



HAL
open science

On the interpretations of joint modelling in community ecology

Giovanni Poggiato, Tamara Münkemüller, Daria Bystrova, Julyan Arbel,
James Clark, Wilfried Thuiller

► **To cite this version:**

Giovanni Poggiato, Tamara Münkemüller, Daria Bystrova, Julyan Arbel, James Clark, et al.. On the interpretations of joint modelling in community ecology. *Trends in Ecology & Evolution*, 2021, 36 (5), pp.391-401. 10.1016/j.tree.2021.01.002 . hal-03153558

HAL Id: hal-03153558

<https://hal.science/hal-03153558>

Submitted on 31 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the interpretations of joint modelling in community ecology

Giovanni Poggiato^{1,2,*}, Tamara Münkemüller¹, Daria Bystrova^{1,2}, Julyan Arbel²,
James Clark^{3,4,5} and Wilfried Thuiller¹

¹Univ. of Grenoble Alpes, CNRS, Univ. Savoie Mont Blanc, LECA, Grenoble, France

²Univ. of Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, France

³Univ. of Grenoble Alpes, Irstea, LESSEM, Grenoble, France

⁴Nicholas School of the Environment, Duke University, Durham, North Carolina
27708 USA

⁵Department of Statistical Science, Duke University, Durham, North Carolina
27708 USA

*Corresponding author: giov.poggiato@gmail.com

Abstract

Explaining and modelling species communities is more than ever a central goal of ecology. Recently, joint species distribution models (JSDMs), which extend species distribution models (SDMs) by considering correlations among species, have been proposed to improve species community analyses and rare species predictions while potentially inferring species interactions. Here, we illustrate the mathematical links between SDMs and JSDMs and their ecological implications and demonstrate that JSDMs, just like SDMs, cannot separate environmental effects from biotic interactions. We provide a guide to the conditions under which JSDMs are (or are not) preferable to SDMs for species community modelling. More generally, we call for a better uptake and clarification of novel statistical developments in the field of biodiversity modelling.

Keywords

Biodiversity modelling, Biotic interactions, Community data, Environmental niche, Joint Species Distribution Models

35 Highlights

36 In an era of global changes, developing reliable biodiversity models has become an important research area.

37 Species distribution models are the common tools to understand and predict the distributions of species across space
38 and time. However, they fail to explicitly account for species interactions.

39 To this aim, joint species distribution models were introduced to tease apart the effect of the environment from that of
40 species interactions, to improve rare species modelling, to account for functional traits, and to improve the predictive
41 power of biodiversity models.

42 Nevertheless, most announced advantages have remained unfulfilled, and there is still a need to better integrate the
43 effect of species interactions in the response of species to environmental change.

44 Glossary

45 **Covariates:** variables used to predict the response variables (see below). In this paper covariates represent the abiotic
46 conditions. A missing covariate is a variable that is not included in the model but has an important effect on the response
47 variables.

48 **Generalized linear model (GLM):** a flexible generalization of ordinary linear regression to predict a response variable
49 from a distribution in the exponential family (Poisson, binomial, etc.), and assuming that some known transformation of
50 the mean response is a linear function of predictor variables.

51 **Hierarchical model:** a statistical model written in multiple levels (hierarchical form). Hierarchical modeling allows sharing
52 information between entities (mostly species here) to facilitate parameter estimation, an advantage commonly referred
53 to as ‘borrowing strength’.

54 **Latent variable:** a variable not directly observed and usually introduced to model correlations between response
55 variables.

56 **Niche (fundamental, sensu Hutchinson):** the physiological dependence of the species on the environment.

57 **Niche (realized, sensu Hutchinson):** the observed relationships between the species and the environment. This is the
58 outcome of both the environmental effect and biotic interactions.

59 **Conditional predictions:** the prediction of the distribution of the value(s) of one or more response variable(s), given the
60 value(s) of one or more other response variable(s). Conditional predictions could be derived through the use of the
61 residual correlation matrix (see below).

62 **Joint predictions:** the prediction of the distribution of the joint values of two or more response variable(s). Joint
63 predictions could be derived through the use of the residual correlation matrix (see below).

64 **Marginal predictions:** the prediction of the distribution of the value(s) of one or more response variable(s), irrespective
65 of the value(s) of one or more other response variable(s). Marginal predictions are the typical output of SDMs and JSDMs.

66 **Regression coefficients:** the parameters that describe the relationships between the response variables and the
67 covariates. In (joint) species distribution models, they are interpreted as descriptions of species’ niches.

68 **Residual correlation matrix:** the correlation matrix between the response variables (see below) after accounting for the
69 effect of the covariates.

70 **Response variables:** the variables of interest to be modelled and predicted. In this article, they mostly represent the
71 presence-absences of species.

72 From ecological theory to biodiversity modelling

73

74 Understanding the ecological processes driving the distribution of life on Earth has always been a central goal in
75 ecology. This is more than ever crucial to project how biodiversity from various ecosystems will respond to global
76 changes. Researchers have long focused on the description of how species are spatially distributed and on the main
77 drivers explaining these distributions (Van Humboldt, early 1800s). It is now clear that three fundamental ecological
78 processes determine whether a species can occupy a site and maintain viable populations: limitation by abiotic
79 conditions, biotic interactions and dispersal limitation (see Box 1, [1–3]).

80 While we theoretically know the complex processes that shape communities, their relative importance is generally
81 unknown, making it difficult to predict how these communities will respond to environmental changes [4]. Statistical
82 ecology has arisen as a discipline that moves away from describing biodiversity patterns towards modelling the output of
83 the ecological processes that generate these patterns [5]. Notably, the so called biodiversity models predict the
84 distribution and abundance of multiple species based on a set of environmental conditions [6]. To properly interpret the
85 parameters of these models, and to guarantee the quality and reliability of their predictions, it is key to understand how
86 they integrate the fundamental ecological processes shaping species ranges and community structure [6].

87 Species distribution models (SDMs, [7]), the most common statistical tool to model species distributions, early on raised
88 debates on how to interpret their parameters in light of ecological processes. SDMs relate the presence-absence or the
89 abundance of a species to environmental **covariates** (see Glossary) and use this relationship to predict its distribution in
90 space and/or time [8]. Originally, most SDMs relied on **generalized linear models** (GLMs, [9]), with the deterministic
91 regression coefficients for the relationship of the species with the environment, and a residual part for the unexplained
92 variation. Usability and increasing data availability have boosted the use of SDMs [10] in ecology and conservation.
93 However, by modelling the observed species-environment relationship for each species independently, they only capture
94 the combined effects of both abiotic and biotic environments (i.e. the so-called **realized niche**, see Glossary). The pure
95 effects of the abiotic environment are not separated from the effects of species interactions and the **fundamental niche**
96 remains unknown [11], which potentially distorts predictions [12]. Despite these issues, SDMs were also used to predict
97 communities by summing over single species predictions (e.g. stacked SDMs,[13,14]), eventually with some additional
98 constraints to account for biotic filters [15,16]. However, this two-step procedure allows neither for error propagation
99 nor for joint parameter estimations and is conceptually flawed as the realized niche estimated from SDMs inherently
100 accounts for biotic constraints.

