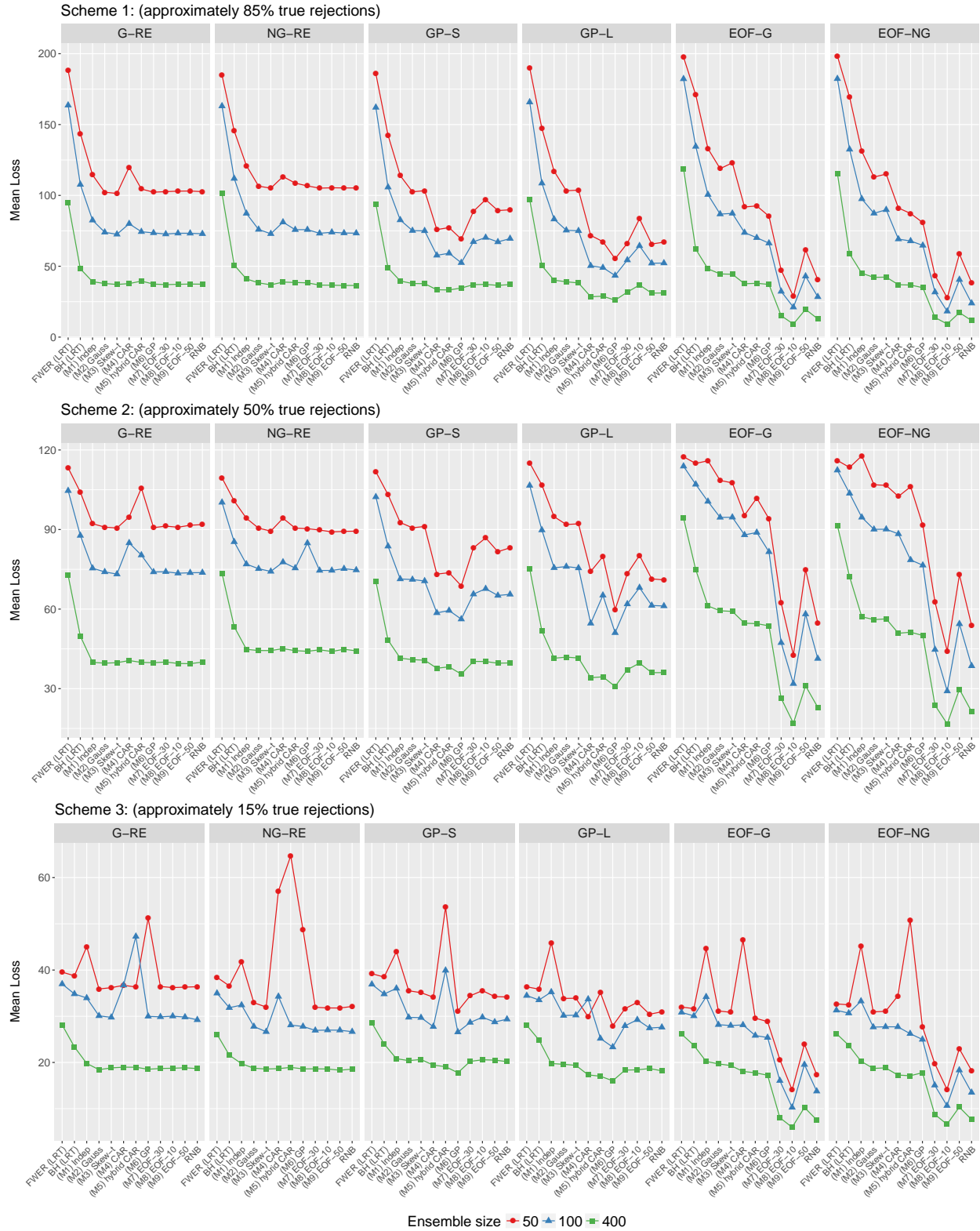
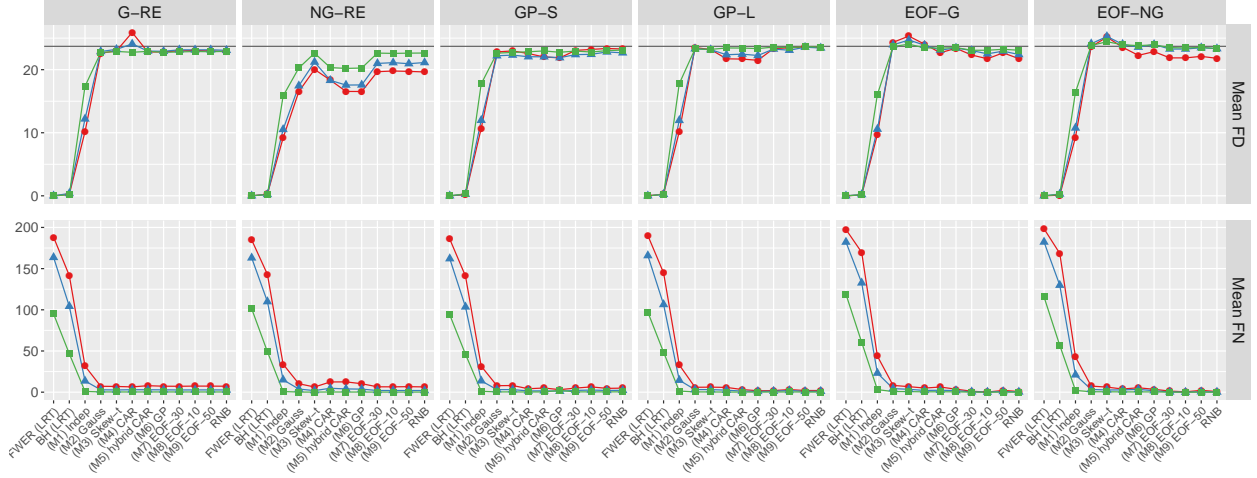
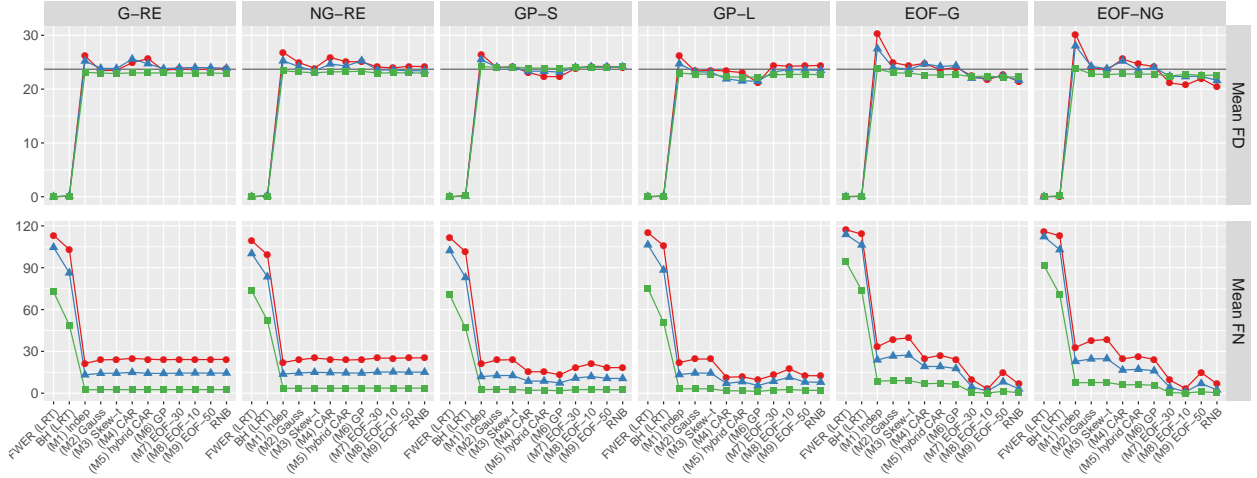


