Towards causal relationships for modelling species distribution

Daniele da Re, Enrico Tordoni, Jonathan Roger Michel Henri Lenoir, Sergio Rubin, Sophie Vanwambeke

Towards causal relationships for modelling species distribution

Daniele Da Re[1,2,*,†], Enrico Tordoni[3,†], Jonathan Lenoir[4],

Sergio Rubin[1], Sophie O. Vanwambeke[1,]

[1]Center for Earth and Climate Research, Earth and Life Institute, UCLouvain, Louvain-la-Neuve, Belgium.
[2]Current address: Center Agriculture Food Environment, University of Trento, San Michele all'Adige, Italy
[3]Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia
[4]UMR CNRS 7058, Ecologie et Dynamique des Systèmes Anthropisés (EDYSAN), Université de Picardie Jules Verne, Amiens, France

[†]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 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.

Abstract

1. Understanding the processes underlying the distribution of species through space and time is fundamental in several research fields spanning from ecology to spatial epidemiology. Correlative species distribution models (SDMs) involve popular statistical tools to infer species geographical distribution thanks to spatiotemporally explicit observations of species occurrences coupled with a set of environmental predictors.

2. So-called SDMs rely on the niche concept to infer or explain the distribution of species, though often focusing only on the abiotic component of the niche (e.g., temperature, precipitation), without clear causal links to the biology of species under investigation. This might result in an over-simplification of the complex niche hypervolume, resulting in a single model formula whose estimates and predictions lack ecological realism.

3. We believe that a causal perspective associated with a finer definition of the modelling target is necessary to develop ecologically more realistic outputs. Here, we propose to infer the geographical distribution of a species by applying the modelling relation approach, a causal conceptual framework developed by the theoretical biologist Robert Rosen, which can be formalized through structural equation modelling (SEM).

4. Implementing the modelling relation into SDMs would improve the inclusion of the causal processes underlying the spatial distribution of species into an inferential formal system, potentially highlighting the methodological steps where uncertainty arises and eventually resulting in model outputs which are tightly linked to the ecology of the target species.

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

1 Introduction

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

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

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

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

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

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

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

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

2 Incorporating hypothesized causal relationships into SDMs

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

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

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

5

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

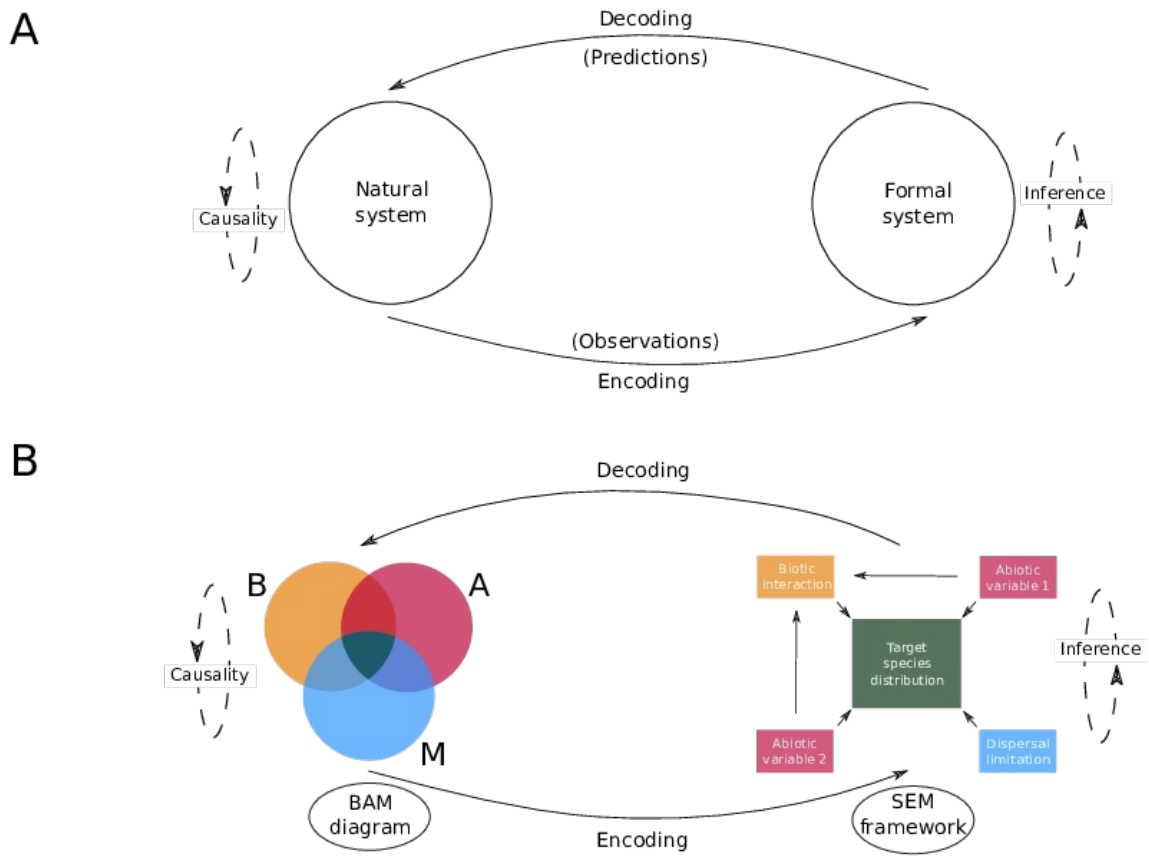
Incorporating causal relations into a model requires a basic knowledge of the study system or organism under investigation in order to formulate specific hypotheses that can later be translated into model equations. In this paper, we define a causal relationship as one for which scientists have a mechanistic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

