Addressing Missing Covariates in Species Distribution Models: Inferential Impacts and Mitigation via Joint Species Distribution Models

Arthur F. ROSSIGNOL

AgroParisTech, 22 Place de l'Agronomie, 91120 Palaiseau, France.

INRAE, UR Écosystèmes forestiers (EFNO), Domaine des Barres,
311 Chemin de la Grande Métairie, 45290 Nogent-sur-Vernisson, France.

arthur.f.rossignol@gmail.com | 0009-0000-1494-377X

Frédéric GOSSELIN

INRAE, UR Écosystèmes forestiers (EFNO), Domaine des Barres,
311 Chemin de la Grande Métairie, 45290 Nogent-sur-Vernisson, France.

frederic.gosselin@inrae.fr | 0000-0003-3737-106X

Abstract

Species distribution models (SDMs) are widely used in ecology to assess the distribution of species populations across space and time. Correlative SDMs, in particular, are used to infer relationships between species records and environmental variables. A classical approach for implementing this type of SDMs is to employ generalized linear mixed models (GLMMs) as a parametric regression method. However, due to the complexity of species-environment relationships, species distributions may depend on unobserved or unmeasurable covariates. In this article, we first recall certain mathematical results showing that such "omitted covariates" typically introduce statistical issues that can bias the inference of observed covariate effects or yield improper confidence intervals. So far, these results have received little attention in ecology. We then present a comprehensive simulation-based investigation of the statistical impact of unobserved covariates on the inference performance of GL(M)Ms for continuous, count, and binary data. We assessed various regression methods, including both frequentist and Bayesian SDMs, and so-called joint species distribution models (JSDMs) used to account for interspecific covariations in presence–absence data. Our work demonstrates that JSDMs provide a robust statistical approach that mitigates inferential issues arising in SDMs due to missing covariates and enables reliable estimates of environmental effects. We further complemented these simulation results by applying JSDMs and SDMs to several ecological datasets, revealing discrepancies between SDM and JSDM estimation of environmental effects and a better predictive capacity for JSDMs than for SDMs. As a general recommendation, we encourage ecologists and practitioners to consider fitting JSDMs when dealing with community data to be able to evaluate whether any information can be extracted from between-species residuals. Ultimately, our results remain broadly applicable to GL(M)Ms in which important variables are suspected of being omitted, in which case generalized linear latent variable models (GLLVMs) could properly correct inference when different entities might share the same omitted important covariate.

Keywords

Missing covariate; Unobserved covariate; Omitted variable bias (OVB); Species distribution model (SDM); Joint species distribution model (JSDM); Generalized linear latent variable model (GLLVM); Model misspecification; Bias mitigation.

Significance

This work addresses a critical, yet insufficiently acknowledged, statistical challenge in ecological statistics: the impact of unobserved environmental predictors on species distribution estimates. Mathematical and simulation results indicate strong bias in estimated environmental effects for presence–absence data when a covariate is omitted. We propose a mitigation strategy based on joint species distribution modelling, which performs very well when the omitted variable is shared across several species. Applications to real ecological datasets demonstrated the widespread prevalence of differences in environmental effects estimation between joint species distribution models and separate species distribution models. Our results extend to generalized linear models and generalized linear latent variable models, which can be applicable across a range of disciplines.

1. Introduction

Species distribution models (SDMs) are a major modeling tool, widely used in ecology and conservation to study the distribution of species populations over space and time (Elith & Leathwick, 2009). They have found several applications in conservation biology (Jung et al., 2021), biodiversity management (Fois et al., 2018), and in assessing the effects of climate change (Santini et al., 2021). Due to their relative ease of implementation, correlative SDMs have been widely employed to infer relationships between species records and environmental variables when assessing habitat suitability and species distributions (e.g., Clark et al., 2017; Guisan et al., 2017). Numerous regression approaches have been developed to analyze abundance or presence–absence, through frequentist, Bayesian, or machine learning methods. Among them, parametric models, such as generalized linear mixed models (GLMMs), and semiparametric models, such as generalized additive mixed models (GAMMs), are classically used (McCullagh & Nelder, 2019; Raudenbush & Bryk, 2002; Stroup et al., 2024).

Nonetheless, SDMs have important conceptual limitations stemming from their underlying assumptions (e.g., M. Austin, 2007; Lee-Yaw et al., 2022; Weber et al., 2017). In practice, an SDM explains a species' distribution by relating it to measured environmental predictors, predominantly corresponding to abiotic conditions. They thus, as such, omit biotic interactions (e.g., competition, facilitation, mutualism, predation, host–parasite dynamics), which yet contribute to shaping species distributions (Wisz et al., 2013). An SDM is built for a single species, effectively assuming independence of the species from the rest of the community; therefore, interspecific interactions are treated as unmeasured noise. Covariate effects inferred from SDMs inevitably capture some of the biotic constraints, thus recovering the *realized niche* — i.e., the observed distribution in the Hutchinsonian sense (Hutchinson, 1957) — rather than the underlying *fundamental niche* (Araújo & Guisan, 2006). This conflation has practical implications: coefficient estimates become harder to interpret causally, and predictions may transfer poorly across space or time when community composition changes. Thus, although SDMs remain valuable for describing current patterns, they are intrinsically limited in their ability to estimate the realized niche and cannot delimit the fundamental niche (Soberón & Nakamura, 2009).

Other limitations stem from the fact that species distributions may also depend on sources of heterogeneity other than biotic interactions, including unmeasured environmental predictors (Wisz et al., 2013), population dynamics (Mielke et al., 2020), or spatial dispersion (Shurin et al., 2009). Specifically, it is often impossible to have all the variables affecting species distributions at one's disposal, due to, for example, ecological complexity or practical sampling limitations. Consequently,

SDMs are instantiated with the observed covariates only, although other unobserved covariates may also have an effect on the species distribution. This results in a *misspecification* of the mean model (WEISBERG, 2014), which can have potential impacts on the reliability of inference outcomes. In particular, it involves the so-called *omitted-variable bias* (OVB), i.e., the bias affecting the estimates of the fixed effects of observed covariates due to missing covariates. This phenomenon is well known in applied statistics, econometrics, and medicine (e.g., CLARKE, 2005; REHM et al., 1992; WILMS et al., 2021; WOOLDRIDGE, 2009). However, we have noticed that this topic remains poorly acknowledged in the context of ecological modeling, despite the important consequences it may have on conclusions drawn (BYRNES & DEE, 2025; KISSLING et al., 2012; RINELLA et al., 2020). Varying denominations have been employed to mention this phenomenon: *omitted*, *unmeasured*, *hidden*, *unobserved*, or *missing* covariates.

Previous mathematical results have established that a missing covariate will affect estimates of the other regression parameters. In Tab. 1, some essential results concerning the inferential consequences of an unobserved variable in a GLM are reviewed. We consider a response variable Y depending on two environmental covariates X_1 and X_2, through the ecological process $Y_i \sim \mathsf{Normal}(z_i, \sigma^2)$ or $\mathsf{Poisson}(z_i)$ or $\mathsf{Bernoulli}(z_i)$ with $f(z_i) = \alpha + \beta X_{1,i} + \gamma X_{2,i}$, at any site i, where $z_i = \mathbb{E}[Y_i \mid X_{1,i}, X_{2,i}]$ and f is a link function. Supposing that covariate X_2 is not observed, the regression model, which has been deliberately misspecified, therefore assumes $f(\tilde{z}_i) = \hat{\alpha} + \hat{\beta} X_{1,i}$ with $\tilde{z}_i = \mathbb{E}[Y_i \mid X_{1,i}]$. One thus wants to know how estimates of $\hat{\alpha}$ and $\hat{\beta}$ would differ from the true values of α and β. With continuous data following a Normal distribution, neither the intercept nor X_1's regression coefficient suffers from bias. With a Poisson distribution for count data, a bias appears on the intercept, but the effect of X_1 remains unbiased even in the absence of X_2. However, in a binary dataset, both parameters become biased, leading to an attenuation (a bias toward zero). Therefore, when analyzing presence–absence data, an underestimation of covariate effects will occur if an important covariate is missing. Notice that these mathematical results hold only if both covariates X_1 and X_2 are independent of each other; otherwise, confounding may occur. We provide, in the Supplementary Material, mathematical proofs of the results claimed in Tab. 1 regarding bias of the optimal estimators (see Sect. A in the SM).

distribution	link function	Bias$[\hat{\alpha}]$	Bias$[\hat{\beta}]$	reference(s)
Normal	identity	$=0$	$=0$	GAIL et al. (1984) CANNER (1991) ISHII et al. (2022)
Poisson	log	$\neq 0$	$=0$	GAIL et al. (1984) PETERSEN & DEDDENS (2000) ISHII et al. (2022)
Bernoulli	logit	$\neq 0$	$\neq 0$	GAIL et al. (1984) ROBINSON & JEWELL (1991) MOOD (2010) CRAMER (2005)
Bernoulli	probit	$\neq 0$	$\neq 0$	YATCHEW & GRILICHES (1985)

Table 1: Synthesis of some mathematical results on the omitted-variable bias. The true simulation model is $Y_i \mid \{X_{1,i}, X_{2,i}\} \sim \text{Normal}(z_i, \sigma^2)$ or $\text{Poisson}(z_i)$ or $\text{Bernoulli}(z_i)$ with $f(z_i) = \alpha + \beta X_{1,i} + \gamma X_{2,i}$, while the misspecified regression model is $Y_i \mid X_{1,i} \sim \text{Normal}(\tilde{z}_i, \hat{\sigma}^2)$ or $\text{Poisson}(\tilde{z}_i)$ or $\text{Bernoulli}(\tilde{z}_i)$ with $f(\tilde{z}_i) = \hat{\alpha} + \hat{\beta} X_{1,i}$, where f is a link function. The covariates X_1 and X_2 are assumed to be independent of each other.

Because ecological datasets often lack important covariates that participate in shaping species distributions, efficient statistical methods to accommodate this limitation are essential for reliable statistical analyses. One popular mitigation strategy has relied on including a site-level random effect intended to capture unmeasured covariates (BOLKER et al., 2009; HARRISON et al., 2018; SCHIELZETH & NAKAGAWA, 2013). Yet, including such a random effect exhibiting a large-range spatial autocorrelation may unintentionally amplify the bias in the estimated effects of the covariates, thereby questioning the usefulness of this solution (REICH et al., 2006). Therefore, the impossibility of classical SDMs to incorporate ecological processes beyond those measured is an important limitation for studies in community ecology (M. P. AUSTIN & VAN NIEL, 2011; WISZ et al., 2013).