Figure B.5: Mean loss using the R_2 criteria, aggregated over the $N_{\text{rep}} = 100$ replicates, for schemes 1, 2, and 3. Note that the x -axis in each subgrid corresponds to the different methods/fitted models.

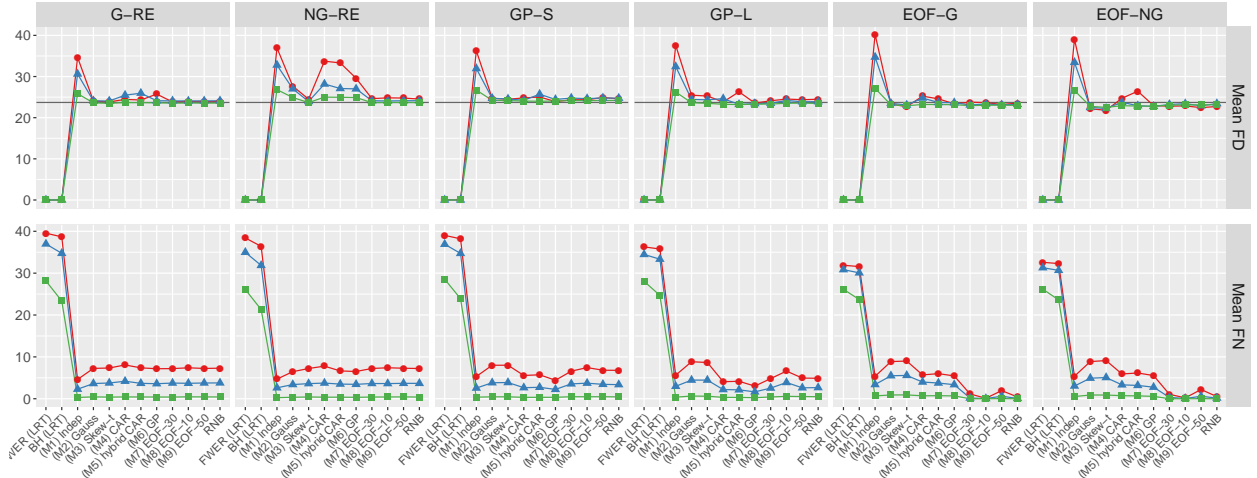
Scheme 1: approximately 85% true rejections



Scheme 2: approximately 50% true rejections



Scheme 3: approximately 15% true rejections



Ensemble size • 50 • 100 • 400

Figure B.6: Mean FD and FN using the R_3 criteria, aggregated over the $N_{\text{rep}} = 100$ replicates, for schemes 1, 2, and 3. Note that the x -axis in each subgrid corresponds to the different methods/fitted models. The target of $\gamma = 0.1M = 23.7$ is plotted for FD.