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

2.1 Rosen's modelling relation

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

6



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

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

7

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

2.2 Applying Rosen's modelling relation

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

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

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

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

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

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

3 Case study

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

3.1 Virtual species

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

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

9

in Bazzichetto et al. (2023).

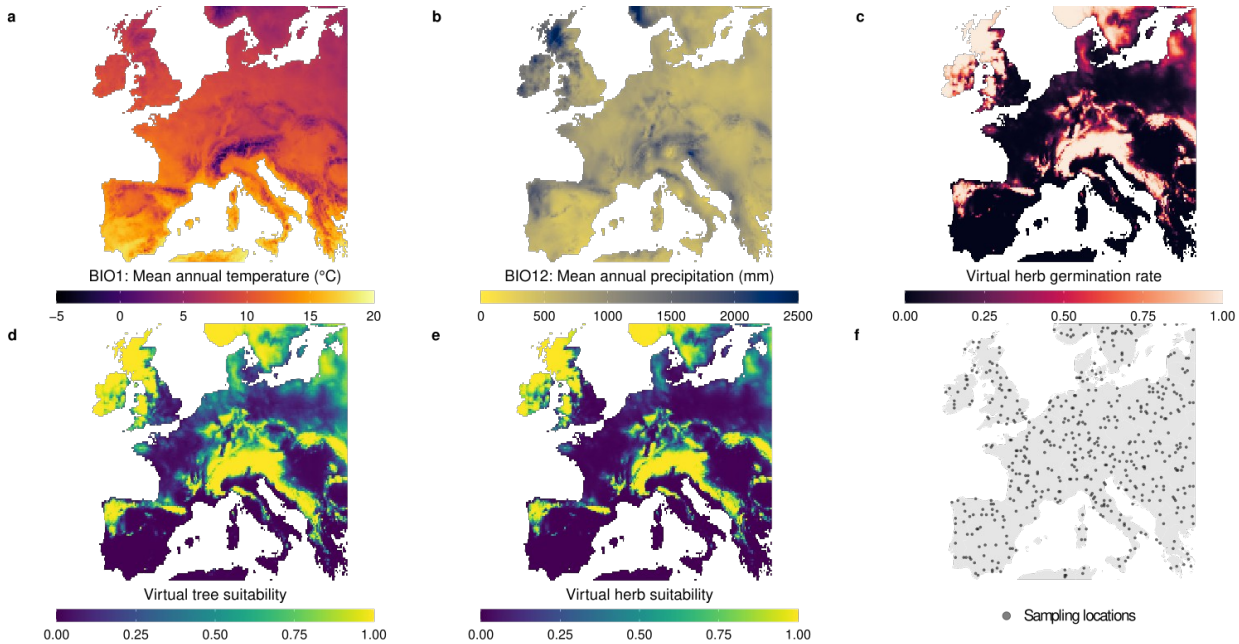
$$\text{logit}(p_i) = \alpha + \beta_{pr} \times \text{precipitations} + \beta_{tm} \times \text{temperature} \quad (2)$$