Over the past fifteen years, community ecology has benefited from advances in statistical methods for species distribution modeling (e.g., GIMENEZ et al., 2014). In particular, *joint species distribution models* (JSDMs) have emerged as a powerful approach to address among-species residual correlations in a multivariate multi-species community (e.g., OVASKAINEN & ABREGO, 2020; VANHATALO et al., 2020; WARTON et al., 2015). Whereas early formulations explicitly estimated the full interspecific covariance matrix, the latent factor approach has reduced the dimensionality of the problem by introducing *latent variables* that serve as additional predictors, accompanied by their own coefficients called *factor loadings* (OVASKAINEN & ABREGO, 2020; WARTON et al., 2015). Moreover, JSDMs were viewed at first as a promising tool for disentangling biotic and abiotic drivers, since they make it possible to infer residual interspecific correlations not explained by the observed environmental co-

variates (POLLOCK et al., 2014; WARTON et al., 2015). However, this claim has been widely debated and even criticized in the literature, primarily because JSDMs are theoretically unable to distinguish true biotic interactions from mere co-occurrences (e.g., BLANCHET et al., 2020; CLARK et al., 2014). More recently, POGGIATO et al. (2021) even stated that JSDMs do not improve estimation of abiotic environmental effects over classical SDMs, thus questioning their interest for estimating the species' realized niches.

Our aim in this article is twofold: (i) to highlight an underappreciated hazard in parametric species distribution modeling arising from important unobserved predictors, and (ii) to clarify some aspects of the ongoing debate on JSDMs concerning their potential to provide more reliable estimates of environmental relationships. We emphasize that JSDMs with a latent factor approach are a suitable tool for handling incomplete datasets, and can even "reconstruct" a missing predictor from residual between-species covariations. In particular, we show that JSDMs reduce inferential issues in estimates of observed environmental effects, especially when it comes to dealing with presence–absence data. In contrast to HUI et al. (2024), we only considered non-spatial JSDMs, in which the latent factors are assumed to be independent across sites, in order to study the effect of covariate omission *per se*. Our results extend the preliminary observations of WARTON et al. (2015) and WILKINSON et al. (2019), and highlight the robustness of JSDMs when facing unobserved predictors. We also argue, contrary to the conclusions of POGGIATO et al. (2021), that JSDMs can yield parameter estimates that significantly differ from those of classical SDMs for presence–absence data. Indeed, we show that JSDMs yield much less biased estimators than SDMs in the presence of important omitted covariates, thus underscoring their interest in community ecology. Globally, our work is intended as a methodological contribution to a better understanding of the ability of JSDMs to estimate species–environment relationships and their robustness to misspecification in species distribution modeling in ecology. Our results remain relevant for binary generalized linear latent variable models (GLLVMs) employed in other scientific disciplines.

2. Results

The steps in this paper are threefold: (1) we use simulated datasets to assess the inferential impact of important missing covariates on single-species SDMs; (2) we apply JSDMs as a statistical mitigation strategy to efficiently handle an unobserved covariate; (3) we illustrate the previous simulated results by applying our strategy to several real ecological datasets.

2.1 Inferential impact assessment with SDMs

First of all, we assessed the statistical behavior of single-species SDMs with an important omitted covariate. Within a predefined simulation framework, we generated multiple synthetic datasets across five distinct scenarios, defined by the type of response data and the link function employed: (1) continuous data (identity link), (2) abundance data (log link), and (3-5) presence–absence data (logit, probit, and complementary log-log links). For each scenario, 1000 datasets were simulated, each comprising two independent covariates (X_1 and X_2) and the corresponding ecological response (Y). In the data-generating processes (DGPs), the response followed a probability distribution (Normal, Poisson, or Bernoulli) whose mean parameter was determined by a linear predictor combining the two covariates through the appropriate link function. All the synthetic datasets were then analyzed with an SDM structurally similar to the DGP, with one key difference — only X_1 was observed while X_2 remained unobserved, thereby generating a misspecified SDM. Thus, only two parameters were systematically estimated: the intercept ($\hat{\alpha}$) and X_1's regression coefficient ($\hat{\beta}$). We investigated several SDM implementations: (i) generalized linear models (GLM); (ii) generalized linear mixed models with a site-level random effect (GLMM); (iii) robust GLMs (GLM-R); (iv) generalized additive mixed models (GAMM); and (v) quasi-likelihood GLMs (GLM-QL). These regression methods were selected within both frequentist and Bayesian frameworks (Tab. 2).

The results obtained regarding the effect of an independent unobserved covariate on SDMs were generally consistent with theoretical predictions (Fig. 1). Complete plots for all investigated SDMs are provided in the Supplementary Material (see Sect. D in the SM, specifically Fig. S5 and Fig. S6). The Gaussian SDMs produced unbiased estimates of both the intercept $\hat{\alpha}$ and the regression coefficient $\hat{\beta}$, with median bias values centered on zero across all methods. In contrast, the Poisson SDMs exhibited a significant positive shift in the intercept $\hat{\alpha}$ when no site-level random effect (SLRE) was included, although the regression coefficient $\hat{\beta}$ remained unbiased across all formulations. Introducing an SLRE effectively eliminated the bias in $\hat{\alpha}$, bringing estimate biases back to zero. For Bernoulli SDMs, both $\hat{\alpha}$ and $\hat{\beta}$ were biased, with $\hat{\beta}$ and $\hat{\alpha}$ exhibiting negative bias (attenuation) across most methods. The bias amplitude depended on the link function, increasing from logit to probit to complementary log-log (cloglog). In particular, Bernoulli probit SDMs yielded biases close to the theoretical value ($1 - \frac{1}{\sqrt{2}} \approx 0.293$, see section A in the SM). In those cases of binary data, introducing an SLRE produced differing effects. The SLRE did not alter the results for most of the frequentist methods (Fig. 1), except for SDM.4, under which $\hat{\alpha}$ and $\hat{\beta}$ became unbiased; however, this improvement came with very large root mean squared random errors (RMSREs) and coverage rates much higher than expected (Tab. S3). By contrast, the Bayesian methods incorporating an SLRE showed heterogeneous

behavior. For example, SDM.6 remained biased even with the SLRE, while SDM.8 produced less biased estimates (Fig. 1), though the remaining bias was still significant and accompanied by very large RMSREs and substantial coverage rate issues (Tab. S3). Notably, neither quasi-likelihood nor robust regression methods were able to prevent the bias; indeed, they consistently exhibited significant bias across all scenarios, even in the Gaussian case, where other methods performed well.

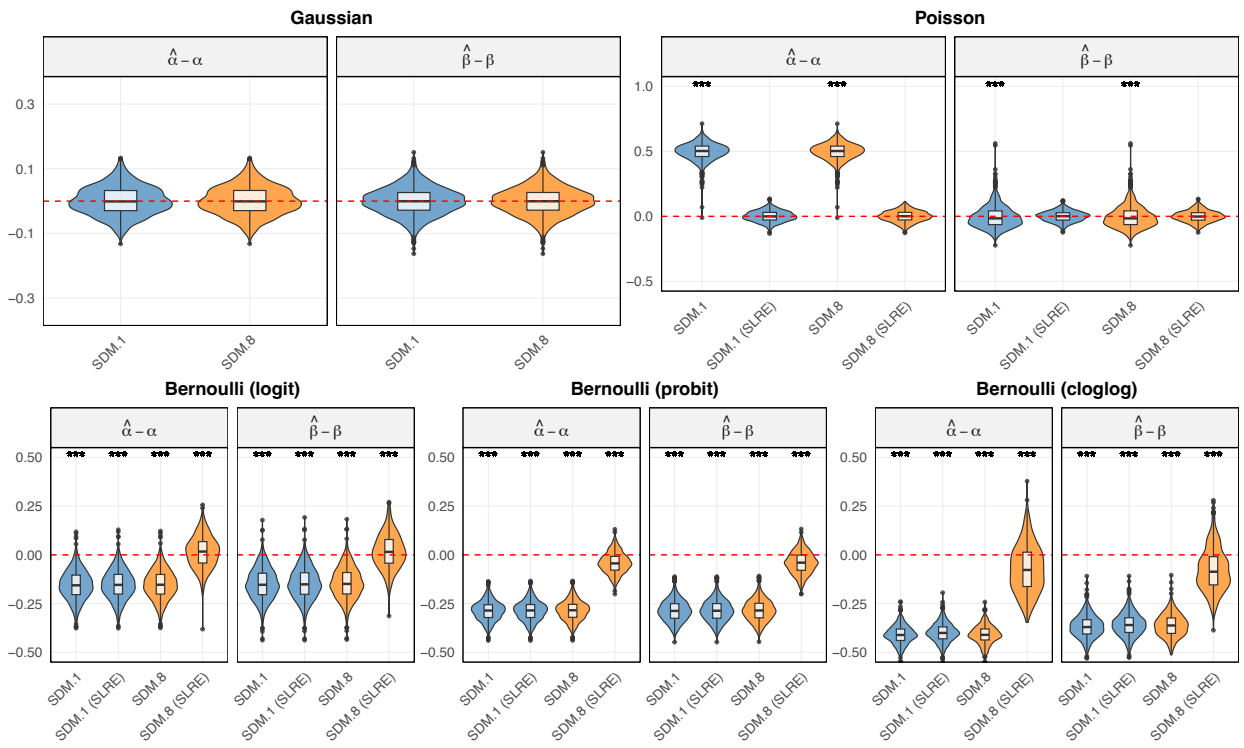

Figure 1: Violin- and box-plots of the estimates of the intercept $\hat{\alpha}$ (upper subplots) and the regression coefficient $\hat{\beta}$ (lower subplots), minus their respective true values. Only the results for SDM.1 and SDM8 (without or with an SLRE) are herein plotted (see Sect. D in the SM for complete results). Each subplot column displays the results for the five data-generation processes (Gaussian, Poisson, and Bernoulli with three different link functions). Colors indicate the type of inference of the SDMs employed (blue for the frequentist method and orange for the Bayesian method). The red dotted line marks zero bias.