B.2 Results from simulation study with $M = 68$ regions

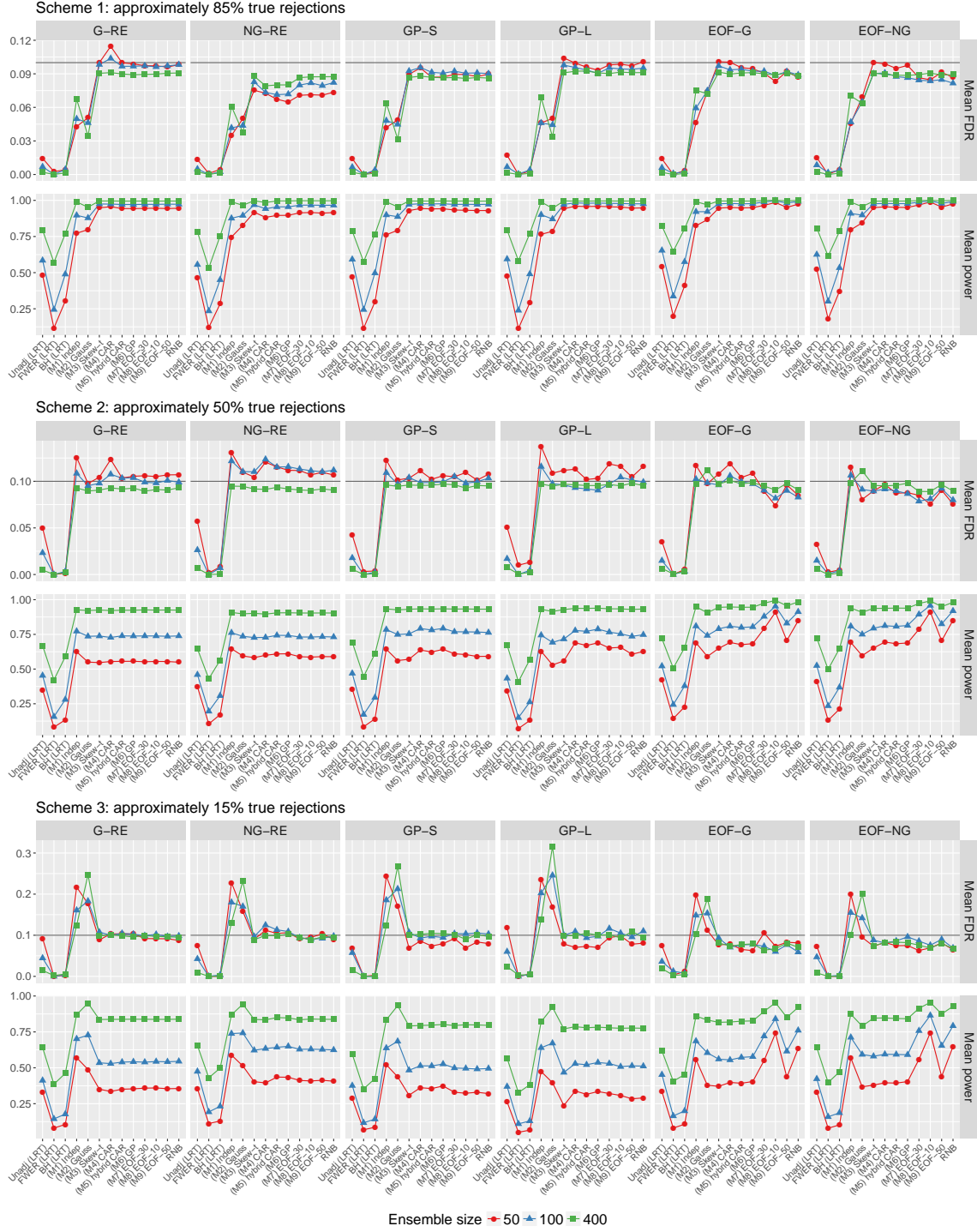
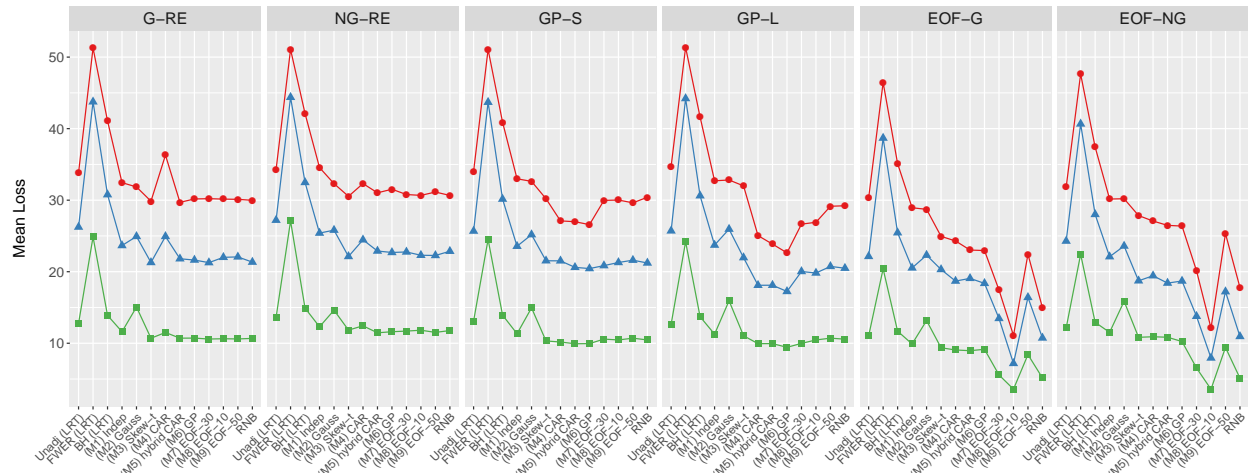
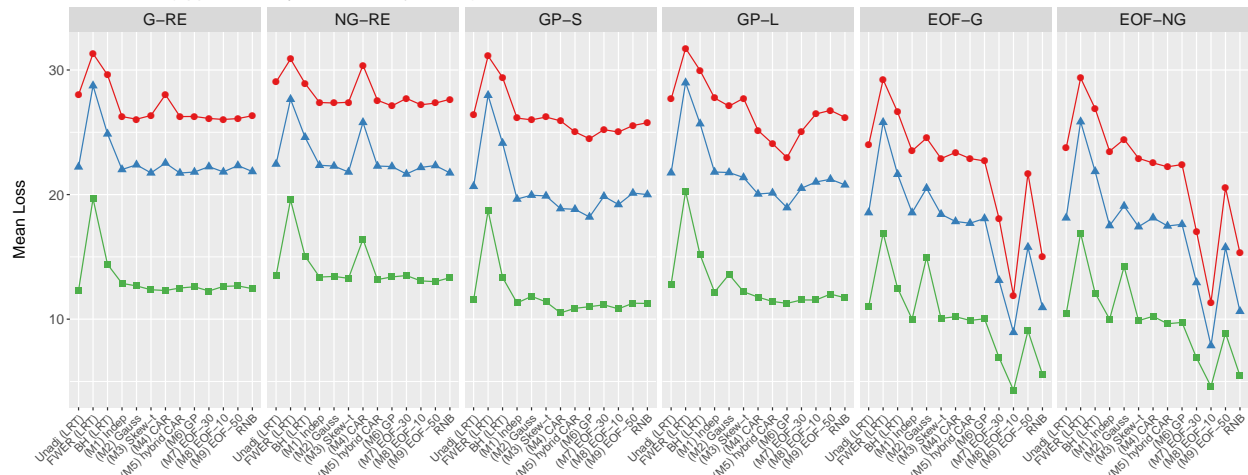


