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

June 2026

Abstract

Species distribution models (SDMs) are widely used in ecology to assess species distributions. Specifically, correlative SDMs are employed to infer relationships between species records and environmental variables. A classical approach to implementing such SDMs is to use parametric statistical models such as generalized linear mixed models (GLMMs). However, since species–environment relationships are complex, species distributions may depend on unobserved covariates. In this article, we first recall mathematical results showing that such "omitted covariates" typically introduce statistical issues that can bias inferences of observed-covariate effects or yield improper confidence intervals. So far, these results have received little attention in ecology. We then present a simulation study of the statistical impact of unobserved covariates on GL(M)M inference for continuous, count, and binary data. We assessed various regression methods, including both frequentist and Bayesian SDMs, as well as joint species distribution models (JSDMs) accounting for interspecific covariations in presence–absence data. In case of shared missing covariates, JSDMs provide an effective approach that mitigates inferential issues arising in presence–absence SDMs: they rightly increase covariate-slope magnitudes relative to single-species SDMs. We complemented these simulations by applying JSDMs and SDMs to real ecological datasets, revealing discrepancies in environmental-effect estimates and better predictive capacity for JSDMs. We recommend that ecologists fit JSDMs when dealing with community data to evaluate whether information can be extracted from between-species residuals. Ultimately, our results apply to any GL(M)M with suspected omitted variables, where generalized linear latent variable models (GLLVMs) could correct inference when jointly monitored entities share an important omitted 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 overlooked, 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. A mitigation strategy based on joint species distribution modelling is proposed, performing well when the omitted variable is shared across several species. These models rightly increase covariate-slope magnitudes compared to single-species models for presence–absence records when missing covariates are shared. Applications to real ecological datasets revealed widespread differences in the estimation of environmental effects between joint and separate species distribution models. Our results extend to generalized linear models and generalized linear latent variable models, applicable across a broader range of disciplines.

1. Introduction

Species distribution models (SDMs) are a major modeling tool widely used in ecology to study and predict the spatial and temporal distribution of species populations (Elith & Leathwick, 2009). Since they relate species occurrence or abundance records to environmental conditions, SDMs provide a quantitative framework for assessing habitat suitability, understanding species–environment relationships, and projecting potential distributional shifts (Franklin, 2010; Guisan & Thuiller, 2005; Guisan et al., 2017). These capacities make SDMs central to both theoretical and applied ecology, with many applications, for example, in conservation planning (Jung et al., 2021), biodiversity management (Fois et al., 2018), invasive species risk assessment (Thuiller et al., 2005), and the assessment of climate-change impacts (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). In the context of climate change, these modeling frameworks have indeed become valuable in contributing to the evaluation and prediction of the future geographical distributions of species under changing environmental conditions (Pearson & Dawson, 2003). Numerous regression approaches have been developed to analyze abundance or presence–absence data, 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 by ecologists or practitioners (McCullagh & Nelder, 2019; Raudenbush & Bryk, 2002; Stroup et al., 2024).

Nonetheless, SDMs have important conceptual limitations stemming from their underlying assumptions (e.g., 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 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 that the species is independent of 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 do not directly estimate 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 (Peyrard & Gimenez, 2022; Wisz et al., 2013), population dynamics (Mielke et al., 2020), spatial dispersion (Shurin et al., 2009), or species detectability (Kellner & Swihart, 2014). 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. 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). 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 important impacts on the reliability of inferential outcomes. In particular, it involves the so-called "omitted-variable bias" (OVB), i.e., the bias that affects the estimates of the fixed effects of observed covariates due to the omission of 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 relatively unexplored in the context of ecological modeling, despite the important consequences it may have on drawn conclusions (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 the omission of a covariate will affect estimates of other regression parameters. In Tab. 1, some essential results concerning the inferential consequences of an unobserved variable in a GLM are synthesized. We herein consider a response variable Y depending on two centered 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 α, β, γ are nonzero, $z_i = \mathbb{E}[Y_i \mid X_{1,i}, X_{2,i}]$, and f is a link function. Supposing that covariate X_2 is not observed, the regression model, which has been misspecified, therefore assumes $f(\tilde{z}_i) = \widehat{\alpha} + \widehat{\beta}\, X_{1,i}$ with $\tilde{z}_i = \mathbb{E}[Y_i \mid X_{1,i}]$. It is thus relevant to quantify the discrepancy between the misspecified-model estimates, $\widehat{\alpha}$ and $\widehat{\beta}$, and

the true parameters, 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 the magnitude of the 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).

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 $f(z_i) = \alpha + \beta X_{1,i} + \gamma X_{2,i}$ (with centered X_1, X_2, and nonzero α, β, γ), while the misspecified regression model is $Y_i \mid X_{1,i} \sim \text{Normal}(\tilde{z}_i, \widehat{\sigma}^2)$ or $\text{Poisson}(\tilde{z}_i)$ or $\text{Bernoulli}(\tilde{z}_i)$ with $f(\tilde{z}_i) = \widehat{\alpha} + \widehat{\beta} X_{1,i}$, where f is a link function. The covariates X_1 and X_2 are assumed to be independent of each other.

distribution	**link function**	Bias$[\widehat{\alpha}]$	Bias$[\widehat{\beta}]$	**reference(s)**
Normal	identity	$= 0$	$= 0$	Gail et al. (1984) Canner (1991) Ishii et al. (2022)
Poisson	log	$\neq 0$ (> 0)	$= 0$	Gail et al. (1984) Petersen & Deddens (2000) Ishii et al. (2022)
Bernoulli	logit	$\neq 0$ $\begin{cases} < 0 & \text{if } \alpha > 0 \\ > 0 & \text{if } \alpha < 0 \end{cases}$	$\neq 0$ $\begin{cases} < 0 & \text{if } \beta > 0 \\ > 0 & \text{if } \beta < 0 \end{cases}$	Gail et al. (1984) Robinson & Jewell (1991) Mood (2010) Cramer (2005)
Bernoulli	probit	$\neq 0$ $\begin{cases} < 0 & \text{if } \alpha > 0 \\ > 0 & \text{if } \alpha < 0 \end{cases}$	$\neq 0$ $\begin{cases} < 0 & \text{if } \beta > 0 \\ > 0 & \text{if } \beta < 0 \end{cases}$	Yatchew & Griliches (1985)