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

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

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

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



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

3.2 Statistical analysis

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

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

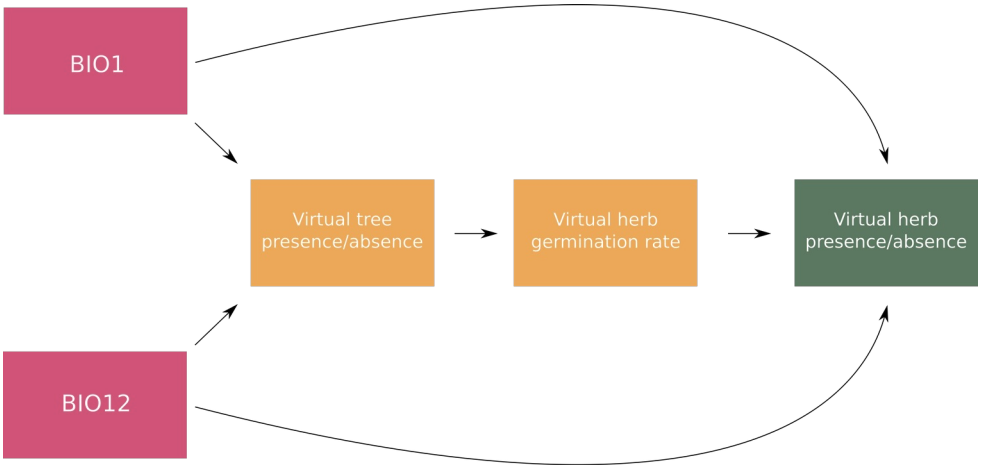
$$\text{Tree} \sim \text{BIO1} + \text{BIO12} \quad (3)$$

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

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

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

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



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

The causal diagram was then converted into a set of candidate models (Eq. 3-5) using the piecewiseSEM and semEff R packages (Lefcheck, 2016; Murphy, 2020). The congruence of the estimated variance-covariance matrix hypothesized in the SEM with the observed variance-covariance matrix in the data was evaluated for each geographic prevalence and cross-validation iterations using a Fisher's C test, whose null hypothesis (H0) is that the model variance-covariance matrix can predict the observed variance-covariance matrix. Hence, a p-value > 0.05 for the Fisher's C test implies that the estimated variance-covariance matrix from the causal diagram mirrors the observed one in the data, therefore validating it (Lefcheck, 2016).