Figure B.7: Mean FDR and power using the R_1 criteria for the WRAF2 regions ($M = 68$), aggregated over the $N_{\text{rep}} = 100$ replicates, for schemes 1, 2, and 3. The target of $\alpha = 0.1$ is plotted for FDR.

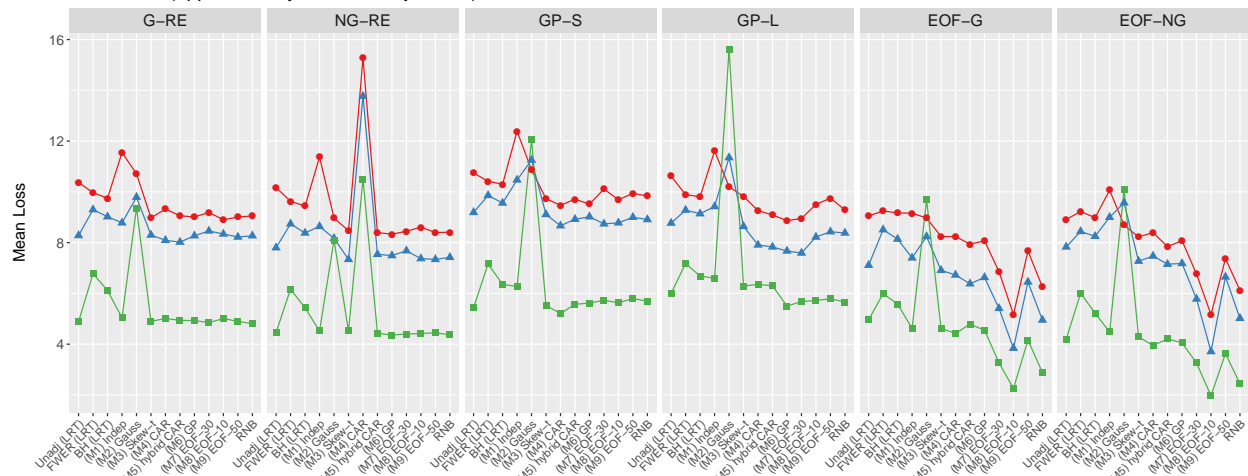
Scheme 1: (approximately 85% true rejections)