2.2 Mitigation strategy with JSDMs

To address and mitigate the inferential consequences of unobserved covariates on fixed effects in the case of presence–absence data, we tested an alternative strategy that leverages shared information among between-species residuals within a community using JSDMs. A simulation framework was again employed, this time involving a community of ten species. We generated 1000 synthetic datasets of presence–absence records through a DGP akin to a Bernoulli GLM with a probit link. As in the SDM simulations, two independent covariates (X_1 and X_2) were considered in the data generation. Regarding the true parameter values of the intercept and slopes, six scenarios (labeled S.1 to S.6) were designed, each defining a different context for species prevalence (see the Methods section). Simu-

lated data were then statistically analyzed using JSDMs, all belonging to a specific class: those that incorporate latent variables. Six different JSDM formulations were evaluated (Tab. 3), varying in their inference framework (frequentist or Bayesian) and constraints on the latent structure. We investigated three configurations (labeled C.1, C.2, and C.3) in terms of model misspecification. Under C.1, X_2 was treated as missing (only X_1 was provided to the model), and a single latent variable was included in the JSDM. Under C.2, X_2 was likewise treated as missing, but no latent variable was included. Under C.3, both X_1 and X_2 were observed, and a single latent variable was included. Consequently, for each species, under C.1 and C.2, the model estimated intercept ($\hat{\alpha}$) and effect of X_1 ($\hat{\beta}$), whereas under C.3, the effect of X_2 ($\hat{\gamma}$) was additionally estimated.

According to configuration C.1's results, using a JSDM to mitigate the impact of a missing predictor led to a successful global reduction in bias, even completely canceling it in some cases (Fig. 2). The six JSDM formulations investigated overall visually recovered, for all scenarios, the true values of both the intercept and X_1's regression coefficient, as they provided an average absolute median bias of 0.016 (Fig. S11, S18, S25, S32, S39, and S46 in the SM). Nevertheless, the statistical significance of the bias depended on the scenario: greater true values of X_2'slope (γ) correlated with a larger proportion of significant biases for the intercept ($\hat{\alpha}$), and low or high species prevalence similarly led to more significant biases for the intercept ($\hat{\alpha}$). Scenarios leading to the most nonsignificant biases for X_1's slope ($\hat{\beta}$) were these assuming a zero intercept in data generation (namely, S.2 and S.4, hence corresponding to small covariate effects and sufficient average species prevalences). Coverage rates were particularly close to the expected value for almost all models (Fig. S12, S19, S26, S33, S40, S47 in the SM). Performance metrics of Bayesian JSDMs were equivalent across scenarios — RMSEs and RMSREs always remained under 0.25. However, for the only frequentist JSDM (JSDM.1), we observed significantly higher RMSEs and RMSREs, especially for scenarios S.3 to S.6. In particular, JSDM.1 still exhibited minor inferential issues with scenario S.3. In general, JSDM.2, JSDM.3, and JSDM.4 yielded the best results for bias and other inferential performance metrics; conversely, JSDM.5 and JSDM.6 globally provided more biased estimates. Regarding the factor loadings ($\hat{\lambda}$), they mainly captured the true value of the missing covariate's coefficient (β), with still some exceptions stemming from formulation-specific JSDM constraints (Fig. S8, S13, S20, S27, S34, S41, S48 in the SM).

Since assumption C.2's results did not incorporate a latent structure, the resulting JSDMs statistically corresponded to stacked single-species SDMs, except for JSDM.3 and JSDM.4, which then defined multi-species distribution models (MSDMs). An MSDM differs from both single-species SDMs and full JSDMs because regression coefficients share information across species through a hierarchical random-effects structure, whereas the model includes no latent structure to capture re-

sidual between-species covariance (see Sect. B in the SM for details about MSDMs). Including this specification subsequently allowed us to verify that a latent variable was necessary to capture the residual correlation arising from the missing covariate. Indeed, configuration C.2's results showed similar inferential issues to those of the single-species SDMs, i.e., the expected bias in intercepts and regression coefficients induced by the missing covariate (Fig. S9, S14, S21, S28, S35, S42, S49 in the SM). Notably, median biases from scenario S.1 (S14) were equal to the median bias from the Bernoulli probit SDMs (1). The other metrics confirmed poor inferential performance in this case: coverage rate issues and inflated RMSEs and RMSREs (Fig. S15, S22, S29, S36, S43, S50 in the SM). Finally, configuration C.3's results demonstrated that incorporating a latent variable, even if all covariates are measured, did not impact the correct inference of the intercept and both covariate effects. Biases were visually very reduced or zero (Fig. S10, S16, S23, S30, S37, S44, S51 in the SM), while associated coverage rates were very close to the expected value, along with RMSEs and RMSREs consistently lower than 0.15. No major differences among the JSDM formulations were observed (Fig. S17, S24, S31, S38, S45, S52 in the SM).

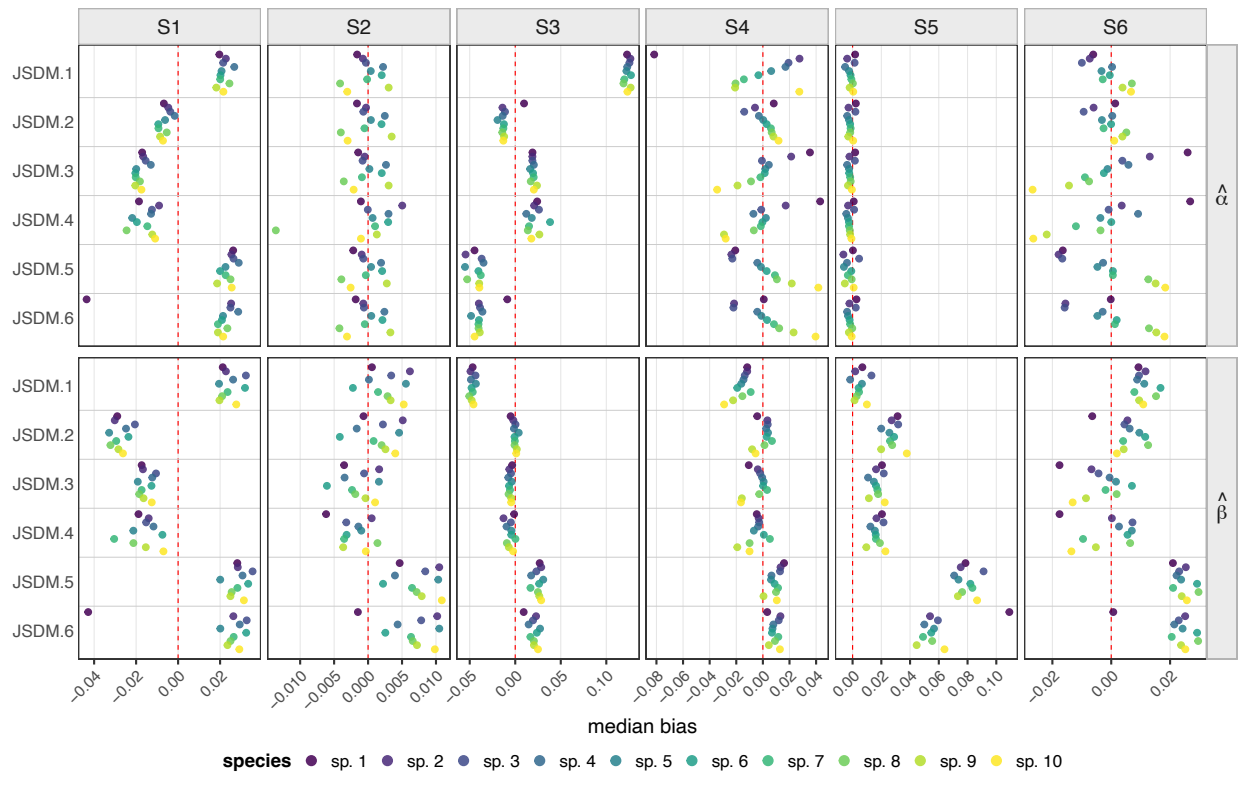

Figure 2: Median bias of the estimates of the intercept $\hat{\alpha}$ (upper panel) and the egression coefficient $\hat{\beta}$ (lower panel) across scenarios S.1 to S.6 (subplots from left to right) and under configuration C.1. Within each subplot, rows correspond to the six JSDM formulations (from JSDM.1 to JSDM.6, see Tab. 3), and each row displays estimates for the 10 species (distinguished by a 10-color scale). The red dotted line marks zero bias.

Furthermore, we evaluated how the number of species strongly related to the missing covariate influences the inference performance of the JSDMs. To this end, we simulated synthetic datasets analogous to the setting of scenario S.1, except that we varied the species' dependence on the unobserved covariate X_2. Among the 10 species, only the last m species was/were assumed to respond to X_2, while the first $10-m$ species did not. We repeated the simulations for $m \in \{1,\dots,10\}$ to assess how well the estimates $\hat{\alpha}$ and $\hat{\beta}$ recovered the true value of the intercept α and the regression coefficient β for species that actually depended on X_2. This investigation highlighted differences among the five JSDM formulations. Here we present the results from JSDM.1 (Fig. 3) while the results from the remaining JSDMs are shown in the Supplementary Material (see Sect. E.3 in the SM).

When only the last species responded to the unobserved covariate, the estimates of $\hat{\alpha}$ and $\hat{\beta}$ for this species exhibited bias due to the unobserved covariate for all JSDM formulations. However, when two species responded to the unobserved covariate, differences emerged between JSDM formulations. The Bayesian models (JSDM.2 to JSDM.5) provided unbiased estimates for $\hat{\alpha}$ and $\hat{\beta}$ (Fig. S55, S56, S57, S58), whereas the frequentist model (JSDM.1) exhibited significant bias for the first two species (Fig. 3), though the bias was smaller than when only a single species depended on the unobserved covariate. When more than two species responded to the unobserved covariate, the Bayesian methods yielded unbiased estimates overall. In contrast, the frequentist method (JSDM.1) retained apparent bias until at least four species responded to the missing covariate. This JSDM notably produced the most expected pattern: the bias caused by the missing covariate decreased progressively in amplitude as the number of responding species increased (Fig. 3).