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

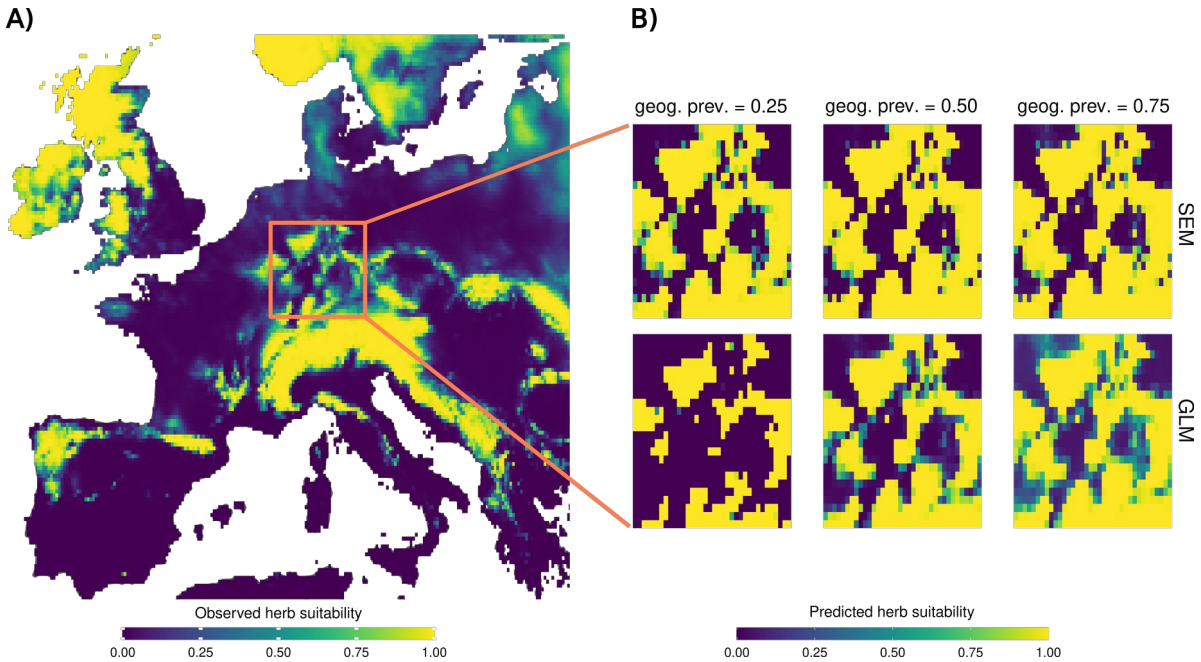
3.3 Results

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

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

12

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



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

4 Discussion

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

13

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

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

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

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

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

Box 1

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

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

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

Fitness: individual reproductive success.

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

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

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

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

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

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

5 Declaration

- Ethics approval and consent to participate: Not applicable.
- Fieldwork permission: Not applicable.
- Competing interests: No conflict of interest has been declared by the authors.
- Funding: This project did not receive specific funding.
- Author's contribution: DDR and SR conceptualized the integration of Rosen’s theory on modelling relation into a species distribution modelling exercise, which was further developed thanks to the suggestions made by SOV and JL on the use of structural equation modelling. DDR and ET performed the data analysis. All the authors critically commented the results and their interpretation; DDR and ET led the writing of the manuscript and produced a

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

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

6 Code availability

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

References

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

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

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

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

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

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

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

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

Bible, R. C. and Peterson, A. T. (2018). Compatible ecological niche signals between biological and archaeological datasets for late-surviving neandertals. American 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 Giljo hann, 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 Zimmer mann, 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., Klonner, G., van Kleunen, M., Essl, F., et al. (2018). Simulating plant
677 invasion dy namics 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 eleva tional 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

report, PeerJ Preprints.

Escobar, L. E. and Craft, M. E. (2016). Advances and limitations of disease biogeography using ecological niche modeling. Frontiers in Microbiology, 7:1174.

Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S. R., Park, H., and Shao, C. (2016). Applications of structural equation modeling (sem) in ecological studies: an updated review. Ecological Processes, 5(1):1–12.

Feng, X., Liang, Y., Gallardo, B., and Papeş, M. (2019). Physiology in ecological niche modeling: using zebra mussel's upper thermal tolerance to refine model predictions through bayesian analysis. Ecography.

Feng, X. and Papeş, M. (2017). Physiological limits in an ecological niche modeling framework: A case study of water temperature and salinity constraints of freshwater bivalves invasive in USA. Ecological Modelling, 346:48–57.

Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. International journal of climatology, 37(12):4302–4315.

Fourcade, Y., Besnard, A. G., and Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Global Ecology and Biogeography, 27(2):245–256.

Fourcade, Y., Engler, J. O., Rödder, D., and Secondi, J. (2014). Mapping species distribu tions with maxent using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. PloS one, 9(5):e97122.

Franklin, J. (2023). Species distribution modelling supports the study of past, present and future biogeographies. Journal of Biogeography.
https://onlinelibrary.wiley.com/doi/10.1111/jbi.14617

Garrido, M., Hansen, S. K., Yaari, R., and Hawlena, H. (2022). A model selection approach to structural equation modelling: A critical evaluation and a road map for ecologists. Methods in Ecology and Evolution, 13(1):42–53.

GBIF: The Global Biodiversity Information Facility (2023) *What is GBIF?*. Available from https://www.gbif.org/what-is-gbif [12 October 2023].