101 In the last decade, multi-species distribution models (MSDMs) and joint species distribution models (JSDMs) were
102 introduced to overcome the assumption of SDMs that species' distributions are independent of each other. MSDMs are
103 extensions of GLMs, where the estimated species-environment relationship between species are connected [17]. By
104 modelling the **regression coefficients hierarchically**, they consider commonalities between species, so that, for instance,
105 species with similar traits respond similarly to the environment [18–21]. As a result, **rare species** could 'borrow strength'
106 from common species if they do not behave fundamentally differently [17]. JSDMs, as a further extension of GLMs (but

107 see [22,23] for other approaches), infer a correlation matrix from the residuals (hereafter **residual correlation matrix**)
108 that reflects species co-occurrence patterns not explained by the environmental predictors [24]. Residual correlations
109 may arise from model mis-specifications, missing covariates or species interactions (Box 2, review in [25–27]). Thus,
110 JSDMs intuitively have been proposed to simultaneously explore, and potentially disentangle, limitations by abiotic
111 conditions and biotic interactions [25]. Although these new statistical models are receiving increasing attention, there is
112 so far a lack of clarification on both the ecological processes they incorporate and on their specific commonalities and
113 advantages with respect to SDMs. Some of the widespread beliefs, such as the idea that JSDMs can "account for biotic
114 interactions in species distribution models" [28], have never been proven.

115 In this paper, we first reunify SDMs, MSDMs and JSDMs under a common notation to better identify their similarities and
116 differences (Box 2). Like MSDMs, JSDMs can also model the regression coefficients hierarchically, but since this is not
117 always implemented (see [28,29]), we consider here JSDMs and MSDMs as two different extensions of SDMs. Second,
118 we tease apart the true advantages of JSDMs from false beliefs and possible misinterpretations, therefore allowing to
119 interpret these models in the light of fundamental ecological processes. Specifically, we address the following questions:

- 120 1. Can JSDMs and MSDMs improve the estimation of species' fundamental niches?
- 121 2. What can the residual correlation matrix tell us about biotic interactions?
- 122 3. When and why do JSDMs outperform SDMs?

123 This opinion piece differs from previous papers on JSDMs in that we neither introduce new methodological developments,
124 nor compare these models with data. Instead, we rigorously and mathematically demonstrate how to interpret MSDMs
125 and JSDMs, providing a guide on why and when these models should be preferred to SDMs. Our aim is to enable users to
126 serenely choose and apply these models to make the best of their potential.

127

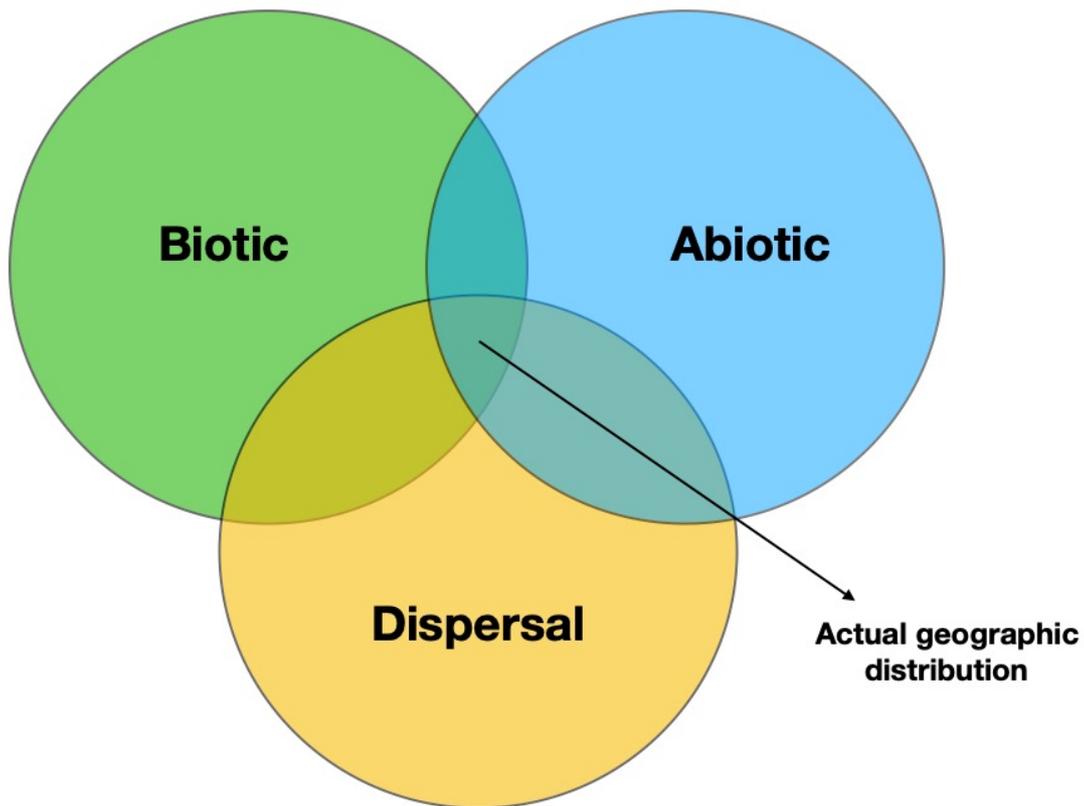
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146

Box 1: The fundamental ecological processes shaping species distribution

Three main conditions need to be met for a species to occupy a site and maintain viable populations (see Figure I, [1–3]):

- the species has to physically reach a site, i.e. to access a region [53];
- the abiotic environmental conditions (e.g. temperature or soil pH) must be physiologically suitable for the species;
- the biotic environment (i.e. interactions with other species) must be suitable for the species.

The first condition is a matter of species' capacity to disperse to a site from other occupied areas. It is related to the biogeographic history of the species, and thus to all factors limiting its distribution from the place where it first originated, such as barriers to migration, biotic and abiotic dispersal vectors or rare long-distance dispersal events. The second condition depends on abiotic conditions, which means that the combination of abiotic environmental variables at the site are within the range of environmental conditions that the species requires to grow and maintain viable populations. These suitable environmental conditions represent the species' fundamental niche [54]. The third condition concerns biotic interactions, i.e. interactions with other organisms, either neutral, positive or negative, symmetric or asymmetric, which themselves are influenced by the environment through their influence on all organisms in the local community. The environmental conditions where a species can therefore survive accounting for other species are called the species' realized niche [54]. This is what we observe when sampling the distribution of a species in the wild. In a given site, these processes influence all species from the regional pool to create local communities that represent a relevant scale to investigate biodiversity distribution (e.g. few square meters for plants, a soil core for microbes) [6, 55].