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 an observation-level random effect intended to capture unmeasured covariates (Bolker et al., 2009; Byrnes & Dee, 2025; Harrison et al., 2018; Schielzeth & Nakagawa, 2013). Yet, including a random effect exhibiting a large-range spatial autocorrelation may unintentionally amplify the bias in the estimated effects of the covariates, thereby casting doubt on the usefulness of this solution (Reich et al., 2006). Therefore, the inability of classical SDMs, including those with random effects, to incorporate ecological variables beyond those measured is an

important limitation for studies in community ecology (Austin & Van Niel, 2011; Hodges & Reich, 2010; 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 species-specific coefficients called factor loadings (Ovaskainen & Abrego, 2020; Warton et al., 2015). In particular, species with similar loadings are expected to exhibit similar responses to the corresponding latent variable, and latent variables can be seen as missing predictors on which species still depend (Bystrova et al., 2022). 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 covariates (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 than SDMs. We especially 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 for presence–absence data yield much less biased estimators, with fewer inferential issues, than SDMs in case of important omitted covariates, thus underscoring their interest in community ecology. We further 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. On the whole, 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 within a simulation framework; (3) we illustrate the previous simulated results by applying our strategy to several real ecological datasets. We recall the main hypotheses underlying the study: independence of data conditionally on the (observed and unobserved) covariates, spatially-independent latent factors in JSDMs, a standardized Normal distribution for the missing covariate, and the same unobserved covariate shared by multiple species within the community.

2.1 Inferential impact assessment with SDMs

First of all, we assessed the statistical behavior of single-species SDMs when an important covariate is omitted. 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, 1,000 datasets were simulated, each comprising two independent covariates (X_1 and X_2) and the corresponding ecological response (Y). In the data-generating processes, the response followed a probability distribution (Normal, Poisson, or Bernoulli) whose mean parameter was determined by a linear predictor that combined the two covariates through the appropriate link function. All the synthetic datasets were then analyzed with an SDM structurally similar to the data-generating process, with one key difference — only X_1 was observed while X_2 remained unobserved, thereby deliberately generating a misspecified SDM. Thus, only two parameters were systematically estimated: the intercept ($\widehat{\alpha}$) and X_1's regression coefficient ($\widehat{\beta}$). For the sake of completeness, we investigated several SDM implementations: (i) generalized linear models (GLM); (ii) generalized linear mixed models with an observation-level random effect (GLMM); (iii) robust GLMs; (iv) generalized additive mixed models (GAMM); and (v) quasi-likelihood GLMs. 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 in Tab. 1 (Fig. 1). Complete plots for all investigated SDMs are provided in the Supplementary Material (see Sect. D in the SM, specifically Fig. S4 and Fig. S5). The Gaussian SDMs produced unbiased estimates of both the intercept $\widehat{\alpha}$ and the regression coefficient $\widehat{\beta}$, with median bias values centered on zero across all methods. In contrast, the Poisson SDMs exhibited a significant positive shift in the intercept $\widehat{\alpha}$ when no observation-level random effect (OLRE) was included, although the regression coefficient $\widehat{\beta}$ remained close to its true value (qualitatively unbiased though still significantly biased) across all formulations; coverage rates, however, always exhibited issues for both estimated parameters. Introducing an OLRE effectively eliminated the bias in $\widehat{\alpha}$, bringing the estimate biases back to zero, and most models yielded unbiased estimates for $\widehat{\alpha}$ and $\widehat{\beta}$, along with the expected coverage rates. For Bernoulli SDMs, both $\widehat{\alpha}$ and $\widehat{\beta}$ were biased, with the biases of $\widehat{\alpha}$ and $\widehat{\beta}$ exhibiting, across most methods, the attenuated magnitude that is expected from the mathematical results (Tab. 1). The absolute 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 Sect. A in the SM). In those cases of binary data, introducing an OLRE produced differing effects. The OLRE did not alter the results for most of the frequentist methods (Fig. 1), except for SDM.4, under which $\widehat{\alpha}$ and $\widehat{\beta}$ 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 OLRE showed heterogeneous behavior. For example, SDM.6 remained biased even with the OLRE, 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). Magnitude analyses also indicated that most Bernoulli probit and cloglog SDMs without OLRE led to moderately to strongly negative biases. 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 (Fig. S4-S5, Tab. S3).

