

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27

# Towards causal relationships for modelling species distribution

Daniele Da Re<sup>1,2,\*,<sup>†</sup></sup>, Enrico Tordini<sup>3,<sup>†</sup></sup>, Jonathan Lenoir<sup>4</sup>,

Sergio Rubin<sup>1</sup>, Sophie O. Vanwambeke<sup>4</sup>.

<sup>1</sup>Center for Earth and Climate Research, Earth and Life Institute, UCLouvain, Louvain-la-Neuve, Belgium.

<sup>2</sup>Current address: Center Agriculture Food Environment, University of Trento, San Michele all'Adige, Italy

<sup>3</sup>Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia

<sup>4</sup>UMR CNRS 7058, Ecologie et Dynamique des Systèmes Anthropisés (EDYSAN), Université de Picardie Jules Verne, Amiens, France

<sup>†</sup>DDR and ET equally contributed to the study.

Corresponding author: daniele.dare@unitn.it

**Open research Statement:** Upon acceptance, the simulated data and code used will be provided via a GitHub repository [https://github.com/danddr/SEM\\_SDMS](https://github.com/danddr/SEM_SDMS) and permanently stored on a Zenodo repository with **DOI:XXXXX**.

**Significance Statement:** In this synthesis paper, we stress the importance of incorporating causal relationships for the modelling of species distribution. Here, we propose the modelling relation as a conceptual framework for modelling complex and hierarchical processes underlying the distribution of living organisms. We provide an application of the modelling relation using a virtual species example and a structural equation modelling approach. The modelling relation allows setting the boundaries of the modelling exercise, increasing model robustness in depicting natural patterns, eventually resulting in clear practical applications tightly linked to the ecology of the target species.

28 **Abstract**

- 29 1. Understanding the processes underlying the distribution of species through  
30 space and time is fundamental in several research fields spanning from  
31 ecology to spatial epidemiology. Correlative species distribution models  
32 (SDMs) involve popular statistical tools to infer species geographical  
33 distribution thanks to spatiotemporally explicit observations of species  
34 occurrences coupled with a set of environmental predictors.
- 35 2. So-called SDMs rely on the niche concept to infer or explain the distribution  
36 of species, though often focusing only on the abiotic component of the niche  
37 (e.g., temperature, precipitation), without clear causal links to the biology of  
38 species under investigation. This might result in an over-simplification of the  
39 complex niche hypervolume, resulting in a single model formula whose  
40 estimates and predictions lack ecological realism.
- 41 3. We believe that a causal perspective associated with a finer definition of the  
42 modelling target is necessary to develop ecologically more realistic outputs.  
43 Here, we propose to infer the geographical distribution of a species by  
44 applying the modelling relation approach, a causal conceptual framework  
45 developed by the theoretical biologist Robert Rosen, which can be  
46 formalized through structural equation modelling (SEM).
- 47 4. Implementing the modelling relation into SDMs would improve the inclusion  
48 of the causal processes underlying the spatial distribution of species into an  
49 inferential formal system, potentially highlighting the methodological steps  
50 where uncertainty arises and eventually resulting in model outputs which  
51 are tightly linked to the ecology of the target species.

52 **Keywords:** Directed Acyclic Graph; Environmental Niche Models; Habitat  
53 Suitability Models; Path Analyses; Process-based Models; Robert Rosen;  
54 Statistical models; Virtual Species.

# 1 Introduction

56 Understanding the processes underlying the distribution of species through space  
57 and time is a fundamental topic in several research fields including ecology,  
58 epidemiology, and biodiversity conservation (Franklin 2023). The geographical  
59 distribution of a species is commonly inferred using the so-called species distribution  
60 models (SDMs). Here we define SDMs as correlative models (e.g., generalized  
61 linear models, random forest, maxent) that establish a statistical relationship  
62 between an observed response variable describing the species distribution in the  
63 geographical space (e.g., presence-absence) and a set of predictors describing the  
64 environmental space occupied by the species over large geographical extents. The  
65 rapid availability of open-access biodiversity data (e.g., BIEN, sPlotOpen, GBIF;  
66 Enquist et al. 2016; Sabatini et al. 2021; GBIF 2023), environmental predictors (e.g.,  
67 WorldClim, Fick and Hijmans, 2017), and open source statistical languages like R,  
68 contributed to the tremendous diffusion of these correlative approaches over the past  
69 two decades (Araújo et al., 2019; Franklin 2023).

70 Nevertheless, numerous authors have raised concerns regarding the capacity  
71 of SDMs to accurately infer species distributions (Kearney and Porter, 2009; Araújo  
72 et al., 2019; Lee-Yaw et al., 2022), expressing specific criticisms about (i) the  
73 conceptual background of correlative SDMs (Kearney, 2006; Austin, 2007), (ii) the  
74 quality of the input data used to train the models (e.g., spatial and temporal biases  
75 when sampling distribution data; Hortal et al., 2008; Fourcade et al., 2014, Rocchini  
76 et al., 2023), (iii) the mismatch between the environmental conditions actually  
77 experienced by the target species and the spatial and temporal resolution of the  
78 abiotic predictors used in SDMs (Urban et al., 2016; Lembrechts et al., 2020), and  
79 the ecological realism of SDMs outputs (e.g., Lee-Yaw et al., 2022). These pitfalls  
80 have been widely discussed in the scientific literature and several methodological  
81 papers on the best practices were proposed (see for instance Araújo et al., 2019;  
82 Zurell et al., 2020; Sillero et al., 2021). The correlative aspect of these modelling  
83 exercises however remains, making SDM predictions often interpreted and  
84 evaluated mostly from a statistical perspective (e.g., models' predictive accuracy)  
85 rather than from their ecological realism (Austin et al., 2006; Merow et al., 2014;  
86 Hellegers et al., 2020).

87 In contrast, many scientists have argued for a causal approach to SDMs,  
88 incorporating biological knowledge into the models, and defining the hierarchical  
89 structure among the various factors influencing the geographical distribution of  
90 species (e.g., Kearney and Porter, 2009; Austin, 2007; Purse and Golding, 2015;  
91 Urban et al., 2016; Chapman et al., 2019). For instance, models based on species  
92 life history traits (i.e., the characteristics influencing individuals' performance or  
93 fitness; Nock et al., 2016; Dawson et al., 2021), have been proposed as an  
94 implementation of classic correlative SDMs, since these life history traits may reflect  
95 the different responses of a species to processes that modulate its distribution  
96 (Regos et al., 2019). These models have the advantage of making explicit the causal  
97 links between the biology of the target species and its environment, although their  
98 complexity and the huge amount of information they require for parameterisation  
99 make them less tractable.

100 The use of Bayesian approaches and the tuning of Bayesian priors, which  
101 entail the incorporation of prior knowledge through the use of Bayes' rule, constitutes  
102 another method to include causal mechanisms while remaining within the framework  
103 of correlative methods (van de Schoot et al., 2021). These approaches proved

104 particularly useful when hierarchical structures had to be incorporated in the models,  
105 as when dealing with complex spatiotemporal dynamics or when sampling efforts  
106 varied (Mäkinen and Vanhatalo, 2018).

107 An alternative approach to account for prior knowledge and hierarchical  
108 structure relies on the use of structural equation modelling (SEM). The SEM  
109 approach provides a comprehensive framework for modelling and analysing complex  
110 systems by incorporating both observed and unobserved variables, allowing  
111 researchers to go beyond simple correlations and examine the underlying structural  
112 relationships among variables (Grace, 2006). A central concept in SEM is the meta-  
113 model, which defines the hierarchical structure among several response and  
114 explanatory variables. This meta-model is essentially a theoretical framework that  
115 represents the researcher's understanding of how the variables are interconnected,  
116 describing the relationships between the variables based on prior knowledge,  
117 theoretical foundations, or empirical evidence. Such a graphical representation of the  
118 links and interconnections among several response and explanatory variables is  
119 borrowed from graph theory and computer science, usually referred as directed  
120 acyclic graphs (DAGs) with a set of rules that can be applied for observational causal  
121 inference in ecology (Arif and MacNeil 2022).

122 Independently from the type of algorithm or statistical approach used in  
123 SDMs, incorporating causal relationships and drawing a DAG diagram for SDMs'  
124 applications requires a deeper understanding of the species biology and the  
125 formulation of clear causal hypotheses about the drivers underlying the geographical  
126 distribution of the focal species. Given the widespread use of SDMs and their critical  
127 role in various research fields, we believe that embracing a causal perspective in  
128 SDMs is not only timely but also essential. Therefore, in this paper, we propose a  
129 conceptual and a technical solution, borrowed from the SEM approach and graph  
130 theory relying on DAG representations, to take causal relationships into account in  
131 SDMs exercises. From a pure conceptual-level perspective, we introduce the Robert  
132 Rosen's modelling relation framework (Rosen 1978; 1986; 1993) as a causal  
133 scheme to guide the design of species distribution models. Robert Rosen (1934 –  
134 1998), a theoretical biologist, introduced the conceptual framework called "modelling  
135 relation" as a fundamental principle in understanding and representing complex  
136 systems like living organisms, arguing that traditional mathematical models often fall  
137 short in capturing their complexity (Rosen, 1978, 1986). The modelling relation  
138 highlights the idea that a model should capture the essential organizational  
139 relationships and constraints of a system, capturing the underlying organizational  
140 principles that guide the system's behaviour rather than merely describing its  
141 components and interactions (Rosen 1993). Rosen's emphasis on organization was  
142 a reaction against reductionist approaches that focus solely on the individual  
143 components of a system without considering a more holistic view of the systemic  
144 interactions and causal constraints that give rise to system's properties.

145 From a more technical viewpoint, we propose to use SEM as the inferential  
146 approach within the modelling relation framework (the formal system in Robert  
147 Rosen's modelling relation scheme; Fig. 1), aiming to better integrate the underlying  
148 causal processes behind the distribution of a species. We highlight the importance of  
149 a carefully constructed conceptual model, using SEM approaches or DAGs that are  
150 built upon the hierarchical nature of the relations linking a species distribution with its  
151 environment, to implement meaningful causal relationships and increase the  
152 ecological realism of SDMs. To illustrate this, we use a set of virtual species,  
153 transferring our hypothesized causal diagram or DAG into a SEM framework and  
154 comparing its results with those of a generalized linear model (GLM), a common  
155 method used in correlative SDMs.

## 156 2 Incorporating hypothesized causal 157 relationships into SDMs

158 The *niche* concept is a fundamental notion in ecology and represent the conceptual  
159 backbone of SDMs. Different definitions of the niche concept have been proposed  
160 (Pocheville et al., 2015; Sales et al., 2021), but, essentially, the niche concept aims  
161 to define the environmental space in which a species could exist, allowing us to  
162 identify the geographical area where those environmental conditions are met, and  
163 the species can persist and reproduce. The design and interpretation of correlative  
164 SDMs is usually framed within the niche concept provided by Soberón and Peterson  
165 (2005), the so-called biotic, abiotic, and movement (BAM) framework. According to  
166 the BAM framework, biotic and abiotic factors, as well as species dispersal  
167 limitations, determine the geographical distribution of a species. The intersection  
168 between the biotic and abiotic components returns the realized niche of the species  
169 (*sensu* Hutchinson, 1957). Consequently, the intersection between the realized  
170 niche and the accessible areas defines the actual or realized geographical  
171 distribution of the species (Soberón and Peterson, 2005). In fact, the BAM  
172 framework provides a way to operationalize the niche concept in the geographical  
173 space, making it appealing for inferring the distribution of a species through SDMs.  
174 Since its introduction in 2005, the BAM framework has become a mainstay in  
175 correlative SDMs exercises and has been applied in multiple scientific fields (e.g.,  
176 Escobar and Craft, 2016; Bible and Peterson, 2018; Franklin 2023).