147
148
149
150

Figure I. The three factors that shape the observed species distribution [3]. The blue circle describes the fundamental niche, while the realized niche is represented by the intersection of the green and blue circle.

151

152 Box 2: Mathematical notations from SDMs to JSDMs

153 Focusing on presence-absence data, the **response variable** $y_{ij} = 1$ if species $j = 1, \dots, S$ is present at site $i = 1, \dots, n$ and
154 0 otherwise. All models relate the S -dimensional vector \mathbf{y}_i to a set of K environmental covariates $\mathbf{x}_i = \{x_{ik}\}_{k=1}^K$.

155 SDMs

156 GLMs can model presence-absence data using a probit link. Probit regression can be described as a **latent variable** model
157 with the probability of presence being modelled as the probability of a latent Gaussian variable to be positive [56]. Each
158 species j is modelled independently, with:

$$\begin{aligned} y_{ij} &= I(z_{ij} > 0) \\ z_{ij} &= \boldsymbol{\beta}_j^T \mathbf{x}_i + \varepsilon_{ij} \\ \varepsilon_{ij} &\stackrel{iid}{\sim} N(0,1) \end{aligned} \quad [I]$$

160 where $I(\cdot)$ is the indicator function and $N(0,1)$ is the standard univariate Gaussian distribution. The variance term is
161 restricted to 1 for identifiability reasons. The regression coefficients $\boldsymbol{\beta}_j \in \mathbb{R}^K$ give the response of species j to the abiotic
162 covariates [26]. The probability of species j to be present at site i is thus $probit(y_{ij} = 1) = \boldsymbol{\beta}_j^T \mathbf{x}_i$.

163 MSDMs

164 MSDMs model the regression coefficients of [I] hierarchically: $\boldsymbol{\beta}_j \stackrel{iid}{\sim} N_K(\boldsymbol{\mu}, \mathbf{V})$, where N_K is the multivariate K -
165 dimensional Gaussian distribution. As a consequence, species' responses to the environment are shared across species,
166 which can be of particular interest for rare species. Coefficients can also be constrained by trait and/or phylogenetic
167 information (by including them in $\boldsymbol{\mu}$ and/or \mathbf{V}).

168 JSDMs

169 Most JSDMs extend GLMs in what is commonly called the multivariate probit model [57]. This model is based on the same
170 latent variable idea as described above, but uses an S -dimensional vector:

$$\begin{aligned} y_{ij} &= I(z_{ij} > 0) \\ \mathbf{z}_i &= \boldsymbol{\beta} \mathbf{x}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\varepsilon}_i &\stackrel{iid}{\sim} N_S(0, \mathbf{R}) \end{aligned} \quad [II]$$

172 where \mathbf{R} is a correlation matrix, and not a covariance matrix, for identifiability reasons. \mathbf{R} describes the residual
173 correlation among taxa, and reflects species co-occurrence patterns not explained by the selected abiotic covariates. $\boldsymbol{\beta}$
174 is a $K \times S$ matrix whose columns $\boldsymbol{\beta}_j$ are the species-specific response to the environment. Importantly, \mathbf{R} does not affect
175 the marginal probability of presence of each species, $probit(y_{ij} = 1) = \boldsymbol{\beta}_j^T \mathbf{x}_i$. Thus, **marginal predictions** only depend
176 on the estimated regression coefficients for both SDMs and JSDMs [57].

177 Many JSDMs use latent factors to reduce the dimension of \mathbf{R} (see Appendix A). JSDMs can also model the regression
178 coefficients hierarchically, therefore integrating the advantages of MSDMs and obtaining highly flexible and complex
179 models [e.g. 26].

180 Reconciling SDMs, MSDMs and JSDMs

181 Model [I] can be written in the same way as [II], but with a diagonal residual correlation $\boldsymbol{\varepsilon}_i \stackrel{iid}{\sim} N(0, \mathbf{I})$. In other words,
182 the only difference is that SDMs and MSDMs assume independent residuals, while JSDMs allow for correlations between
183 them.

184

185 Question 1: Can JSDMs and MSDMs improve the estimation of species' 186 fundamental niches?

187 Characterizing the fundamental niches with observational data, teasing apart the effects of abiotic and biotic ecological
188 processes on species distributions and community assembly, is a critical challenge to predict the future of biodiversity
189 [6,12]. Since they model multiple species together, we may believe that MSDMs and JSDMs can better fit the response
190 of each species to environmental covariates by using information on the other species, and thus ultimately, may allow

191 to retrieve the fundamental niche of species. JSDMs, in particular, have been repeatedly suggested to separate abiotic
192 and biotic conditions and -if this suggestion was right- should allow to approach species' fundamental niches [25]. But
193 can the models hold these promises? Below, we outline why this is not the case, neither for JSDMs nor for MSDMs.

194 In both SDMs and JSDMs, the species' niche (approximated by the regression coefficients) is estimated through
195 minimizing species-specific regression residuals. In other words, should we infer a residual correlation matrix from the
196 residuals (JSDMs) or not (SDMs), the estimated niches coincide. In Appendix B, we demonstrate that the estimates of
197 the regression coefficients are identical for JSDMs and SDMs, at least for Gaussian data. The uncertainty around these
198 estimates might differ, but it is difficult to prove whether one is always greater or lower than the other one. Extending
199 this analytical proof to other data types is challenging. However, empirical comparisons for presence-absence data also
200 showed no differences in the regression coefficients estimates between a comparable SDM and a JSDM approach
201 (same package, same inference, only the estimation of correlation matrix differed, Box 3). Indeed, since JSDMs model
202 the expected distribution of species as exclusively dependent on the environmental conditions (through the regression
203 term), while all the other factors potentially influencing species' distributions (e.g., missing predictors, biotic
204 interactions) can only impact the (co)variation (given by the residual correlations) around this expected value. In
205 consequence, JSDMs, just like SDMs, do not control for the effect of other species when inferring species niches, and
206 thus only retrieve the realized niches (see Appendix B for a further discussion). Importantly, it also means that for a set
207 of modelled species, the correlations between the residuals of independent SDMs closely approximate the residual
208 correlation matrix of a JSDM (Box 3), with the advantage of the latter to propagate model uncertainties in a more
209 correct way and the former to be easier to apply ([30], page 11).