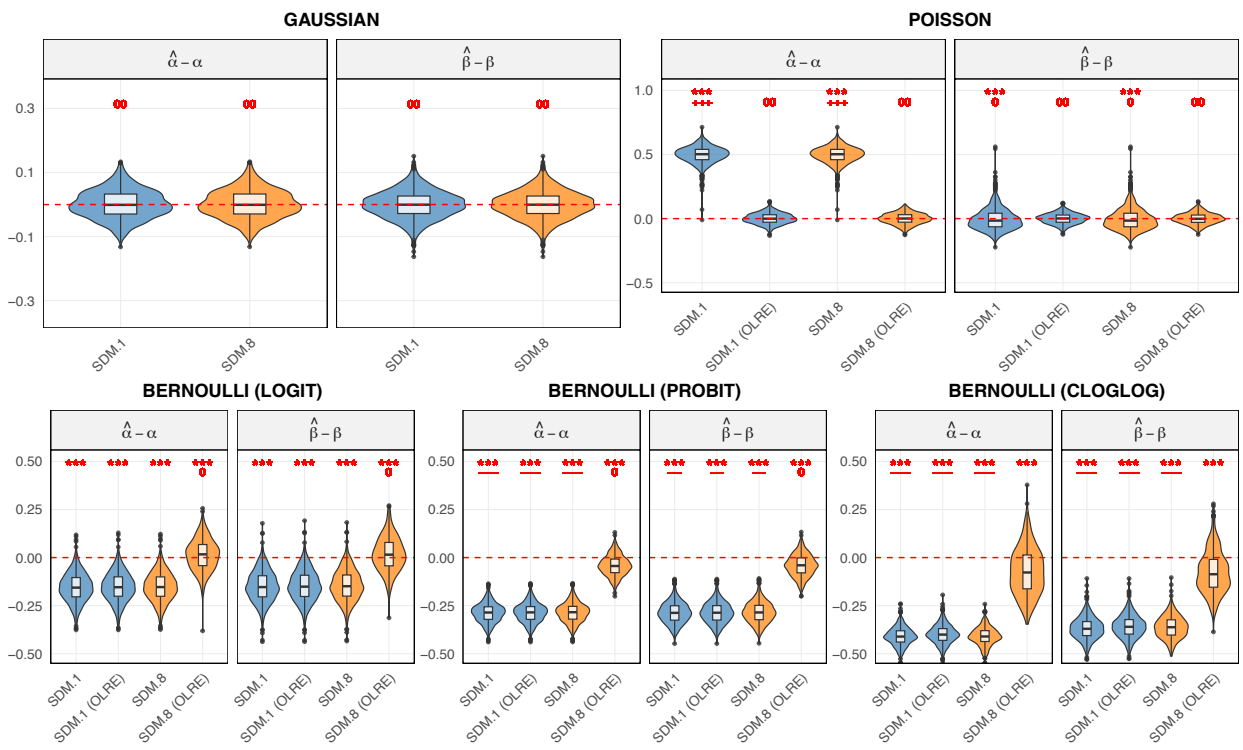

Figure 1: Violin- and box-plots of the deviations of the estimates of the intercept $\widehat{\alpha}$ and the regression coefficient $\widehat{\beta}$ from their respective true values. Only the results for SDM.1 and SDM.8, without or with an observation-level random effect (OLRE), are herein plotted (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. Above the corresponding box-plots, both bias significance and bias magnitude were reported. Bias significance was assessed with a p-value from a robust linear model: *** if $p \leq 0.001$, ** if $0.001 < p \leq 0.01$, * if $0.01 < p \leq 0.05$, no symbol if $p > 0.05$. Bias magnitude was "conclusive" if 95% of the $(\widehat{\theta}_k - \theta)$, with $k \in \{1, ..., N\}$ and $N = 1000$ the number of replicates, were in the interval: $[-0.05, 0.05]$, denoted by 000 and referred to as "strongly negligible", $[-0.1, 0.1]$, denoted by 00 and referred to as "moderately negligible", $[-0.5, 0.5]$, denoted by 0 and referred to as "weakly negligible", $[0.05, +\infty]$, denoted by $+$ and referred to as "weakly positive", $[0.1, +\infty]$, denoted by $++$ and referred to as "moderately positive", $[0.5, +\infty]$, denoted by $+++$ and referred to as "strongly positive", $[-\infty, -0.05]$, denoted by $-$ and referred to as "weakly negative", $[-\infty, -0.1]$, denoted by $--$ and referred to as "moderately negative", $[-\infty, -0.5]$, denoted by $---$ and referred to as "strongly negative", and "inconclusive" otherwise (no symbol).

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 1,000 synthetic datasets of presence–absence records through a data-generating process 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 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). Simulated data were then statistically analyzed using a specific type of JSDMs: JSDMs incorporating 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 ($\widehat{\alpha}$) and effect of X_1 ($\widehat{\beta}$), whereas under C.3, the effect of X_2 ($\widehat{\gamma}$) 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 qualitatively 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. S10, S14, S18, S22, S26, S30 in the SM). Nevertheless, the statistical significance of the bias depended on the scenario: greater true values of X_2's slope (γ) correlated with a larger proportion of significant biases for the intercept ($\widehat{\alpha}$), and low or high species prevalence similarly led to more significant biases for the intercept ($\widehat{\alpha}$). Scenarios leading to the smallest biases for X_1's slope ($\widehat{\beta}$) were those assuming a zero intercept in data generation (namely, S.2 and S.4, hence corresponding to small covariate effects and sufficient average species prevalences). In terms of magnitude analysis, biases were weakly negligible in all scenarios for almost all JSDMs. Coverage rates were particularly close to the expected value for almost all models (Tab. S4, S7, S10, S13, S16, S19 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 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 produced more biased estimates globally. Regarding the factor loadings ($\widehat{\lambda}$), they mainly captured the true value of the missing covariate's coefficient (β), with still some exceptions stemming from formulation-specific JSDM constraints (Fig. S7, S11, S15, S19, S23, S27, S31 in the SM).

Since configuration C.2's results did not incorporate a latent structure, the resulting JSDMs under C.2 statistically corresponded to stacked single-species SDMs (Schmitt et al., 2017), 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 residual between-species covariance (see Sect. B in the SM for details about MSDMs). Including this specification subsequently enabled 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. S8, S12, S16, S20, S24, S28, S32 in the SM). Notably, median biases from scenario S.1 (S12) were equal to the median bias from the Bernoulli probit SDMs (Fig. 1), and they were moderately negative according to magnitude analyses. The other metrics confirmed poor inferential performance in this case: coverage rate issues and inflated RMSEs and RMSREs (Tab. S5, S8, S11, S14, S17, S20 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 qualitatively very reduced or zero (Fig. S9, S13, S17, S21, S25, S29, S33 in the SM), along with weakly negligible bias magnitudes, 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. S6, S9, S12, S15, S18, S21 in the SM).