Scheme 2: (approximately 50% true rejections)



Scheme 3: (approximately 15% true rejections)



Ensemble size 50 100 400

Figure B.8: Mean loss using the R_2 criteria for the WRAF2 regions ($M = 68$), aggregated over the $N_{\text{rep}} = 100$ replicates, for schemes 1, 2, and 3.

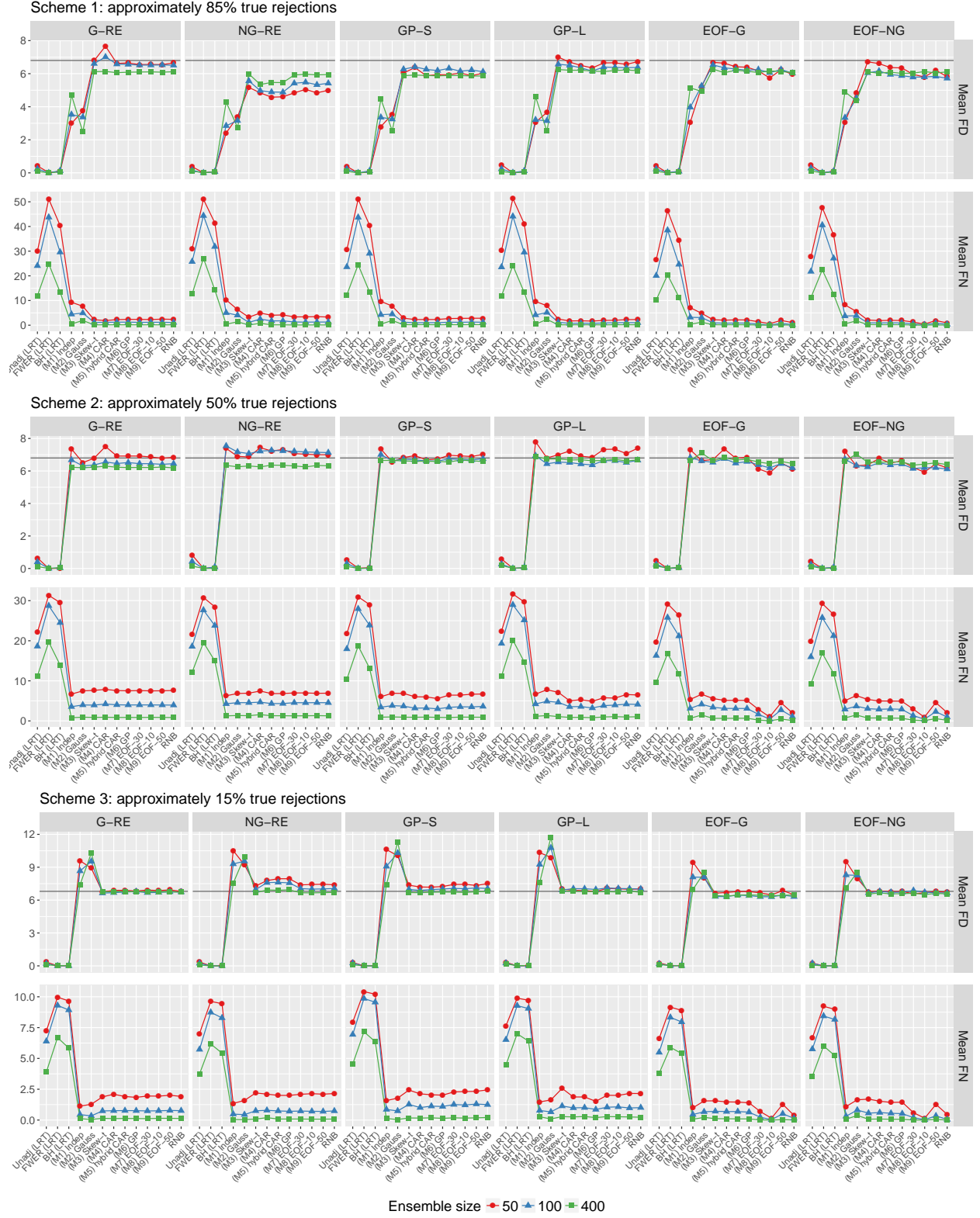


Figure B.9: Mean FD and FN using the R_3 criteria for the WRAF2 regions ($M = 68$), aggregated over the $N_{\text{rep}} = 100$ replicates, for schemes 1, 2, and 3. The target of $\gamma = 0.1M = 6.8$ is plotted for FD.