210

211 In contrast, MSDMs (and JSDMs with hierarchical coefficients) estimate different species niches than SDMs,
212 especially for rare species. This is, however, not linked to species interactions. Thanks to the hierarchical part of the
213 model, MSDMs share information between species [17], and can constrain, for example, two phylogenetically or
214 functionally closely-related species to respond similarly to the environment (i.e. similar niches) [18]. Taking phylogeny
215 and/or functional traits into account allows to test their importance in shaping species distribution [26]. MSDMs have
216 been considered as a great improvement for modelling rare species for which niche estimates are difficult to obtain due
217 to low sample size. Forcing niche estimates to resemble those of closely related common species circumvents this
218 problem. However, this advantage only holds if rare and common species respond in the same way to the environment
219 and leads to false estimates if this assumption is wrong. While the assumption may hold for hardly detectable species,
220 there are many ecological reasons why truly rare species differ from common species in their response to the
221 environment. Species can be rare because they are specialized to specific conditions, or because they are relicts [31].
222 Consistently, studies examining the predictive performances of SDMs and MSDMs for rare species suggest that gains in
223 performance are context dependent [32].

224

225

226

227 Question 2: What can the residual correlation matrix tell us about biotic
228 interactions?

229 Inferring biotic interactions from co-occurrence patterns is a particularly hot topic in current ecological research
230 [33–35]. In this context, some seminal articles have emphasized the potential of JSDMs to capture the signal of biotic
231 interactions in the residual correlation matrix [36]. Although other authors entirely rejected this proposition [37], many
232 are still left with the idea that the residual correlation matrix may ‘hint at a biological interaction between species’ [24]
233 or ‘inform about biotic constraints’ [28]. Ongoing discussions turn around the scale mismatch between the true
234 interactions and the modelled environment [37], the influence of missing predictors [38] and the symmetric constraint
235 of correlation matrices [39] as important limitations of JSDMs, while others object that the signal that biotic interactions
236 leave on co-occurrence data prevents any inference, whatever the method used [40,41]. Here, our argument focuses on
237 a more fundamental limitation of JSDMs. Indeed, if the regression coefficients only estimate species’ realized niches
238 (question 1), not much of the signal of biotic interactions can remain in the residuals (even without any of the above-
239 mentioned problems) and what remains strongly depends on the characteristics of these interactions.

240 When considering two species A and B with overlapping fundamental niches (Figure 1.a) and assuming that A is
241 the strongest competitor, then B will be excluded from the overlapping area (Figure 1.b). The famous Barnacles in the
242 low tide area are a typical example where *Balanus* (species A) excludes *Chtamalus* (species B) from large parts of its
243 fundamental niche [42]. Applied to this data, SDMs, MSDMs and JSDMs will (wrongly) attribute the absence of species
244 B to the abiotic conditions. Since the realized niches entirely explain the negative correlation between the two species,
245 no information on biotic interactions is left in the residuals, preventing JSDMs (and SDMs and MSDMs when correlating
246 their residuals) to suggest a competitive interaction from the residual correlation matrix (the same logic applies for
247 facilitation).

248 In contrast, let’s assume that species A and B compete symmetrically, excluding each other about half the time in
249 the overlapping region. An example is the unshaded reaches of Augusta Creek, Michigan (USA) (see [43], for a
250 terrestrial example), where, at high velocity sites, the likelihood that a site will be dominated by the macroalga
251 *Cladophora glomerata* or by epilithic microalgal lawn inhabited by several species of sessile grazers (e.g. the caddisflies
252 *Leucotrichia pictipes*) is determined by who establishes first [44]. In this case, biotic interactions do not only affect the
253 realized niches (that is decreased in magnitude, Figure 1.c), but also the species covariation around the expected
254 distributions. So, the realized niches cannot fully explain the negative correlation between the species, and this part will
255 appear in the residuals. Under the assumption of a well-specified model, JSDMs will identify the negative residual
256 correlation between the species, which can truly be attributed to the competitive interaction between A and B. Finally,
257 notice that a common response to an unmeasured environmental covariate (e.g. both species prefer a warm climate)
258 might lead to a positive correlation even if the two species do not interact [38,45].

259 While abundance data may provide more informative than presence-absence to detect variations around the
260 realized niches, environmental and biotic effects will still be confounded in the estimation of species responses to the

261 environment (i.e. in β). Therefore, when partitioning species covariance into shared environmental preferences and
262 residual co-occurrence patterns [24], one has to remember that the former are due to the realized niche and not due to
263 the fundamental one, with the consequences that the latter only reflects a small part of the signal of biotic interactions.
264 In conclusion, even if biotic interactions are an important process, their signal on co-distributions will be either fully or
265 partly hidden in β .

266

267 As a statistical side note, we need to keep in mind that, even in the specific case that the residual correlation matrix
268 R really captures an imprint of species interactions (which is unlikely for real data [40]), it represents the marginal
269 correlations among the residuals and thus mixes the direct (e.g. competition) and indirect (e.g. a shared predator)
270 associations between species. To conclude on direct associations between two species, we need to calculate the precision
271 matrix instead $\Omega = R^{-1}$, that represents the (residual) partial correlation between species while controlling for the
272 effects of the other species [46,47].