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'

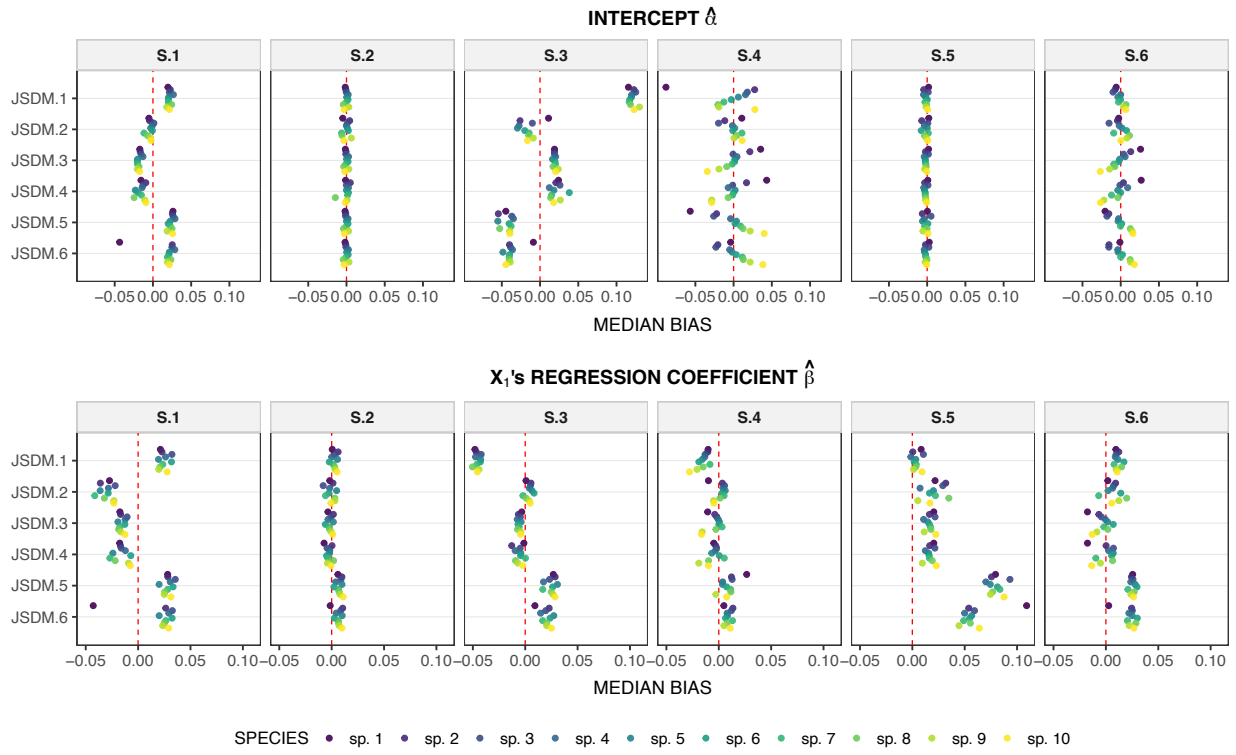

Figure 2: Median bias of the estimates of the intercept $\widehat{\alpha}$ (upper panel) and the regression coefficient $\widehat{\beta}$ (lower panel) across scenarios S.1 to S.6 (subplots from left to right) and under configuration `C.1`. Within each subplot, rows correspond to the six JSDM formulations (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.

dependence on the unobserved covariate X_2. Among the 10 species, only the first m species were assumed to respond to X_2, while the remaining $10-m$ species did not. We repeated the simulations for $m \in \{1, \ldots, 10\}$ to assess how well the estimates $\widehat{\alpha}$ and $\widehat{\beta}$ 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 six 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 first species responded to the unobserved covariate, the estimates of $\widehat{\alpha}$ and $\widehat{\beta}$ 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.4 provided unbiased estimates for $\widehat{\alpha}$ and $\widehat{\beta}$ (Fig. S36, S37, S38), whereas unexpected bias was noticeable for the last two Bayesian JSDMs (Fig. S39 and S40). 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, all 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 a somewhat logical pattern in which the bias caused by the missing covariate decreased progressively in amplitude as the number of responding species increased (Fig. 3).

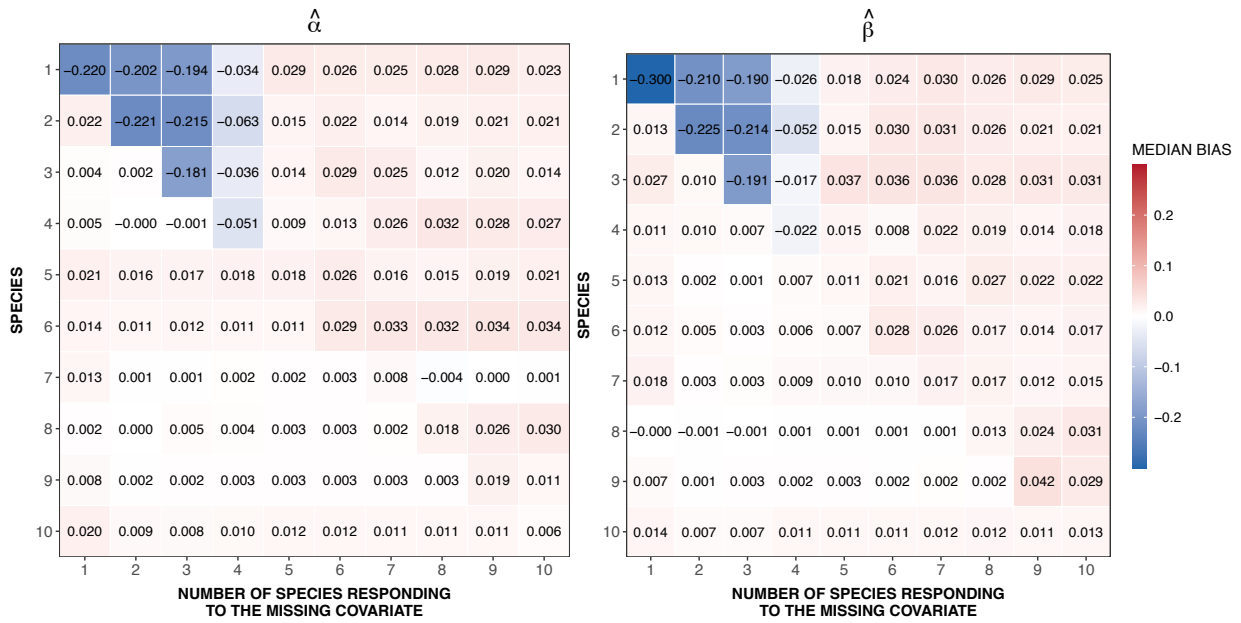

Figure 3: Heatmaps of the median bias of the estimated intercept $\widehat{\alpha}$ (left subplot) and the estimated regression coefficient $\widehat{\beta}$ (right subplot) with JSDM.1. In both subplots, each column represents a scenario where an increasing number of species (from only 1 to all 10 species) 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. For column m, only the first m species respond to the unobserved covariate. 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 (named dataset.1 to dataset.6). Each dataset comprised presence–absence records for multiple taxa along with environmental covariates measured at sampled sites. We fitted two models featuring a probit link: 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 it could correspond to an important missing covariate. We thus interpreted the corresponding factor loadings as the species-specific slopes associated with this hypothetical missing covariate.