Getz, W. M., Marshall, C. R., Carlson, C. J., Giuggioli, L., Ryan, S. J., Romañach, S. S., Boettiger, C., Chamberlain, S. D., Larsen, L., D'Odorico, P., et al. (2018). Making ecological models adequate. Ecology letters, 21(2):153–166.

Grace, J. (2022). General guidance for custom-built structural equation models. One Ecosystem, 7:e72780.

Grace, J.B. and Irvine, K.M., 2020. Scientist's guide to developing explanatory statistical models using causal analysis principles. Ecology, 101(4), p.e02962.

Grace, J. B. (2006). Structural equation modeling and natural systems. Cambridge University Press.

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 distri bution 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 establish ment 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 Correl ative 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.

Kineman, J. J. and Wessman, C. A. (2021). Relational systems ecology: The anticipatory niche and complex model coupling. Handbook of Systems Sciences, pages 871–916.

Kraemer, M. U., Reiner Jr, R. C., and Bhatt, S. (2019). Causal inference in spatial mapping. Trends in parasitology, 35(10):743–746.

Larter, M., Pfautsch, S., Domec, J.-C., Trueba, S., Nagalingum, N., and Delzon, S. (2017). Aridity drove the evolution of extreme embolism resistance and the radiation of conifer genus callitris. New Phytologist, 215(1):97–112.

Lee-Yaw, J., L. McCune, J., Pironon, S., and N. Sheth, S. (2021). Species distribution models rarely predict the biology of real populations. Ecography, n/a(n/a).

Lefcheck, J. S. (2016). piecewisesem: Piecewise structural equation modelling in r for ecology, evolution, and systematics. Methods in Ecology and Evolution, 7(5):573–579.

Lefcheck, J.S., Byrnes, J.E.K. and Grace, J.B., (2020). piecewiseSEM: Piecewise Structural Equation Modeling (2.1.2)[Computer software].

Lembrechts, J. J., Aalto, J., Ashcroft, M. B., De Frenne, P., Kopeck`y, M., Lenoir, J., Luoto, M., Maclean, I. M., Roupsard, O., Fuentes-Lillo, E., et al. (2020). Soiltemp: A global database of near-surface temperature. Global Change Biology, 26(11):6616– 6629.

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., & Bellard, C. (2018). Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, 45(9), 1994–2002.

Leroy, B., Meynard, C. N., Bellard, C., and Courchamp, F. (2016). virtualspecies, an r package to generate virtual species distributions. Ecography, 39(6):599–607.

Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1), 103–114.

Lozier, J. D., Aniello, P., & Hickerson, M. J. (2009). Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. Journal of Biogeography, 36(9), 1623-1627.

Mäkinen, J. and Vanhatalo, J. (2018). Hierarchical bayesian model reveals the distribu tional shifts of arctic marine mammals. Diversity and Distributions, 24(10):1381–1394.

Marchetto, E., Da Re, D., Tordoni, E., Bazzichetto, M., Zannini, P., Celebrin, S., Chieffallo, L., Malavasi, M., & Rocchini, D. (2023). Testing the effect of sample prevalence and sampling methods on probability-and favourability-based SDMs. *Ecological Modelling*, 477, 110248.

Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., and Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? Ecography, 37(12):1267–1281.

Metcalf, G. S. (2019). Design and the modeling relation. She Ji: The Journal of Design, Economics, and Innovation, 5(4):373–376.

Meisner, J., Kato, A., Lemerani, M. M., Mwamba Miaka, E., Ismail Taban, A., Wakefield, J., ... & Rabinowitz, P. M. (2022). The effect of livestock density on Trypanosoma brucei gambiense and *T. b. rhodesiense*: A causal inference-based approach. PLoS neglected tropical diseases, 16(8), e0010155.

Meynard, C. N., Leroy, B., and Kaplan, D. M. (2019). Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? Ecography, 42(12):2021–2036.

Meynard, C. N., & Kaplan, D. M. (2012). The effect of a gradual response to the environment on species distribution modeling performance. Ecography, 35(6), 499-509.

Mikulecky, D. C. (2001). Robert rosen (1934-1998): a snapshot of biology's newton. Computers and Chemistry, 4(25):317–327.