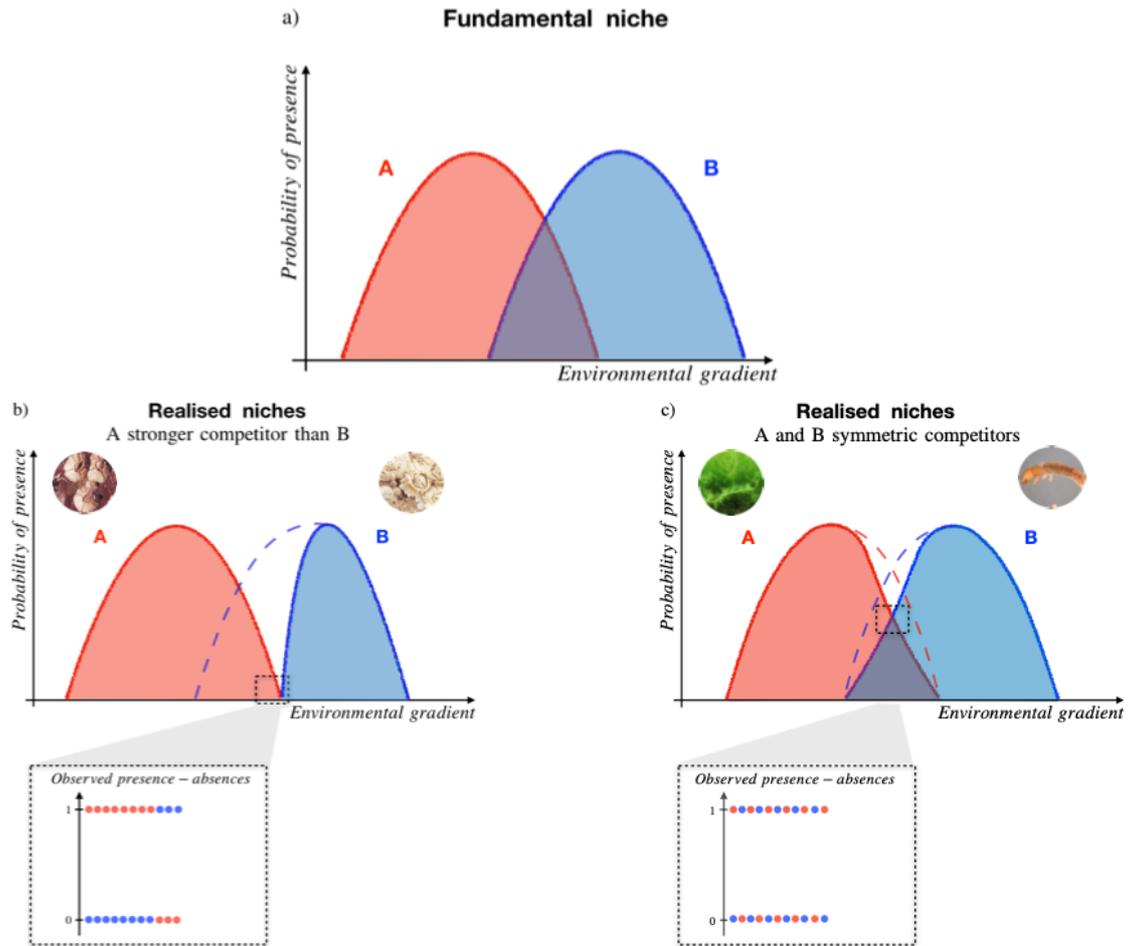


Figure 1. Effects of biotic interactions on species' niches. The top panel (a) shows the fundamental niche of two species (A and B). The bottom panels show two extreme scenarios of competition and the resulting realized niches (the fundamental niches are represented with the dashed line). SDMs, MSDMs and JSDMs retrieve the realized niches only. On the left (b), wherever the fundamental niches of A and B overlap, A excludes B, even under weak abiotic conditions but still suitable for both species (e.g. *Balanus* and *Chtamalus* in [32]). The observed presences and absences in the interaction zone (the dashed rectangle) reflect this dichotomy due to competition exclusion, with little or no variation around the expected distribution where A is present, and B is absent. Since the realized niches entirely explain the negative correlation between A and B, JSDMs will not identify a negative residual correlation. On the right (c), species A and B compete in a symmetric way, by excluding each other about half of the times where their niches overlap (*Cladophora glomerata* and *Psychomyia flavida* in [34]). If the expected distribution is the same for both species (their observed probability of occurrence in the conflict region is 0.5), their covariation around it is highly significant in terms of interactions, since the two species never co-occur. Here, JSDMs (but also MSDMs and SDMs when correlating their residuals) will detect a negative residual correlation since the realized niches do not fully explain the negative correlation between species.

273

274

Box 3: An empirical example

275

276

277

278

279

280

281

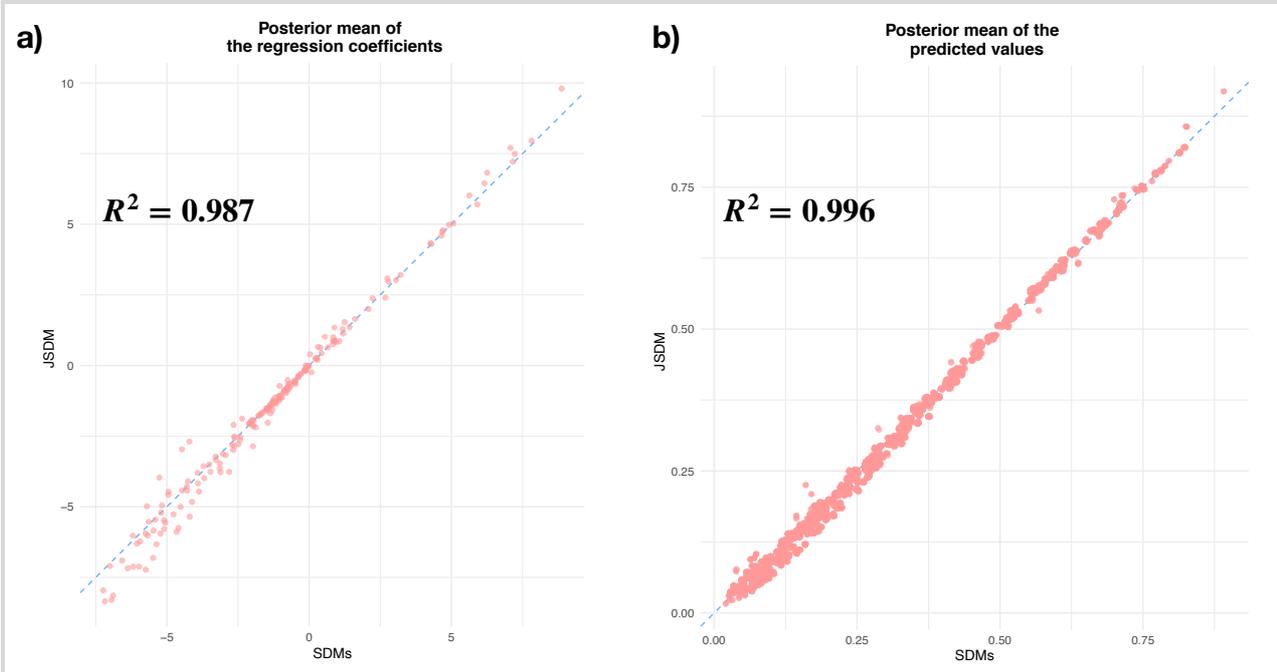
282

283

To elucidate the differences between SDMs and JSDMs in an empirical case-study, we focus on the response of alpine plants to snowmelt dates in Aravo (French Alps, [58]), as also done by [25]. We considered 65 species (all with more than 4 occurrences) at 75 sites, with snowmelt dates as the environmental covariate (linear and quadratic term, using orthogonal polynomials to reduce correlation among the covariates). The data are available from the R package *ade4* [59]. To strictly focus on the effect of the residual correlation matrix on the estimates of the model, we avoid the confounding effects that can affect our results (e.g. choice of priors, different inference strategy, different implementation) by using the R package *BayesComm* [60], that allows us to choose whether residuals are considered as independent (multiple SDMs) or not (JSDM) and does not model the regression coefficients hierarchically (see Appendix C for the code and further details).