Results for dataset.5 are displayed in Fig. 4. To analyze them, we compared, for the intercept and every included explanatory variable, each species-specific regression coefficient estimate from the JSDM (y-axis) with the equivalent regression coefficient estimate from the corresponding single-species SDM (x-axis). We checked whether the shift between SDM and JSDM estimates was compatible with the bias that is mathematically predicted from a missing covariate in an SDM (the so-called omitted-variable bias). More precisely, for each couple of estimates of SDM and JSDM, the corresponding dot was colored based on the bias expected in the SDM according to mathematical results (see Sect. A.3 in the SM) from estimated species-specific loadings in the JSDM. The background color corresponds to the exact mathematical bias so that when the color of a dot matches the background color, the difference between SDM and JSDM estimates was consistent in magnitude with the expected bias caused by a missing covariate captured by the latent variable.

We obtained similar results for the six real datasets (see Sect. F in the SM, including Fig. S41, S42, S43, S44, S45, S46). Overall, outcomes were consistent with what was predicted from a missing covariate, since we clearly see a shift of multiple estimates from SDM to JSDM (some differences are very strong and very significant). We were thus able to interpret most of the shifts between JSDM and SDM estimates within the missing-covariate framework. However, we also found a few shifts that cannot be explained by a missing covariate (namely, points falling outside the background-colored areas). We hence conducted additional analyses to evaluate the potential causes of these differences, but we could not find a convincing elucidation (details are available in Section F.2 in the SM). 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 smaller.

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. 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

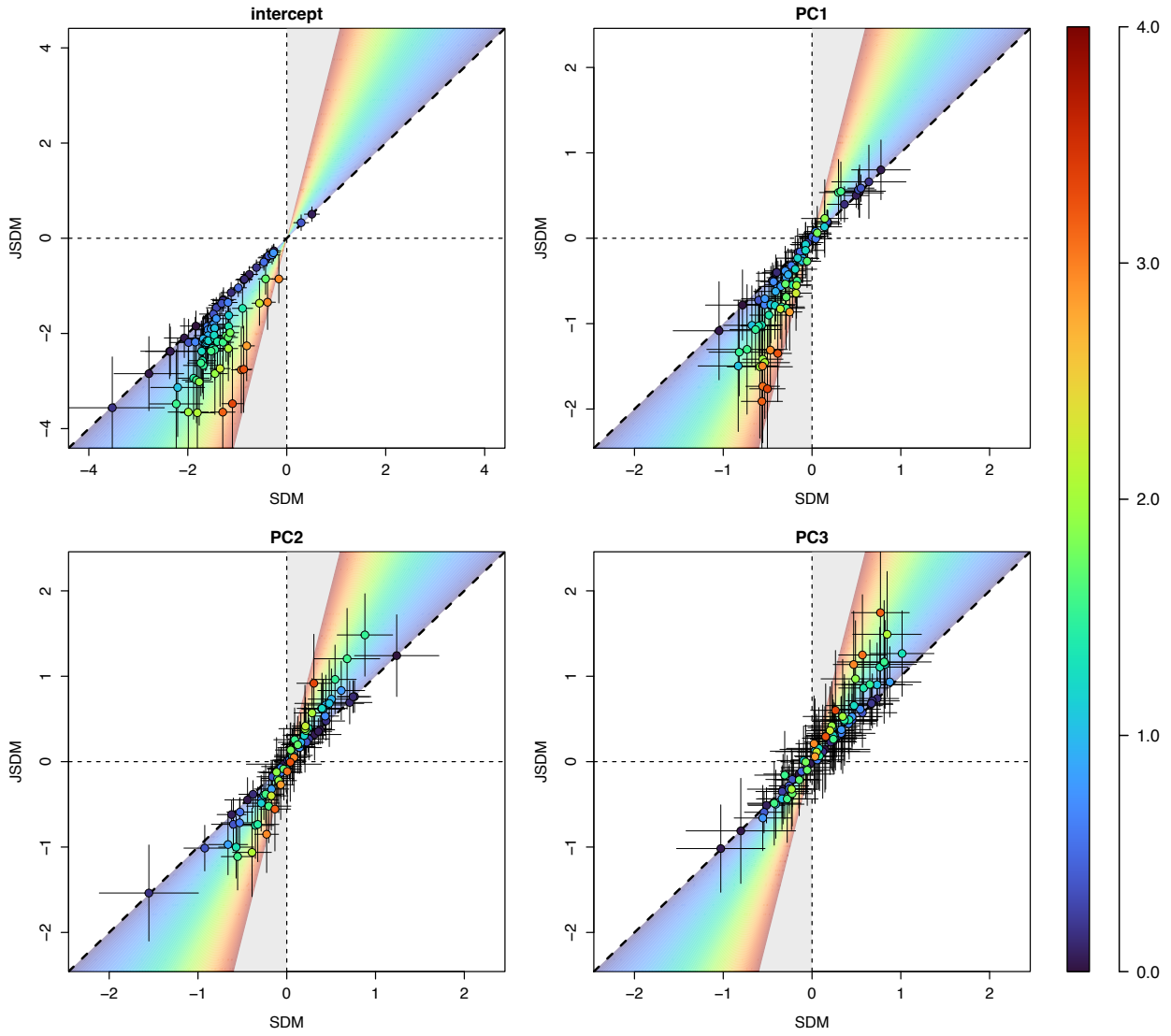

Figure 4: Comparison of SDM and JSDM estimates of the intercept and regression coefficients for dataset.5. Each subplot corresponds to the intercept or to one of the three principal components (PCs) of the covariate set. In each subplot, the horizontal axis (x-axis) shows the SDM estimates across species, and the vertical axis (y-axis) the JSDM estimates across species, with each point representing a species. Black line segments indicate 95% confidence intervals for the SDM (x-axis) and JSDM (y-axis) estimates. For each pair of estimates of SDM and JSDM, the corresponding point was colored based on the bias expected in the SDM according to mathematical results (see Sect. A.3 in the SM) from estimated species-specific loadings in the JSDM. The background color corresponds to the exact mathematical bias so that when the color of a dot matches the background color, the difference between SDM and JSDM estimates was consistent in magnitude with the expected bias caused by a missing covariate captured by the latent variable.