177 Correlative SDMs' outputs depict (and synthesise) the distribution of a species  
178 as a detailed and spatially contiguous map representing an index of  
179 environmental/habitat suitability (Guisan et al., 2017), with the maximum values of  
180 this index typically interpreted as the areas that are most suitable for the target  
181 species. These maps are often visually attractive and are assumed to be  
182 straightforward to read and interpret, thus contributing to the promotion and  
183 dissemination of SDMs. These outputs, however, are primarily assessed from a  
184 statistical perspective (e.g., the models' predictive accuracy) rather than in terms of  
185 their ecological realism. Many efforts have been devoted to solve various  
186 methodological issues of SDMs, mainly dealing with: statistical techniques; spatial  
187 and temporal autocorrelation in the data; spatial and temporal sampling bias of the  
188 response variable; variable selection; model selection; and predictive accuracy. The  
189 scientific literature is very rich in that respect (e.g., Muscarella et al., 2014; Fourcade  
190 et al., 2014; Varela et al., 2014; Aiello-Lammens et al., 2015; Qiao et al., 2015, 2019;  
191 Hallgren et al., 2019; Brun et al., 2020; Simmonds et al., 2020; Bazzichetto et al.,  
192 2023; see Sillero and Barbosa, 2020 for a summary of common methodological  
193 pitfalls of SDMs and Sillero et al., 2021 for a step by step methodological guide to  
194 SDMs).

195 However, the conceptual background necessary for generating meaningful  
196 and hypothesis-driven SDMs has been much less discussed (but see Araujo and  
197 Guisan 2006; Austin 2007; Thuiller et al. 2013). Interest in alternative modelling  
198 approaches looking for deeper causal relationships between the distribution of a  
199 species and its potential determinants has been growing (Kearney and Porter, 2009;  
200 Hartemink et al., 2011; Urban et al., 2016; Feng., 2017; Staniczenko et al., 2017;  
201 Briscoe et al., 2019; Kraemer et al., 2019; Arif and MacNeil, 2023). Indeed, a  
202 modelling perspective based on the biology of the target organism and associated

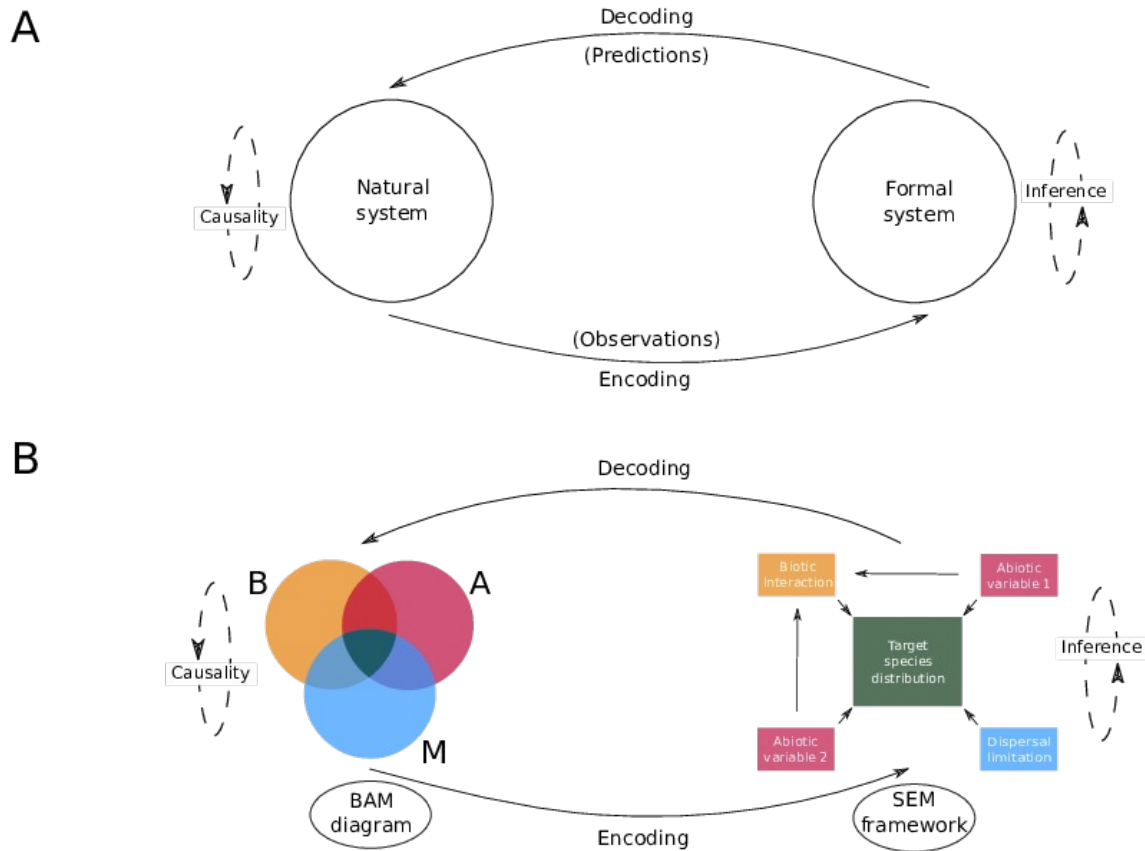
203 with a finer definition of the objective of the model might help to develop more  
204 ecologically realistic outputs with explicit causal links. This would help to avoid  
205 correlative SDMs outputs biased by spurious correlative spatial structure underlying  
206 both response variable and predictors, especially when the predictors have no direct  
207 causal links with the response variable (Lozier, Aniello and Hickerson, 2009;  
208 Fourcade et al., 2018; Journé et al., 2020), and to foster more meaningful and scale-  
209 appropriate interpretation of the results.

210 Incorporating causal relations into a model requires a basic knowledge of the  
211 study system or organism under investigation in order to formulate specific  
212 hypotheses that can later be translated into model equations. In this paper, we  
213 define a causal relationship as one for which scientists have a mechanistic basis for  
214 expecting that variations induced in a driver variable can lead to a change in the  
215 distribution of a response variable. This definition corresponds to the general  
216 scientific definition employed in the natural sciences and is the definition associated  
217 with the enterprise of causal modelling (Grace and Irvine 2020). We recognize that  
218 the alternative enterprise of inferring causal relations from data in the absence of  
219 mechanistic knowledge, a common situation in the social sciences, introduces  
220 additional requirements.

221 Several authors have proposed practical suggestions or guidelines to clarify  
222 the model assumptions and increase model's biological realism (e.g., Araujo et al.,  
223 2019; Chapman et al., 2019; Zurell et al., 2020; Srivastava et al., 2021).  
224 Conceptually speaking, we believe the so-called modelling relation framework  
225 developed by Robert Rosen in the 1980s (Rosen, 1985) could be especially relevant  
226 to incorporate causal relationships into SDMs.

## 227 2.1 Rosen's modelling relation

228 Robert Rosen's modelling relation framework is a conceptual framework designed to  
229 understand how a biological system could be coded into an inferential mathematical  
230 system through causal inference (Mikulecky, 2001). The modelling relation can be  
231 defined as a process of relating two structures, a material one governed by causality,  
232 and a mathematical one governed by inferential rules (see Chapt. 2-3 in Rosen,  
233 1986). The former is the *natural system*, hence the *causal* system of investigation,  
234 while the latter is the *formal system* used to infer the *natural* one (Fig. 1A). The  
235 relation between these two structures is given by 'encoding' the causality of the  
236 *natural system* into a *formal system* of inference and by 'decoding' such inference  
237 back to the causal phenomenon. The encoding arrow drawn from left to right of Fig.  
238 1A, represents the observations and measurements of the *natural systems* aiming to  
239 capture its causality, while the arrow from the *formal system* toward the natural one  
240 represents the decoding operation of the prediction into the *natural system* made by  
241 the mathematical *formal system*.



243 [double column] **Figure 1:** (A) Robert Rosen's modelling relation. (B) Example of application of the modelling relation to model  
 244 the distribution of a species (natural system, depicted in green within the Biotic Abiotic Movement (BAM; conceptual framework)  
 245 by means of a Structural Equation Model (SEM; formal system).

246 Though the view of an inferential model in Rosen's modelling relation is not  
 247 completely new (Pattee, 2007) and shares the same rationale of the backdoor  
 248 criteria used when building DAGs (i.e., it uses domain knowledge, above all else, to  
 249 determine the best causal model for a given causal query; see Arif and MacNeil,  
 250 2022), the modelling relation framework represents a valid epistemological tool to  
 251 guide (and refine) the incorporation of ecological knowledge into more biologically  
 252 realistic SDMs. To design the inferential model structure, the encoding section  
 253 requires that the user summarizes the main assumptions and the uncertainties about  
 254 the natural system (e.g., the main determinants of the distribution of a given species  
 255 following the niche theory, such as the BAM diagram; Fig. 1A), and to define them as  
 256 mathematical equations and relations (e.g., translating the BAM diagram into a  
 257 causal and mathematical diagram; Fig. 1B). Clearly, if these assumptions are wrong  
 258 or imprecise, we would obtain biased predictions, eventually resulting in a lack of  
 259 ecological realism. In this view, Siekmann (2018) proposes Rosen's modelling  
 260 relation as a type of process-based model where the model outputs from the formal  
 261 system can be compared to the natural system and used to validate the  
 262 assumptions. Similarly, an ecological process-based model generally focuses on a  
 263 particular aspect of the natural system such as a given life history trait of the target  
 264 species, thus providing a possible explanation according to the underlying  
 265 assumptions of the formal system (Siekmann, 2018). It follows that various models  
 266 can be built under different assumptions (e.g., different and competing causal  
 267 diagrams), and their results compared and interpreted in the light of the ecological

268 assumptions they respectively made on the natural system (Fudge and Turko, 2020).  
269 Rosen’s modelling relation can thus be used to design and compare different  
270 competitive hypotheses about the investigated natural system, therefore treating  
271 modelling as an experimental exercise (Siekmann, 2018; Metcalf, 2019).

## 272 2.2 Applying Rosen’s modelling relation

273 To date, few attempts have been made to include the modelling relations into SDMs  
274 exercises. For instance, Kineman (2007, 2009) as well as Kineman and Wessman  
275 (2021) applied a correlative approach where response curves between the predicted  
276 habitat suitability and the environmental factors were mostly tuned by visual  
277 interpretation and expert-based assessment. In particular, Kineman (2007)  
278 highlighted how his approach was mainly designed as an exploratory tool to learn  
279 about ecological relationships and test ecological hypotheses. However, we could  
280 not find a broader application of Rosen’s modelling relation aiming at modelling  
281 species distribution. As a conceptual framework, the modelling relation is  
282 independent from the statistical method used (Siekmann, 2018; Metcalf, 2019), but  
283 we suggest that the rationale behind the SEM approach (Grace, 2006) fits well within  
284 the modelling relation *formal system*.