284 Environmental niche and prediction

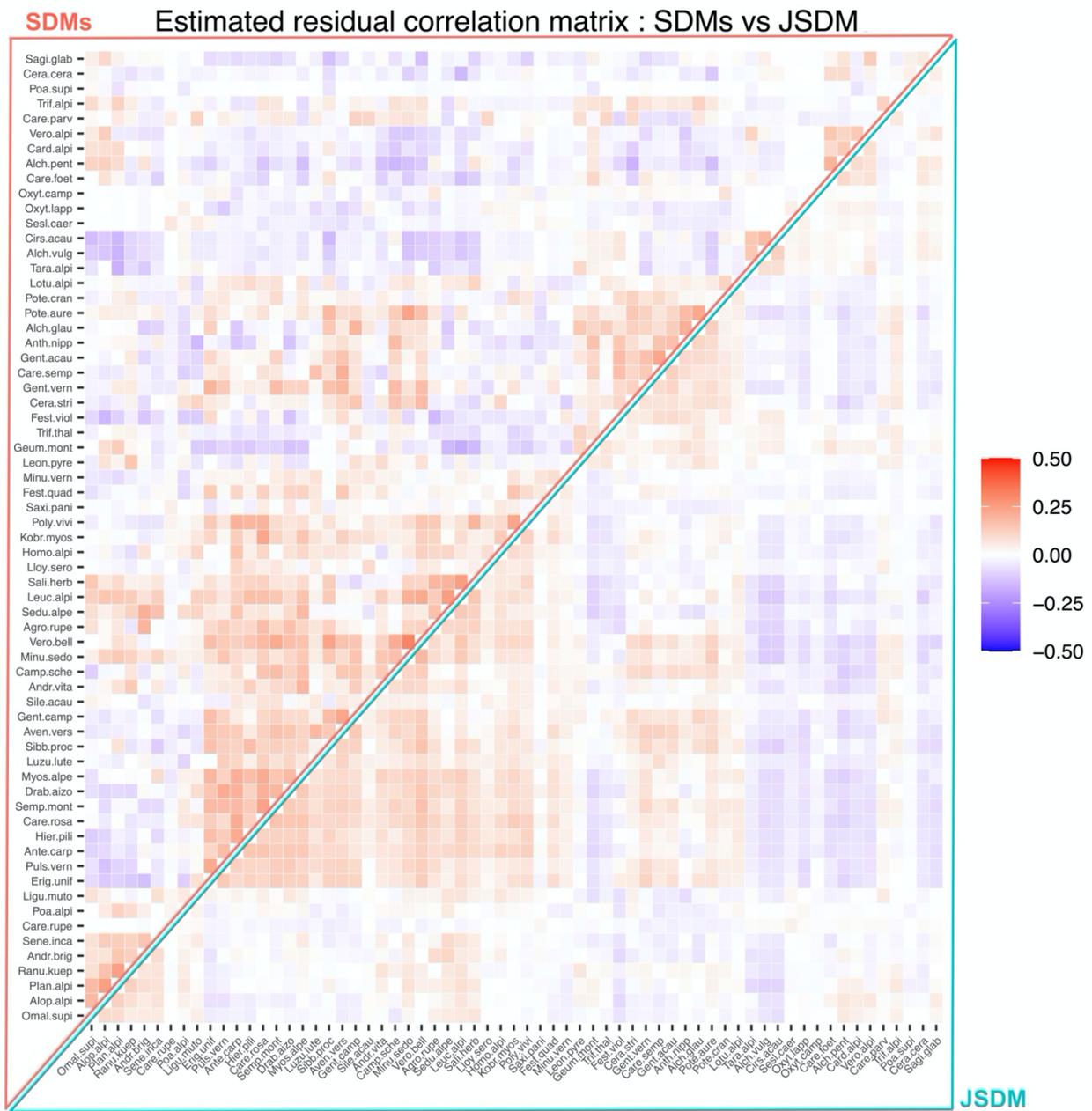
285 SDMs and JSDM estimated the same environmental niches. We can see almost no difference between the regression
286 coefficients (Figure 1a, $R^2 = 0.987$ between the posterior means of the two models), and, in this case, the credible
287 intervals are also very similar (see Appendix D). As a natural consequence, the marginal predictions are extremely close
288 too ($R^2 = 0.996$ between the posterior predictive means of the two models Figure 1b).



289
290 Figure 1. On the left (a), relationship between the posterior means of the regression coefficients for all species estimated
291 by SDMs on the x-axis and JSDM on the y-axis. Each point corresponds to a single coefficient (i.e. intercept, linear and
292 quadratic term for snowmelt date) for a single modelled species. On the right (b), relationship between the posterior
293 means of the predicted probability of presence. Each point corresponds to the predicted probability of presence of a
294 single species at a single site. The blue dashed lines correspond to the 1:1 line.

295 Residual correlation matrix

296 We compared the correlation between the residuals of SDMs and the residual correlation matrix inferred by JSDM. Since
297 a JSDM is a probabilistic model that allows error propagation, it is clearly preferable over multiple SDMs to infer a
298 correlation matrix from the residuals. Here, we carry out this computation only to show the similarity between the two
299 approaches. The residuals of the SDMs are calculated as the difference between the latent variables and the regression
300 term, to stick to JSDMs definition of residuals (see Appendix D for other kinds of residuals). The residual correlation
301 matrices estimated by SDMs and JSDM are very similar ($R^2 = 0.862$ between the estimates of the two models, 95%
302 credible intervals match in 98% of the cases, Figure II).



303

304 Figure 1. Comparison of residual correlation matrices from multiple independent SDMs (post-hoc calculated) and a JSDM.
 305 SDMs residual correlations are represented in the upper triangular matrix, JSDM correlations in the lower triangular matrix.
 306 $R^2 = 0.862$ between the estimates of the two models, 95% credible intervals match in 98% of the cases (either both positive, or
 307 both negative, or both overlapping zero).

308

309 **Question 3: When and why do JSDMs outperform SDMs?**

310 One of the major objectives of species distribution models is to predict community compositions under new, eventually
 311 future, abiotic conditions. For SDMs, MSDMs and JSDMs, the marginal prediction of each species (i.e. unconditionally
 312 on the others) is only driven by whether the new environmental conditions are suitable for the species, even if the
 313 marginal predictions of MSDMs (and JSDMs with hierarchical coefficients) can differ, for the reasons highlighted in
 314 question 1. However, and importantly, this implies that all methods will project future species distributions without

315 accounting for biotic interactions, although they are likely to play a critical role in the reorganization of communities as
316 a result of global changes [4]. Since the estimated regression coefficients do not change whether species are modelled
317 jointly or not, the marginal predictions do not change either, but have different uncertainties. In other words, fitting
318 and predicting each species independently (SDMs), or with a JSDM, will lead to the same marginal predictions (as
319 shown in Box 2, and see also Figure 2 of [48]). This explains why [29], [49] and [50] did not find clear differences in the
320 predictive performance between JSDMs and SDMs neither at the species nor at the community level.
321 As a consequence, species richness predictions, that sum the mean marginal probabilities of SDMs vs. JSDMs, will
322 inevitably coincide [51]. However, since the variance of a sum of correlated variables is not merely the sum of the
323 variances, the residual correlation matrix does affect the uncertainty around the predicted richness. This is highlighted in
324 the third box of [25], where the authors show that if the residual correlation across species was ignored (SDMs), the
325 credible intervals were too narrow to capture the observed value of species richness.