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 attenuated, as confirmed by both mathematical and simulation results. This means that, for presence–absence SDMs with omitted variables and without spatial structure in the covariates, estimates will be of lower magnitude than what they should be — which is of practical concern. Conversely, in count data SDMs, only the intercept is biased. Although introducing an observation-level random effect (OLRE) 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 OLRE 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 presence–absence, 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 since 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. Inferential issues manifested by SDMs (notably regarding parameter recovery and coverage rate) were reduced as soon as we employed JSDMs. 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 multi-species distribution models 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 of the slope of observed explanatory variables 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 species-specific regression parameters, 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 values of these loadings for better estimation of the regression coefficients of the environmental covariates — with stronger departures between JSDM and SDM estimates and stronger biases for SDM estimates when these absolute values are strong. 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. S34).

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 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% credible intervals, 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 credible intervals for the covariate effects. Yet, when no such discrepancies were present, the JSDM produced narrower credible intervals 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.

To address residual spatial dependence left unexplained by an SDM (which may arise from unmeasured spatially structured predictors or processes), many authors have often relied upon introducing a spatial random effect in GLMMs and hierarchical models, when studying a single species (Dormann et al., 2007). As most covariates are themselves spatially structured in reality, a spatial random effect would be able to (partially) capture such an omitted covariate. Interestingly, an SDM with a spatial random effect and a JSDM with non-spatial latent factors can be viewed as conceptually parallel modeling approaches. The former captures residual spatial covariation among sites, whereas the latter captures residual interspecific covariation among species. In both cases, the model exploits covariation along a specific dimension — space in the spatial SDM and species in the JSDM — to account for latent structure that may reflect unmeasured covariates that influence the response variable. We hypothesize that a spatial random effect can capture a spatially-structured missing covariate more efficiently than a non-spatial random effect by leveraging information shared across sites through spatial autocorrelation. However, employing spatial random effects often introduces another hazard: the so-called spatial confounding, which arises when the spatial random effect is correlated with covariates included in the model (e.g., Pim et al., 2026). It is unclear, however, which phenomena would be the most prevalent in any ecological dataset.

Admittedly, 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 straightforward with nonparametric approaches — average marginal effects are equivalent to the parametric fixed-covariate effects but require intensive calculation. Second, we consistently assumed that the missing covariate was spatially unstructured and independent of the other observed environmental covariates. This assumption is clearly debatable, as covariates are actually almost always (spatially auto)correlated. Making this assumption

enabled us to restrict the study to a specific source of bias that arises even when the confounder is not spatially structured. Yet, the issues of spatial or community confounding in JSDMs have been explored, highlighting the potential inferential risks these phenomena may pose, such as increased bias in covariate effect (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 welcome to check whether our results can extend to more spatially realistic settings. Third, we consistently assumed perfect detection, although imperfect species detectability is well known to require explicit modelling of the observation process. Still, it remains unclear how imperfect detection would interact with unobserved covariates, and whether JSDMs incorporating an explicit observation layer would be able to address it. Fourth, 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 additional species that share the missing covariate to reduce bias in environmental-effect estimates. Fifth, 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 that JSDMs would allow us to correct, as was found with an omitted covariate. Sixth, there is currently no tool to detect whether other omitted covariates remain when only a single species responds to each of them. In such cases, the magnitude would still be underestimated with JSDMs, and the bias would be attenuated with JSDMs compared to SDMs. Seventh, a major limitation of JSDMs — beyond the intrinsic limitations of our study — lies in their cost: they are simultaneously data-intensive, as they require a substantial number of records for multiple species across several sites (although this requirement is tempered by the limited number of species needed to recover the covariate effect, see Fig. 3), and computationally intensive, because of the high number of parameters that must be inferred through the latent structure.

In conclusion, this study demonstrates the existence of bias in parametric presence–absence SDMs when (at least) one important covariate remains unmeasured. Our findings indicate a systematic underestimation of the magnitude of included predictors' effect under independence of covariates and without spatial structure. 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 in these conditions that ecologists and practitioners targeting species distribution estimations systematically fit JSDMs, as they allow for correctly estimating the observed environmental effects when an important missing covariate is shared by several species. 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. Moreover, ecologists interested in detecting situations where important covariates shared by several species are missing could also fit corresponding single-species SDMs and then compare the resulting environmental effects with those of the JSDM. Attenuation in the environmental effects in SDMs compared to JSDMs would suggest the existence of unacknowledged environmental predictors shared by several species 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. Ultimately and crucially, our results remain valid in other scientific disciplines making use of similar statistical tools where important covariates are suspected to be omitted in binary GL(M)Ms. One could think, for instance, of epidemiology and medicine (with studies involving multiple patients affected by several diseases and subgroups of patients influenced by unrecorded patterns), econometrics (via modeling households' purchases of several product categories with an unobserved effect of consumer segments), social and political sciences (through analyzing whether parties received votes in different municipalities with an unmeasured influence of social stratification), psychometrics, genomics. In these cases, generalized linear latent variable models (GLLVMs) could properly correct inference of binary data in situations where different entities monitored together 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 1,000 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 $i = 1, \ldots, n$. 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

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

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,

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

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,

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

with Φ the cumulative distribution function of the standard Normal distribution (i.e., 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 replicated dataset, 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 (OLRE); (iii) a robust GLM; (iv) a generalized additive mixed model (GAMM); and (v) a quasi-likelihood GLM. Specifically, the misspecified models for implementations (i) and (ii) are expressed as follows:

(i) GLM:

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

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 an observation-level random effect (OLRE):

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

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 OLRE.

We therefore estimated the intercept $\widehat{\alpha}$ and the single regression coefficient $\widehat{\beta}$ 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`; it enabled us to monitor and control MCMC convergence and the number of effective values (MCMC iteration autocorrelation) (Gosselin, 2026).