C Centered parameterization for the skew- t distribution

Note: the parameter symbols used in this section do not correspond to the symbols used in the main draft of the text.

[Azzalini and Capitanio \(2003\)](#) introduced the skew- t family of distributions, with probability density function

$$f_{ST}(y; \xi, \omega, \alpha, \nu) = \frac{2}{\omega} t_\nu \left(\frac{y - \xi}{\omega} \right) T_{\nu+1} \left(\frac{\alpha(y - \xi)}{\omega} \sqrt{\frac{\nu + 1}{\nu + (y - \xi)/\omega}} \right), \quad (\text{C.1})$$

where t_ν and T_ν denote the probability density and cumulative distribution function, respectively, of a standard t distribution with ν degrees of freedom. In (C.1), $\xi \in \mathbb{R}$ is a location parameter, $\omega \in \mathbb{R}^+$ is a scale parameter, $\alpha \in \mathbb{R}$ controls the skewness, and $\nu \in \mathbb{R}^+$ controls the tail behavior. Unfortunately, as noted by [Arellano-Valle and Azzalini \(2008\)](#) (and others), using the “direct” parameterization $\theta_D = (\xi, \omega, \alpha, \nu)$ has both theoretical and practical problems: for example, the likelihood behaves strangely for a neighborhood of $\alpha = 0$, in that the profile likelihood for α has a stationary point at 0. Furthermore, at $\alpha = 0$, the expected Fisher information is singular, even though all of the parameters are identifiable. In practical terms, this means that the parameter estimates (especially ξ and ω) can trade off with one another to give qualitatively similar results for an individual data set.

To address this problem, [Arellano-Valle and Azzalini \(2008\)](#) discuss a “centered” parameterization (for the skew-normal distribution; a corresponding result holds for the skew- t), originally introduced by [Azzalini and Capitanio \(2003\)](#). Instead of θ_D , the centered parameterization involves $\theta_C = (\mu, \sigma, \delta, \nu)$, where

$$\mu = \xi + \omega \sqrt{2/\pi} \frac{\alpha}{\sqrt{1 + \alpha^2}}, \quad -\infty < \mu < \infty,$$

$$\sigma = \omega \sqrt{1 - \frac{2}{\pi} \frac{\alpha^2}{1 + \alpha^2}}, \quad 0 < \sigma < \infty,$$

and

$$\delta = \frac{\alpha}{\sqrt{1 + \alpha^2}}, \quad -1 < \delta < 1,$$

with inverse transformations

$$\xi = \mu - \frac{\sigma}{\sqrt{1 - \frac{2}{\pi}\delta^2}} \sqrt{2/\pi}\delta, \quad \omega = \frac{\sigma}{\sqrt{1 - \frac{2}{\pi}\delta^2}}, \quad \alpha = \frac{\delta}{\sqrt{1 - \delta^2}}. \quad (\text{C.2})$$

(Note: ν is the same in both parameterizations.) Using θ_C avoids the problems associated with θ_D ; in practice, the likelihood associated with θ_C is given by (C.1), after substituting in (C.2).

D Prior specification for the parametric Bayesian models

In general, the priors used for all parameters will be proper but diffuse, with fixed hyperparameters. The details for each model are as follows; all of the priors below are for both $k \in \{F, C\}$.

M1 Beta-binomial, independent across regions

The only parameters in M1 are the probabilities themselves, which have already been assigned beta priors. The hyperparameters are set to $a_p = b_p = 1$, i.e., the probabilities are given an uniform prior.

M2 Exchangeable Gaussian prior

The parameters in M2 are the scenario-specific mean μ_k and variance τ_k^2 , with priors

$$\mu_k \sim N(0, 10^2), \quad \tau_k \sim U(0, 100),$$

where $N(a, b)$ is the Gaussian distribution with mean a and variance b and $U(c, d)$ is the uniform distribution on the interval (c, d) .

M3 Exchangeable skew- t prior

Following [Arellano-Valle and Azzalini \(2008\)](#), M3 involves the scenario-specific “centered” parameters (see Appendix C) location μ_k , scale σ_k , skewness δ_k , and degrees of freedom ν_k . The prior distributions used are

$$\mu_k \sim N(0, 10^2), \quad \sigma_k \sim U(0, 100), \quad \delta_k \sim U(-1, 1), \quad 1/\nu_k \sim U(0, 1).$$

M4 CAR prior

The parameters in M4 are the scenario-specific mean μ_k and variance τ_k^2 ; however, because the CAR prior is improper, we fix $\mu_k = 0$ (see Appendix E). As before, $\tau_k \sim U(0, 100)$.