285 The SEM approach provides a comprehensive framework for analysing  
286 complex relationships (both direct and indirect) among variables by combining  
287 elements of factor analysis, regression analysis, and path analysis (Grace, 2006). A  
288 structural equation model begins with a causal diagram, a graphical representation  
289 of the hypothesized causal structure of the studied system (Fan et al., 2016; Garrido  
290 et al., 2022). One effective approach is the utilization of DAGs (Greenland et al.,  
291 1999; Pearl et al., 2016), which are constructed to represent researchers'  
292 hypotheses regarding how explanatory variables influence the response variable(s).  
293 Each variable can be defined as exogenous, endogenous or mediator. Exogenous  
294 variables are only independent variables (i.e., only pointed towards other variables).  
295 Endogenous variables are dependent variables (i.e., pointed at by other variables),  
296 but can also be used as independent variables pointing towards other endogenous  
297 variables in more complex structures, playing a mediating effect (i.e., mediators). For  
298 instance, variable A may affect variable C either directly or indirectly via a mediating  
299 effect from variable B, which means that variable A is exogenous while B and C are  
300 endogenous. Through SEM, DAGs can unveil confounding factors that must be  
301 considered in regression analysis to obtain unbiased coefficients. Moreover, they  
302 can reveal mediation pathways or situations involving multiple response variables  
303 (Grace, 2006).

304 The strength of SEM relies on testing different hypotheses (i.e., different causal  
305 diagrams that can be used as candidates and competing “meta-models”) about the  
306 causal relationships between the variables considered in the studied system. Recent  
307 advances in SEM allow us to deal with a wide range of error distributions (e.g.,  
308 Poisson and binomial families) and data structures (e.g., hierarchical or longitudinal  
309 dataset), thanks to the piecewiseSEM R package (Lefcheck, 2016; Lefcheck, Byrnes  
310 and Grace 2020). Indeed, the hypothesized set of causal pathways can be validated  
311 only if the proposed model is consistent with the observations. In other words, if the  
312 model-estimated variance-covariance matrix can predict the variance-covariance  
313 matrix of the observational dataset:

$$314 \quad \Sigma = \Sigma(\Phi) \quad (1)$$

315 where  $\Sigma$  is the observed variance-covariance matrix, and  $\Sigma(\Phi)$  is the model-



316 estimated covariance matrix expressed in terms of  $\Phi$ , the matrix of model-estimated  
317 parameters (i.e., coefficients). Austin (2007) was one of the very first scientists  
318 proposing the application of SEM to SDMs, advocating the importance of including  
319 and evaluating a causal structure into the modelling exercise. However, due to  
320 technical limitations such as the application of SEM to data not fitting a Gaussian  
321 error distribution and the estimate of only linear relationships prevented a broader  
322 application of this methodology to data types commonly found in ecological studies  
323 (Lefcheck, 2016; Grace, 2022). Recent technical developments overcome some of  
324 these limitations (e.g., Chu et al., 2019; Carvalho-Rocha et al., 2021; Cerqueira et  
325 al., 2021; Quiroga et al., 2021), but their application into SDMs remains surprisingly  
326 low.

## 327 3 Case study

328 To illustrate the potential of using SEM directly embedded into Rosen's modelling  
329 relation (cf. the *formal system*) and rooted in the BAM framework of the niche theory  
330 used in most SDM studies (cf. the *natural system*), we used a virtual species  
331 approach (Leroy et al., 2016; Meynard et al., 2019). We first simulated the  
332 geographical distribution of two virtual species. The first one is fully dependent on  
333 the abiotic conditions while the second one is influenced by both the abiotic  
334 conditions and the presence of the first species. Then, we provided a causal diagram  
335 or DAG aiming to explain the spatial distribution of the second virtual species by  
336 means of both direct and indirect (mediating) effects from both abiotic and biotic (the  
337 first virtual species) constraints.

### 338 3.1 Virtual species

339 The virtual species approach provides the great advantage of knowing exactly the  
340 species' ecological niche and its predicted distribution into the geographical space  
341 (Meynard et al., 2019). Here, for the sake of simplicity, we considered only two  
342 bioclimatic variables retrieved from the WorldClim2 database (BIO1 for mean annual  
343 temperature and BIO12 for mean annual precipitation; Fick and Hijmans, 2017). The  
344 spatial extent of the area of interest (AOI; spatial resolution of ~10 minutes, ~18.6  
345 km at the Equator) was cropped to match that of Central and Southern Europe to  
346 reduce the computational effort of this illustrative application (Fig. 2A-B).

347 Specifically, we created a virtual tree species whose geographical distribution  
348 depends on its response to both BIO1 (thermal range: 5-13°C) and BIO12  
349 (precipitation range: 526-1257 mm; Fig. S1.1A-B). This results in a tree species  
350 mostly distributed in the mountainous area of Europe (Fig. 2D), displaying a  
351 continentality gradient (East-West macroclimatic gradient) coupled with higher  
352 suitability at the cold end of the BIO1 gradient. The geographical distribution of the  
353 second virtual species, a shade-tolerant herbaceous species, is driven by the same  
354 abiotic variables as the virtual tree species, but favoured by a warmer range of mean  
355 annual temperature conditions (thermal range: 11-20°C) and a drier range of mean  
356 annual precipitations (precipitation range: 255-739 mm; Fig. S1.1AB), resulting in a  
357 wider potential geographical distribution compared to the three species if considering  
358 abiotic component only. The true species habitat suitability ( $p$ ) across the AOI was  
359 generated using binomial generalised linear models (GLMs), or logistic regressions,  
360 assuming sigmoid (i.e., non-quadratic) response curves between the occurrence of  
361 the species and the chosen predictors (Eq. 2), and following the approach described

362 in Bazzichetto et al. (2023).

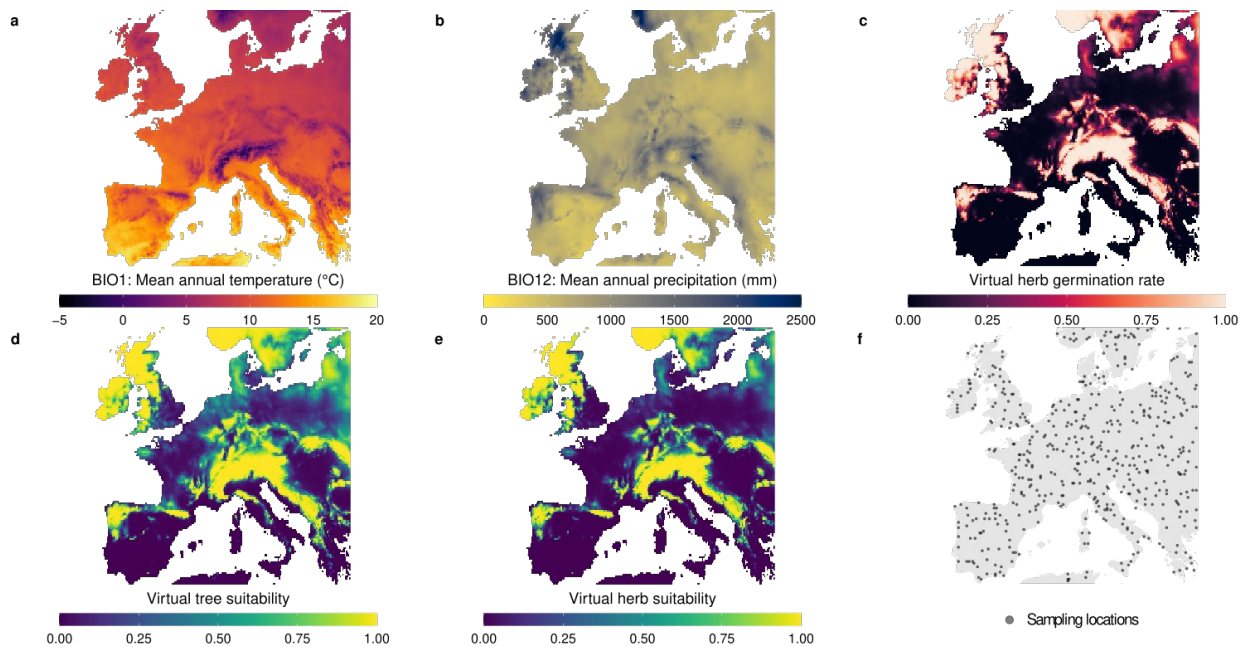
363 
$$\text{logit}(p_i) = \alpha + \beta_{pr} \times \text{precipitations} + \beta_{tm} \times \text{temperature} \quad (2)$$

364 where  $\text{logit}(p_i)$  is the natural logarithm of the odd ratio  $p_i/(1-p_i)$ ,  $\alpha$  is the model  
365 intercept,  $\beta_{pr}$  is the regression parameter for the linear term (i.e., sigmoid shape) of  
366 precipitation,  $\beta_{tm}$  is the regression parameter for the linear term (i.e., sigmoid shape )  
367 of temperature. Regression parameters for the tree species were set to 1 ( $\alpha$ ), 0.01  
368 ( $\beta_{pr}$ ), and -1 ( $\beta_{tm}$ ), whilst for the herb species, they were set to 1 ( $\alpha$ ), 0.015 ( $\beta_{pr}$ ), and -  
369 0.85 ( $\beta_{tm}$ ). Logit-transformed probabilities were turned to the unit interval [0,1] using  
370 the logistic function available through the `plgis` function in the `stats R` package (R  
371 Core Team, 2023).

372 We decided to constrain the geographical distribution of the herb species by  
373 the occurrence of the virtual tree species, to simulate an obligate biotic interaction  
374 (i.e., the herbaceous species benefits from growing in the shade of the virtual tree  
375 species). To simulate this biotic constraint, we computed the germination rate of the  
376 virtual herbaceous species as a function of the habitat suitability of the virtual tree  
377 species: namely, the germination rate of the virtual herbaceous species increased  
378 logarithmically with the habitat suitability provided by the virtual tree species (Fig.  
379 S1.1C).

380 Eventually, the resulting geographical distribution of the virtual herbaceous  
381 species (Fig. 2E) was defined by the intersection between its climatic niche and the  
382 biotic constraint of its germination rate depending on the habitat suitability of the  
383 virtual tree species (Fig. 2A-C). The obtained habitat suitability maps of the two  
384 virtual species (Fig. 2D-E) were then converted into presence-absence maps using  
385 the function `convertToPA` of the `virtualspecies R` package.

386 To add stochasticity in this simulation exercise, we generated three different  
387 scenarios for the dispersal capacity of the virtual herb species, by varying its  
388 geographical prevalence (the number of pixels actually occupied by the species out of  
389 the total number of pixels available in the geographical space), while keeping fixed the  
390 virtual tree species geographical prevalence. As a result, we assigned a fixed  
391 geographical prevalence equals to 0.4 to the virtual tree species, while for the  
392 herbaceous species we simulated three dispersal scenarios (low, medium, high) whose  
393 underlying geographical prevalence was set to 0.25, 0.50, and 0.75, respectively (Fig.  
394 S1.2). We then randomly sampled 500 locations across the AOI to extract information  
395 on the presence-absence of each of the two virtual species, the value of the germination  
396 rate of the virtual herbaceous species, as well as the values of BIO1 and BIO12 (Fig.  
397 2F). We repeated this operation 10 times, the predictive accuracy of each simulation  
398 was estimated using a spatial cross-validation with 15 spatial folds retaining 80% of the  
399 observations for training and 20% for testing. This allowed us to generate a toy dataset  
400 to calibrate our SEM models built within the Rosen's modelling relation. A detailed  
401 description of the virtual species simulation, the sampling methodology and the R codes  
402 used to generate this modelling exercise are available on GitHub  
403 [https://github.com/danddr/SEM\\_SDMs](https://github.com/danddr/SEM_SDMs).