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 1,000 replicates, yielding 1,000 datasets of presence–absence records for a community of $S = 10$ species over 1,000 sites. At each replicate, the data were simulated independently from the same Bernoulli GLMM with a probit link, as follows,

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

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 are 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, $\beta_1 = \beta_2 = 0.5$); S.5, a community at intermediate prevalence where covariate X_1 exerts a strong effect and covariate X_2 a weak effect ($\beta_0 = 0$, $\beta_1 = 1.5$, $\beta_2 = 0.2$); and S.6, a community spanning

Table 2: List of regression methods from R packages employed for fitting SDMs. The columns "without OLRE" and "with OLRE" indicate whether the SDM could be instantiated only without OLRE, only with OLRE, or in either configuration.

model label	built-in function	R package(s)	inference	model type	without OLRE	with OLRE	reference(s)
SDM.1	`lm` | `glm` | `lmer` | `glmer`	`stats` | `lme4`	frequentist	GL(M)M	■	■	Bates et al. (2015) R Core Team (2025)
SDM.2	`glmmTMB`	`glmmTMB`	frequentist	GL(M)M	■	■	Kristensen et al. (2016) Thygesen et al. (2017)
SDM.3	`fitme`	`spaMM`	frequentist	GL(M)M	■	■	Rousset & Ferdy (2014) Lee & Nelder (2001)
SDM.4	`mixed_model`	`GLMMadaptive`	frequentist	GLMM		■	Rizopoulos (2025)
SDM.5	`brm`	`brms`	Bayesian	GL(M)M	■	■	Bürkner (2018)
SDM.6	`inla`	`INLA`	Bayesian	GL(M)M	■	■	Lindgren et al. (2011) Rue et al. (2009)
SDM.7	—	`nimble`	Bayesian	GL(M)M	■	■	De Valpine et al. (2017)
SDM.8	—	`runjags`	Bayesian	GL(M)M	■	■	Plummer (2003) Denwood (2016)
SDM.9	`lmrob` | `glmrob`	`robustbase`	frequentist	robust GL(M)M	■		Maechler et al. (2026)
SDM.10	`glm`	`stats`	frequentist	quasi-likelihood GLM	■		R Core Team (2025)
SDM.11	`gam` | `gamm`	`mgcv`	frequentist	GA(M)M	■	■	Wood (2017)

rare to common species that respond positively to one covariate but negatively to another (intercepts evenly spaced from -1 to 1, $\beta_1 = 1$, $\beta_2 = -0.5$). 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 a 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,

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

where $\tilde{z}_i = \mathbb{E}\left[Y_i^{[s]} \mid X_{1,i}\right]$, $\widehat{\alpha}^{[s]}$ is the intercept of species s, $\widehat{\beta}^{[s]}$ is the regression coefficient of species s with respect to the covariate X_1, $\widehat{\lambda}^{[s]}$ is the factor loading of species s with respect to the latent variable $\widehat{W} \sim \text{Normal}(\mathbf{0}, \mathbf{I})$. We emphasize that the latent variable $\widehat{W}$ 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 $\widehat{\lambda}^{[s]} = 0$ for all species, thereby removing the term $\widehat{\lambda}^{[s]} \widehat{W}_i$ in Eq. (7). Under configuration C.3, the fitted JSDMs followed Eq. (7) with an additional term $\widehat{\gamma}^{[s]} X_{2,i}$ 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, 2026).

Note that setting the number of latent factors 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.

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 first 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 $\widehat{\alpha}$ and $\widehat{\beta}$ 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:

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

where $\alpha^{[1]} = \cdots = \alpha^{[S]} = \beta^{[1]} = \cdots = \beta^{[S]} = \gamma^{[1]} = \cdots = \gamma^{[m]} = 1$ and $\mathbb{1}\{\cdot\}$ is an indicator function. Statistical analyses, assuming X_2 unobserved, were also systematically performed on the six 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:

$$\text{Bias}\left[\widehat{\theta}\right] = \text{median}\{\widehat{\theta}_1, \dots, \widehat{\theta}_N\} - \theta$$

- the (empirical) coverage rate (CR):

$$\text{CR}\left[\widehat{\theta}\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left\{\theta \in \left[\ \widehat{\theta}_i - 1.96\, \widehat{\text{SE}}[\widehat{\theta}_i],\ \widehat{\theta}_i + 1.96\, \widehat{\text{SE}}[\widehat{\theta}_i]\ \right]\right\}$$

Table 3: List of regression methods from R packages employed for fitting JSDMs. The column "equivalent model with no latent variables" indicates the model to which the JSDM is equivalent when the number of latent variables is set to zero: MSDM stands for multi-species distribution model and SSDM stands for stacked species distribution model.

model label	R package	regression coefficients	constraint on latent variables	constraint on loading matrix	inference	equivalent model at no latent variable	reference(s)
JSDM.1	`gllvm`	fixed effects	unit variances	positive diagonal zero upper diagonal free lower diagonal	frequentist	SSDM	Niku et al. (2019) Kristensen et al. (2016)
JSDM.2	`jSDM`	fixed effects	unit variances	positive diagonal zero upper diagonal free lower diagonal	Bayesian	SSDM	Vieilledent & Clément (2025)
JSDM.3	`Hmsc`	random effects (hierarchical structure)	unit variances	free matrix	Bayesian	MSDM	Tikhonov et al. (2020) Ovaskainen & Abrego (2020)
JSDM.4	`nimble`	random effects (hierarchical structure)	unit variances	free matrix	Bayesian	MSDM	De Valpine et al. (2017)
JSDM.5	`runjags`	fixed effects	unit variances	positive diagonal zero upper diagonal free lower diagonal	Bayesian	SSDM	Plummer (2003) Denwood (2016)
JSDM.6	`runjags`	fixed effects	free variances	unit diagonal zero upper diagonal free lower diagonal	Bayesian	SSDM	Plummer (2003) Denwood (2016)

(based on Wald intervals)

- the root mean square error (RMSE):

$$\mathrm{RMSE}\left[\widehat{\theta}\right] = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\theta}_i - \theta\right)^2}$$

- the root mean square random error (RMSRE):

$$\mathrm{RMSRE}\left[\widehat{\theta}\right] = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(\widehat{\theta}_i - \theta\right)^2 + \widehat{\mathrm{SE}}[\widehat{\theta}_i]^2\right)}$$

(allowing us to account simultaneously for both the shift of the optimal estimate from the true value and variability around this optimal estimate)

where $\widehat{\theta}_1, \ldots, \widehat{\theta}_N$ are the estimates of any parameter whose true value is θ, $\mathbb{1}\{\cdot\}$ is an indicator function, and $\widehat{\mathrm{SE}}[\widehat{\theta}_i]$ stands for the standard error of the corresponding optimal estimate $\widehat{\theta}_i$.