M5 Hybrid CAR/exchangeable prior

The parameters in M5 are the scenario-specific mean μ_k , variance τ_k^2 , and mixture parameter λ_k , with priors

$$\mu_k \sim N(0, 10^2), \quad \tau_k \sim U(0, 100), \quad \lambda_k \sim U(0, 1).$$

M6 Gaussian process prior

The parameters in M6 are the scenario-specific mean μ_k , variance τ_k^2 , and spatial “range” parameter ϕ_k , with priors

$$\mu_k \sim N(0, 10^2), \quad \tau_k \sim U(0, 100), \quad \phi_k \sim U(0, c_\phi),$$

where $c_\phi = (1/2) \max\{||\mathbf{s}_i - \mathbf{s}_j||\}$, since the range of the Gaussian process would not be expected to exceed one-half of the maximum distance between the region centroids. Note that the smoothness parameter for the Matérn correlation function will be considered fixed, at 0.5 (corresponding to an exponential correlation function).

M7/M8/M9 EOF-based structure with a Gaussian prior for a fixed number of coefficients

The parameters in these three models are the scenario-specific mean μ_k , EOF coefficients α_k , scale σ_k , skewness δ_k , and degrees of freedom ν_k . As with the robust nonparametric Bayesian model,

$$\mu_k \sim N(0, 10^2), \quad \sigma_k \sim U(0, 100^2), \quad \delta_k \sim U(-1, 1), \quad 1/\nu_k \sim U(0, 1).$$

In a more standard approach, the EOF coefficients (across $p = 30$, $p = 10$, and $p = 50$) now have an exchangeable Gaussian prior:

$$\alpha_{kl} \stackrel{\text{iid}}{\sim} N(0, \sigma_\alpha^2), \quad l = 1, \dots, p,$$

where $\sigma_\alpha \sim U(0, 100)$.

E Markov chain Monte Carlo

The posterior distribution for each of the hierarchical models M2-M9 and RNB is not available in closed form, so we resort to Markov chain Monte Carlo (MCMC) methods to obtain samples from the joint posterior distribution for each model. All models are fit using the `nimble` software for R (de Valpine et al., 2017). While the MCMC is straightforward for RNB, M2, M3, M5, M6, M7, M8, and M9 (using standard Gibbs sampling with Metropolis Hastings steps), model M4 requires an adjustment to the standard MCMC (see the next section). The code used to fit these models are available in the online reproducibility documents.

E.1 Computational details for the CAR parameterization

Recall that computation for the CAR model is hindered by the fact that the intrinsic CAR prior is improper. This results in two problems: first, the random effects are identifiable only up to an additive constant; second, the CAR prior is undefined for the full random effects vector. While more sophisticated solutions to the first problem are possible, for the purposes of this work we simply set $\mu_k = 0$ to fix the identifiability problem.

Rue and Held (2005) outline steps to address the second problem. The CAR prior is

$$p(\beta_k | \mathbf{Q}_k, \tau_k^2) \propto |\tau_k^{-2} \mathbf{Q}|^{1/2} \exp \left\{ -\frac{1}{2} \beta_k^\top \mathbf{Q}_k \beta_k \right\};$$

however, the rank of \mathbf{Q} is $M - 1$ ($\mathbf{1}^\top \mathbf{Q} = 0$), so the determinant $|\tau_k^{-2} \mathbf{Q}| = 0$. While the CAR prior is improper for an M -dimensional space, it is proper for a $(M - 1)$ -dimensional subspace. Following Rue and Held (2005), the prior contribution to the posterior is actually

$$\tilde{p}(\beta_k | \mathbf{Q}_k, \tau_k^2) = (2\pi\tau_k^2)^{-\frac{(M-1)}{2}} \left(\prod_{i=1}^{M-1} \lambda_{ki} \right)^{1/2} \exp \left\{ -\frac{1}{2} \beta_k^\top \mathbf{Q}_k \beta_k \right\},$$

where $\{\lambda_{ki} : i = 1, \dots, M - 1\}$ are the non-zero eigenvalues of \mathbf{Q}_k .

F Further details on the simulation study

F.1 Simulation scheme for each true state

The six true states used as population distributions for the simulation study are listed in Table 1 of the main text. The actual sampling procedure for each true state is now outlined.

First, for the Gaussian random effects (G-RE), the logit probabilities are simply draws from a Gaussian distribution:

$$\text{logit}(p_k) \stackrel{\text{iid}}{\sim} N(m_k, v_k^2).$$

Next, for the gamma random effects (NG-RE), the logit probability anomalies (i.e., deviations from the means m_k) are draws from a shifted gamma distribution:

$$\text{logit}(p_k) \stackrel{\text{iid}}{\sim} G(a_k, b_k) - c_k,$$

where a_k and b_k are the shape and scale parameters, respectively. The Gaussian process samples (GP-S and GP-L) are first drawn collectively from

$$\text{logit}(\mathbf{p}_k) \sim N_M(m_k \mathbf{1}_M, \mathbf{S}),$$

where the elements of \mathbf{S} are $S_{ij} = v_k^2 \mathcal{M}_{g_k}(\|\mathbf{s}_i - \mathbf{s}_j\|/r_k)$ (where $\mathcal{M}_g(\cdot)$ is the Matérn correlation function and \mathbf{s}_i is the centroid of region i) and then centered to have an empirical mean of zero.