Murphy, M. (2020). semeff: Automatic calculation of effects for piecewise structural equation models. R package.

Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., and Anderson, R. P. (2014). Enm eval: An r package for conducting spatially indepen dent evaluations and estimating optimal model complexity for maxent ecological niche models. Methods in Ecology and Evolution, 5(11):1198–1205.

Naimi, A. I., Cole, S. R., & Kennedy, E. H. (2017). An introduction to g methods. International journal of epidemiology, 46(2), 756-762.

Nock, C. A., Vogt, R. J., and Beisner, B. E. (2016). Functional Traits, pages 1–8. American Cancer Society.

O'Grady, C. (2020). Psychology's replication crisis inspires ecologists to push for more reliable research. ScienceMag.org.

Pattee, H. H. (2007). Laws, constraints, and the modeling relation–history and interpretations. Chemistry & biodiversity, 4(10):2272–2295.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). Causal inference in statistics: A primer. John Wiley & Sons.

Pennington, P. T., Cronk, Q. C. B., Richardson, J. A., Woodward, F. I., Lomas, M. R., and Kelly, C. K. (2004). Global climate and the distribution of plant biomes. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359(1450):1465–1476.

Pigliucci, M. (2002). Are ecology and evolutionary biology" soft" sciences? In Annales Zoologici Fennici, pages 87–98. JSTOR.

Pocheville, A. (2015). The ecological niche: history and recent controversies. In Handbook of evolutionary thinking in the sciences, pages 547–586. Springer.

Purse, B. V., & Golding, N. (2015). Tracking the distribution and impacts of diseases with biological records and distribution modelling.Biological Journal of the Linnean

Society, 115(3), 664-677.

Qiao, H., Feng, X., Escobar, L. E., Peterson, A. T., Soberón, J., Zhu, G., and Papeş, M. (2019). An evaluation of transferability of ecological niche models. Ecography, 42(3):521–534.

Qiao, H., Soberón, J., and Peterson, A. T. (2015). No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. Methods in Ecology and Evolution, 6(10):1126–1136.

Quiroga, R. E., Premoli, A. C., and Fernández, R. J. (2021). Niche dynamics in am phitropical desert disjunct plants: Seeking for ecological and species-specific influences. Global Ecology and Biogeography, 30(2):370–383.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Regos, A., Gagne, L., Alcaraz-Segura, D., Honrado, J. P., and Domínguez, J. (2019). Ef fects of species traits and environmental predictors on performance and transferability of ecological niche models. Scientific reports, 9(1):1–14.

Robins, J., & Hernan, M. (2008). Estimation of the causal effects of time-varying exposures. Chapman & Hall/CRC Handbooks of Modern Statistical Methods, 553-599.

Rocchini, D., Tordoni, E., Marchetto, E., Marcantonio, M., Barbosa, A. M., Bazzichetto, M., ... & Malavasi, M. (2023). A quixotic view of spatial bias in modelling the distribution of species and their diversity. npj Biodiversity, *2*(1), 10.

Rosen, R. (1978). Fundamentals of measurement and representation of natural systems. Elsevier North-Holland. New York.

Rosen, R. (1986). Anticipatory systems: Philosophical, mathematical and methodological foundations. In Anticipatory systems. Pergamon, Oxford.

Rosen, R. (1993). On models and modeling. Applied mathematics and computation, 56(2-3), 359-372.

Sabatini, F. M., Lenoir, J., Hattab, T., Arnst, E. A., Chytr`y, M., Dengler, J., De Ruffray, P., Hennekens, S. M., Jandt, U., Jansen, F., et al. (2021). splotopen–an environmen tally balanced, open-access, global dataset of vegetation plots. Global Ecology and Biogeography, 30(9):1740–1764.

Sales, L. P., Hayward, M. W., and Loyola, R. (2021). What do you mean by “niche”? modern ecological theories are not coherent on rhetoric about the niche concept. Acta Oecologica, 110:103701.

Siekmann, I. (2018). An applied mathematician’s perspective on rosennean complexity. Ecological Complexity, 35:28–38.

