When and How to Use Restricted Spatial Regression to Separate Environmental Effects from Spatial Confounding

Raquel Ruiz Diaz[1] and James T. Thorson[2]

[1] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, 98195, USA

[2] Resource Ecology and Fisheries Management Division, Alaska Fisheries Science Center, Seattle, WA, 98144, USA

*Corresponding author: raquelrd@uw.edu

Abstract

Aim: To provide practical guidance for ecologists on when to use standard spatial generalized linear mixed models (SGLMMs) versus Restricted Spatial Regression (RSR). We reframe the debate by arguing the choice depends on whether the total effect or direct effect of covariates will be more transferable across space.

Innovation: Our study's primary innovation is to introduce a causal framework to this debate. We distinguish between the SGLMM, which estimates the direct effect by controlling for unmeasured spatial confounders, and the RSR, which estimates the total effect by incorporating pathways through unmeasured spatial mediators. We also provide the first implementation of the RSR estimator in the widely-used tinyVAST R package.

Main conclusions: The choice of model must be driven by the ecological goal and underlying assumptions. 1) The SGLMM is the appropriate tool for hypothesis testing and conditional "in-sample" prediction (interpolation), where it accounts for lower degrees of freedom due spatial autocorrelation. 2) The RSR estimator, despite its highly inflated Type I error, is superior for unconditional prediction (forecasting). This is particularly true when the unmeasured spatial process can be assumed to have a fixed (stationary) sample correlation with the covariates over space. The RSR estimator implicitly incorporates this confounding relationship, making it more effective for predicting total effects in a new, unobserved area where the spatial pattern of confounding is expected to persist. We recommend a two-step conceptual approach: use the SGLMM for robust

variable selection, then, if forecasting is the goal and the unmeasured process has a fixed correlation with covariates, use the RSR estimator for total effect predictions.

keywords: spatial generalized linear mixed model, Gaussian Markov random field, climate forecast, species distribution model.

Introduction

Understanding the spatial distribution of species is fundamental to ecology. Species distributions are influenced by physical and environmental conditions, as well as ecological factors (e.g., competition, predation, dispersal) and evolutionary history (local adaptation), which in turn shape community dynamics [1]. Species distribution models (SDMs) are essential tools for linking environmental covariates to species abundance and distribution, providing critical insights for conservation planning and management [2, 3]. However, ecologists using these models face two persistent challenges: spatial autocorrelation, where response values at nearby locations are correlated [4], and spatial confounding, where a covariate is collinear with the underlying spatial structure in the data [5]. While often discussed separately, these issues are deeply intertwined and can compromise model inference [6, 7].

To mitigate spatial autocorrelation, researchers have developed various statistical approaches, including spatial autoregressive (SAR) models and geostatistical models [8, 9]. More recently, Gaussian Markov Random Fields (GMRFs) have gained prominence in spatial ecology for their computational efficiency and flexibility in modeling spatial dependencies within both Bayesian and likelihood-based frameworks [10, 11]. GMRFs, particularly when integrated into frameworks such as Integrated Nested Laplace Approximation (INLA) and stochastic partial differential equations (SPDE), offer robust tools for approximating spatially structured processes in ecological models [12].

The standard approach to managing spatial autocorrelation is to fit a spatial generalized linear mixed model (SGLMM), which includes a spatial random field to account for unexplained spatial patterns [13, 14, 15]. While this successfully addresses autocorrelation [16, 17], it creates a new dilemma. When a covariate is correlated with the spatial random effect, the model can struggle to partition their shared influence, often leading to biased estimates of the species-environment

relationships we aim to understand [18]. This has fueled a contentious debate. One proposed solution is Restricted Spatial Regression (RSR), an estimator designed to "deconfound" the model by ensuring the spatial random effect is orthogonal to the fixed-effect covariates [19, 20, 21, 22]. However, RSR has faced significant criticism for its unreliable uncertainty estimates and highly inflated Type I error rates [23, 24, 22]. As a result, ecologists are left uncertain about which method is appropriate for common ecological goals, like mapping current distributions or forecasting future changes.

We argue that this debate can be reframed to provide more clarity for ecological practitioners. The choice between an SGLMM and RSR is not a matter of which model is more statistically "correct," but rather a deliberate choice of the quantity the researcher aims to estimate (the "estimand"). This choice is critical because it determines the transferability of the estimated relationship (i.e., its ability to be applied to new locations or future time periods), which is a core challenge in predictive ecology. We propose that these models estimate two fundamentally different assumptions about system dynamics:

- The SGLMM estimates the *direct effect* of a covariate. By partitioning variance between the measured covariate and the spatial random field, the SGLMM treats the spatial field as a statistical proxy for unmeasured, spatially-structured confounding variables. This approach aims to isolate the direct relationship between the covariate and the response, controlling for these unmeasured confounders.

- The RSR estimates the *total effect*. By re-attributing variance from the spatial field back to the covariate, the RSR estimates the combined influence of the measured covariate and any unmeasured spatial processes that are correlated with it.

These two approaches can be viewed as alternative structural causal models and visualized as directed acyclic graphs (Fig. 1). Importantly, these two DAGS can result in the same variance among variables, but they correspond to fundamentally different assumptions about how a change in a covariate X would affect the response Y. This distinction between describing correlations and predicting counterfactual responses has been acknowledged in recent ecological papers [25, 26, 27], but has not been discussed in the context of spatial autocorrelation.

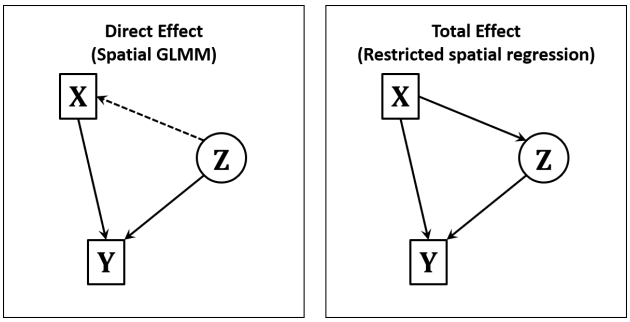

Figure 1: Directed acyclic graphs representing two alternative assumptions regarding the relationship between covariate X, response Y, and latent spatial variable Z, showing (left panel) that estimating the direct effect using a spatial generalized linear mixed model corresponds to the direct path (black arrow) $X \to Y$ while controlling for the path from confounder $Z \to Y$ (and ignores the potential path from $Z \to X$), while the total effect (right panel) corresponds to both the direct effect as well as the indirect effect $X \to Z \to Y$ mediated by Z.

This paper aims to provide a conceptual framework for ecologists to choose between SGLMM and RSR based on their scientific goal (e.g., hypothesis testing for direct effects vs. forecasting total effects) and their assumptions about the underlying ecological processes. Using simulations and a case study, we demonstrate which model estimates which quantity and provide practical guidance for navigating the complex trade-offs of spatial modeling. We also introduce a novel implementation of RSR in the widely used *tinyVAST* R package [28], making this technique more accessible to the ecological community.

Methods

In the following, we introduce a restricted spatial regression (RSR) estimator for slope parameters that relate the response of spatial generalized linear mixed model (SGLMM) to specified covariates. We then use simulation and case-study examples to compare the performance of four estimators:

1. *GLM*: Including covariates without any spatial latent variables using a standard generalized linear model;

2. *SGLMM*: Including covariates while also estimating a spatial GMRF to account for spatial autocorrelation;

3. *RSR-SGLMM*: Including covariates and a spatial GMRF, while then transforming the estimated slopes to compute the RSR estimator;

4. *SGLMM-selected*: Starting with the SGLMM, we then perform backwards model selection by eliminating all covariates for which a two-sided Wald test indicates that the covariate is not significant at a $p < 0.05$ level;

We include the SGLMM-selected because the SGLMM is expected to usefully identify covariates that are not statistically significant, such that model selection might improve performance relative to the SGLMM by serving as a simple form of regularization [29].