It is slightly less straightforward to generate samples from EOF-G and EOF-NG, especially because the generated data needs to have properties comparable to the other simulations (in terms of the correct proportion of true rejections and empirical variance of the true log risk ratio). The following (somewhat complicated) scheme made this possible (the k subscript has been omitted).

1. For $j = 1, \dots, p$ (where we use $p = 30$ basis functions for the “truth”), draw $\alpha_j \sim N(0, s_j^2)$.
2. Draw $x_j \stackrel{\text{iid}}{\sim} N(0, v^2)$ (for EOF-G) or $x_j \stackrel{\text{iid}}{\sim} k[G(b, c) - d]$ (for EOF-NG).
3. Calculate the probabilities as $\mathbf{p} = \text{logit}^{-1} [m \mathbf{1}_M + \mathbf{H} \mathbf{a} + \mathbf{x}]$.

F.2 Fixed hyperparameter values for the true states

Tables F.1-F.5 contain the fixed hyperparameters used to sample draws from the fixed population distributions across the N_{rep} replicates. The values were determined after trial and error, and were set according to two criteria: first, that the true proportion of rejections would match up with the corresponding scheme, and second, that the variance of the true log risk ratio (empirically, over many replicates) would be approximately 0.9.

Table F.1: Fixed hyperparameter values used for simulations from the Gaussian random effects (G-RE), across Schemes 1–3.

	Scheme 1	Scheme 2	Scheme 3
m_C	logit(0.08)	logit(0.08)	logit(0.08)
m_F	logit(0.03)	logit(0.08)	logit(0.19)
v_C^2, v_F^2	0.72^2	0.74^2	0.775^2

Table F.2: Fixed hyperparameter values used for simulations from the shifted gamma random effects (NG-RE), across Schemes 1–3. Note: a is the shape parameter and b is the scale parameter.

	Scheme 1	Scheme 2	Scheme 3
m_C	logit(0.08)	logit(0.08)	logit(0.08)
m_F	logit(0.03)	logit(0.08)	logit(0.18)
a_C, a_F	4	3.75	3.5
b_C, b_F	0.375	0.4	0.4286
c_C, c_F	1.5	1.5	1.5

Table F.3: Fixed hyperparameter values used for simulations from the spatial Gaussian process effects (GP-S and GP-L), across Schemes 1–3. Note: the distances in \mathbb{R}^3 are re-scaled to have a maximum of 1 unit.

	Scheme 1	Scheme 2	Scheme 3
m_C	logit(0.08)	logit(0.08)	logit(0.08)
m_F	logit(0.03)	logit(0.08)	logit(0.18)
v_C^2, v_F^2	0.6	0.6	0.6
r_C, r_F (short)	0.06	0.06	0.06
r_C, r_F (long)	0.10	0.10	0.10
g_C, g_F	2	2	2

Table F.4: Fixed hyperparameter values used for simulations from the EOF effects with Gaussian discrepancy (EOF-G), across Schemes 1–3.

	Scheme 1	Scheme 2	Scheme 3
m_C	logit(0.08)	logit(0.08)	logit(0.08)
m_F	logit(0.03)	logit(0.08)	logit(0.19)
$s_j^2, j = 1, \dots 5$	3.5^2	3.5^2	3.5^2
$s_j^2, j = 5, \dots 10$	1^2	1^2	1^2
$s_j^2, j = 10, \dots 30$	0.05^2	0.05^2	0.05^2
v_C^2, v_F^2	0.01^2	0.01^2	0.01^2

Table F.5: Fixed hyperparameter values used for simulations from the EOF effects with gamma discrepancy (EOF-NG), across Schemes 1–3.

	Scheme 1	Scheme 2	Scheme 3
m_C	logit(0.08)	logit(0.08)	logit(0.08)
m_F	logit(0.03)	logit(0.08)	logit(0.19)
$s_j^2, j = 1, \dots 5$	3.5^2	3.5^2	3.5^2
$s_j^2, j = 5, \dots 10$	1^2	1^2	1^2
$s_j^2, j = 10, \dots 30$	0.05^2	0.05^2	0.05^2
k_C, k_F	0.02	0.02	0.02
b_C, b_F	5	5	5
c_C, c_F	0.4	0.4	0.4
d_C, d_F	2	2	2

References

- Arellano-Valle, R. B. and Azzalini, A. (2008). The centred parametrization for the multivariate skew-normal distribution. *Journal of Multivariate Analysis*, 99(7):1362 – 1382. Special Issue: Multivariate Distributions, Inference and Applications in Memory of Norman L. Johnson.
- 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 (Statistical Methodology)*, 65(2):367–389.
- 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 (Methodological)*, 57(1):pp. 289–300.
- Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):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(4):1165–1188.
- 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(2):403–413.
- Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103.
- 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(456):1151–1160.
- Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

- Schwartzman, A. and Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika*, 98(1):199–214.
- Shu, H., Nan, B., and Koeppe, R. (2015). Multiple testing for neuroimaging via hidden Markov random field. *Biometrics*, 71(3):741–750.
- 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(479):901–912.
- Sun, W. and Cai, T. T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):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 (Statistical Methodology)*, 77(1):59–83.