Bias significance was assessed with a p-value computed with the function `lmrob` from `R` package `robustbase` (Maechler et al., 2026) 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 $p > 0.05$. To complement significance analyses, whose relevance in the context of replicated simulation studies can be debatable (White et al., 2014), we also conducted bias magnitude analyses. Bias magnitude was "conclusive" if 95% of the $(\widehat{\theta}_k - \theta)$, with $k \in \{1, \ldots, N\}$ and $N = 1000$ the number of replicates, were in the interval: $[-0.05, 0.05]$, denoted by 000 and referred to as "strongly negligible", $[-0.1, 0.1]$, denoted by 00 and referred to as "moderately negligible", $[-0.5, 0.5]$, denoted by 0 and referred to as "weakly negligible", $[0.05, +\infty]$, denoted by $+$ and referred to as "weakly positive", $[0.1, +\infty]$, denoted by $++$ and referred to as "moderately positive", $[0.5, +\infty]$, denoted by $+++$ and referred to as "strongly positive", $[-\infty, -0.05]$, denoted by - and referred to as "weakly negative", $[-\infty, -0.1]$, denoted by - - and referred to as "moderately negative", $[-\infty, -0.5]$, denoted by - - - and referred to as "strongly negative", and "inconclusive" otherwise (no symbol) (Dixon & Pechmann, 2005).

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). Among model outputs, we used the coefficient estimates associated with each of the retained principal components, along with the loadings estimated by the JSDM. 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 with this article.

Data Availability

The R codes that support the investigations in this article are available in the following GitHub repository: `https://github.com/arthur-f-rossignol/article-004`. The real datasets are all already publicly available (Abrego et al., 2025; Choler, 2005; Harris, 2015; Nabe-Nielsen et al., 2025; Ovaskainen et al., 2016; Pollock et al., 2014).

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. The authors thank Vicki Moore for polishing the English and Giuliana Brambilla for helpful remarks on a previous draft, as well as two anonymous referees whose comments significantly improved the manuscript. The AI-based writing assistant Grammarly was used to improve the language and readability of the manuscript and the appendices; the generative AI tool Claude Opus 4.8 was used to help format the revised manuscript into the PNAS LaTeX template; 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, Formal analysis, 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. P. (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

Bystrova, D., Poggiato, G., Arbel, J., & Thuiller, W. (2022). Latent factor models: A tool for dimension reduction in joint species distribution models. In N. Peyrard & O. Gimenez (Eds.), *Statistical Approaches for Hidden Variables in Ecology* (Wiley, pp. 135–156). https://doi.org/https://doi.org/10.1002/9781119902799.ch7

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

Dixon, P. M., & Pechmann, J. H. K. (2005). A statistical test to show negligible trend. *Ecology*, 86(7), 1751–1756. https://doi.org/10.1890/04-1343

Dormann, C. F., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., R. Peres-Neto, P., Reineking, B., Schröder, B., M. Schurr, F., & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, 30(5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

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

Franklin, J. (2010). *Mapping Species Distributions: Spatial Inference and Prediction* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511810602

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. (2026). runMCMCbtadjust: Runs Monte Carlo Markov Chain – With Either 'JAGS', 'nimble' or 'greta' – While Adjusting Burn-in and Thinning Parameters [R package version 1.1.3].

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, 8(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

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. *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

Hodges, J. S., & Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4), 325–334. https://doi.org/10.1198/tast.2010.10052

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

Kellner, K. F., & Swihart, R. K. (2014). Accounting for Imperfect Detection in Ecology: A Quantitative Review. *PLoS ONE*, 9(10), e111436. https://doi.org/10.1371/journal.pone.0111436

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

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., & Anna di Palma, M. (2026). *Robustbase: Basic robust statistics* [R package version 0.99-7].

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. *Methods in Ecology and Evolution*, 7(4), 428–436. https://doi.org/10.1111/2041-210X.12502

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? *Global Ecology and Biogeography*, 12(5), 361–371. https://doi.org/10.1046/j.1466-822X.2003.00042.x

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

Peyrard, N., & Gimenez, O. (Eds.). (2022). *Statistical Approaches for Hidden Variables in Ecology*.

Pim, I. V. M., de Carvalho, L. M. F., & Prates, M. O. (2026). Spatial Confounding: A review of concepts, challenges, and current approaches. https://doi.org/10.48550/ARXIV.2602.17792

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 species 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. *Methods in Ecology and Evolution*, 4(1), 14–24. https://doi.org/10.1111/j.2041-210x.2012.00251.x

Schmitt, S., Pouteau, R., Justeau, D., De Boissieu, F., & Birnbaum, P. (2017). SSDM: An R package to predict distribution of species richness and composition based on stacked species distribution models. *Methods in Ecology and Evolution*, 8(12), 1795–1803. https://doi.org/10.1111/2041-210X.12841

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

Thuiller, W., Richardson, D. M., Pyšek, P., Midgley, G. F., Hughes, G. O., & Rouget, M. (2005). Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, 11(12), 2234–2250. https://doi.org/10.1111/j.1365-2486.2005.001018.x

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].

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.

White, J. W., Rassweiler, A., Samhouri, J. F., Stier, A. C., & White, C. (2014). Ecologists should not use statistical significance tests to interpret simulation model results. *Oikos*, 123(4), 385–388. https://doi.org/10.1111/j.1600-0706.2013.01073.x

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. (2023). 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