We investigate the following questions:

1. *False discovery rates*: When fitting "false" covariates that are not associated with the simulated response, do these models identify false covariates as statistically significant at the intended Type-1 error rate, or do they instead have an elevated "false discovery rate"?
2. *Performance for slope estimates*: Which model has the lowest bias or imprecision when estimating the slope for covariates?
3. *Conditional predictive performance*: Which model has lowest predictive error when conditioning predictions upon both the estimated slope and GMRF values?
4. *Unconditional predictive performance*: Which of the models has lowest predictive error when conditioning predictions only upon the estimated slope and ignoring the GMRF values?
5. *Model selection*: Does model selection affect the performance of the SGLMM estimator?

We expect that conditional prediction will have lower predictive error than unconditional prediction within the spatial domain of data. However, we are also interested in the performance of unconditional prediction for two reasons:

- *Transferability among studies*: Unconditional prediction corresponds to the component of the model that is easily "transferable" among studies. For example, ecologists reading a paper will often interpret the estimated value for the slope (e.g., by comparing it with other experimental or observational measurements), and may use the reported value for slope estimates to then parameterize a subsequent model. When used this way, readers will often ignore the estimated

GMRF values (which may not be reported, or only visualized on a map of model output), such that future uses of study results will correspond to unconditional prediction. This practice of interpreting regression slope parameters and ignoring GMRF estimates implicitly assumes that the covariate and GMRF are independent, so no information about the slope remains in the estimated GMRF.

- *Model projections*: As a model is projected beyond the range of the data, a spatial GMRF will tend to revert to zero (or remain at its value at the geographic boundary). Therefore, when projecting models using end-of-century climate conditions or across a larger geographic domain, the model performance will revert to unconditional prediction.

Spatial linear mixed model

To begin, we first introduce the spatial generalized linear mixed model (SGLMM). This involves fitting response y_i for each sample $i \in \{1, 2, ...I\}$, using $j \in \{1, 2, ..., J\}$ covariates with values x_{ij} in matrix $\mathbf{X}$, as well as the two-dimensional geographical coordinates $\mathbf{s}_i$:

$$y_i \sim f(\mu_i, \theta) \tag{1}$$

$$g(\mu_i) = \mathbf{X}_i\beta + \mathbf{A}_i\omega \tag{2}$$

$$\omega \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{R}) \tag{3}$$

where f is the probability distribution with parameters θ, g is the link function, β is the estimated slope parameters, ω is the vector of spatial random effects with mean zero, spatial correlation $\mathbf{R}$ and pointwise variance σ^2, and $\mathbf{A}_i$ projects from random effects to the location of data based on geographical coordinates (where $\mathbf{A}$ is an indicator matrix when applying an areal model, and an interpolation matrix when approximating a continuous spatial function). The spatial variable ω is included to represent spatial correlation arising from unmeasured processes (i.e., missing covariates or endogenous spatial patterns), and its inclusion is intended to ensure that residual errors (represented by f) are independent among samples i. After identifying maximum likelihood estimates $\hat{\beta}$ and empirical Bayes predictions $\hat{\omega}$, conditional prediction using the plug-in estimator

records $\hat{\mu}_i = g^{-1}(\mathbf{X}_i\hat{\beta} + \mathbf{A}_i\hat{\omega})$ for each sample, and unconditional prediction defines $\hat{\mu}_i' = g^{-1}(\mathbf{X}_i\hat{\beta})$ while neglecting the impact of random effects $\mathbf{A}_i\hat{\omega}$. Finally, unconditional RSR prediction records $\hat{\mu}_i^* = g^{-1}(\mathbf{X}_i\beta^*)$, where $\hat{\beta}^*$ is the RSR estimator for slopes as we define in Section .

This SGLMM results in "confounding" between the covariates $\mathbf{X}$ and the projected value of the spatial variable $\mathbf{A}\omega$. This confounding arises when any covariate j is correlated with the spatial variable, i.e., $\sum_{i=1}^{I}(x_{ij}\mathbf{A}_i\omega) \neq 0$. In this case, the estimated value for the spatial random effect will "soak up" some portion of variation that would otherwise be attributed to slope β_j. Conceptually, the SGLMM is "correcting for" the correlation between covariate and random effect that is expected purely by chance, given the estimated parameters controlling the estimated spatial correlation. Comparing the estimate $\hat{\beta}_j$ from the SGLMM with the estimate $\hat{\beta}_j^*$ from a standard GLM (i.e., when fixing $\omega = 0$), we see that $\hat{\beta}_j$ from the SGLMM will approach $\hat{\beta}_j^*$ from the GLM as the spatial variation σ^2 approaches zero. Alternatively, as σ^2 increases, $\hat{\beta}_j$ and $\hat{\beta}_j^*$ can diverge, and this divergence may even result in a change in the sign of the estimated slope [20].

Restricted spatial regression

To address the difference between GLM and SGLMM estimates of the slope, we next introduce the restricted spatial regression (RSR) estimator. To calculate the RSR, we first fit the SGLMM (Eq. 1) and extract the maximum-likelihood estimate of slopes $\hat{\beta}$ and the empirical Bayes estimate of the spatial random effect $\hat{\omega}$. We then calculate the RSR estimator $\hat{\beta}^*$ by adjusting $\hat{\beta}$ to add back in the portion of variation that was "soaked up" by the spatial random effect:

$$\hat{\beta}^* = \hat{\beta} + \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\hat{\omega}}_{\text{regression of } \mathbf{A}\omega \text{ on } \mathbf{X}} \tag{4}$$

In this expression, $(\mathbf{X}^T\mathbf{X})^{-1}$ is a $J \times J$ matrix representing the inverse-covariance of the random effects, and this can be inverted even for large sample sizes. Usefully, the RSR estimator (and associated standard errors) can be calculated post-hoc without any change in the structure of the SGLMM. The associated standard error for $\hat{\beta}^*$ is calculated using the generalized delta method. Similarly, the interpolated value of random effects $\mathbf{A}\omega$ must be adjusted to also correct for this change:

$$\mathbf{A}\hat{\omega}^* = \mathbf{A}\hat{\omega} - \mathbf{X}\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\hat{\omega}}_{\text{regression of } \mathbf{A}\omega \text{ on } \mathbf{X}} \tag{5}$$

This equation makes clear that the predicted value $\hat{\mu}_i$ remains the same whether using the original or RSR estimator for the slopes and random effects, i.e., $\mathbf{X}\hat{\beta} + \mathbf{A}\hat{\omega} = \mathbf{X}\hat{\beta}^* + \mathbf{A}\hat{\omega}^*$ (see Supporting Information S1 for further discussion).

The principle behind this algebraic adjustment can be understood by considering the underlying structure of a spatial process, as illustrated conceptually in fig. 2. A spatial process, like a GMRF, can be thought of as being composed of many underlying spatial patterns, or "basis functions," at different scales. The formula in Equation 4 is mathematically equivalent to projecting these basis functions to be orthogonal to the covariates, thereby removing their shared correlation by design. While the RSR estimator is calculated directly without explicitly constructing these basis functions, the figure provides the statistical intuition for how it achieves deconfounding.