405 [double column] **Figure 2:** (A-B) The set of abiotic variables (BIO1 and BIO1) used to create the two virtual species. (C) The  
 406 germination rate of the virtual herb species computed as a function of the habitat suitability of the virtual tree species. (D) The  
 407 habitat suitability of the virtual tree species. (E) The habitat suitability of the virtual herb species. (F) Sampling locations. The  
 408 geographic projection used is the WGS84 - World Geodetic System 1984, EPSG: 4326.

## 409 3.2 Statistical analysis

410 The main goal of this modelling exercise is to demonstrate the applicability of the  
 411 SEM approach (cf. causal diagrams) within Rosen's modelling relation and to  
 412 compare its predictive accuracy along with the stability of model's coefficients with  
 413 respect to a traditional SDM algorithm not relying on causal diagrams such as GLMs.  
 414 By presenting the modelling relation as a hypothesis testing conceptual exercise, we  
 415 hypothesized a causal diagram aiming to describe the distribution of the target forest  
 416 herb species (Fig. 3), whereby the geographical distribution of the forest herba  
 417 species represents the *natural system* and the causal diagram from the SEM  
 418 approach represents the *formal system*. In the causal diagram or DAG (Fig. 3):

- 420 • BIO1 and BIO12 (abiotic components) have a direct effect on both the virtual tree  
 421 and the virtual herb species distribution (Eq. 3, 5);

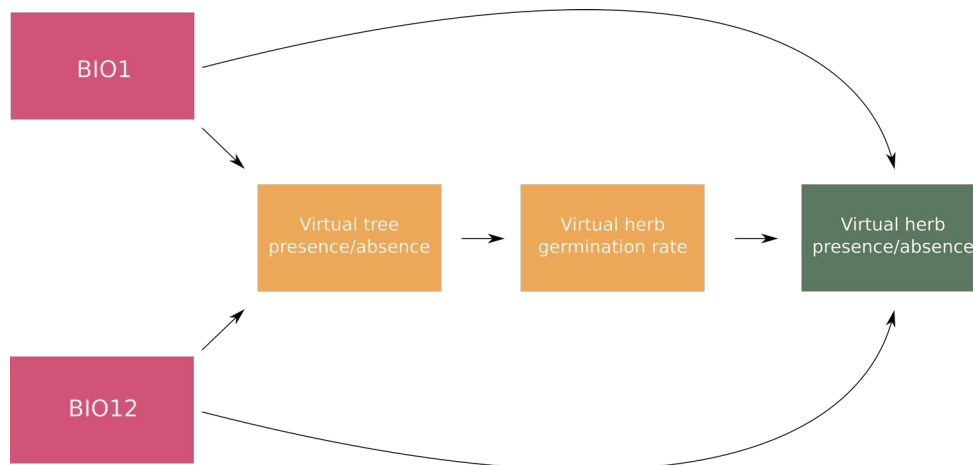
$$422 \text{Tree} \sim \text{BIO1} + \text{BIO12} \quad (3)$$

- 423 • the occurrence of the virtual tree species has a direct effect on the germination  
 424 rate of the herb species and an indirect (*via* the germination rate) effect on the  
 425 actual distribution of the virtual herb species (Eq. 4);

$$426 \text{Germination rate} \sim \text{Tree} \quad (4)$$

- 427 • the germination rate (biotic component) of the virtual herb species has a direct  
 428 effect on the actual distribution of the virtual herb species (Eq. 5).

$$429 \text{Herb} \sim \text{BIO1} + \text{BIO12} + \text{Germination rate} \quad (5)$$



431 [single column] **Figure 3:** Hypothesized causal diagram explaining the distribution of the virtual herb species. Purple boxes  
 432 indicate abiotic variables, orange boxes indicate biotic variables while green box displays the response variable.  
 433

434 The causal diagram was then converted into a set of candidate models (Eq. 3-  
 435 5) using the `piecewiseSEM` and `semEff` R packages (Lefcheck, 2016; Murphy,  
 436 2020). The congruence of the estimated variance-covariance matrix hypothesized in  
 437 the SEM with the observed variance-covariance matrix in the data was evaluated for  
 438 each geographic prevalence and cross-validation iterations using a Fisher's C test,  
 439 whose null hypothesis ( $H_0$ ) is that the model variance-covariance matrix can predict  
 440 the observed variance-covariance matrix. Hence, a  $p$ -value  $> 0.05$  for the Fisher's C  
 441 test implies that the estimated variance-covariance matrix from the causal diagram  
 442 mirrors the observed one in the data, therefore validating it (Lefcheck, 2016).

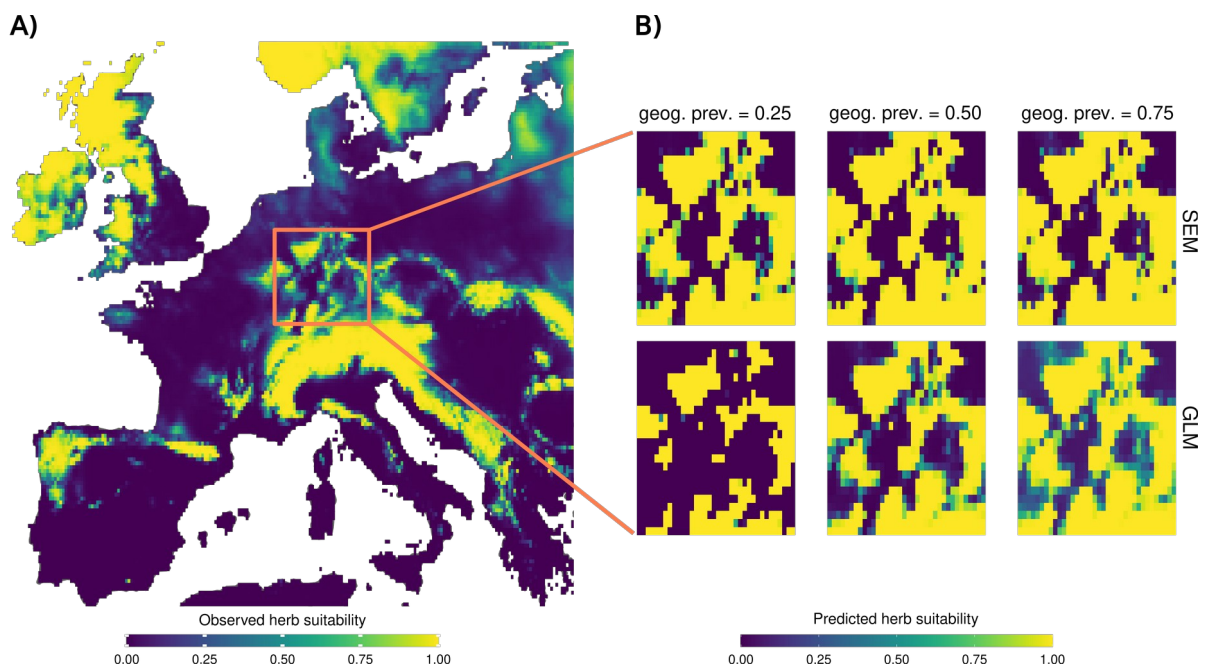
443 Finally, for comparison purposes and as an example of a classic non-  
 444 hierarchical SDM, we computed a binomial GLM, where the presence-absence of  
 445 the virtual herb species (cf. the only response variable) was modelled as a function  
 446 of three predictor variables: BIO1, BIO12, and the germination rate. We also  
 447 computed a set of metrics routinely used to assess the predictive performance of  
 448 SDMs: (i) the area under the ROC curve (AUC); (ii) sensitivity; (iii) specificity; (iv) the  
 449 true skill statistic (TSS); (v) the coefficient of determination ( $R^2$ , here to be intended  
 450 as a pseudo- $R^2$  computed using the Nagelkerke approach) ; (vi) and the root mean  
 451 squared error (RMSE). The  $R^2$  and the RMSE were computed by comparing the true  
 452 (i.e., simulated) habitat suitability of the virtual herb species with the one predicted by  
 453 each combination of models and geographical prevalence (Meynard and Kaplan,  
 454 2012). A detailed description of the validation metrics is available in Guisan et al.  
 455 (2017).

### 456 3.3 Results

457 The Fisher's C test did not support the causal diagram proposed in Fig. 3 as the  
 458 hypothetical causal structure representing the variance-covariance matrix observed  
 459 in the training dataset ( $p < 0.05$ ), suggesting the inclusion of direct effects for both  
 460 BIO1 and BIO12 on the germination rate of the herb species (Eq. 4). Once these two  
 461 additional direct effects were integrated, the Fischer's C test supported the updated  
 462 causal diagram ( $p > 0.05$ ).

463 The predictive accuracy metrics computed for the models of the virtual herb  
 464 species on the testing dataset showed comparable outcomes for both SEM and  
 465 GLM, whose variation was mainly related to the geographical prevalence of the  
 466 virtual herb species rather than to the modelling technique used (Fig. S1.3). The

467 RMSE values of the SEM, in particular, showed a rather stable behaviour across the  
 468 different geographical prevalence values, whereas in the GLM these RMSE values  
 469 tended to increase with the geographical prevalence. Furthermore, the SEM showed  
 470 more stable coefficient estimates with different geographic prevalences compared to  
 471 the GLM: whilst the coefficients estimated by the SEM are stable and always  
 472 significant, coefficients estimated by the GLM varied greatly across the cross-  
 473 validation iterations and geographical prevalences (Fig. S1.4). The variation in the  
 474 estimated coefficients affected the spatial predictions: the inclusion of a mediating  
 475 effect may lead to more stable spatial predictions of the SEM across the three  
 476 dispersal scenarios compared to the spatial predictions of the GLM (Fig. 4). As a  
 477 consequence, also the spatial variability of the RMSE computed between the  
 478 observed (i.e., simulated) herb suitability and the median of predicted cross-validated  
 479 iterations for each geographical prevalence and models showed similar spatial  
 480 pattern, but the magnitude of the RMSE tended to increase across the different  
 481 geographical prevalences more for the GLM than for the SEM (Tab. S1.5).  
 482



484 [double column] **Figure 4:** The observed (A) and predicted (B) habitat suitability values for the virtual herb species in a subset  
 485 of the study area under different combinations of geographic prevalences and models. The geographic projection used is the  
 486 WGS84 - World Geodetic System 1984, EPSG: 4326.

## 487 4 Discussion

488 In this paper, we introduced the Rosen's modelling relation and proposed its  
 489 application for SDMs by means of causal diagrams or DAGs borrowed from the SEM  
 490 approach. Based on the results of our virtual species exercise, the modelling relation  
 491 and SEM approach are valuable tools to incorporate biological knowledge and the  
 492 hierarchical structure of the links between variables into correlative SDMs, by  
 493 encoding the assumptions related to the distribution of a species (natural system)  
 494 into the formal system of Rosen's modelling relation. Our findings suggest that  
 495 building a model relying on a strong conceptual basis improves the stability of the  
 496 estimated model's coefficients, without necessarily increasing the predictive  
 497 accuracy metrics of the model. We speculate that the hierarchical structure of the  
 498 causal diagram helped to reveal the relationships between the virtual herb species

499 and its determinant, independently of the sampling (cross-validation iteration) and  
500 the geographic prevalence of the species. Despite the generally favourable results in  
501 terms of predictive performance for both modelling approaches, we argue that  
502 comparing predictive accuracy metrics may not be the most effective way to assess  
503 how appropriate different models are. In fact, prior studies demonstrated that these  
504 metrics are influenced by a variety of factors, such as sample prevalence (Guisan et  
505 al., 2017; Leroy et al., 2018; Marchetto et al., 2023), sample location bias (Fourcade  
506 et al., 2018, Jiménez-Valverde, 2021 Dubos et al., 2022; Rocchini et al., 2023) and  
507 the size of the study region (Lobo et al., 2008).