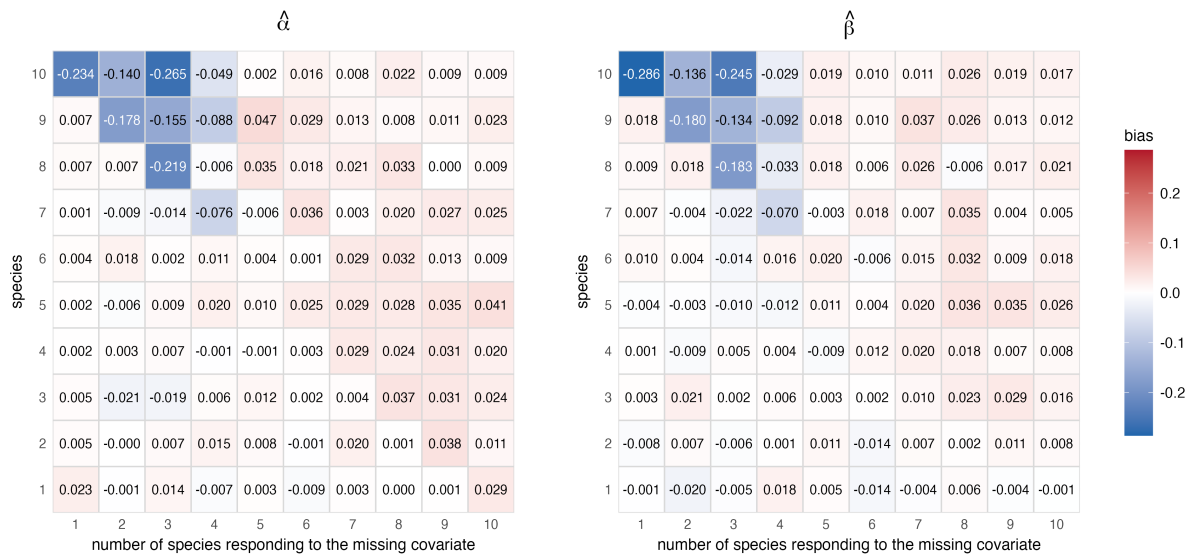

Figure 3: Heatmaps of the median bias of the estimated intercept $\hat{\alpha}$ (left subplot) and the estimated regression coefficient $\hat{\beta}$ (right subplot) with JSDM.1. In both subplots, each column represents a scenario where an increasing number of species (from only 1 to all 10 species) respond to the unobserved covariate X_2, while all species respond to the covariate X_1. Each row corresponds to one of the 10 species in the community. Cell colors indicate the magnitude of the median bias (blue = negative bias, red = positive bias), with values displayed numerically.

2.3 Application to real ecological datasets

To complement our simulation results, we analyzed six real ecological datasets. Each dataset comprised presence–absence records for multiple taxa along with environmental covariates measured at sampled sites. We fitted two models: an SDM and a JSDM with a single latent variable. As a consequence, we hypothesized that a single latent variable could account for residual between-species covariances, and we also hypothesized that could correspond to an important missing covariate. We thus interpreted the corresponding factor loadings as the species-specific regression coefficients for this hypothetical missing covariate. We obtained results that were overall consistent with what was predicted from a missing covariate. In the case reported herein (Fig. 4), we clearly see a shift of several estimates from SDM to JSDM (some differences are very strong and very significant). These differences align with the mathematical expectations regarding the impact of an unobserved covariate. Specifically, in Fig. 4, when the color of a point matches the background, the difference between SDM and JSDM estimates is consistent with the bias that is mathematically expected from a missing covariate (the so-called OVB) based on the expectation of the associated loading coefficient for that species. We obtained similar results for the five other datasets (see Sect. F in the SM), enabling explaining the shifts between JSDM and SDM estimates with the missing covariate framework. However, we also found shifts that cannot be explained by an unobserved covariate — points falling outside

the background-colored areas. We therefore conducted additional analyses to evaluate the potential causes of these differences, but we were unable to obtain satisfactory elucidation. In terms of predictive performance, JSDMs consistently outperformed SDMs, with notably stronger gains over datasets exhibiting significant shifts in coefficient estimates, whereas the improvement was more modest, yet still present, over datasets with shifts that were less/not significant.

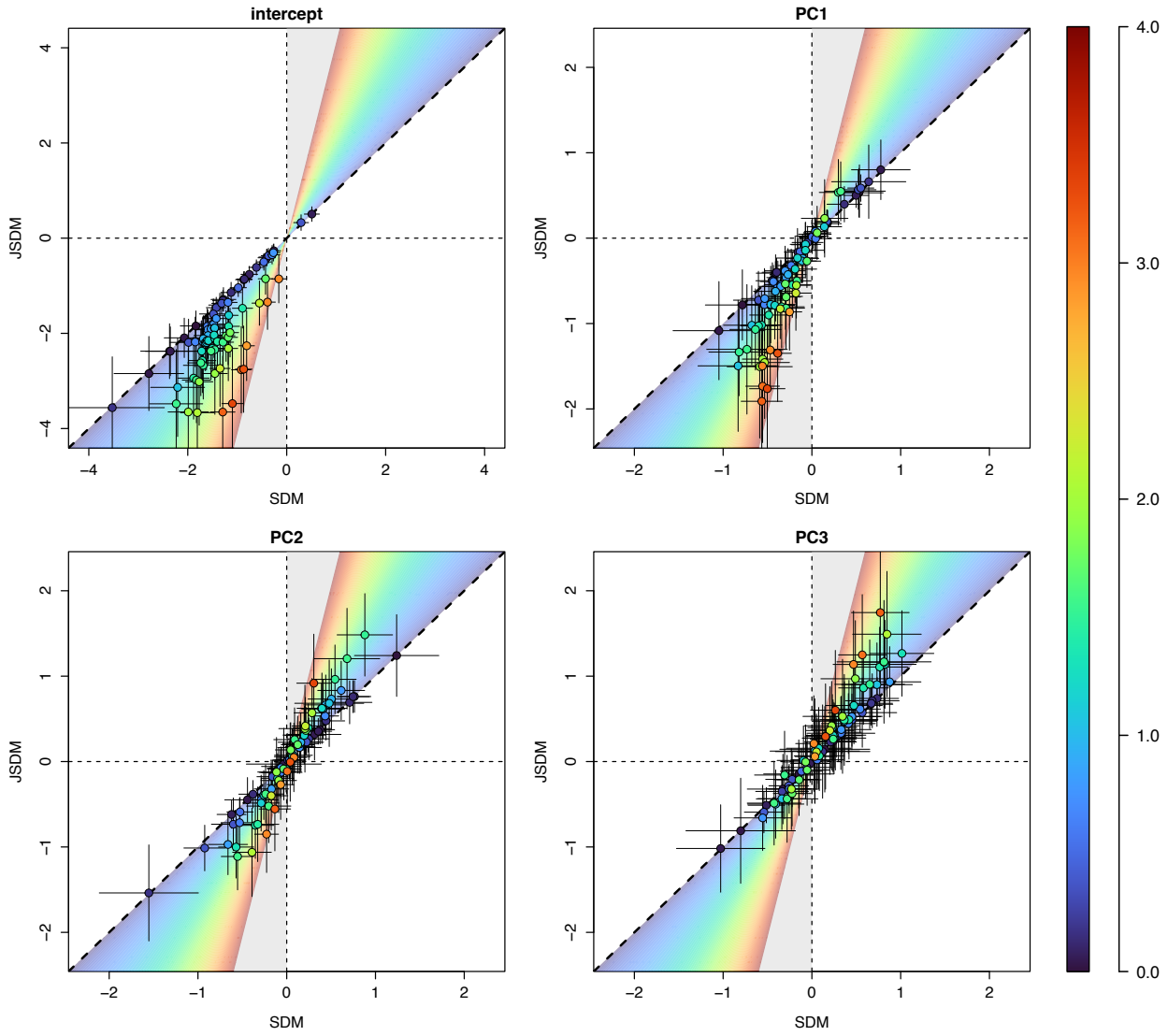

Figure 4: Comparison of SDM and JSDM estimates of the intercept and regression coefficients for dataset.5. Each subplot corresponds to the intercept or to one of the retained principal components (PCs). In each subplot, the horizontal axis shows the SDM estimates and the vertical axis the JSDM estimates, with each point representing a species. Black line segments indicate 95% confidence intervals for the SDM (horizontal) and JSDM (vertical) estimates. Dot colors indicate the estimated loadings of the single latent variable for the corresponding species. Background colors indicate the mathematical OVB, with the same color scale as for the dots. When the color of a point matches the background, the difference between the SDM and JSDM estimates is consistent with the bias expected from a missing covariate.

3. Discussion

In this study, we first investigated the impact of an unobserved predictor on statistical parametric species distribution models (SDMs), combining mathematical and simulation-based results with statistical analyses of real datasets. A missing covariate in an SDM may arise either because it is erroneously assumed to have no effect on species distribution, or because it is considered relevant but cannot be measured in practice (TIKHONOV et al., 2025). Although mathematical results for the bias caused by the omission of an important covariate in GLMs are already well-established (see Sect. A in the SM), to the best of our knowledge, these results are not acknowledged in ecology in the context of SDMs. Therefore, one primary aim of this article is to raise ecologists' and practitioners' awareness of the potential inferential effects of such an unobserved covariate, which could lead to misleading conclusions (OVASKAINEN et al., 2016). Especially in the case of presence–absence data — and even in the absence of spatial structure — omitting an important covariate can alter the estimates for the remaining environmental effects in terms of bias, accuracy, and coverage rate. As a mitigation strategy for multi-species presence–absence data, we found that random effects were not fully efficient, whereas joint species distribution models (JSDMs) exhibited a remarkable ability to capture unacknowledged predictors through among-species residuals when the omitted covariate had an effect on multiple species.

With respect to inferential problems in SDMs, when a species responds to a covariate that is unobserved in presence–absence models, the estimated intercept and the environmental effects become biasedly attenuated, as confirmed by both mathematical and simulation results. Conversely, in count data SDMs, the bias issue is manifest only in the intercept. Although introducing a site-level random effect (SLRE) in a Poisson model corrected all the inferential issues overall, this was not the case for Bernoulli presence–absence SDMs. Even in cases where bias was reduced, other inference performance metrics such as coverage rate and root mean squared random error (RMSRE) deteriorated markedly in our simulations. Consequently, adding SLRE cannot be relied upon to resolve all inferential problems for presence–absence data, although this option is often employed in ecology (e.g., HARRISON et al., 2018). This important finding directly challenges classical parametric SDM approaches, which are often applied to presence–absence records, emphasizing the need to explore alternative strategies that avoid such estimation issues.

Regarding the performance of JSDMs, we obtained, on the whole, substantially less bias with JSDMs than with the corresponding SDMs. Most performance metrics were comparable across our different JSDM formulations, which was somewhat expected as the JSDMs were based on the same overall assumption: they all included a single latent variable used to model between-species residual

covariances not accounted for by environmental effects. However, when scrutinizing bias significance, we noticed that JSDM formulations without positivity or nullity constraints on some of the loadings (JSDM.2 and JSDM.3) yielded overall unbiased estimates of the intercept and regression coefficient. In contrast, the other formulations continued to yield estimators with significant — though visually small — bias (JSDM.1, JSDM.4, JSDM.5). These results are in accordance with the statements of WARTON et al. (2015) and the results of WILKINSON et al. (2019) (Appendix S5), who found that a presence–absence JSDM with a non-spatial latent factor caused no apparent bias for environmental effects. We complemented their findings by systematically demonstrating that the latent factor can resolve inference issues arising from unobserved covariates in the presence–absence SDMs. The ability of JSDMs to mitigate OVB by leveraging shared information across species depended primarily on the number of species and on the statistical inference method (frequentist or Bayesian). Additionally, testing MSDMs (i.e., JSDMs with no latent variable) clearly demonstrated the necessity of accounting for residual interspecific covariance via a latent variable to successfully reduce bias in environmental effects, since mere information sharing between species within a hierarchical formulation is insufficient.