Simulation experiment

To explore the performance of the GLM and spatial SGLMM with the RSR estimator, we conduct a simulation experiment with 9 scenarios and 50 replicates per scenario. To do this, we generated one "true" covariate with a known, non-zero effect on the response variable, alongside a varying number of "false" covariates with no true effect. Scenarios are formed from a $3 \times 3 = 9$ factorial design across three sample sizes ($N = \{50, 100, 200\}$) and a three numbers of "false covariates" ($K = \{1, 3, 5\}$). The false covariates allow us to determine whether models identify them as statistically significant above the intended Type-1 error ("false discovery") rate of $p = 0.05$, such that a model with elevated false-discovery rate will identify a false covariate as significant in more than 5% of simulation replicates. The "true" covariate allows us to assess bias in parameter estimates and accuracy in prediction (RMSE). Models are fitted using *tinyVAST* [28] release 1.2.0 in the R statistical environment (R-Core-2025), which estimates parameters using gradients computed using Template Model Builder [30] while implementing the Laplace approximation to marginalize across the distribution for random effects [31].

Within each simulation replicate, we simulate data over a 1×1 square spatial domain, where N

samples occur at random locations that arise from a Poisson disk process. Given these N sampling locations, we then approximate a Matérn correlation function using the SPDE method [10], i.e., defining a finite-element mesh (FEM) that has vertices at each each sampling location as well as additional interior and boundary vertices. We then calculate the inverse-covariance (precision) of the Gaussian Markov random field that results from approximating spatial diffusion using local stiffness of the FEM:

$$\mathbf{Q} = \tau^2 \left(\kappa^4 \mathbf{M}_0 + 2\kappa^2 \mathbf{M}_1 + \mathbf{M}_0 \right) \tag{6}$$

where κ represents the decorrelation rate and τ controls the variance of the forcing spatial process prior to diffusion. We then use this precision matrix to simulate one "true" covariate $\mathbf{z}$ that is associated with a linear predictor, K "false covariates" $\mathbf{w}_k$ that are not associated with the linear predictor, and one spatial latent variable ω that represents residual autocorrelation. We also including normally distributed measurement errors in each recorded sample:

$$y_i \sim \text{Normal}(\mu_i, \sigma_y^2) \tag{7}$$

$$\mu_i = z_i + \omega_i \tag{8}$$

All spatial terms ("true" covariate, false covariates, and latent variable) are simulated independently (i.e., have an expected correlation of zero), although their simulated values will end up being correlated based only on chance. When their simulated values are correlated within the domain, this then results in confounding among model terms. This design allows for a direct test of each model's ability to either control for this confounding (the SGLMM) or absorb it into the fixed effect estimate (the RSR). In particular, we specify $\kappa = 3.67$ (i.e., the distance with a 10% correlation is 0.75), $\tau = 0.384$ (i.e., each covariate and residual variable has a pointwise standard deviation of 0.2), and $\sigma_y = 0.2$. For each replicate, we record the simulated data $\mathbf{y}$ as well as the true density μ.

We then fit these simulated data using the GLM and SGLMM (Eq. 1) while including $\mathbf{X} = (\mathbf{z}, \mathbf{w}_i, ..., \mathbf{w}_k)$ as the matrix of covariates. After estimation, we record GLM prediction,

the conditional and unconditional prediction from the SGLMM ($\hat{\mu}_i$ and $\hat{\mu}'_i$) as well as the unconditional RSR prediction ($\hat{\mu}^*_i$). We also record the GLM, SGLMM and RSR estimators for the slopes ($\hat{\beta}$ and $\hat{\beta}^*$), and extract the standard error for both estimators calculated using the generalized delta method from the Hessian matrix. Finally, we also fit the GLM including $\mathbf{X}$ as covariates without any spatial latent variable, and record estimated slopes and predictions.

To evaluate performance, we:

- compare $\hat{\mu}_i$, $\hat{\mu}'_i$ and $\hat{\mu}^*_i$ with the true value μ, and calculate root-mean-squared error;
- compare $\hat{\beta}$ and $\hat{\beta}^*$ with the known values ($\beta_z = 1$ and $\beta_w = 0$);
- identify whether the estimated slope is statistically significant for both the true covariate β_z and the false covariates β_w, calculating significance with a two-sided Wald test using the estimated standard error for each slope.

Case study demonstration

We also demonstrate model performance by fitting the GLM and SGLMM to bottom trawl samples of numerical abundance for seafloor-associated species in the eastern and northern Bering Sea: Walleye pollock (*Gadus chalcogrammus*), tanner crab (*Chionoecetes bairdi*) and Pacific cod (*Gadus macrocephalus*). We specifically fit a log-linked Tweedie distribution, using a quadratic response to bottom temperature, and also including year as a factor (i.e., a separate intercept for each year). We use samples from 1982-2019, and there are approximately 330-380 samples in the eastern Bering Sea in each year as well as an additional 150-300 samples in the northern Bering Sea in a smaller subset of years. We specifically record the partial effect of temperature for each model (with confidence intervals obtained by sampling from the sparse joint precision of fixed and random effects). We then compare this temperature effect across models, to demonstrate the real-world differences that arise between GLM, SGLMM, and RSR estimators.

Results

Performance estimating covariates

The simulation experiment revealed critical differences in how the models handle hypothesis testing and parameter estimation (Figure 3). For the three null covariates (False 1, False 2 and False 3), the SGLMM approach maintained Type I error rates near the nominal 5% level, making it a reliable tool for hypothesis testing. the GLM and RSR-SGLMM showed dramatically elevated false discovery rates, exceeding 50% (Figure 3 A). This confirms that RSR is unsuitable for determining the statistical significance of a covariate.

When estimating the "true" covariate effect (real covariate with $\beta = 1$), the results illustrate the fundamental difference between the estimators. As shown in Figure 3 B, the SGLMM provides a less biased estimate of the "true" *direct effect* parameter. The RSR-SGLMM estimate, however, is predictably biased away from the true *direct effect*. This occurs because our simulation design intentionally introduced confounding between the covariate and the spatial field; the RSR estimator correctly estimates the total effect by absorbing this confounding, which is by definition different from the *direct effect* parameter we simulated.

0.1 Conditional prediction

For conditional prediction (i.e., predictions based upon covariates and spatial random effects), the SGLMM and RSR-SGLMM performed identically (as expected), and this was very similar to the performance of the SGLMM-selected, with median RMSE values clustered around 0.08 (Figure 3 C). The SGLMM, RSR-SGLMM, and SGLMM-selected methods showed comparable performance with overlapping interquartile ranges. Notably, the GLM approach was excluded from this comparison as it lacks spatial random effects and therefore cannot provide conditional predictions that account for spatial autocorrelation. The similar performance across spatial methods suggests that the underlying spatial structure is being captured equivalently, regardless of the estimator used.

0.2 Unconditional prediction

The evaluation of unconditional prediction (predicting the response from covariate effects alone) highlights the practical consequences of estimating a *direct* versus a *total effect* (Figure 4). The

GLM and RSR-SGLMM achieved identical and superior performance, with the lowest mean RMSE values across all scenarios (ranging from 0.201 at n=50 to 0.188 at n=200). In contrast, the SGLMM and SGLMM-selected consistently showed higher RMSE. All methods showed improvement with increasing sample size, with GLM and RSR-SGLMM demonstrating the most substantial reduction in RMSE (0.013 units from n=50 to n=200). When examining the effect of false covariates, GLM and RSR-SGLMM again showed identical performance improvements as the number of false covariates increased from 1 to 5 (RMSE decreasing from 0.203 to 0.182), while SGLMM showed modest improvement (from 0.213 to 0.202) and SGLMM-selected remained relatively stable across the range of false covariates.

Case study examples

The application of the three modeling approaches to the groundfish data of the Eastern Bering Sea revealed substantial differences in the estimated temperature-density relationships between species (Figure 5). A consistent pattern emerged whereby the Restricted Spatial Regression (RSR-SGLMM) model produced temperature response curves similar to the non-spatial GLM, while the standard spatial SGLMM often yielded markedly different estimates.