326 The inferred residual correlation matrix still provides information on co-occurrence patterns that can be used to
327 improve predictions. Indeed, JSDMs can leverage on the shared residual structure (that does not need be related to biotic
328 interactions) to better estimate the probability of species co-occurrences and to provide **joint and conditional predictions**
329 [52]. In other words, we should not interpret the residual correlation matrix, but rather exploit it.

330 When we commonly observe two co-occurring species, our expectation to see one when we see the other increases. This
331 is what is called conditional prediction: the probability of presence of one (or more) species, given the presence, or
332 absence, of one (or more) other species. JSDMs can exploit the residual correlation matrix to provide such predictions,
333 where the observed species are basically used as predictor of the unobserved species. Conditional predictions can be of
334 a great asset in several ecological applications. For instance, in invasion ecology, we could use JSDMs to determine the
335 probability of invasive species to be present given the distribution of native species. Not only can they improve
336 predictions, but they can also provide a better understanding of the studied system [49]. Studying how co-occurrence
337 probabilities vary along environmental gradients, can also provide important knowledge on communities. Under the
338 independence assumption of SDMs, the probability of co-occurrence is simply the product of marginal occurrence
339 probabilities, but this estimate fails to integrate interspecific correlations. JSDMs are instead a potentially suitable tool
340 for this task, since the probability of co-occurrence also depends on the residual correlations: positively correlated
341 residuals lead to higher probability of co-occurrence than SDMs and vice-versa. Importantly, accounting for residual
342 correlations to predict species co-occurrences inherently requires to have meaningful residuals that reflect underlying
343 mechanisms (e.g. dispersal limitations, biotic interactions). In the extreme case of residual correlations completely driven
344 by model error and/or misspecification, joint and conditional predictions might not improve, or even worsen, co-
345 occurrence probabilities, especially when extrapolating in space and time.

346

347 Concluding remarks

348 The recent emergence of MSDMs and JSDMs has raised expectations to integrate some fundamental ecological processes
349 in species distribution modelling, in particular to disentangle biotic interaction effects from environmental effects on

350 species co-distributions. However, we show that these models do not account for biotic interactions when predicting
351 distribution patterns, instead they infer correlations among taxa after accounting for environmental covariates.
352 Therefore, they can only infer species' realized niches, and marginal predictions are not improved. We emphasize that
353 we should not interpret the residual correlation matrix from a pure interaction perspective (whose ability to infer biotic
354 interactions is strongly context dependent), but leverage on it, using conditional predictions, the under-exploited
355 advantage of JSDMs. Hierarchical models, like MSDMs (or JSDMs with hierarchical effects) allow to test for the
356 importance of traits and/or phylogeny and might bring interesting information notably for species that are difficult to
357 detect, but the assumption behind these hierarchical effects need to be clearly understood by users.

358 Outstanding Questions

359 To what extent do biotic interactions leave an imprint in co-occurrence patterns to enable them to be distinguished
360 from environmental effects? Under what conditions or types of interactions are these imprints detectable and what prior
361 information would be needed to help the inference?

362
363 How can we better harness temporal data from multiple sources to exploit theory-based temporally-dynamic joint species
364 distributions? Can dynamic JSDMs model species rich communities or would they be restricted to specific cases?

365
366 How can conditional dependencies in JSDMs or related graphical models be better used to provide conditional predictions
367 for invasion risk assessment, re-introduction analyses or rare species modelling?

368
369 How can we account for biotic interactions when predicting species distribution and community compositions? How can
370 we make best use of prior information on forbidden or known interactions?

371 Acknowledgments

372 We thank Frederic Gosselin, Frederic Mortier, Laura Pollock, Bjorn Reineking, and Stephane Robin for the insightful
373 discussions on the properties of JSDMs. G.P., D.B., W.T., and T.M. were supported by the GAMBAS project funded by
374 the Agence Nationale pour la Recherche (ANR-18-CE02-0025). W.T. and J.S.C. also acknowledge support from the Pro-
375 gramme d'Investissement d'Avenir under project FORBIC (18-MPGA-0004). This work also received funding from the
376 ERA-Net BiodivERsA - Belmont Forum, with the national funder Agence Nationale pour la Recherche (FutureWeb: ANR-
377 18-EBI4-0009) to W.T. and the National Science Foundation (NSF grant: 1854976) to J.S.C.

378 References

- 379 1 Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecol. Lett.* 3, 349–361
- 380 2 Lortie, C.J. *et al.* (2004) Rethinking plant community theory. *Oikos* 107, 433–438
- 381 3 Soberón, J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecol. Lett.* 10, 1115–
382 1123
- 383 4 Tylianakis, J.M. *et al.* (2008) Global change and species interactions in terrestrial ecosystems. *Ecol. Lett.* 11, 1351–
384 1363
- 385 5 Gimenez, O. *et al.* (2014) Statistical ecology comes of age. *Biol. Lett.* 10, 20140698
- 386 6 Thuiller, W. *et al.* (2013) A road map for integrating eco-evolutionary processes into biodiversity models. *Ecol.*
387 *Lett.* 16, 94–105

- 388 7 Guisan, A. and Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* 8, 993–1009
389
- 390 8 Guisan, A. *et al.* (2017) *Habitat Suitability and Distribution Models: With Applications in R*, Cambridge University Press.
391
- 392 9 McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models, Second Edition*, Chapman & Hall.
- 393 10 Yates, K.L. *et al.* (2018) Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.* 33, 790–802
394
- 395 11 Araújo, M.B. and Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *J. Biogeogr.* 33, 1677–1688
396
- 397 12 Wisz, M. *et al.* (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biol. Rev. Camb. Philos. Soc.* 88, 15–30
398
- 399 13 Guisan, A. and Rahbek, C. (2011) SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *J. Biogeogr.* 38, 1433–1444
400
- 401 14 Calabrese, J.M. *et al.* (2014) Stacking species distribution models and adjusting bias by linking them to macroecological models. *Glob. Ecol. Biogeogr.* 23, 99–112
402
- 403 15 Staniczenko, P.P.A. *et al.* (2017) Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. *Ecol. Lett.* 20, 693–707
404
- 405 16 D’Amen, M. *et al.* (2018) Improving spatial predictions of taxonomic, functional and phylogenetic diversity. *J. Ecol.* 106, 76–86
406
- 407 17 Ovaskainen, O. and Soininen, J. (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology* 92, 289–295
408
- 409 18 Pollock, L.J. *et al.* (2012) The role of functional traits in species distributions revealed through a hierarchical model. *Ecography (Cop.)* 35, 716–725
410
- 411 19 Jamil, T. *et al.* (2013) Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *J. Veg. Sci.* 24, 988–1000
412
- 413 20 Brown, A.M. *et al.* (2014) The fourth-corner solution – using predictive models to understand how species traits interact with the environment. *Methods Ecol. Evol.* 5, 344–352
414
- 415 21 Carboni, M. *et al.* (2018) Functional traits modulate the response of alien plants along abiotic and biotic gradients. *Glob. Ecol. Biogeogr.* 27, 1173–1185
416
- 417 22 Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods Ecol. Evol.* 6, 465–473
418
- 419 23 Vanhatalo, J. *et al.* (2020) Additive multivariate Gaussian processes for joint species distribution modeling with heterogeneous data. *Bayesian Anal.* 15, 415–447
420
- 421 24 Pollock, L.J. *et al.* (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol. Evol.* 5, 397–406
422
- 423 25 Warton, D.I. *et al.* (2015) So Many Variables: Joint Modeling in Community Ecology. *Trends Ecol. Evol.* 30, 766–779
424
- 425 26 Ovaskainen, O. *et al.* (2017) How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* 20, 561–576
426
- 427 27 Clark, J.S. *et al.* (2017) Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. *Ecol. Monogr.* 87, 34–56
428