We also observed in our simulations that the latent factor closely matched the omitted covariate and that the associated loadings were consistent with the omitted regression parameter, as expected. Since the omitted parameter drives the bias magnitude in SDMs, as revealed by mathematical and simulation results, if one interprets the factor loadings as the missing regression parameters for each species, particular attention must be paid to the absolute value of these loadings for better estimation of the regression coefficients of the environmental covariates. We therefore caution against interpreting the results of the latent structure of JSDMs solely in "relative" terms of residual correlations between species — as is often practiced in ecology — because the "absolute" information carried by the associated variances is lost. We further warn against restricting the range of values that loadings can take: an example of such restrictions occurs in the model used by TOBLER et al. (2019) (loadings were restricted to the interval $[-1, 1]$). Our simulations show that such restrictions incur important persistent bias in the estimation of the JSDM parameters for the remaining explanatory variables (see Sect. E.2 in the SM, especially Fig. S53).

When analyzing real presence–absence datasets, we observed substantial differences between estimates from single-species SDMs and those from the associated JSDM. These shifts were generally consistent with the expected bias expected from a missing predictor in parametric SDMs, based on the species loading interpreted as the coefficient of the omitted covariate. When using datasets with a small number of sites (fewer than 500), these differences were accompanied by wide 95% credibility intervals (CIs), which often made them non-significant. However, larger datasets produced

more pronounced and significant shifts. Interestingly, for species showing important discrepancies between SDM and JSDM estimates, the JSDM tended to yield wider CIs for the covariate effects. Yet, when no such discrepancies were present, the JSDM produced narrower CIs than the corresponding single-species SDMs. This latter finding appears somewhat intuitive, as JSDMs can exploit shared information across species. Our conclusions, therefore, contradict those of POGGIATO et al. (2021), who argued that JSDMs, compared to SDMs, usually do not improve estimates.

Our results were consistent with those of NORBERG et al. (2019) — although using different metrics — who reported superior predictive performance of JSDMs on presence–absence records compared to nonparametric and parametric SDMs, and aligned with the joint predictive performance reported in WILKINSON et al. (2023). More specifically, we found that JSDMs performed slightly better than SDMs for communities whose coefficient estimates across species were close between the two approaches, and markedly outperformed SDMs for communities whose environmental effect estimates across species diverged significantly between the two approaches (according to our missing covariate interpretative framework). These results are consistent with our hypothesis that the ability of JSDMs to account for such a missing covariate would not only correct inferential problems in parametric SDMs (see the previous paragraph) but would also improve the predictive capacity of JSDMs relative to both nonparametric and parametric SDMs.

Nonetheless, the present study has several limitations, and challenges remain for future work. First, in the SDM literature, many models employed are nonparametric, such as machine learning or deep learning frameworks (e.g., ELITH et al., 2006), whereas our work primarily used parametric statistical models. Still, our restricted focus on parametric methods is justified by the emphasis on model inference properties and statistical inference, which is not possible with nonparametric approaches. Second, we consistently assumed that the missing covariate was spatially unstructured and independent of the other observed environmental covariates. This assumption can be somewhat debatable, as covariates are actually often (spatially auto)correlated. Yet, the issues of spatial or community confounding in JSDMs have been explored, highlighting the potential inferential risks these phenomena may pose (e.g., HUI et al., 2024; VAN EE et al., 2022). As statistical complications may arise in real data beyond our ideal case of independent missing predictors, a follow-up simulation study accounting for correlations between observed and unobserved covariates and explicitly acknowledging spatial structures would be welcomed to check whether our results can extend to more spatially-realistic settings. Third, our conclusions might, to some degree, depend on the specific simulation framework used, although we assessed different scenarios of species prevalence and dependence on the unobserved predictor. Specifically, for fewer sites and/or rarer species and/or communities with species richness greater than 10, it might be necessary to include more species sharing the missing covariate

to remove the bias in environmental effect estimates. Fourth, throughout the study, we focused exclusively on missing environmental predictors, although JSDMs are also known to capture symmetric biotic interactions (POGGIATO et al., 2021; ZURELL et al., 2018). It is unclear to us whether symmetric biotic interactions would generate similar inferential problems in SDMs, which JSDMs would allow us to correct, as found with an omitted covariate.

In conclusion, this study demonstrates the existence of bias in parametric presence–absence SDMs when (at least) one important covariate remains unmeasured. When we have presence–absence records for all species of the community, and provided that several species within the community share this missing covariate and that the covariate is independent of the other environmental predictors, JSDMs are then able to resolve the inferential issues faced by SDMs. In light of our inferential findings, together with our results on real data indicating better predictive capacity of JSDMs, we recommend that ecologists and practitioners targeting species distribution estimations systematically fit JSDMs as they allow to correctly estimate the observed environmental effects when an important missing covariate is shared by several species. Ecologists interested in detecting situations where important covariates are missing could also fit corresponding single-species SDMs and then compare the resulting environmental effects with those of the JSDM. Discrepancies in the environmental effects would suggest the existence of unacknowledged environmental predictors or perhaps specific kinds of biotic interactions (KISSLING et al., 2012). Hence, we emphasize the primary role of JSDMs in improving the inference of species–environment relationships over their *potential* ability to capture biotic interactions. Our investigations indeed offered valuable insights into the statistical interest of JSDMs, which we encourage researchers to use more frequently when studying ecological community presence–absence data. Ultimately, our results remain valid in other scientific disciplines where important covariates are suspected to be omitted in binary GL(M)Ms, in which case generalized linear latent variable models (GLLVMs) could properly correct inference in situations where different entities might share the same omitted important covariate.

4. Methods

All computations were carried out on the `R` software version 4.5.1 (R CORE TEAM, 2025). References to the `R` packages employed are given throughout the text.

4.1 Species distribution models

To assess the impact of a missing covariate on SDMs and illustrate the corresponding mathematical results for the OVB, we used a simulation framework that enabled us to control the true parameter values used to generate the data. We considered three types of data: continuous (responses in $\mathbb{R}$), count (responses in $\mathbb{N}$), and binary (responses in $\{0,1\}$). All three cases are common in ecological datasets: plant surface coverage measurements are continuous data, the number of animals at a site is count data, and presence–absence records of a species at a location are binary data. The simulation included 1000 replicates (i.e., different synthetic datasets) to approximate asymptotic behavior as closely as possible.

We focused on a set of $n = 1000$ distinct sites labeled $i = 1,\dots,n$. The sites were sampled only once. Let Y denote the response variable and Y_i denote its value at site i. We considered two environmental covariates X_1 and X_2, whose values at site i were respectively denoted $X_{1,i}$ and $X_{2,i}$. The two covariates were assumed to be independent of each other, verifying

$$\forall i \in \{1,\dots,n\}, \qquad X_{1,i} \sim \mathsf{Normal}(0,\sigma^2_{X_1}) \qquad \text{and} \qquad X_{2,i} \sim \mathsf{Normal}(0,\sigma^2_{X_2}) \tag{1}$$

where $\sigma^2_{X_1}$ and $\sigma^2_{X_2}$ are fixed variances (WILKINSON et al., 2019). In our simulation, the response variable was then drawn from a distribution depending on its type (continuous, count, binary), as follows,

$$\forall i \in \{1,\dots,n\}, \quad \begin{cases} \begin{cases} Y_i \mid \{X_{1,i},X_{2,i}\} \sim \mathsf{Normal}(z_i,\sigma^2) \\ z_i = \alpha + \beta X_{1,i} + \gamma X_{2,i} \end{cases} & \text{for continuous data,} \\ \begin{cases} Y_i \mid \{X_{1,i},X_{2,i}\} \sim \mathsf{Poisson}\,(z_i) \\ \log(z_i) = \alpha + \beta X_{1,i} + \gamma X_{2,i} \end{cases} & \text{for count data,} \\ \begin{cases} Y_i \mid \{X_{1,i},X_{2,i}\} \sim \mathsf{Bernoulli}\,(z_i) \\ f(z_i) = \alpha + \beta X_{1,i} + \gamma X_{2,i} \end{cases} & \text{for binary data,} \end{cases} \tag{2}$$

where $z_i = \mathbb{E}[Y_i \mid \{X_{1,i},X_{2,i}\}]$, σ^2 is the variance of the Normal distribution, and f is either the probit,

the logit, or the complementary-log-log (cloglog) function,

$$\forall p \in]0,1[, \qquad \begin{cases} \text{probit}(p) = \Phi^{-1}(p) \\ \text{logit}(p) = \log\left(\dfrac{p}{1-p}\right) \\ \text{cloglog}(p) = \log\left(-\log(1-p)\right) \end{cases} \tag{3}$$

with Φ the cumulative distribution function of the standard Normal distribution (i.e., $\text{Normal}(0,1)$). For parameter values, we chose $\alpha = \beta = \gamma = 1$ and $\sigma^2_{X_1} = \sigma^2_{X_2} = 1$, along with $\sigma^2 = 1$ for the variance of the Normal distribution. The chosen parameter values for the coefficients corresponded to intermediate realistic covariate effects. At each replicate, the covariates and the response variable were resampled from a distinct seed.

We then statistically analyzed the simulated data under a misspecified scenario in which only the response Y and the covariate X_1 were observed at each sampled site, while X_2 was unobserved (without presuming the reason for its unobservedness). We evaluated five SDM implementations: (i) GLMs; (ii) GLMMs featuring a site-level (spatially unstructured) random effect (SRLE); (iii) a robust GLM (denoted GLM-R); (iv) a generalized additive mixed model (GAMM); and (v) a quasi-likelihood GLM (denoted GLM-QL). Specifically, the misspecified models for implementations (i) and (ii) are expressed as follows:

(i) GLM:

$$\forall i \in \{1,\ldots,n\}, \qquad \begin{cases} Y_i \mid X_{1,i} \sim \text{Normal}(\tilde{z}_i, \sigma^2) \text{ or } \text{Poisson}(\tilde{z}_i) \text{ or } \text{Bernoulli}(\tilde{z}_i) \\ f(\tilde{z}_i) = \hat{\alpha} + \hat{\beta}\, X_{1,i} \end{cases} \tag{4}$$

where $\tilde{z}_i = \mathbb{E}[Y_i \mid X_{1,i}]$, f denotes the link function (among identity, log, probit, logit, cloglog), and σ^2 is the variance of the Normal distribution,

(ii) GLMM with a site-level random effect (SRLE):

$$\forall i \in \{1,\ldots,n\}, \qquad \begin{cases} Y_i \mid X_{1,i} \sim \text{Poisson}(\tilde{z}_i) \text{ or } \text{Bernoulli}(\tilde{z}_i) \\ f(\tilde{z}_i) = \hat{\alpha} + \hat{\beta}\, X_{1,i} + \varepsilon_i \\ \varepsilon_i \overset{\text{i.i.d.}}{\sim} \text{Normal}\left(0, \sigma^2\right) \end{cases} \tag{5}$$

where $\tilde{z}_i = \mathbb{E}[Y_i \mid X_{1,i}]$, f denotes the link function (among log, probit, logit, cloglog), and σ^2 is the variance of the SLRE.