For adult Pollock, both the GLM and RSR-SGLMM identified a thermal optimum between 2–3 , while the SGLMM predicted a slightly warmer peak at 4 . A similar pattern was observed for adult Pacific Cod, where the GLM and RSR-SGLMM indicated peak densities around 3 , whereas the SGLMM placed the optimum at 4 . In the case of Tanner crab, the models produced more divergent results. The GLM and RSR-SGLMM estimated a single optimum between 3–4 , whereas the SGLMM suggested highest densities in much colder waters below 0 , and a decline with increasing temperatures.

Discussion

Spatial confounding presents a critical challenge in ecology, where the collinearity between environmental covariates and unmodeled spatial processes (e.g., population diffusion), can distort model inference and predictions. Our study compared GLMs, spatial GLMMs, and the RSR estimator, and reframed the debate around a critical, often-overlooked question: what is the ecological quan-

tity of interest? We argue the choice between SGLMM and RSR cannot be decided based on the fit to data, but instead is based on an explicit choice between estimating a covariate's *direct effect* versus its *total effect*. This choice must be guided by the scientific goal and explicit assumptions about the nature of unmeasured spatial processes.

The spatial random field in an SGLMM is a proxy for real, unmeasured ecological processes (e.g., latent variables and individual movement) that are spatially structured [10, 11] . The central issue is how to interpret the collinearity between these unmeasured processes and a measured covariate. This collinearity can arise from an unmeasured process being a confounders (a common cause of both the measured covariate and the response) or mediators (a variable on the causal pathway between the covariate and the response). The SGLMM estimates the effect of a covariate after correcting for the correlation expected to occur by chance, effectively treating the spatial field as a confounder to be controlled for [32, 33]. By partialing out this shared variance, the SGLMM estimates the *direct effect*. This makes the SGLMM suitable for hypothesis testing and estimating standard errors due to its reliable control over Type I error. In contrast, the RSR approach is designed to estimate the *total effect* (the full influence of a covariate) as if no spatial term were present, while still accounting for spatial autocorrelation in the residuals [19, 20]. It achieves this by re-attributing the shared variance back to the covariate, thereby treating the spatial field as a mediator whose influence should be included.

For unconditional prediction (predicting the response from covariate effects alone) or model transferability, the RSR estimator is often a more appropriate tool, particularly if the relationship between latent variable and covariate is fixed (stationary) [17, 18]. This is especially relevant for climate-related covariates, where many spatial processes (e.g., local adaptations, biotic interactions, habitat modifications) may be consequences rather than causes of climatic conditions. However, the utility of RSR for transferability hinges on the critical assumption that the relationship between the covariate and these mediating processes remains stationary across the space or time to which the model is being transferred [34]. If a local mediating process in the training data (e.g., a specific prey species or biotic interaction) is absent or behaves differently in the new context, the total effect estimated by RSR will likely not be transferable. Therefore, RSR is most appropriate when there is strong ecological evidence that mediating processes are both stable and generalizable across the intended domain of application.

This distinction is not merely a statistical nuance; it can lead to profoundly different ecological conclusions. As demonstrated in our case study, the inferred temperature responses for key marine species like Tanner crab varied considerably, with the SGLMM estimating higher density at low temperatures (¡ 0) while the RSR, similar to a simple GLM, estimated a higher densities at temperature around 3–4 . Such discrepancies directly impact our understanding of species' fundamental niches and can yield conflicting forecasts for distribution shifts under climate change [18].

It is important to note that although we recommend RSR for unconditional prediction, its inflated Type-I error rate makes it unsuitable for variable selection or formal hypothesis testing. Our results, which align with recent statistical literature [35, 23, 24], confirm that using RSR to decide whether a covariate has a "significant" effect would lead to an unacceptably high rate of false discoveries.

We therefore recommend a three-step conceptual approach for practitioners:

1. State the ecological goal and assumptions. First, determine whether the research question requires estimating a direct effect (controlling for confounders) or a total effect (including mediators).

2. Use the standard SGLMM for hypothesis testing and variable selection. Its reliable control over Type I error makes it the only appropriate tool to test whether a covariate has a meaningful association with the response after accounting for spatial structure.

3. If the goal is forecasting, use the RSR-SGLMM to estimate the total effect magnitude. For covariates identified as important in the first step, and assuming the unmeasured spatial processes are stable mediators, the RSR can then be used to estimate the full, transferable slope for use in unconditional predictions or long-term forecasting.

Distinguishing between confounders and mediators requires ecological reasoning and a priori knowledge of the system to justify their assumptions about the underlying causal structure, often using tools like Directed Acyclic Graphs (DAGs) to make these assumptions explicit [27, 26, 36]. When uncertainty exists about these causal relationships, a sensitivity analysis using both the SGLMM and RSR approaches can help assess the robustness of conclusions to different assumptions

about unmeasured spatial processes.

Our findings open important avenues for future research. While we hypothesize that RSR's superior unconditional predictive performance will translate to more accurate long-term forecasting, this remains to be fully tested. References [18] and [37] demonstrated the utility of the RSR estimator for correctly partitioning effects in joint species distribution models, but empirical tests of long-term forecasting accuracy are still needed. Studies using long-term ecological monitoring data could conduct leave-future-out cross-validation experiments (retrospective skill testing), where the model model is fitted to an early period and the predicted species distributions decades later is then evaluated [38].

In conclusion, this study provides clear guidance on the trade-offs between standard and restricted spatial regression in spatial modeling. The RSR estimator is the preferred tool when the goal is unconditional prediction and the relationship between latent variable and covariates is stationary, such that the total effect is transferable across space or time. For conditional prediction, such as mapping a species' current distribution, a standard SGLMM provides a more statistically conservative approach due to its better control of false discoveries. By implementing the RSR estimator in the widely used tinyVAST R package, we have made this powerful technique more accessible to the ecological community, facilitating more nuanced and targeted applications of spatial modeling to address pressing questions in conservation and management.

Acknowledgments

We thank Anthony Charsley, Jennifer Bigman, Andrew Allyn, and Julia Indivero for the insightful discussions that helped inspire this research. We also thank Arnaud Grüss and Sean Anderson for a revision of a previous version of this manuscript. Finally, we thank Kasper Kristensen for their ongoing development of Template Model Builder, without which this work would not be practical.

References

[1] Malin L. Pinsky, Rebecca L. Selden, and Zoë J. Kitchel. Climate-Driven Shifts in Marine Species Ranges: Scaling from Organisms to Communities. *Annual Review of Marine Science*, 12(Volume 12, 2020):153–179, January 2020.

ISSN 1941-1405, 1941-0611. doi: 10.1146/annurev-marine-010419-010916. URL https://www.annualreviews.org/content/journals/10.1146/annurev-marine-010419-010916. Publisher: Annual Reviews.

[2] Jane Elith and John R. Leathwick. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40(Volume 40, 2009):677–697, December 2009. ISSN 1543-592X, 1545-2069. doi: 10.1146/annurev.ecolsys.110308.120159. URL https://www.annualreviews.org/content/journals/10.1146/annurev.ecolsys.110308.120159. Publisher: Annual Reviews.

[3] Malin L. Pinsky, Gabriel Reygondeau, Richard Caddell, Juliano Palacios-Abrantes, Jessica Spijkers, and William W. L. Cheung. Preparing ocean governance for species on the move. *Science*, 360(6394):1189–1191, June 2018. doi: 10.1126/science.aat2360. URL https://www.science.org/doi/full/10.1126/science.aat2360. Publisher: American Association for the Advancement of Science.

[4] W. R. Tobler. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46:234–240, 1970. ISSN 0013-0095. doi: 10.2307/143141. URL https://www.jstor.org/stable/143141. Publisher: [Clark University, Wiley].

[5] Emiko Dupont, Isa Marques, and Thomas Kneib. Demystifying Spatial Confounding, November 2024. URL http://arxiv.org/abs/2309.16861. arXiv:2309.16861 [stat].