508 Essentially, predictive models and causal inference are two different tools, the  
509 former attempting to find the best model predicting the response variable and the  
510 latter attempting to disentangle the effects of the predictors on the response variable  
511 (Arif and MacNeil, 2022). Therefore, our SEM application for SDMs might be used to  
512 assess causal relationships between variables affecting the geographical  
513 distributions of species (i.e. attribution) but may not always be the most appropriate  
514 tool for generating accurate predictions on the actual species distribution. In other  
515 words, model prediction and model attribution are two different applications that may  
516 prove complementary but one cannot replace the other.

517 In our view, one of the most interesting aspect of SEM application to SDMs is  
518 the capacity of discovering unanticipated mechanisms through conditional  
519 independence testing, e.g., that there are direct effects between species that were  
520 not considered before, or revealing the effect of a latent variable not yet measured or  
521 discovered (Lefcheck, 2016; Lefcheck, Byrnes and Grace 2020; Arif and MacNeil,  
522 2022).

523 Whilst the natural-to-formal systems relationships presented in Rosen's  
524 modelling relation is made explicit in the SEM rationale (causal diagrams), the  
525 modelling relation can be applied in any correlative method to introduce causality into  
526 ecological modelling. Rosen's modelling relation can help modellers in their  
527 conceptual definition of a causal model, which can then be put into practice using  
528 different modelling approaches (correlative and process-based). However, other  
529 methodological approaches aiming to include biological realism or accounting for  
530 causality in correlative models exist, even though their application in ecology is  
531 extremely limited. For instance, the parametric g-formula proposed by Robins and  
532 Hernán (2009) employs a causal diagram to account for time-varying factors and  
533 time-varying confounder effects. Specifically, the g-formula allows for estimating the  
534 causal effects of sustained treatment strategies from observational data with time-  
535 varying treatments and has been applied prevalently in epidemiological studies (Keil  
536 et al., 2014; Naimi et al., 2017; Meisner et al., 2022). Bayesian SDMs are another  
537 way of introducing hypothesized causality by adding ecological or physiological  
538 knowledge in the model using informative priors, representing a prior belief regarding  
539 the probability distribution of an unknown parameter. For instance, Feng et al. (2019)  
540 gathered thermal limits and survival information for the zebra mussel *Dreissena*  
541 *polymorpha* from the literature and used these to calibrate correlative Bayesian  
542 models.

543 Unlike correlative models, process-based models are usually independent of  
544 geographical observations of the taxa under investigation. These typically express  
545 biological (or other) processes by a mathematical equation (e.g., ordinal differential  
546 equation or matrix population models) relating an indicator of the process (e.g., a life  
547 history trait such as the number of offsprings) to different factors affecting its  
548 performance (e.g., environmental conditions) (Kearney et al., 2010; Da Re et al.,  
549 2022). For instance, Larter et al. (2017) showed how a single plant functional trait  
550 (xylem resistance to cavitation) displayed a strong statistical relationship with its

551 species distribution in relation to aridity across the climatic range of the species.  
552 Process-based SDMs have also been successfully used in invasion ecology to  
553 simulate and forecast invasion risk under different global change scenarios (Carboni  
554 et al., 2018; Strubbe et al., 2023). Within the family of process-based models, Agent  
555 based models (ABMs) aim to predict species population or community dynamics by  
556 modelling multiple individuals (agents) that interact with their environment and  
557 among each other. For each agent, ABMs require the specification of state variables,  
558 which can include age, size, and spatial location, as well as physiological and  
559 behavioural traits (Zhang and DeAngelis, 2020).

560 Rosen's modelling relation coupled with the SEM approach, as advocated  
561 here, is one of the methods allowing to design and refine ecological hypotheses,  
562 thus treating modelling as an experimental exercise. Within the field of SDMs, the  
563 modelling relation can represent a wider conceptual tool to model species  
564 distribution based on causal and ecologically-based assumptions, potentially  
565 resulting in an increase of the ecological realism of SDMs. Inferring the spatial  
566 distribution of a species of high interest (e.g., a vector-borne species, a species of  
567 conservation concern, an invasive alien species) using a correlative approach and  
568 bioclimatic variables only, not accounting for uncertainty in the data and without a  
569 solid causal approach, may ultimately lead to ecological inconsistencies and  
570 subsequently to inaccurate estimates, with strong ecological and even socio-  
571 economic repercussions (Escobar and Craft, 2016; Hellegers et al., 2020).  
572 Furthermore, such inconsistencies in the outcomes generated by ecological models  
573 may undermine the trust in ecological research (Currie, 2019; O'Grady, 2020; Lee-  
574 Yaw et al., 2021). Certainly, when knowledge on the target organism is scarce, a  
575 correlative approach may be the only option available, but a causal-oriented  
576 definition of the modelling exercise is crucial to enhance the ecological realism of the  
577 models (Getz et al., 2018) and to ensure the models' transferability to novel  
578 conditions.

579 Ecologists aspire to foster knowledge on global environmental changes  
580 induced by human activities, such as climate change, biological invasions and  
581 habitat loss. To efficiently tackle such challenges, clear, robust, and well-defined  
582 epistemological premises about the main determinants of species distribution and  
583 species distribution change are needed to design realistic experiments (Pigliucci,  
584 2002; Currie, 2019). Epistemological premises are not just philosophical murmuring  
585 but allow us to set the boundaries of the modelling exercise, increasing model  
586 robustness in depicting natural patterns and resulting in clear practical applications  
587 (Currie, 2019; Dawson et al., 2023). Rosen's modelling relation and its  
588 implementation by means of the SEM approach requires to clearly define the *natural*  
589 *system* (the key response variable of interest), such as the *niche*, *habitat* or *biome*  
590 (see Box 1), which inherently define different biological entities and cannot be used  
591 interchangeably. It may also help to identify when model assumptions are causal or  
592 not and to develop a suite of model comparisons (hypothesis-driven modelling) that  
593 can robustly explain the variation in the data while accounting for ecological  
594 observations.

595 **Box 1**

596 Biotic Abiotic Movement (BAM): heuristic framework which defines the species  
597 population distribution as those areas where abiotic, biotic and accessible areas  
598 intersect.

599 Biome: a large cluster of plant species that are defined in terms of the  
600 recognizable physiognomy of the dominant species (e.g. savanna, *sensu*  
601 Pennington et al., 2004)

602 Ecophysiology: a branch of biology studying how the environment surrounding an  
603 organism (both abiotic and biotic component) interacts with its physiology.

604 Fitness: individual reproductive success.

605 Functional trait: those characteristics influencing performance or fitness of an  
606 individual (*sensu* Nock et al., 2016)

607 Fundamental niche: the region of the  $n$ -dimensional space (Hutchinsonian  
608 hypervolume) where the biotic interactions are excluded, and thus only the abiotic  
609 conditions affect the fitness..

610 Habitat: the actual spatio-temporal configuration of environmental conditions  
611 where an organism either actually or potentially lives (*sensu* Kearney, 2006)

612 Hutchinsonian niche concept:  $n$ -dimensional space (hypervolume), where each  
613 dimension is an abiotic or biotic condition and the relations among them allow the  
614 species to exist in a self-maintained population without immigration.

615 Mechanistic niche: those sets of environmental conditions that allow an organism  
616 to complete its life cycle and successfully reproduce (*sensu* Kearney, 2006)

617 Realized niche: a smaller fraction of the fundamental niche constrained by biotic  
618 interactions.

## 619 **5 Declaration**

620 • Ethics approval and consent to participate: Not applicable.

621 • Fieldwork permission: Not applicable.

622 • Competing interests: No conflict of interest has been declared by the authors.

623 • Funding: This project did not receive specific funding.

624 • Author's contribution: DDR and SR conceptualized the integration of Rosen's  
625 theory on modelling relation into a species distribution modelling exercise,  
626 which was further developed thanks to the suggestions made by SOV and JL  
627 on the use of structural equation modelling. DDR and ET performed the data  
628 analysis. All the authors critically commented the results and their  
629 interpretation; DDR and ET led the writing of the manuscript and produced a



630 first draft, which was further improved by all other authors.

631 • Acknowledgments: The authors are grateful to Dr. Francesco Petruzzellis,  
632 Prof. Julianne Meisner, Dr. Bethan Purse, Prof. Caroline Nieberding and Prof.  
633 Eric Lambin who provide constructive feedback and commented on a previous  
634 version of this manuscript. DDR was supported by a FRS-FNRS ASP Belgian  
635 grant (Grant No. 34766961), ET is supported by the Estonian Research  
636 Council grant (MOBJD1030).

## 637 **6 Code availability**

638 The codes used are fully operational under R 4.3 (R Core Team, 2023). The scripts  
639 used for the analyses presented in this paper is available in the GitHub repository  
640 [https://github.com/danddr/SEM\\_SDMs](https://github.com/danddr/SEM_SDMs).

## 641 **References**

642 Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., and Anderson, R.  
643 P. (2015). sptthin: an r package for spatial thinning of species occurrence records  
644 for use in ecological niche models. *Ecography*, 38(5):541–545.

645 Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early,  
646 R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., et al. (2019). Standards for  
647 distribution models in biodiversity assessments. *Science Advances*,  
648 5(1):eaat4858.

649 Araujo, M. B. and Guisan, A. (2006). Five (or so) challenges for species distribution  
650 modelling. *Journal of biogeography*, 33(10):1677–1688.

651 Arif, S., & MacNeil, M. A. (2023). Applying the structural causal model framework for  
652 observational causal inference in ecology. *Ecological Monographs*, 93(1), e1554.

653 Arif, S., & MacNeil, M. A. (2022). Predictive models aren't for causal inference.  
654 *Ecology Letters*, 25(8), 1741-1745.

655 Austin, M. (2007). Species distribution models and ecological theory: a critical  
656 assessment and some possible new approaches. *Ecological modelling*, 200(1-  
657 2):1–19.

658 Austin, M., Belbin, L., Meyers, J. a. A., Doherty, M., and Luoto, M. (2006). Evaluation  
659 of statistical models used for predicting plant species distributions: role of artificial  
660 data and theory. *Ecological modelling*, 199(2):197–216.

661 Bazzichetto, M., Lenoir, J., Da Re, D., Tordoni, E., Rocchini, D., Malavasi, M., ... &  
662 Sperandii, M. G. (2023). Sampling strategy matters to accurately estimate  
663 response curves' parameters in species distribution models. *Global Ecology and*  
664 *Biogeography*. 32, 1717–1729.

665 Bible, R. C. and Peterson, A. T. (2018). Compatible ecological niche signals between  
666 biological and archaeological datasets for late-surviving neandertals. *American*  
667 *journal of physical anthropology*, 166(4):968–974.