Sillero, N., Arenas-Castro, S., Enriquez-Urzelai, U., Vale, C. G., Sousa-Guedes, D., Martínez-Freiría, F., Real, R., and Barbosa, A. M. (2021). Want to model a species niche? a step-by-step guideline on correlative ecological niche modelling. Ecological Modelling, 456:109671.

Sillero, N. and Barbosa, A. M. (2020). Common mistakes in ecological niche models. International Journal of Geographical Information Science, pages 1–14.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., and O'Hara, R. B. (2020). Is more data always better? a simulation study of benefits and limitations of integrated distribution models. Ecography.

Soberón, J. and Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics, 2 (January). https://doi.org/10.17161/bi.v2i0.4.

Staniczenko, P. P., Sivasubramaniam, P., Suttle, K. B., & Pearson, R. G. (2017). Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. Ecology letters, 20(6), 693-707.

Strubbe, D., Jiménez, L., Barbosa, A. M., Davis, A. J., Lens, L., & Rahbek, C. (2023). Mechanistic models project bird invasions with accuracy. Nature Communications, 14(1), 2520.

Thuiller, W., Münkemüller, T., Lavergne, S., Mouillot, D., Mouquet, N., Schiffers, K., and Gravel, D. (2013). A road map for integrating eco-evolutionary processes into biodiversity models. Ecology letters, 16:94–105.

Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J.-B., Pe'er, G., Singer, A., Bridle, J., Crozier, L., De Meester, L., Godsoe, W., et al. (2016). Improving the forecast for biodiversity under climate change. Science, 353(6304):aad8466.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. Nature Reviews Methods Primers, 1(1):1–26.

Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F. (2014). En vironmental filters reduce the effects of sampling bias and improve predictions of eco logical niche models. Ecography, 37(11):1084–1091.

Zhang, B., & DeAngelis, D. L. (2020). An overview of agent-based models in plant biology and ecology. Annals of Botany, 126(4), 539-557.

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., et al. (2020). A standard protocol for reporting species distribution models. Ecography.

Supplementary Materials 1

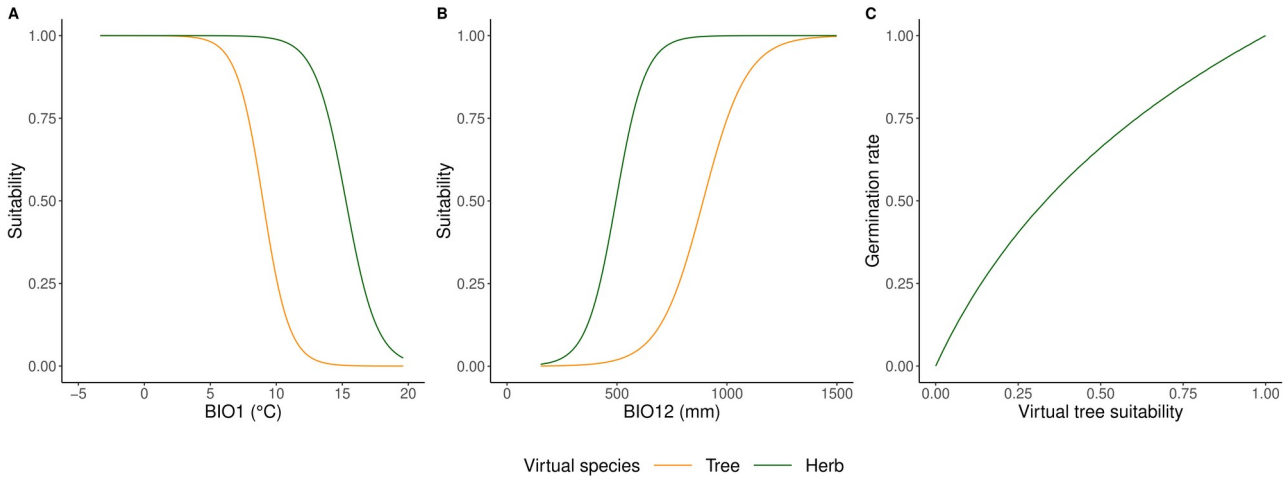


Figure S1.1 Simulated response curves for the tree (orange) and herb (green) virtual species along the temperature (A) and precipitation (B) gradients. Herb virtual species germination rate along a gradient of the virtual tree species suitability (c).