[6] D G Clayton, L Bernardinelli, and C Montomoli. Spatial Correlation in Ecological Analysis. *International Journal of Epidemiology*, 22(6):1193–1202, December 1993. ISSN 0300-5771. doi: 10.1093/ije/22.6.1193. URL https://doi.org/10.1093/ije/22.6.1193.

[7] Jérôme Guélat and Marc Kéry. Effects of spatial autocorrelation and imperfect detection on species distribution models. *Methods in Ecology and Evolution*, 9(6):1614–1625, 2018. ISSN 2041-210X. doi: 10.1111/2041-210X.12983. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12983. _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12983.

[8] Carsten F. Dormann, Jana M. McPherson, Miguel B. Araújo, Roger Bivand, Janine Bolliger, Gudrun Carl, Richard G. Davies, Alexandre Hirzel, Walter Jetz, W. Daniel Kissling, Ingolf Kühn, Ralf Ohlemüller, Pedro R. Peres-Neto, Björn Reineking, Boris Schröder, Frank M. Schurr, and Robert Wilson. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628, 2007. ISSN 1600-0587. doi: 10.1111/j.2007.0906-7590.05171.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2007.0906-7590.05171.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2007.0906-7590.05171.x.

[9] Colin M. Beale, Jack J. Lennon, Jon M. Yearsley, Mark J. Brewer, and David A. Elston. Regression analysis of spatial data. *Ecology Letters*, 13 (2):246–264, 2010. doi: https://doi.org/10.1111/j.1461-0248.2009.01422.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1461-0248.2009.01422.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1461-0248.2009.01422.x.

[10] Finn Lindgren, Håvard Rue, and Johan Lindström. An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(4): 423–498, September 2011. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2011.00777.x. URL `https://doi.org/10.1111/j.1467-9868.2011.00777.x`.

[11] Noel Cressie and Christopher K. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, April 2011. ISBN 978-0-471-69274-4.

[12] Finn Lindgren and Håvard Rue. Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, 63:1–25, February 2015. ISSN 1548-7660. doi: 10.18637/jss.v063.i19. URL `https://doi.org/10.18637/jss.v063.i19`.

[13] Owen R. Liu, Eric J. Ward, Sean C. Anderson, Kelly S. Andrews, Lewis A K. Barnett, Stephanie Brodie, Gemma Carroll, Jerome Fiechter, Melissa A. Haltuch, Chris J. Harvey, Elliott L. Hazen, Pierre-Yves Hernvann, Michael Jacox, Isaac C. Kaplan, Sean Matson, Karma Norman, Mercedes Pozo Buil, Rebecca L. Selden, Andrew Shelton, and Jameal F. Samhouri. Species redistribution creates unequal outcomes for multispecies fisheries under projected cli-

mate change. *Science Advances*, 9(33):eadg5468, August 2023. doi: 10.1126/sciadv.adg5468. URL https://www.science.org/doi/full/10.1126/sciadv.adg5468. Publisher: American Association for the Advancement of Science.

[14] Raquel Ruiz-Diaz, Mariano Koen-Alonso, Frédéric Cyr, Jonathan A. D. Fisher, Sherrylynn Rowe, Katja Fennel, Lina Garcia-Suarez, and Tyler D. Eddy. Climate models drive variation in projections of species distribution on the Grand Banks of Newfoundland. *PLOS Climate*, 3(11):e0000520, November 2024. ISSN 2767-3200. doi: 10.1371/journal.pclm.0000520. URL https://journals.plos.org/climate/article?id=10.1371/journal.pclm.0000520. Publisher: Public Library of Science.

[15] Maurice C. Goodman, Jonathan C. P. Reum, Cheryl L. Barnes, Andre E. Punt, James N. Ianelli, Elizabeth A. McHuron, Giulio A. De Leo, and Kirstin K. Holsman. Climate Covariate Choice and Uncertainty in Projecting Species Range Shifts: A Case Study in the Eastern Bering Sea. *Fish and Fisheries*, 26(2):219–239, 2025. ISSN 1467-2979. doi: 10.1111/faf.12875. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/faf.12875. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/faf.12875.

[16] James T. Thorson, Mark D. Scheuerell, Andrew O. Shelton, Kevin E. See, Hans J. Skaug, and Kasper Kristensen. Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*, 6(6):627–637, 2015. ISSN 2041-210X. doi: 10.1111/2041-210X.12359. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12359. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12359.

[17] Trevor J. Hefley, Kristin M. Broms, Brian M. Brost, Frances E. Buderman, Shannon L. Kay, Henry R. Scharf, John R. Tipton, Perry J. Williams, and Mevin B. Hooten. The basis function approach for modeling autocorrelation in ecological data. *Ecology*, 98(3):632–646, 2017. ISSN 1939-9170. doi: 10.1002/ecy.1674. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ecy.1674. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecy.1674.

[18] Francis K. C. Hui, Quan Vu, and Mevin B. Hooten. Spatial confounding in joint

species distribution models. *Methods in Ecology and Evolution*, 15(10):1906–1921, October 2024. ISSN 2041-210X, 2041-210X. doi: 10.1111/2041-210X.14420. URL https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14420.

[19] Brian J. Reich, James S. Hodges, and Vesna Zadnik. Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models. *Biometrics*, 62(4):1197–1206, December 2006. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2006.00617.x. URL https://doi.org/10.1111/j.1541-0420.2006.00617.x.

[20] James S. Hodges and Brian J. Reich. Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. *The American Statistician*, 64(4):325–334, November 2010. ISSN 0003-1305. doi: 10.1198/tast.2010.10052. URL https://doi.org/10.1198/tast.2010.10052. Publisher: ASA Website _eprint: https://doi.org/10.1198/tast.2010.10052.

[21] John Hughes and Murali Haran. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):139–159, 2013. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2012.01041.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2012.01041.x. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2012.01041.x.

[22] Jérémy Lamouroux, Alizée Geffroy, Sébastien Leblond, Caroline Meyer, and Isabelle Albert. Addressing spatial confounding in geostatistical regression models: An R-INLA approach. *Methods in Ecology and Evolution*, 16(9):2082–2097, 2025. ISSN 2041-210X. doi: 10.1111/2041-210X.70106. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.70106. _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.70106.

[23] Dale L. Zimmerman and Jay M. Ver Hoef. On Deconfounding Spatial Confounding in Linear Models. *The American Statistician*, 76(2):159–167, April 2022. ISSN 0003-1305. doi: 10.1080/00031305.2021.1946149. URL https://doi.org/10.1080/00031305.2021.1946149. Publisher: ASA Website _eprint: https://doi.org/10.1080/00031305.2021.1946149.

[24] Kori Khan and Catherine A. Calder. Restricted Spatial Regression Methods: Implications for Inference. *Journal of the American Statistical Association*, 117(537):482–494, 2022. doi:

10.1080/01621459.2020.1788949. URL https://doi.org/10.1080/01621459.2020.1788949. Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2020.1788949.

[25] James B. Grace and Kathryn M. Irvine. Scientist's guide to developing explanatory statistical models using causal analysis principles. *Ecology*, 101(4):e02962, 2020. ISSN 1939-9170. doi: 10.1002/ecy.2962. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ecy.2962. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecy.2962.

[26] James B. Grace. An integrative paradigm for building causal knowledge. *Ecological Monographs*, 94(4):e1628, 2024. ISSN 1557-7015. doi: 10.1002/ecm.1628. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1628. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecm.1628.

[27] Suchinta Arif and M. Aaron MacNeil. Applying the structural causal model framework for observational causal inference in ecology. *Ecological Monographs*, 93(1):e1554, 2023. ISSN 1557-7015. doi: 10.1002/ecm.1554. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1554. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecm.1554.