We therefore estimated the intercept $\hat{\alpha}$ and the single regression coefficient $\hat{\beta}$ associated with the observed covariate X_1. We assessed the behavior of eleven regression methods, across various `R`

packages, as listed in Tab. 2 (SDM.1 to SDM.11). Models SDM.7 and SDM.8 cannot make use of any built-in function for fitting a Bayesian GL(M)M, and thus require a customized definition of regression models. To run models based on NIMBLE or JAGS, we employed the R package `runMCMCbtadjust`; this enabled us to monitor and control MCMC convergence and the number of effective values (MCMC iteration autocorrelation) (GOSSELIN, 2024).

Table 2: List of regression methods from R packages employed for fitting SDMs.

model label	built-in function	R package	inference	implementation(s)	description	reference(s)
SDM.1	`glmer`	`lme4`	frequentist	GLM | GLMM	Laplace approximation and adaptive Gauss–Hermite quadrature	Bates et al. (2015)
SDM.2	`glmmTMB`	`glmmTMB`	frequentist	GLM | GLMM	Laplace approximation via TMB using automatic differentiation	Kristensen et al. (2016) Thygesen et al. (2017)
SDM.3	`spaMM`	`spaMM`	frequentist	GLM | GLMM	H-likelihood framework with Laplace approximations for random effects	Rousset & Ferdy (2014) Lee & Nelder (2001)
SDM.4	`GLMMadaptive`	`GLMMadaptive`	frequentist	GLM | GLMM	maximum likelihood via adaptive Gauss–Hermite quadrature (default) or Laplace approximation	Rizopoulos (2025)
SDM.5	`brm`	`brms`	Bayesian	GLM | GLMM	Bayesian MCMC via Stan (HMC/NUTS) with weakly to strongly informative priors	Bürkner (2018)
SDM.6	`inla`	`INLA`	Bayesian	GLM | GLMM	integrated nested Laplace approximation (INLA) for latent Gaussian models	Lindgren et al. (2011) Rue et al. (2009)
SDM.7	–	`nimble`	Bayesian	GLM | GLMM	Bayesian MCMC via NIMBLE	De Valpine et al. (2017)
SDM.8	–	`runjags`	Bayesian	GLM | GLMM	Bayesian MCMC via JAGS	Plummer (2003) Denwood (2016)
SDM.9	`glmRob`	`robust`	frequentist	GLM-R	robust M-estimation via IRLS with sandwich (Huber–White) covariance for SEs	Wang et al. (2006)
SDM.10	`gamm`	`mgcv`	frequentist	GAMM	penalized likelihood for smooths with smoothing selection	Wood (2017)
SDM.11	`glm`	`stats`	frequentist	GLM-QL	Laplace approximation and adaptive Gauss–Hermite quadrature	

4.2 Joint species distribution models

We investigated *joint species distribution models* (JSDMs) as a strategy to address the inferential consequences of unobserved covariates on fixed-effect estimates for presence–absence data. JSDMs are multivariate statistical models that allow inference of interspecific residual covariations not explained by observed environmental predictors, typically driven by biotic interactions or unmeasured covariates influencing species' distributions. To this end, we conducted a simulation study comprising six data-generating scenarios. Each scenario involved 1000 replicates, yielding 1000 datasets of presence–absence records for a community of $S = 10$ species. At each replicate, the data were simulated independently from the same Bernoulli GLMM with a probit link, as follows,

$$\forall s \in \{1,\dots,S\}, \quad \forall i \in \{1,\dots,n\}, \qquad \begin{cases} Y_i^{[s]} \mid \{X_{1,i}, X_{2,i}\} \sim \mathsf{Bernoulli}\left(z_i^{[s]}\right) \\ \Phi^{-1}\left(z_i^{[s]}\right) = \alpha^{[s]} + \beta^{[s]} X_{1,i} + \gamma^{[s]} X_{2,i} \end{cases} \tag{6}$$

where $z_i^{[s]} = \mathbb{E}\left[Y_i^{[s]} \mid X_{1,i}, X_{2,i}\right]$ denotes the conditional expectation of the response, $\alpha^{[s]}$ is the intercept for species s, and $\beta^{[s]}$ and $\gamma^{[s]}$ are the regression coefficients of species s associated with covariates X_1 and X_2, respectively.

The six data generating scenarios (labeled S.1 to S.6), mimicking different ecological contexts of species prevalence over space: S.1, a community where all species have common, while responding uniformly and strongly to both covariates ($\beta_0 = \beta_1 = \beta_2 = 1$); S.2, a community of moderately prevalent species (50%) with weaker responses to both covariates ($\beta_0 = 0$, $\beta_1 = \beta_2 = 0.5$); S.3, a community dominated by rare species, with the same moderate covariate effects as S.2 ($\beta_0 = -1.5$, $\beta_1 = \beta_2 = 0.5$); S.4, a heterogeneous community in which species range from rare to common (intercepts evenly spaced from -1.5 to 1.5); S.5, a community at intermediate prevalence where covariate X_1 exerts a strong effect and covariate X_2 a weak effect ($\beta_1 = 1.5$, $\beta_2 = 0.2$); and S.6, a community spanning rare to common species that respond positively to one covariate but negatively to another ($\beta_1 = 1$, $\beta_2 = -0.5$). Parameter values for simulation are given in Tab. S2. We did not include any additional interspecific residual correlations in the simulated responses.

For the data statistical analysis, the covariate X_1 was assumed to be always measured and included in the statistical model. Concerning the covariate X_2, we focused on three distinct configurations: C.1 assumed that X_2 was unobserved and the JSDM implemented with a single latent variable; C.2 assumed that X_2 was unobserved and the JSDM implemented without latent variable; C.3 assumed that X_2 was measured and the JSDM implemented with a single latent variable. We employed the latent factor approach to implement JSDMs with a single latent variable (see Sect. B in the SM for

a more detailed description). Therefore, for a given species s and at a given site i, the JSDM under configuration C.1 was formulated as follows,

$$\forall s \in \{1,\dots,S\}, \quad \forall i \in \{1,\dots,n\}, \qquad \begin{cases} Y_i^{[s]} \mid X_{1,i} \sim \text{Bernoulli}\left(\tilde{z}_i^{[s]}\right) \\ \Phi^{-1}\left(\tilde{z}_i^{[s]}\right) = \hat{\alpha}^{[s]} + \hat{\beta}^{[s]} X_{1,i} + \hat{\lambda}^{[s]} \hat{W}_i \end{cases} \tag{7}$$

where $\tilde{z}_i = \mathbb{E}\left[Y_i^{[s]} \mid X_{1,i}\right]$, $\hat{\alpha}^{[s]}$ is the intercept of species s, $\hat{\beta}^{[s]}$ is the regression coefficient of species s with respect to the covariate X_1, and $\hat{\lambda}^{[s]}$ is the factor loading of species s with respect to the latent variable $\hat{W}$. We emphasize that the latent variable $\hat{W}$ and its species-species factor loadings are also inferred by the model, with implementation-specific constraints. Under configuration C.2, fitting JSDMs without latent factor reduced to imposing $\hat{\lambda}^{[s]} = 0$ for all species, thereby removing the term $\hat{\lambda}^{[s]}\hat{W}_i$ in Eq. 7. Under configuration C.3, the fitted JSDMs followed Eq. 7 with an additional the term $\hat{\gamma}^{[s]} X_{2,i}$ in the linear predictor.

We considered six JSDM implementations (Tab. 3), whose technical details are provided in the Supplementary Material (Sect. C in the SM). The first implementation (JSDM.1) uses the `R` package `gllvm`, leveraging the `R` package `TMB`. The second implementation (JSDM.2) is based on the `R` package `jSDM`, while the third (JSDM.3) employs the widely used `R` package `Hmsc`, both within a Bayesian framework. The other three Bayesian implementations (JSDM.4, JSDM.5, and JSDM.6) were customized using NIMBLE or JAGS, interfaced through the `R` packages `nimble` and `runjags`, respectively, along with the `R` package `runMCMCbtadjust` used to monitor and control MCMC convergence and the number of effective values (MCMC iteration autocorrelation) (GOSSELIN, 2024).

Note that setting the number of latent factor to zero in a JSDM yielded independent single-species SDMs fitted together, except for JSDM.3 and JSDM.4 as they used a hierarchical structure for the regression coefficients. Thereby, JSDM.3 or JSDM.4 without latent variable consisted of a *multi-species distribution model* (MSDM). Specifically, an MSDM did not account for residual between-species covariances through a latent structure, but assumed a hierarchical structure in which the regression coefficients were treated as random effects with estimated mean vector and variance-covariance matrix (allowing shared information between species). See POGGIATO et al. (2021) and Sect. B in the Supplementary Material for details about MSDMs.