- 668 Briscoe, N. J., Elith, J., Salguero-Gómez, R., Lahoz-Monfort, J. J., Camac, J. S.,  
669 Giljohann, K. M., Holden, M. H., Hradsky, B. A., Kearney, M. R., McMahon, S. M.,  
670 et al. (2019). Forecasting species range dynamics with process-explicit models:  
671 matching methods to applications. *Ecology letters*, 22(11):1940–1956.
- 672 Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wüest, R. O., Wang, Z., and  
673 Zimmermann, N. E. (2020). Model complexity affects species distribution  
674 projections under climate change. *Journal of Biogeography*, 47(1):130–142.
- 675 Carboni, M., Guéguen, M., Barros, C., Georges, D., Boulangeat, I., Douzet, R.,  
676 Dullinger, S., Klöner, G., van Kleunen, M., Essl, F., et al. (2018). Simulating plant  
677 invasion dynamics in mountain ecosystems under global change scenarios.  
678 *Global change biology*, 24(1):e289–e302.
- 679 Carvalho-Rocha, V., Peres, C. A., and Neckel-Oliveira, S. (2021). Habitat amount  
680 and ambient temperature dictate patterns of anuran diversity along a subtropical  
681 elevational gradient. *Diversity and Distributions*, 27(2):344–359.
- 682 Chapman, D., Pescott, O. L., Roy, H. E., & Tanner, R. (2019). Improving species  
683 distribution models for invasive non-native species with biologically informed  
684 pseudo-absence selection. *Journal of Biogeography*, 46(5), 1029-1040.
- 685 Cerqueira, R. C., de Rivera, O. R., Jaeger, J. A., and Grilo, C. (2021). Direct and  
686 indirect effects of roads on space use by jaguars in Brazil. *Scientific reports*,  
687 11(1):1–9.
- 688  
689 Chu, C., Lutz, J. A., Král, K., Vrška, T., Yin, X., Myers, J. A., Abiem, I., Alonso, A.,  
690 Bourg, N., Burslem, D. F., et al. (2019). Direct and indirect effects of climate on  
691 richness drive the latitudinal diversity gradient in forest trees. *Ecology letters*,  
692 22(2):245–255.
- 693 Currie, D. J. (2019). Where Newton might have taken ecology. *Global ecology and  
694 biogeography*, 28(1):18–27.
- 695 Da Re, D., Van Bortel, W., Reuss, F., Müller, R., Boyer, S., Montarsi, F., ... &  
696 Marcantonio, M. (2022). dynamAedes: a unified modelling framework for invasive Aedes  
697 mosquitoes. *Parasites & Vectors*, 15(1), 1-18.
- 698 Dawson, M. N., Mainali, K., Meyer, R., Noonan, M., Papeş, M., Parenti, L. R., &  
699 Villalobos, F. (2023). Reshaping biogeography: Perspectives on the past, present and  
700 future. *Journal of Biogeography*, 50(8), 1405-1408.
- 701 Dubos, N., Préau, C., Lenormand, M., Papuga, G., Monsarrat, S., Denelle, P., Le  
702 Louarn, M., Heremans, S., Roel, M., Roche, P., & Luque, S. (2022). Assessing the effect  
703 of sample bias correction in species distribution models. *Ecological Indicators*, 145,  
704 109487.
- 705 Dawson, S. K., Carmona, C. P., González-Suárez, M., Jönsson, M., Chichorro, F.,  
706 Mallen Cooper, M., Melero, Y., Moor, H., Simaika, J. P., and Duthie, A. B. (2021).  
707 The traits of “trait ecologists”: An analysis of the use of trait and functional trait  
708 terminology. *Ecology and evolution*, 11(23):16434–16445.
- 709 Enquist, B. J., Condit, R., Peet, R. K., Schildhauer, M., and Thiers, B. M. (2016).  
710 Cyberinfrastructure for an integrated botanical information network to investigate  
711 the ecological impacts of global climate change on plant biodiversity. Technical

- 712 report, PeerJ Preprints.
- 713 Escobar, L. E. and Craft, M. E. (2016). Advances and limitations of disease  
714 biogeography using ecological niche modeling. Frontiers in Microbiology, 7:1174.
- 715 Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S. R., Park, H., and Shao, C. (2016).  
716 Applications of structural equation modeling (sem) in ecological studies: an  
717 updated review. Ecological Processes, 5(1):1–12.
- 718 Feng, X., Liang, Y., Gallardo, B., and Papeş, M. (2019). Physiology in ecological  
719 niche modeling: using zebra mussel's upper thermal tolerance to refine model  
720 predictions through bayesian analysis. Ecography.
- 721 Feng, X. and Papeş, M. (2017). Physiological limits in an ecological niche modeling  
722 framework: A case study of water temperature and salinity constraints of  
723 freshwater bivalves invasive in USA. Ecological Modelling, 346:48–57.
- 724 Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution  
725 climate surfaces for global land areas. International journal of climatology,  
726 37(12):4302–4315.
- 727 Fourcade, Y., Besnard, A. G., and Secondi, J. (2018). Paintings predict the  
728 distribution of species, or the challenge of selecting environmental predictors and  
729 evaluation statistics. Global Ecology and Biogeography, 27(2):245–256.
- 730  
731 Fourcade, Y., Engler, J. O., Rödder, D., and Secondi, J. (2014). Mapping species  
732 distributions with maxent using a geographically biased sample of presence data:  
733 a performance assessment of methods for correcting sampling bias. PLoS one,  
734 9(5):e97122.
- 735  
736 Franklin, J. (2023). Species distribution modelling supports the study of past, present  
737 and future biogeographies. Journal of Biogeography.  
738 <https://onlinelibrary.wiley.com/doi/10.1111/jbi.14617>
- 739 Garrido, M., Hansen, S. K., Yaari, R., and Hawlena, H. (2022). A model selection  
740 approach to structural equation modelling: A critical evaluation and a road map for  
741 ecologists. Methods in Ecology and Evolution, 13(1):42–53.
- 742 GBIF: The Global Biodiversity Information Facility (2023) *What is GBIF?*. Available  
743 from <https://www.gbif.org/what-is-gbif> [12 October 2023].
- 744 Getz, W. M., Marshall, C. R., Carlson, C. J., Giuggioli, L., Ryan, S. J., Románach, S.  
745 S., Boettiger, C., Chamberlain, S. D., Larsen, L., D'Odorico, P., et al. (2018).  
746 Making ecological models adequate. Ecology letters, 21(2):153–166.
- 747 Grace, J. (2022). General guidance for custom-built structural equation models. One  
748 Ecosystem, 7:e72780.
- 749 Grace, J.B. and Irvine, K.M., 2020. Scientist's guide to developing explanatory  
750 statistical models using causal analysis principles. Ecology, 101(4), p.e02962.
- 751 Grace, J. B. (2006). Structural equation modeling and natural systems. Cambridge  
752 University Press.
- 753 Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic

- 754 research. Epidemiology, 37-48.
- 755 Guisan, A., Thuiller, W., and Zimmermann, N. E. (2017). Habitat suitability and  
756 distribution models: with applications in R. Cambridge University Press.
- 757 Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y., and Mackey, B. (2019). Species  
758 distribution models can be highly sensitive to algorithm configuration. Ecological  
759 Modelling, 408:108719.
- 760 Hartemink, N., Vanwambeke, S. O., Heesterbeek, H., Rogers, D., Morley, D.,  
761 Pesson, B., Davies, C., Mahamdallie, S., and Ready, P. (2011). Integrated  
762 mapping of establishment risk for emerging vector-borne infections: a case study  
763 of canine leishmaniasis in southwest france. PloS one, 6(8).
- 764 Hellegers, M., Ozinga, W. A., Hinsberg van, A., Huijbregts, M. A., Hennekens, S. M.,  
765 Schaminée, J. H., Dengler, J., and Schipper, A. M. (2020). Evaluating the  
766 ecological realism of plant species distribution models with ecological indicator  
767 values. Ecography, 43(1):161–170.
- 768 Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M., & Baselga, A. (2008).  
769 Historical bias in biodiversity inventories affects the observed environmental niche  
770 of the species. Oikos, 117(6), 847-858.  
771
- 772 Hutchinson, G. (1957). Concluding remarks cold spring harbor symposia on  
773 quantitative biology, 22: 415–427. GS SEARCH.  
774
- 775 Jiménez-Valverde, A. (2021). Prevalence affects the evaluation of discrimination  
776 capacity in presence-absence species distribution models. Biodiversity and  
777 Conservation, 30(5), 1331–1340.
- 778 Journé, V., Barnagaud, J.-Y., Bernard, C., Crochet, P.-A., and Morin, X. (2020).  
779 Correlative climatic niche models predict real and virtual species distributions  
780 equally well. Ecology, 101(1):e02912.
- 781 Kearney, M. (2006). Habitat, environment and niche: what are we modelling? Oikos,  
782 115(1):186–191.
- 783 Kearney, M. and Porter, W. (2009). Mechanistic niche modelling: combining  
784 physiological and spatial data to predict species' ranges. Ecology letters,  
785 12(4):334–350.
- 786 Kearney, M., Simpson, S. J., Raubenheimer, D., and Helmuth, B. (2010). Modelling  
787 the ecological niche from functional traits. Philosophical Transactions of the Royal  
788 Society B: Biological Sciences, 365(1557):3469–3483.
- 789 Keil, A. P., Edwards, J. K., Richardson, D. R., Naimi, A. I., & Cole, S. R. (2014). The  
790 parametric G-formula for time-to-event data: towards intuition with a worked  
791 example. Epidemiology (Cambridge, Mass.), 25(6), 889.
- 792 Kineman, J. J. (2007). Relational complexity in natural science and the design of  
793 ecological informatics. PhD thesis, Citeseer.
- 794 Kineman, J. J. (2009). Relational theory and ecological niche modelling. In  
795 Proceedings of the 53rd Annual Meeting of the ISSS-2009, Brisbane, Australia.

- 796 Kineman, J. J. and Wessman, C. A. (2021). Relational systems ecology: The  
797 anticipatory niche and complex model coupling. Handbook of Systems Sciences,  
798 pages 871–916.
- 799 Kraemer, M. U., Reiner Jr, R. C., and Bhatt, S. (2019). Causal inference in spatial  
800 mapping. Trends in parasitology, 35(10):743–746.
- 801 Larter, M., Pfautsch, S., Domec, J.-C., Trueba, S., Nagalingum, N., and Delzon, S.  
802 (2017). Aridity drove the evolution of extreme embolism resistance and the  
803 radiation of conifer genus callitris. New Phytologist, 215(1):97–112.
- 804 Lee-Yaw, J., L. McCune, J., Pironon, S., and N. Sheth, S. (2021). Species  
805 distribution models rarely predict the biology of real populations. Ecography,  
806 n/a(n/a).
- 807 Lefcheck, J. S. (2016). `piecewisem`: Piecewise structural equation modelling in R  
808 for ecology, evolution, and systematics. Methods in Ecology and Evolution,  
809 7(5):573–579.
- 810 Lefcheck, J.S., Byrnes, J.E.K. and Grace, J.B., (2020). `piecewiseSEM`: Piecewise  
811 Structural Equation Modeling (2.1.2)[Computer software].
- 812 Lembrechts, J. J., Aalto, J., Ashcroft, M. B., De Frenne, P., Kopecký, M., Lenoir, J.,  
813 Luoto, M., Maclean, I. M., Roupsard, O., Fuentes-Lillo, E., et al. (2020). Soiltemp:  
814 A global database of near-surface temperature. Global Change Biology,  
815 26(11):6616– 6629.
- 816 Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M.,  
817 & Bellard, C. (2018). Without quality presence–absence data, discrimination metrics  
818 such as TSS can be misleading measures of model performance. Journal of  
819 Biogeography, 45(9), 1994–2002.
- 820 Leroy, B., Meynard, C. N., Bellard, C., and Courchamp, F. (2016). `virtualespecies`, an  
821 R package to generate virtual species distributions. Ecography, 39(6):599–607.
- 822 Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences  
823 and their importance in species distribution modelling. Ecography, 33(1), 103–114.
- 824 Lozier, J. D., Aniello, P., & Hickerson, M. J. (2009). Predicting the distribution of  
825 Sasquatch in western North America: anything goes with ecological niche  
826 modelling. Journal of Biogeography, 36(9), 1623-1627.
- 827 Mäkinen, J. and Vanhatalo, J. (2018). Hierarchical bayesian model reveals the  
828 distributional shifts of arctic marine mammals. Diversity and Distributions,  
829 24(10):1381–1394.
- 830 Marchetto, E., Da Re, D., Tordoni, E., Bazzichetto, M., Zannini, P., Celebrin,  
831 S., Chieffallo, L., Malavasi, M., & Rocchini, D. (2023). Testing the effect of sample  
832 prevalence and sampling methods on probability-and favourability-based  
833 SDMs. Ecological Modelling, 477, 110248.
- 834 Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand,  
835 S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., and Elith, J. (2014). What do  
836 we gain from simplicity versus complexity in species distribution models?  
837 Ecography, 37(12):1267–1281.