[28] James T. Thorson, Sean C. Anderson, Pamela Goddard, and Christopher N. Rooper. tinyVAST: R Package With an Expressive Interface to Specify Lagged and Simultaneous Effects in Multivariate Spatio-Temporal Models. *Global Ecology and Biogeography*, 34(4):e70035, 2025. ISSN 1466-8238. doi: 10.1111/geb.70035. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/geb.70035. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/geb.70035.

[29] M. B. Hooten and N. T. Hobbs. A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28, 2015. ISSN 1557-7015. doi: 10.1890/14-0661.1. URL https://onlinelibrary.wiley.com/doi/abs/10.1890/14-0661.1. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/14-0661.1.

[30] Kasper Kristensen, Anders Nielsen, Casper W. Berg, Hans Skaug, and Bradley M. Bell. TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statisti-*

cal Software, 70:1–21, April 2016. ISSN 1548-7660. doi: 10.18637/jss.v070.i05. URL https://doi.org/10.18637/jss.v070.i05.

[31] Hans J. Skaug and David A. Fournier. Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics & Data Analysis*, 51 (2):699–709, November 2006. ISSN 0167-9473. doi: 10.1016/j.csda.2006.03.005. URL https://www.sciencedirect.com/science/article/pii/S0167947306000764.

[32] Nina K. Lany, Phoebe L. Zarnetske, Andrew O. Finley, and Deborah G. McCullough. Complementary strengths of spatially-explicit and multi-species distribution models. *Ecography*, 43(3):456–466, 2020. ISSN 1600-0587. doi: 10.1111/ecog.04728. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.04728. _eprint: https://nsojournals.onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.04728.

[33] Jussi Mäkinen, Elina Numminen, Pekka Niittynen, Miska Luoto, and Jarno Vanhatalo. Spatial confounding in Bayesian species distribution modeling. *Ecography*, 2022(11):e06183, 2022. ISSN 1600-0587. doi: 10.1111/ecog.06183. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.06183. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.06183.

[34] Katherine L. Yates, Phil J. Bouchet, M. Julian Caley, Kerrie Mengersen, Christophe F. Randin, Stephen Parnell, Alan H. Fielding, Andrew J. Bamford, Stephen Ban, A. Márcia Barbosa, Carsten F. Dormann, Jane Elith, Clare B. Embling, Gary N. Ervin, Rebecca Fisher, Susan Gould, Roland F. Graf, Edward J. Gregr, Patrick N. Halpin, Risto K. Heikkinen, Stefan Heinänen, Alice R. Jones, Periyadan K. Krishnakumar, Valentina Lauria, Hector Lozano-Montes, Laura Mannocci, Camille Mellin, Mohsen B. Mesgaran, Elena Moreno-Amat, Sophie Mormede, Emilie Novaczek, Steffen Oppel, Guillermo Ortuño Crespo, A. Townsend Peterson, Giovanni Rapacciuolo, Jason J. Roberts, Rebecca E. Ross, Kylie L. Scales, David Schoeman, Paul Snelgrove, Göran Sundblad, Wilfried Thuiller, Leigh G. Torres, Heroen Verbruggen, Lifei Wang, Seth Wenger, Mark J. Whittingham, Yuri Zharikov, Damaris Zurell, and Ana M. M. Sequeira. Outstanding Challenges in the Transferability of Ecological Models. *Trends in Ecology*

& Evolution, 33(10):790–802, October 2018. ISSN 0169-5347. doi: 10.1016/j.tree.2018.08.001. URL https://www.sciencedirect.com/science/article/pii/S0169534718301812.

[35] Ephraim M. Hanks, Erin M. Schliep, Mevin B. Hooten, and Jennifer A. Hoeting. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26(4):243–254, 2015. ISSN 1099-095X. doi: 10.1002/env.2331. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2331. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.2331.

[36] Jarrett E. K. Byrnes and Laura E. Dee. Causal Inference With Observational Data and Unobserved Confounding Variables. *Ecology Letters*, 28(1):e70023, 2025. ISSN 1461-0248. doi: 10.1111/ele.70023. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.70023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.70023.

[37] Justin J. Van Ee, Jacob S. Ivan, and Mevin B. Hooten. Community confounding in joint species distribution models. *Scientific Reports*, 12(1): 12235, July 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-15694-6. URL https://www.nature.com/articles/s41598-022-15694-6. Publisher: Nature Publishing Group.

[38] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. ISSN 1600-0587. doi: 10.1111/ecog.02881. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.02881. _eprint: https://nsojournals.onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.02881.

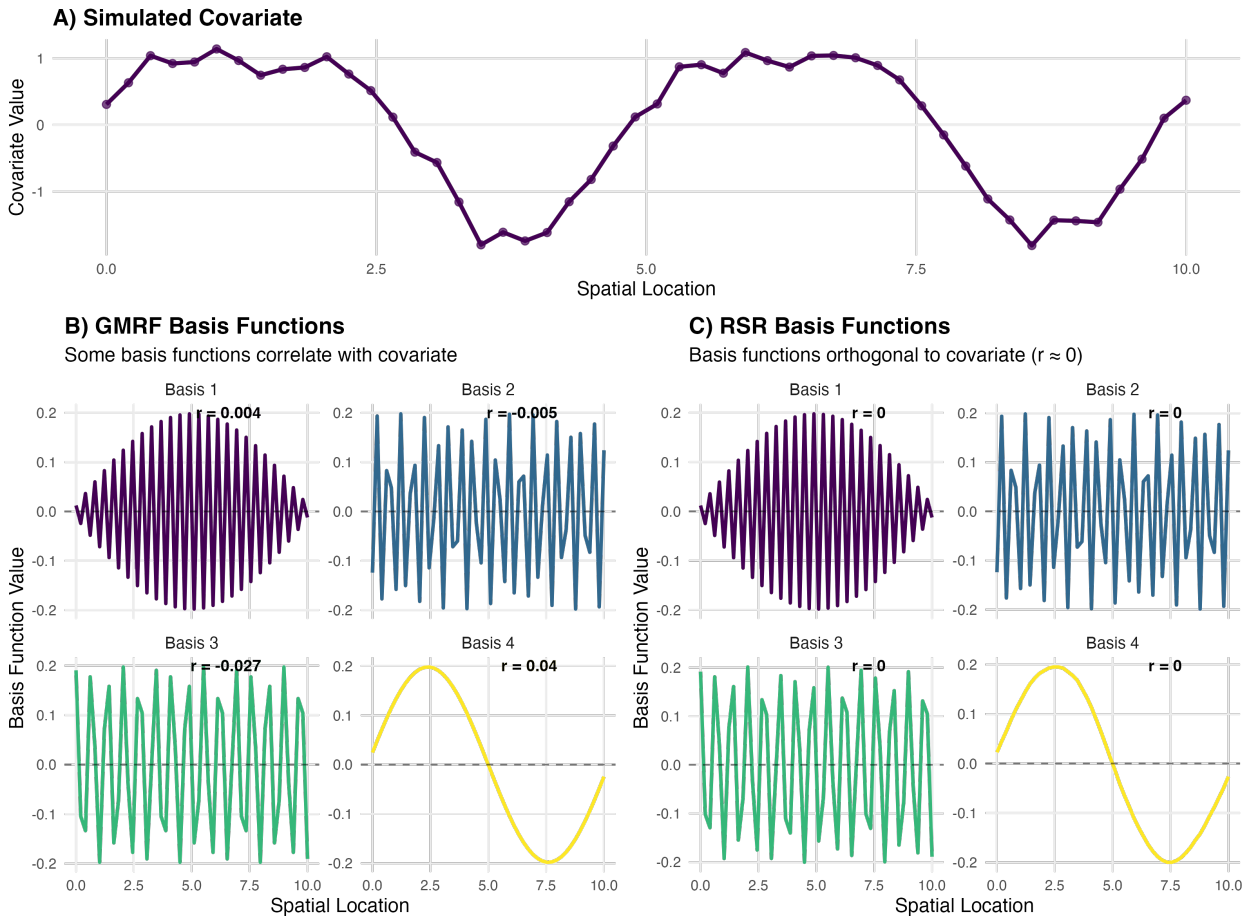

Figure 2: A conceptual illustration of the Restricted Spatial Regression (RSR) principle. A) A simulated covariate is shown across a generic one-dimensional spatial domain (e.g., latitude or distance along a transect). B) A spatial process like a GMRF is composed of many underlying spatial patterns, or "basis functions," at different scales; four representative examples are shown. By chance, some of these basis functions will be correlated with the covariate (note the non-zero correlation, r). The total confounding in a model is the cumulative effect of these chance correlations across all basis functions. C) The RSR procedure is mathematically equivalent to projecting these basis functions into a space where they are orthogonal to (uncorrelated with) the covariate.