Table 3: List of regression methods from R packages employed for fitting JSDMs.

model label	R package	regression coefficients	constraint on latent variables	constraint on loading matrix	inference	reference(s)
JSDM.1	`gllvm`	fixed effects	Ø	unit diagonal and zero upper diagonal	frequentist	NIKU et al. (2019) KRISTENSEN et al. (2016)
JSDM.2	`jSDM`	fixed effects	Ø	positive diagonal and zero upper diagonal	Bayesian	VIEILLEDENT & CLÉMENT (2025)
JSDM.3	`Hmsc`	random effects	unit variance	Ø	Bayesian	TIKHONOV et al. (2020) OVASKAINEN & ABREGO (2020)
JSDM.4	`nimble`	random effects	unit variance	Ø	Bayesian	DE VALPINE et al. (2017)
JSDM.5	`runjags`	fixed effects	unit variance	positive diagonal and zero upper diagonal	Bayesian	PLUMMER (2003) DENWOOD (2016)
JSDM.6	`runjags`	fixed effects	Ø	unit diagonal and zero upper diagonal	Bayesian	PLUMMER (2003) DENWOOD (2016)

Additionally, to evaluate how the number of species responding to the missing covariate would influence the performance of the JSDMs, we simulated synthetic datasets analogous to the previous setting, except that we varied species' dependence on the unobserved covariate X_2. Specifically, among the $S = 10$ species, only the last m species were assumed to respond to X_2, while the remaining $10 - m$ species did not. We repeated the simulations for $m \in \{1, \dots, 10\}$ to assess how well the estimates $\hat{\alpha}$ and $\hat{\beta}$ recovered the true value of the intercept α and the regression coefficient β for species that actually did depend on X_2. That is, presence–absence data were generated as follows,

$$\forall i \in \{1, \dots, n\}, \quad \begin{cases} Y_i^{[1]} \mid \{X_{1,i}, X_{2,i}\} \sim \text{Bernoulli}\left(\Phi\left(\alpha^{[1]} + \beta^{[1]} X_{1,i}\right)\right) \\ \qquad \vdots \\ Y_i^{[S-m]} \mid \{X_{1,i}, X_{2,i}\} \sim \text{Bernoulli}\left(\Phi\left(\alpha^{[S-m]} + \beta^{[S-m]} X_{1,i}\right)\right) \\ Y_i^{[S-m+1]} \mid X_{1,i} \sim \text{Bernoulli}\left(\Phi\left(\alpha^{[S-m+1]} + \beta^{[S-m+1]} X_{1,i} + \gamma^{[S-m+1]} X_{2,i}\right)\right) \\ \qquad \vdots \\ Y_i^{[S]} \mid X_{1,i} \sim \text{Bernoulli}\left(\Phi\left(\alpha^{[S]} + \beta^{[S]} X_{1,i} + \gamma^{[S]} X_{2,i}\right)\right) \end{cases} \tag{8}$$

where $\alpha^{[1]} = \cdots = \alpha^{[S]} = \beta^{[1]} = \cdots = \beta^{[S]} = \gamma^{[S-m+1]} = \cdots = \gamma^{[S]} = 1$ and $\gamma^{[1]} = \cdots = \gamma^{[S-m]} = 0$. Statistical analyses were also systematically performed on the five JSDM formulations from Tab. 3.

4.3 Metrics for post-inference analysis

Post-inference analyses were performed on all the results obtained from the statistical models (both SDMs and JSDMs). We systematically calculated:

- the median bias: $\text{Bias}\left[\hat{\theta}\right] = \text{median}\{\hat{\theta}_1, \dots, \hat{\theta}_N\} - \theta$
- the coverage rate (CR): $\text{CR}\left[\hat{\theta}\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left\{\theta \in \left[\, \hat{\theta}_i - 1.96\, \widehat{\text{SE}}\left[\hat{\theta}_i\right],\ \hat{\theta}_i + 1.96\, \widehat{\text{SE}}\left[\hat{\theta}_i\right] \,\right]\right\}$,
- the root mean square error (RMSE): $\text{RMSE}\left[\hat{\theta}\right] = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\hat{\theta}_i - \theta\right)^2}$,
- the root mean square random error: $\text{RMSRE}\left[\hat{\theta}\right] = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\left(\hat{\theta}_i - \theta\right)^2 + \widehat{\text{SE}}\left[\hat{\theta}_i\right]^2\right)}$,

where $\hat{\theta}_1, \dots, \hat{\theta}_N$ are the estimates of any parameter whose true value is θ, and $\mathbb{1}\{\cdot\}$ is an indicator function. Bias significance was assessed with a p-value computed with the function `lmRob` from `R` package `robust` (WANG et al., 2006) from a robust linear model and is indicated with *** if $p \leq 0.001$, ** if $0.001 < p \leq 0.01$, * if $0.01 < p \leq 0.05$, and no symbol if $p > 0.05$.

4.4 Study of real ecological datasets

The six real multivariate ecological datasets come from previously published works (ABREGO et al., 2025; CHOLER, 2005; HARRIS, 2015; NABE-NIELSEN et al., 2025; OVASKAINEN et al., 2016; POLLOCK et al., 2014). Each dataset encompasses presence–absence records for several taxa, along with environmental covariates measured at the sampled sites. Additional methods and results are provided in the Supplementary Material (see Sect. F in the SM). For each dataset, a principal component analysis (PCA) was performed on the environmental covariates (NORBERG et al., 2019). When fewer than five covariates were available, all the corresponding principal components were retained; when more than five were available, only the five leading principal components (i.e., those with the largest eigenvalues) were retained. This procedure yielded an orthogonal set of independent environmental predictors and effectively mitigated multicollinearity among the covariates. Furthermore, species occurring at more than 95% of sites or fewer than 5% of sites were systematically excluded from the datasets. We systematically fitted SDM.8 (Tab. 2) and JSDM.6 (Tab. 3). Assessment and comparison of predictive capacity was performed according to NORBERG et al. (2019)'s splitting procedures between training and validation subsets, and by computing per-site log-likelihoods. Specifically, for each splitting procedure, after having inferred regression parameters on the training datasets, we computed each model's per-site log-likelihood on the validation datasets. Details about splitting procedures and likelihood computation are given in the Supplementary Material (see Sect. F in the SM).

Supplementary Material

Supplementary Material is provided to this article.

Data Availability

The `R` codes that support the investigations in this article are available in the following GitHub repository: `https://github/arthur-f-rossignol/article-004`. The real datasets are all already publicly available (see the corresponding references in the Supplementary Material).

Acknowledgments

A.F.R. is grateful to the French Ministry of Agriculture and the 'Corps des IPEF' for a PhD fellowship supporting this research. The INRAE MIGALE Bioinformat-

ics facility (https://migale.inrae.fr) and the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr), part of the 'Institut Français de Bioinformatique' (ANR-11-INBS-0013), are acknowledged for providing the necessary computational resources. Both authors thank Vicki MOORE for polishing the English and Giuliana BRAMBILLA for helpful comments on a previous version of the manuscript. The AI-based writing assistant Grammarly was also used to improve the manuscript's language and readability; the authors remain fully responsible for the content.

Author Contributions

A.F.R.: Methodology, Investigation, Formal analysis, Software, Visualization, Writing – original draft, Writing – review and editing. F.G.: Conceptualization, Supervision, Methodology, Validation, Writing – review and editing.

Competing Interests

The authors declare no competing interests.

References

ABREGO, N., NIITTYNEN, P., KEMPPINEN, J., & OVASKAINEN, O. (2025). Joint species-trait distribution modeling: The role of intraspecific trait variation in community assembly. *Ecology*, 106(9), e70174. https://doi.org/10.1002/ecy.70174

ARAÚJO, M. B., & GUISAN, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10), 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

AUSTIN, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200(1), 1–19. https://doi.org/10.1016/j.ecolmodel.2006.07.005

AUSTIN, M. P., & VAN NIEL, K. P. (2011). Improving species distribution models for climate change studies: Variable selection and scale: Species distribution models for climate change studies. *Journal of Biogeography*, 38(1), 1–8. https://doi.org/10.1111/j.1365-2699.2010.02416.x

BATES, D., MÄCHLER, M., BOLKER, B., & WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). https://doi.org/10.18637/jss.v067.i01

BLANCHET, F. G., CAZELLES, K., & GRAVEL, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, 23(7), 1050–1063. https://doi.org/10.1111/ele.13525

BOLKER, B. M., BROOKS, M. E., CLARK, C. J., GEANGE, S. W., POULSEN, J. R., STEVENS, M. H. H., & WHITE, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

BÜRKNER, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, 10(1), 395. https://doi.org/10.32614/RJ-2018-017

BYRNES, J. E. K., & DEE, L. E. (2025). Causal Inference With Observational Data and Unobserved Confounding Variables. *Ecology Letters*, 28(1), e70023. https://doi.org/10.1111/ele.70023

CANNER, P. L. (1991). Covariate adjustment of treatment effects in clinical trials. *Controlled Clinical Trials*, 12(3), 359–366. https://doi.org/10.1016/0197-2456(91)90016-F

CHOLER, P. (2005). Consistent Shifts in Alpine Plant Traits along a Mesotopographical Gradient. *Arctic, Antarctic, and Alpine Research*, 37(4), 444–453. https://doi.org/10.1657/1523-0430(2005)037[0444:CSIAPT]2.0.CO;2

CLARK, J. S., GELFAND, A. E., WOODALL, C. W., & ZHU, K. (2014). More than the sum of the parts: Forest climate response from joint species distribution models. *Ecological Applications*, 24(5), 990–999. https://doi.org/10.1890/13-1015.1

CLARK, J. S., NEMERGUT, D., SEYEDNASROLLAH, B., TURNER, P. J., & ZHANG, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, 87(1), 34–56. https://doi.org/10.1002/ecm.1241

CLARKE, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22(4), 341–352. https://doi.org/10.1080/07388940500339183

CRAMER, J. S. (2005). Omitted variables and misspecified disturbances in the logit model. *Tinbergen Institute Discussion Paper*, 05-084/4.

DE VALPINE, P., TUREK, D., PACIOREK, C. J., ANDERSON-BERGMAN, C., LANG, D. T., & 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. https://doi.org/10.1080/10618600.2016.1172487

DENWOOD, M. J. (2016). Runjags: An r package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, 71(9). https://doi.org/10.18637/jss.v071.i09

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., ... E. Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Fois, M., Cuena-Lombraña, A., Fenu, G., & Bacchetta, G. (2018). Using species distribution models at local scale to guide the search of poorly known species: Review, methodological issues and future directions. *Ecological Modelling*, 385, 124–132. https://doi.org/10.1016/j.ecolmodel.2018.07.018

Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71(3), 431–444. https://doi.org/10.1093/biomet/71.3.431

Gimenez, O., Buckland, S. T., Morgan, B. J. T., Bez, N., Bertrand, S., Choquet, R., Dray, S., Etienne, M.-P., Fewster, R., Gosselin, F., Mérigot, B., Monestiez, P., Morales, J. M., Mortier, F., Munoz, F., Ovaskainen, O., Pavoine, S., Pradel, R., Schurr, F. M., ... Rexstad, E. (2014). Statistical ecology comes of age. *Biology Letters*, 10(12), 20140698. https://doi.org/10.1098/rsbl.2014.0698

Gosselin, F. (2024). runMCMCbtadjust: Runs Monte Carlo Markov Chain – With Either 'JAGS', 'nimble' or 'greta' – While Adjusting Burn-in and Thinning Parameters (r package).