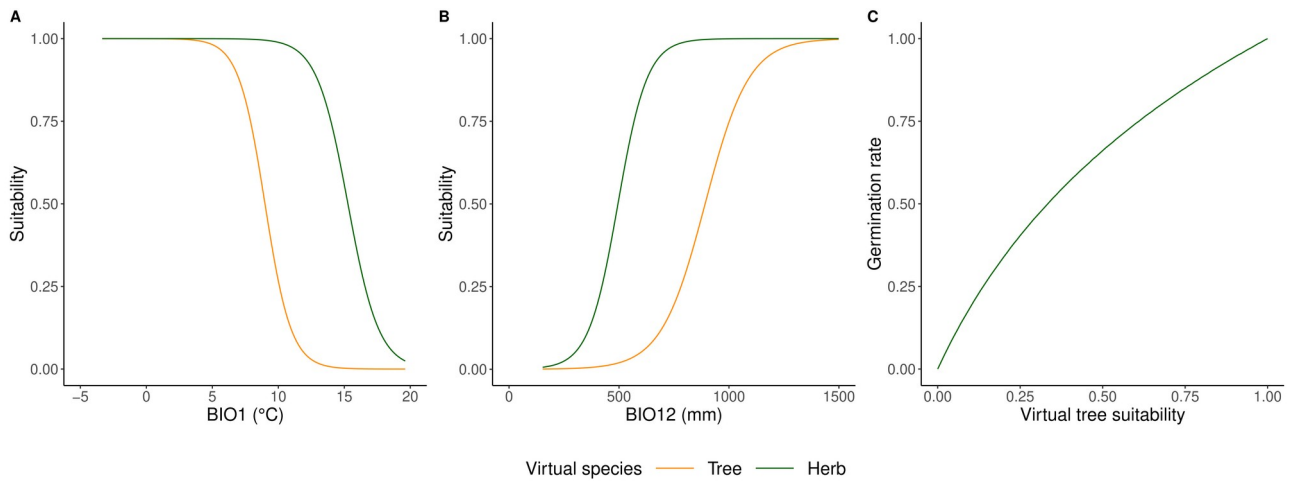
- 838 Metcalf, G. S. (2019). Design and the modeling relation. She Ji: The Journal of  
839 Design, Economics, and Innovation, 5(4):373–376.
- 840 Meisner, J., Kato, A., Lemerani, M. M., Mwamba Miaka, E., Ismail Taban, A.,  
841 Wakefield, J., ... & Rabinowitz, P. M. (2022). The effect of livestock density on  
842 *Trypanosoma brucei gambiense* and *T. b. rhodesiense*: A causal inference-based  
843 approach. PLoS neglected tropical diseases, 16(8), e0010155.
- 844 Meynard, C. N., Leroy, B., and Kaplan, D. M. (2019). Testing methods in species  
845 distribution modelling using virtual species: what have we learnt and what are we  
846 missing? Ecography, 42(12):2021–2036.
- 847 Meynard, C. N., & Kaplan, D. M. (2012). The effect of a gradual response to the  
848 environment on species distribution modeling performance. Ecography, 35(6),  
849 499-509.
- 850 Mikulecky, D. C. (2001). Robert rosen (1934-1998): a snapshot of biology's newton.  
851 Computers and Chemistry, 4(25):317–327.
- 852 Murphy, M. (2020). semeff: Automatic calculation of effects for piecewise structural  
853 equation models. R package.
- 854 Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte,  
855 M., and Anderson, R. P. (2014). Enm eval: An r package for conducting spatially  
856 independent evaluations and estimating optimal model complexity for maxent  
857 ecological niche models. Methods in Ecology and Evolution, 5(11):1198–1205.
- 858 Naimi, A. I., Cole, S. R., & Kennedy, E. H. (2017). An introduction to g methods.  
859 International journal of epidemiology, 46(2), 756-762.
- 860 Nock, C. A., Vogt, R. J., and Beisner, B. E. (2016). Functional Traits, pages 1–8.  
861 American Cancer Society.
- 862 O'Grady, C. (2020). Psychology's replication crisis inspires ecologists to push for  
863 more reliable research. ScienceMag.org.
- 864
- 865 Pattee, H. H. (2007). Laws, constraints, and the modeling relation—history and  
866 interpretations. Chemistry & biodiversity, 4(10):2272–2295.
- 867 Pearl, J., Glymour, M., & Jewell, N. P. (2016). Causal inference in statistics: A  
868 primer. John Wiley & Sons.
- 869 Pennington, P. T., Cronk, Q. C. B., Richardson, J. A., Woodward, F. I., Lomas, M.  
870 R., and Kelly, C. K. (2004). Global climate and the distribution of plant biomes.  
871 Philosophical Transactions of the Royal Society of London. Series B: Biological  
872 Sciences, 359(1450):1465–1476.
- 873 Pigliucci, M. (2002). Are ecology and evolutionary biology" soft" sciences? In  
874 Annales Zoologici Fennici, pages 87–98. JSTOR.
- 875 Pocheville, A. (2015). The ecological niche: history and recent controversies. In  
876 Handbook of evolutionary thinking in the sciences, pages 547–586. Springer.
- 877 Purse, B. V., & Golding, N. (2015). Tracking the distribution and impacts of diseases  
878 with biological records and distribution modelling. Biological Journal of the Linnean

- 879 Society, 115(3), 664-677.
- 880 Qiao, H., Feng, X., Escobar, L. E., Peterson, A. T., Soberón, J., Zhu, G., and Papeş,  
881 M. (2019). An evaluation of transferability of ecological niche models. Ecography,  
882 42(3):521–534.
- 883 Qiao, H., Soberón, J., and Peterson, A. T. (2015). No silver bullets in correlative  
884 ecological niche modelling: insights from testing among many potential algorithms  
885 for niche estimation. Methods in Ecology and Evolution, 6(10):1126–1136.
- 886 Quiroga, R. E., Premoli, A. C., and Fernández, R. J. (2021). Niche dynamics in an  
887 phytropical desert disjunct plants: Seeking for ecological and species-specific  
888 influences. Global Ecology and Biogeography, 30(2):370–383.
- 889 R Core Team (2023). R: A Language and Environment for Statistical Computing. R  
890 Foundation for Statistical Computing, Vienna, Austria.
- 891 Regos, A., Gagne, L., Alcaraz-Segura, D., Honrado, J. P., and Domínguez, J.  
892 (2019). Effects of species traits and environmental predictors on performance and  
893 transferability of ecological niche models. Scientific reports, 9(1):1–14.
- 894 Robins, J., & Hernan, M. (2008). Estimation of the causal effects of time-varying  
895 exposures. Chapman & Hall/CRC Handbooks of Modern Statistical Methods, 553-  
896 599.
- 897 Rocchini, D., Tordoni, E., Marchetto, E., Marcantonio, M., Barbosa, A. M.,  
898 Bazzichetto, M., ... & Malavasi, M. (2023). A quixotic view of spatial bias in  
899 modelling the distribution of species and their diversity. npj Biodiversity, 2(1), 10.
- 900 Rosen, R. (1978). Fundamentals of measurement and representation of natural  
901 systems. Elsevier North-Holland. New York.
- 902 Rosen, R. (1986). Anticipatory systems: Philosophical, mathematical and  
903 methodological foundations. In Anticipatory systems. Pergamon, Oxford.
- 904 Rosen, R. (1993). On models and modeling. Applied mathematics and computation,  
905 56(2-3), 359-372.
- 906 Sabatini, F. M., Lenoir, J., Hattab, T., Arnst, E. A., Chytr`y, M., Dengler, J., De  
907 Ruffray, P., Hennekens, S. M., Jandt, U., Jansen, F., et al. (2021). splotopen—an  
908 environmen tally balanced, open-access, global dataset of vegetation plots. Global  
909 Ecology and Biogeography, 30(9):1740–1764.
- 910
- 911 Sales, L. P., Hayward, M. W., and Loyola, R. (2021). What do you mean by “niche”?  
912 modern ecological theories are not coherent on rhetoric about the niche concept.  
913 Acta Oecologica, 110:103701.
- 914 Siekmann, I. (2018). An applied mathematician’s perspective on rosennean  
915 complexity. Ecological Complexity, 35:28–38.
- 916 Sillero, N., Arenas-Castro, S., Enriquez-Urzelai, U., Vale, C. G., Sousa-Guedes, D.,  
917 Martínez-Freiría, F., Real, R., and Barbosa, A. M. (2021). Want to model a species  
918 niche? a step-by-step guideline on correlative ecological niche modelling.  
919 Ecological Modelling, 456:109671.