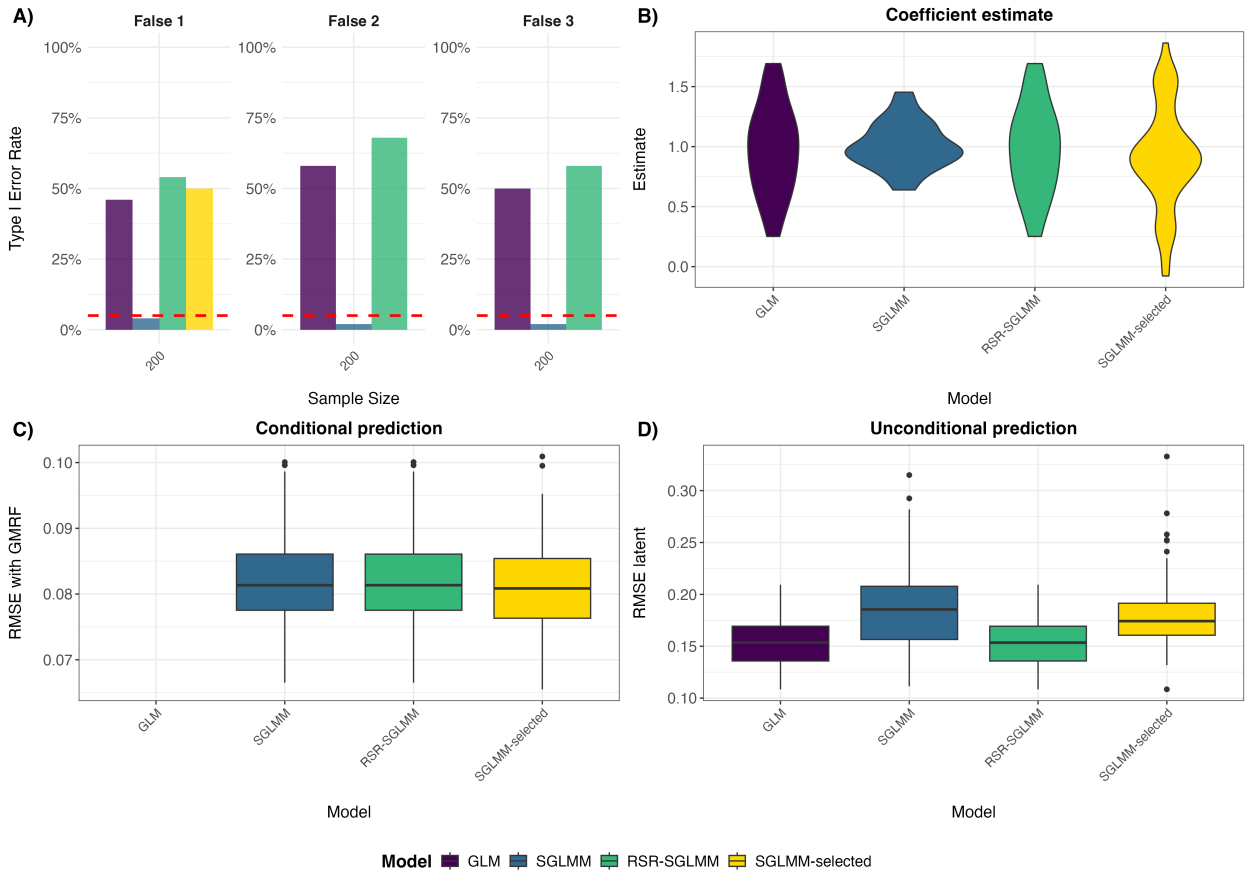

Figure 3: Performance comparison of spatial modeling approaches across simulation scenarios (n = 200 simulations). A): Type I error rates for three false discovery scenarios, with the red dashed line indicating the nominal 5% significance level. B): Distribution of parameter estimates across models, showing bias and variance in covariate effect estimation. C): Root Mean Square Error (RMSE) for latent field estimation, demonstrating spatial prediction accuracy. D): RMSE comparison when using GMRF basis functions, showing relative performance of each approach for spatial structure recovery. Models compared include: Covariate Only (purple), SGLMM (blue), RSR-SGLMM (green), and SGLMM-selected (yellow).