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat Suitability and Distribution Models: With Applications in R* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781139028271

Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model (D. Warton, Ed.). *Methods in Ecology and Evolution*, 6(4), 465–473. https://doi.org/10.1111/2041-210X.12332

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, e4794. https://doi.org/10.7717/peerj.4794

Hui, F. K. C., Vu, Q., & Hooten, M. B. (2024). Spatial confounding in joint species distribution models. *Methods in Ecology and Evolution*, 15(10), 1906–1921. https://doi.org/10.1111/2041-210X.14420

Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 22(0), 415–427. https://doi.org/10.1101/SQB.1957.022.01.039

Ishii, R., Maruo, K., & Gosho, M. (2022). Effect of Covariate Omission in Randomised Controlled Trials: A Review and Simulation Study. *International Statistical Review*, 90(1), 100–117. https://doi.org/10.1111/insr.12468

Jung, M., Arnell, A., De Lamo, X., García-Rangel, S., Lewis, M., Mark, J., Merow, C., Miles, L., Ondo, I., Pironon, S., Ravilious, C., Rivers, M., Schepaschenko, D., Tallowin, O., Van Soesbergen, A., Govaerts, R., Boyle, B. L., Enquist, B. J., Feng, X., … Visconti, P. (2021). Areas of global importance for conserving terrestrial biodiversity, carbon and water. *Nature Ecology & Evolution*, 5(11), 1499–1509. https://doi.org/10.1038/s41559-021-01528-7

Kissling, W. D., Dormann, C. F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G. J., Montoya, J. M., Römermann, C., Schiffers, K., Schurr, F. M., Singer, A., Svenning, J.-C., Zimmermann, N. E., & O'Hara, R. B. (2012). Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, 39(12), 2163–2178. https://doi.org/10.1111/j.1365-2699.2011.02663.x

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, 70(5). https://doi.org/10.18637/jss.v070.i05

Lee, Y., & Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88(4), 987–1006. https://doi.org/10.1093/biomet/88.4.987

Lee-Yaw, J. A., L. McCune, J., Pironon, S., & N. Sheth, S. (2022). Species distribution models rarely predict the biology of real populations. *Ecography*, 2022(6), e05877. https://doi.org/10.1111/ecog.05877

Lindgren, F., Rue, H., & Lindström, J. (2011). 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. https://doi.org/10.1111/j.1467-9868.2011.00777.x

McCullagh, P., & Nelder, J. (2019). *Generalized Linear Models* (2nd ed.). Routledge. https://doi.org/10.1201/9780203753736

Mielke, K. P., Claassen, T., Busana, M., Heskes, T., Huijbregts, M. A. J., Koffijberg, K., & Schipper, A. M. (2020). Disentangling drivers of spatial autocorrelation in species distribution models. *Ecography*, 43(12), 1741–1751. https://doi.org/10.1111/ecog.05134

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, 26(1), 67–82. https://doi.org/10.1093/esr/jcp006

Nabe-Nielsen, J., Nabe-Nielsen, L. I., & Ovaskainen, O. (2025). Drivers of plant community composition and diversity in low Arctic western Greenland. *Ecography*, 2025(8), e07816. https://doi.org/10.1002/ecog.07816

Niku, J., Hui, F. K. C., Taskinen, S., & Warton, D. I. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, 10(12), 2173–2182. https://doi.org/10.1111/2041-210X.13303

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., ... Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3), e01370. https://doi.org/10.1002/ecm.1370

Ovaskainen, O., & Abrego, N. (2020). *Joint Species Distribution Modelling: With Application in R* (Cambridge University Press). https://doi.org/https://doi.org/10.1017/9781108591720

Ovaskainen, O., Roy, D. B., Fox, R., & Anderson, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models (D. Orme, Ed.). *Methods in Ecology and Evolution*, 7(4), 428–436. https://doi.org/10.1111/2041-210X.12502

Petersen, M. R., & Deddens, J. A. (2000). Effects of omitting a covariate in poisson models when the data are balanced. *Canadian Journal of Statistics*, 28(2), 439–445. https://doi.org/10.2307/3315990

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 1–10.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J. S., & Thuiller, W. (2021). On the Interpretations of Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, 36(5), 391–401. https://doi.org/10.1016/j.tree.2021.01.002

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., & McCarthy, M. A. (2014). Understanding co-occurrence by modelling spe-

cies simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, 5(5), 397–406. https://doi.org/10.1111/2041-210X.12180

R Core Team. (2025). R: A Language and Environment for Statistical Computing (v. 4.5.1).

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods* (2nd ed., Vol. 1). SAGE Publications Inc.

Rehm, J., Arminger, G., & Kohlmeier, L. (1992). Using follow-up data to avoid omitted variable bias: An application to cardiovascular epidemiology. *Statistics in Medicine*, 11(9), 1195–1208. https://doi.org/10.1002/sim.4780110906

Reich, B. J., Hodges, J. S., & Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4), 1197–1206. https://doi.org/10.1111/j.1541-0420.2006.00617.x

Rinella, M. J., Strong, D. J., & Vermeire, L. T. (2020). Omitted variable bias in studies of plant interactions. *Ecology*, 101(6), e03020. https://doi.org/10.1002/ecy.3020

Rizopoulos, D. (2025). GLMMadaptive: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature (R package).

Robinson, L. D., & Jewell, N. P. (1991). Some Surprising Results about Covariate Adjustment in Logistic Regression Models. *International Statistical Review / Revue Internationale de Statistique*, 59(2), 227. https://doi.org/10.2307/1403444

Rousset, F., & Ferdy, J.-B. (2014). Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography*, 37(8), 781–790. https://doi.org/10.1111/ecog.00566

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., & Huijbregts, M. A. J. (2021). Assessing the reliability of species distribution projections in climate change research. *Diversity and Distributions*, 27(6), 1035–1050. https://doi.org/10.1111/ddi.13252

Schielzeth, H., & Nakagawa, S. (2013). Nested by design: Model fitting and interpretation in a mixed model era (R. Freckleton, Ed.). *Methods in Ecology and Evolution*, 4(1), 14–24. https://doi.org/10.1111/j.2041-210x.2012.00251.x

Shurin, J. B., Cottenie, K., & Hillebrand, H. (2009). Spatial autocorrelation and dispersal limitation in freshwater organisms. *Oecologia*, 159(1), 151–159. https://doi.org/10.1007/s00442-008-1174-z

SOBERÓN, J., & NAKAMURA, M. (2009). Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, 106(supplement_2), 19644–19650. https://doi.org/10.1073/pnas.0901637106

STROUP, W. W., PTUKHINA, M., & GARAI, J. (2024). *Generalized linear mixed models: Modern concepts, methods and applications* (Second edition). CRC Press Taylor & Francis Group. https://doi.org/https://doi.org/10.1201/9780429092060

THYGESEN, U. H., ALBERTSEN, C. M., BERG, C. W., KRISTENSEN, K., & NIELSEN, A. (2017). Validation of ecological state space models using the Laplace approximation. *Environmental and Ecological Statistics*, 24(2), 317–339. https://doi.org/10.1007/s10651-017-0372-4

TIKHONOV, G., OPEDAL, Ø. H., ABREGO, N., LEHIKOINEN, A., DE JONGE, M. M. J., OKSANEN, J., & OVASKAINEN, O. (2020). Joint species distribution modelling with the R-package Hmsc. *Methods in Ecology and Evolution*, 11(3), 442–447. https://doi.org/10.1111/2041-210X.13345

TIKHONOV, G., OPEDAL, Ø. H., ABREGO, N., LEHIKOINEN, A., de JONGE, M. M. J., OKSANEN, J., & OVASKAINEN, O. (2025). Hmsc 3.0: Getting started with Hmsc: Low-dimensional multivariate models.

TOBLER, M. W., KÉRY, M., HUI, F. K. C., GUILLERA-ARROITA, G., KNAUS, P., & SATTLER, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, 100(8), e02754. https://doi.org/10.1002/ecy.2754

VAN EE, J. J., IVAN, J. S., & HOOTEN, M. B. (2022). Community confounding in joint species distribution models. *Scientific Reports*, 12(1), 12235. https://doi.org/10.1038/s41598-022-15694-6

VANHATALO, J., HARTMANN, M., & VENERANTA, L. (2020). Additive Multivariate Gaussian Processes for Joint Species Distribution Modeling with Heterogeneous Data. *Bayesian Analysis*, 15(2). https://doi.org/10.1214/19-BA1158

VIEILLEDENT, G., & CLÉMENT, J. (2025). *Jsdm: Joint species distribution models* [R package version 0.2.7].

WANG, J., ZAMAR, R., MARAZZI, A., YOHAI, V., SALIBIAN-BARRERA, M., MARONNA, R., ZIVOT, E., ROCKE, D., MARTIN, D., MAECHLER, M., KONIS, K., & TODOROV, V. (2006). Robust: Port of the S+ "Robust Library" [Institution: Comprehensive R Archive Network Pages: 0.7-5]. https://doi.org/10.32614/CRAN.package.robust

WARTON, D. I., BLANCHET, F. G., O'HARA, R. B., OVASKAINEN, O., TASKINEN, S., WALKER, S. C., & HUI, F. K. (2015). So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, 30(12), 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Weber, M. M., Stevens, R. D., Diniz-Filho, J. A. F., & Grelle, C. E. V. (2017). Is there a correlation between abundance and environmental suitability derived from ecological niche modelling? A meta-analysis. *Ecography*, 40(7), 817–828. https://doi.org/10.1111/ecog.02125

Weisberg, S. (2014). *Applied linear regression* (4th ed.). Wiley.

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. (2023, November). A comparison of predictive performance of joint species distribution models for presence-absence data. https://doi.org/10.32942/X2G31C

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2019). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, 10(2), 198–211. https://doi.org/10.1111/2041-210X.13106

Wilms, R., Mäthner, E., Winnen, L., & Lanwehr, R. (2021). Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology*, 5, 100075. https://doi.org/10.1016/j.metip.2021.100075

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J.-A., Guisan, A., Heikkinen, R. K., Høye, T. T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Öckinger, E., Schmidt, N. M., … Svenning, J.-C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological Reviews*, 88(1), 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781315370279

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach* (4th ed.). South Western cengage learning.

Yatchew, A., & Griliches, Z. (1985). Specification error in probit models. *The Review of Economics and Statistics*, 67(1), 134. https://doi.org/10.2307/1928444

Zurell, D., Pollock, L. J., & Thuiller, W. (2018). Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography*, 41(11), 1812–1819. https://doi.org/10.1111/ecog.03315