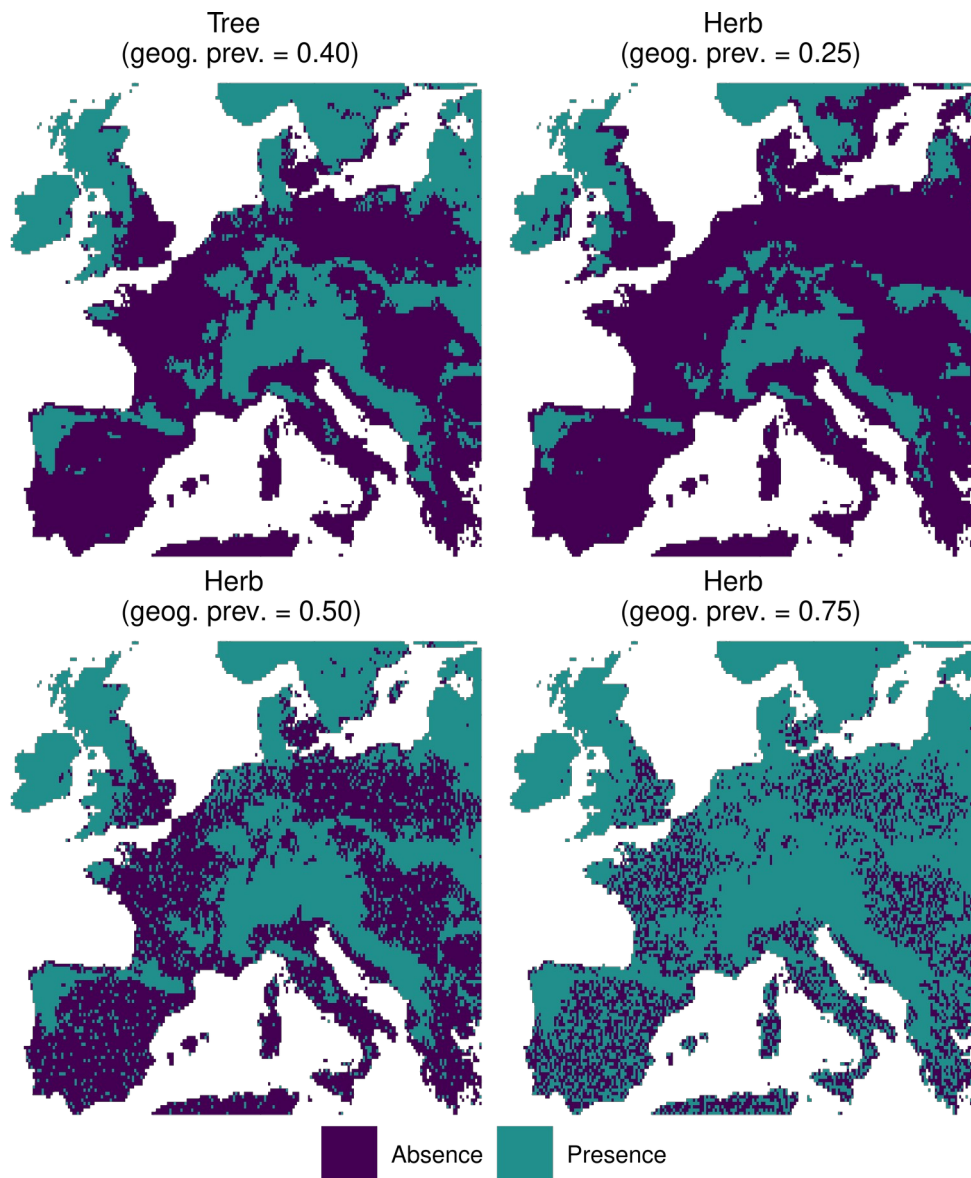
- 920 Sillero, N. and Barbosa, A. M. (2020). Common mistakes in ecological niche models.  
921 International Journal of Geographical Information Science, pages 1–14.
- 922 Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., and O’Hara, R. B. (2020).  
923 Is more data always better? a simulation study of benefits and limitations of  
924 integrated distribution models. Ecography.
- 925 Soberón, J. and Peterson, A. T. (2005). Interpretation of models of fundamental  
926 ecological niches and species’ distributional areas. Biodiversity Informatics, 2  
927 (January). <https://doi.org/10.17161/bi.v2i0.4>.
- 928 Staniczenko, P. P., Sivasubramaniam, P., Suttle, K. B., & Pearson, R. G. (2017).  
929 Linking macroecology and community ecology: refining predictions of species  
930 distributions using biotic interaction networks. Ecology letters, 20(6), 693-707.  
931
- 932 Strubbe, D., Jiménez, L., Barbosa, A. M., Davis, A. J., Lens, L., & Rahbek, C. (2023).  
933 Mechanistic models project bird invasions with accuracy. Nature Communications,  
934 14(1), 2520.
- 935 Thuiller, W., Münkemüller, T., Lavergne, S., Mouillot, D., Mouquet, N., Schiffrers, K.,  
936 and Gravel, D. (2013). A road map for integrating eco-evolutionary processes into  
937 biodiversity models. Ecology letters, 16:94–105.
- 938 Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J.-B., Pe’er, G., Singer, A., Bridle,  
939 J., Crozier, L., De Meester, L., Godsoe, W., et al. (2016). Improving the forecast  
940 for biodiversity under climate change. Science, 353(6304):aad8466.
- 941 van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G.,  
942 Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian  
943 statistics and modelling. Nature Reviews Methods Primers, 1(1):1–26.
- 944 Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2014).  
945 Environmental filters reduce the effects of sampling bias and improve predictions  
946 of ecological niche models. Ecography, 37(11):1084–1091.
- 947 Zhang, B., & DeAngelis, D. L. (2020). An overview of agent-based models in plant  
948 biology and ecology. Annals of Botany, 126(4), 539-557.
- 949 Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos,  
950 G., Feng, X., Guillera-Arroita, G., Guisan, A., et al. (2020). A standard protocol for  
951 reporting species distribution models. Ecography.



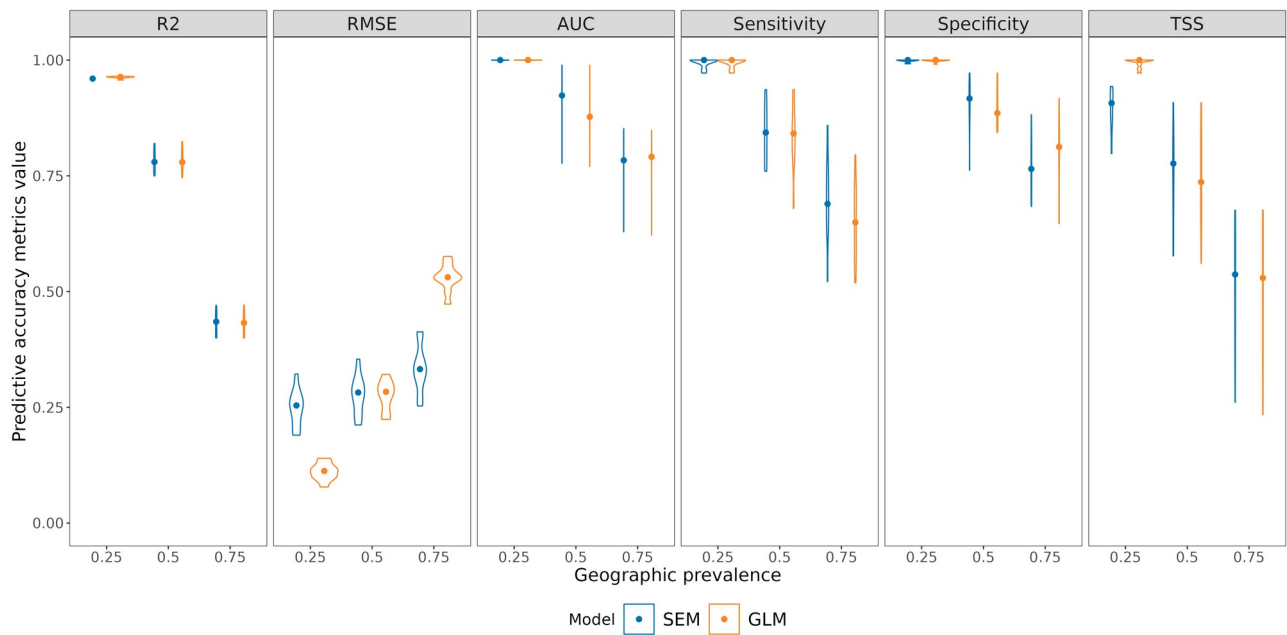
## Supplementary Materials 1



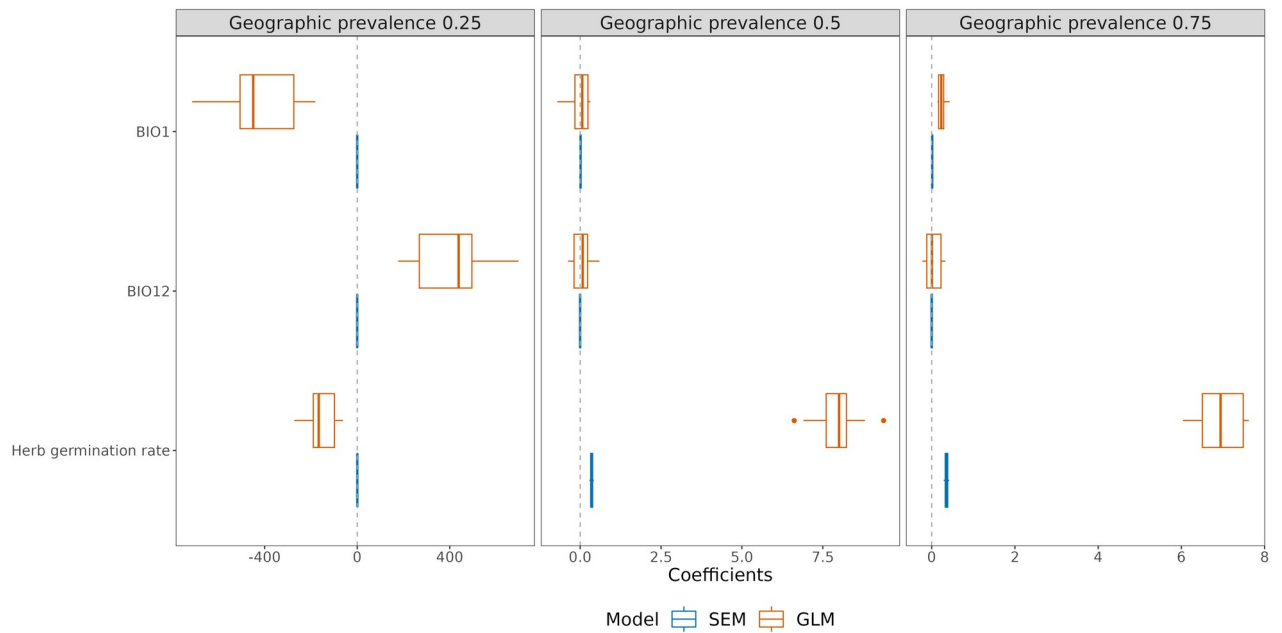
**Figure S1.1** Simulated response curves for the tree (orange) and herb (green) virtual species along the temperature (A) and precipitation (B) gradients. Herb virtual species germination rate along a gradient of the virtual tree species suitability (c).



**Figure S1.2** Tree and herb virtual species presence-absence distribution along different geographical prevalences.



**Figure S1.3** Violin plots reporting the distribution of the values of the metrics of predictive performance for the virtual herb species habitat suitability modeled as a function of the tree virtual species presence-absence and virtual herb species germination rate, and varying the geographical prevalence of the herb species (x axis). Dots represent median values of the metrics of predictive accuracy, while columns indicate the different performance metrics: R2 = coefficient of determination; RMSE = root mean squared error; AUC = area under the curve; TSS = true skill statistic. Colours are associated with the three modeling approaches tested (structural equation modelling, SEM, in blue; generalised linear models, GLM, in yellow).



**Figure S1.4** Boxplots reporting the distribution of the values of coefficients estimates of the virtual herb species habitat suitability modeled as a function of BIO1, BIO12 and virtual herb species germination rate, and varying the geographical prevalence of the herb species. Colours are associated with the three modeling approaches tested (structural equation modelling, SEM, in blue; generalised linear models, GLM, in yellow).

**Table S1.5** RMSE computed between the median of predicted cross-validated iterations for each geographical prevalence and models and the observed (i.e., simulated) herb suitability.

Model	geog.prev	RMSE
SEM	0.25	0.35
SEM	0.5	0.38
SEM	0.75	0.39
GLM	0.25	0.26
GLM	0.5	0.29
GLM	0.75	0.37