- 429 28 Wilkinson, D.P. *et al.* (2019) A comparison of joint species distribution models for presence-absence data.
430 *Methods Ecol. Evol.* 10, 198–211
- 431 29 Norberg, A. *et al.* (2019) A comprehensive evaluation of predictive performance of 33 species distribution models
432 at species and community levels. *Ecol. Monogr.* 89, 834–848
- 433 30 Tikhonov, G. (2018) , Bayesian latent factor approaches for modeling ecological species communities. , Helsinki :
434 Helsingin yliopisto,
- 435 31 Gaston, K.J. (1994) What is rarity? In *Rarity* pp. 1–21, Springer
- 436 32 Nieto-Lugilde, D. *et al.* (2018) Multiresponse algorithms for community-level modelling: Review of theory,
437 applications, and comparison to species distribution models. *Methods Ecol. Evol.* 9, 834–848
- 438 33 Morales-Castilla, I. *et al.* (2015) Inferring biotic interactions from proxies. *Trends Ecol. Evol.* 30, 347–356
- 439 34 Sander, E. *et al.* (2017) Ecological Network Inference From Long-Term Presence-Absence Data. *Sci. Rep.* 7, 7154
- 440 35 Freilich, M. *et al.* (2018) Species co-occurrence networks: Can they reveal trophic and non-trophic interactions in
441 ecological communities? *Ecology* 99, 690–699
- 442 36 Ovaskainen, O. *et al.* (2016) Using latent variable models to identify large networks of species-to-species
443 associations at different spatial scales. *Methods Ecol. Evol.* 7, 549–555
- 444 37 Clark, J.S. *et al.* (2014) More than the sum of the parts: forest climate response from joint species distribution
445 models. *Ecol. Appl.* 24, 990–999
- 446 38 Kissling, W.D. *et al.* (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages
447 at large spatial extents. *J. Biogeogr.* 39, 2163–2178
- 448 39 Dormann, C.F. *et al.* (2018) Biotic interactions in species distribution modelling: 10 questions to guide
449 interpretation and avoid false conclusions. *Glob. Ecol. Biogeogr.* 27, 1004–1016
- 450 40 Blanchet, F.G. *et al.* (2020) Co-occurrence is not evidence of ecological interactions. *Ecol. Lett.* 23, 1050–1063
- 451 41 Holt, R.D. (2020) Some thoughts about the challenge of inferring ecological interactions from spatial data.
452 *Biodivers. Informatics* 17, 61–85
- 453 42 Connell, J.H. (1961) The Influence of Interspecific Competition and Other Factors on the Distribution of the
454 Barnacle *Chthamalus Stellatus*. *Ecology* 42, 710–723
- 455 43 Palmer, T.M. *et al.* (2003) Competition and Coexistence: Exploring Mechanisms That Restrict and Maintain
456 Diversity within Mutualist Guilds. *Am. Nat.* 162, S63–S79
- 457 44 Hart, D. (1992) Community organization in streams: the importance of species interactions, physical factors, and
458 chance. *Oecologia* 91, 220–228
- 459 45 Momal, R. *et al.* (2020) Accounting for missing actors in interaction network inference from abundance data.
460 *arXiv* DOI: stat.AP/2007.14299
- 461 46 Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods Ecol. Evol.* 6,
462 465–473
- 463 47 Popovic, G. *et al.* (2019) Untangling direct species associations from indirect mediator species effects with
464 graphical models. *Methods Ecol. Evol.* 10, 1571–1583
- 465 48 Chen, D. *et al.* (2017) , Deep multi-species embedding. , in *IJCAI International Joint Conference on Artificial
466 Intelligence*, pp. 3639–3646
- 467 49 Zurell, D. *et al.* (2020) Testing species assemblage predictions from stacked and joint species distribution models.
468 *J. Biogeogr.* 47, 101–113

- 469 50 Caradima, B. *et al.* (2019) From individual to joint species distribution models: A comparison of model complexity
470 and predictive performance. *J. Biogeogr.* 46, 2260–2274
- 471 51 Gelfand, A.E. and Shirota, S. (2019) Clarifying species dependence under joint species distribution modeling.
472 *bioRxiv* DOI: 10.1101/744359
- 473 52 Wilkinson, D.P. *et al.* (2020) Defining and evaluating predictions of joint species distribution models. *Methods*
474 *Ecol. Evol.* <https://doi.org/10.1111/2041-210X.13518>
475
- 476 53 Barve, N. *et al.* (2011) The crucial role of the accessible area in ecological niche modeling and species distribution
477 modeling. *Ecol. Modell.* 222, 1810–1819
- 478 54 Hutchinson (1957) Population studies: Animal ecology and demography. *Bull. Math. Biol.* 53, 193–213
- 479 55 Ricklefs, R.E. (2010) Life-history connections to rates of aging in terrestrial vertebrates. *Proc. Natl. Acad. Sci.*
480 107, 10314–10319
- 481
- 482 56 Albert, J.H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.*
483 88, 669–679
- 484 57 Chib, S. and Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika*, 85, 347–361
485
- 486 58 Choler, P. (2005) Consistent shifts in alpine plant traits along a mesotopographical gradient. *Arct. Antarct. Alp.*
487 *Res.* 37: 444–453
488
- 489 59 Dray S., and Dufour, A. (2007) The ade4 package: Implementing the duality diagram for ecologists. *J. Stat.*
490 *Softw.* 22:1–20.
491
- 492 60 Golding, N. and Harris, D.J. (2015) BayesComm: Bayesian Community Ecology Analysis. R package version 0.1-2