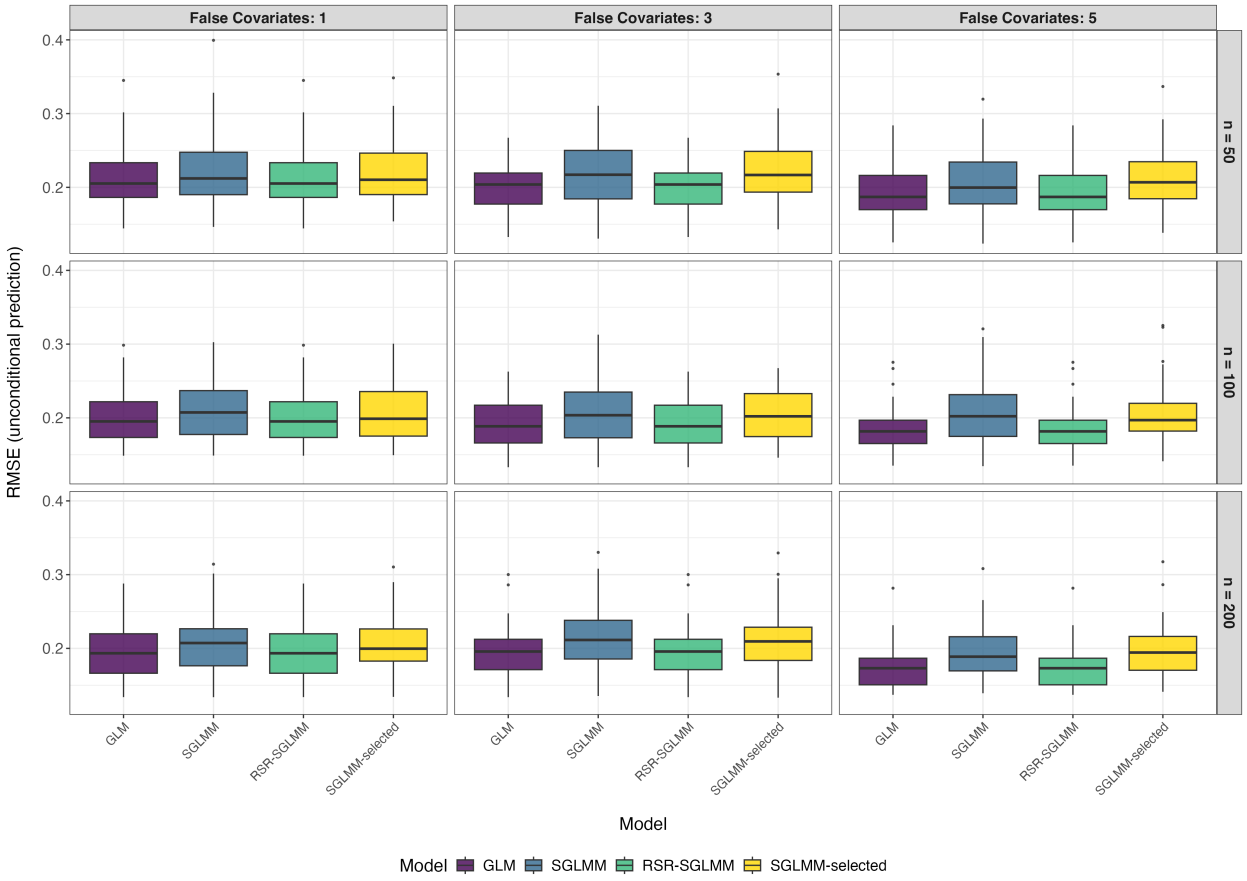

Figure 4: RMSE performance (lower is better) for latent field estimation across varying numbers of false covariates and sample sizes. Root Mean Square Error (RMSE) for predicting the latent spatial field is shown across three scenarios with different numbers of false covariates (columns: 1, 3, and 5 false covariates) and three sample sizes (rows: n = 50, 100, and 200). Models compared include: GLM (purple), SGLMM (blue), RSRSGLMM (green), and SGLMMselected (yellow).

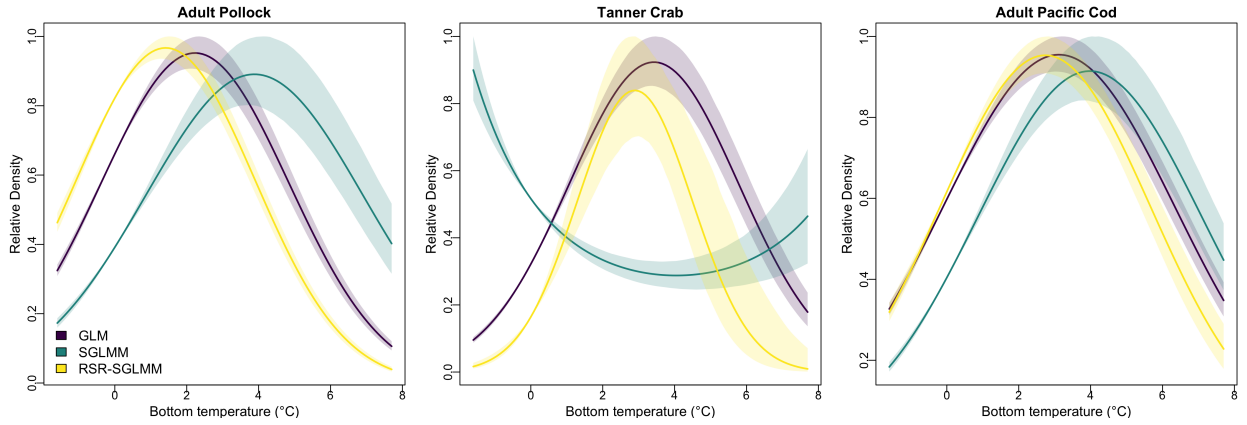

Figure 5: Temperature response curves for three marine species from the Bering Sea. Predicted log density responses to bottom temperature () for Adult Pollock, Tanner Crab, and Adult Pacific cod using three modeling approaches: GLM (purple), SGLMM (blue), and RSR-SGLMM (yellow). Relative density scaled to max = 1.

1 Supporting Information S1

1.1 RSR adjustment for random-effect coefficients

In the main text, we present the estimators:

$$\hat{\beta}^* = \hat{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\hat{\omega} \tag{1}$$

and:

$$\mathbf{A}\hat{\omega}^* = \mathbf{A}\hat{\omega} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\hat{\omega} \tag{2}$$

where $g^{-1}(\mu) = \mathbf{X}\beta^*$ approximates the GLM estimator and $g^{-1}(\mu) = \mathbf{X}\beta^* + (\mathbf{A}\omega)^*$ is equivalent to the SGLMM estimator.

However, the analyst might instead seek to calculate the adjusted value for random effects ω' themselves. In this case, Eq. 5 can be approximated as:

$$\hat{\omega}' = \hat{\omega} - \delta' \tag{3}$$

$$\delta' = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\hat{\omega} \tag{4}$$

where $\mathbf{Z} = \mathbf{A}^T\mathbf{X}$. This approximation is exact when $\mathbf{A}$ is an indicator matrix (i.e., using an areal model, or the FEM for the SPDE method has a vertex for each sample location), because $\mathbf{A}^T\mathbf{A}$ is then a diagonal matrix of 1s and 0s, with 1s corresponding to vertices at sample locations and 0 otherwise. We use $\hat{\omega}'$ to indicate that it is a different (lower-resolution) estimator for the RSR adjustment $\hat{\omega}^*$. Similarly, the lower-resolution RSR estimator is updated as:

$$\hat{\beta}' = \hat{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\hat{\omega}' \tag{5}$$

This then allows us to identify the pair of vectors $\{\hat{\beta}', \hat{\omega}'\}$ that are adjusted to be decorrelated at a shared scale. However, identifying this pair requires working on the lower resolution implied by interpolation matrix $\mathbf{A}$.

1.2 Projecting beyond data

When projecting beyond the range of data, an analyst might then seek an estimator that uses conditional prediction "near the data", and unconditional-RSR prediction "away from the data", while using the estimated distribution of random effects (i.e., the precision matrix $\mathbf{Q}$) to bridge between the two estimators. For example, this arises when applying the RSR estimator to a spatio-temporal model, when the analyst might want to use conditional prediction during years with available data, the unconditional-RSR prediction when projecting far into the future (such that data are uninformative about random effects), and bridging between them for short-term projections (years immediately following the most recent data).

To express this formally, we augment our previous notation by defining random effects within the range of fitted data ω_A and covariates $\mathbf{X}_A$, and then additional random effects ω_B and covariates $\mathbf{X}_B$ beyond the range of the data. We use RSR-adjusted covariates $\hat{\beta}'$ for both sets of prediction, and the RSR-adjusted random effects $\hat{\omega}'_A$ within the range of data, such that we obtain the conditional predictions within this range. However, we must decide how to bridge the RSR-correction δ'_A beyond the data to calculate the RSR adjustment δ'_B beyond the range of data. We further define the joint precision:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{\mathrm{AA}} & \mathbf{Q}_{\mathrm{AB}} \\ \mathbf{Q}_{\mathrm{BA}} & \mathbf{Q}_{\mathrm{BB}} \end{pmatrix} \tag{6}$$

where $\mathbf{Q}_{\mathrm{AA}}$ is the precision within the data (among ω_A), $\mathbf{Q}_{\mathrm{BB}}$ is the precision beyond the data (for ω_B), and $\mathbf{Q}_{\mathrm{AB}}$ is the cross-precision.

We then envision calculating the RSR adjustment for ω_{B} using the conditional krigging formula:

$$\delta'_{\mathrm{B}} = \mathbf{Q}_{\mathrm{BB}}^{-1}\mathbf{Q}_{\mathrm{BA}}\delta'_{\mathrm{A}} \tag{7}$$

$$\hat{\omega}'_{\mathrm{B}} = \hat{\omega}_{\mathrm{B}} - \delta'_{\mathrm{B}} \tag{8}$$

where this RSR adjustment δ'_{B} will naturally revert towards zero well beyond the range of data. We recommend further research testing the application of the RSR estimator when bridging between conditional and RSR-unconditional predictors in spatio-temporal forecasts.