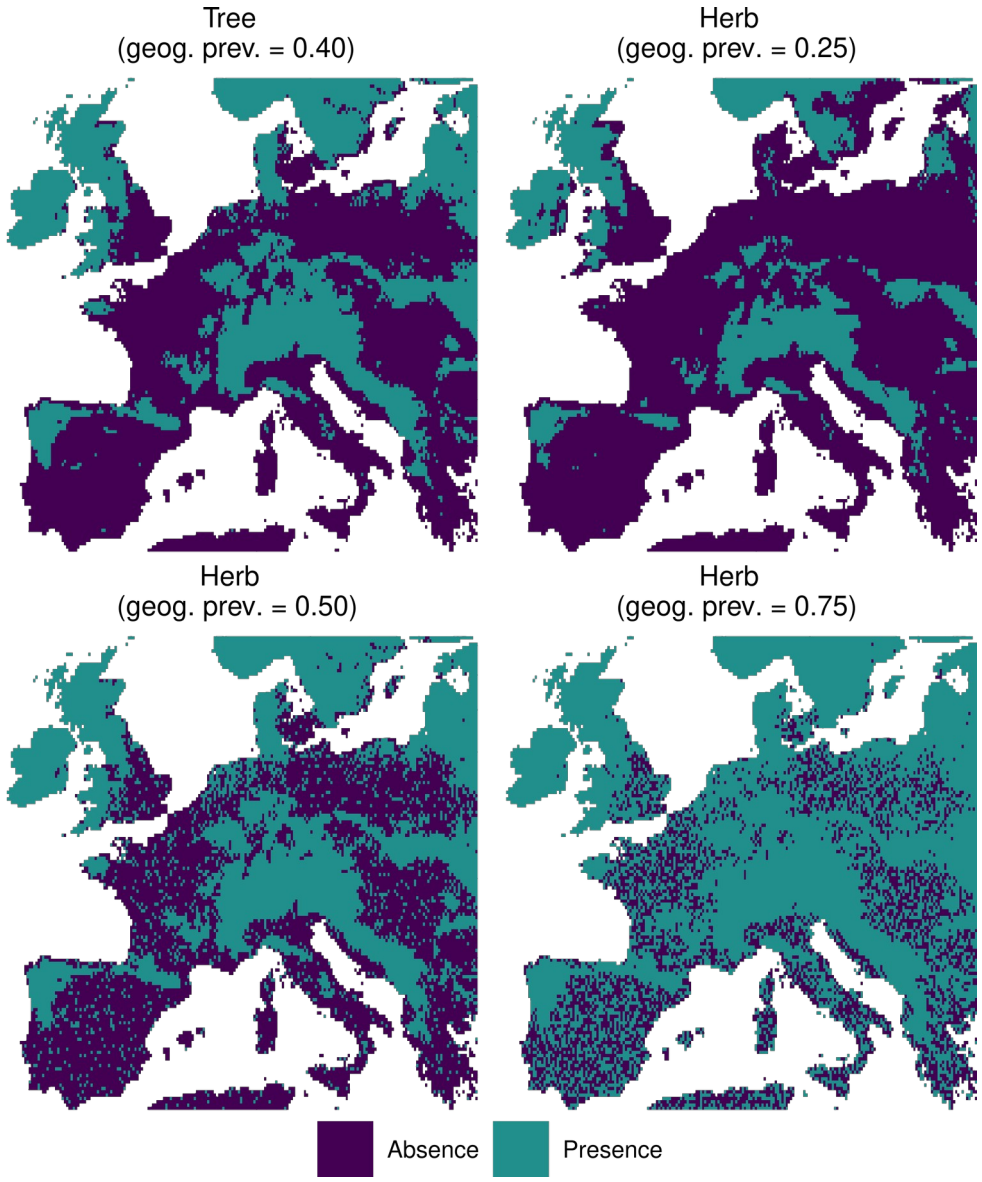


Figure S1.2 Tree and herb virtual species presence-absence distribution along different geographical prevalences.

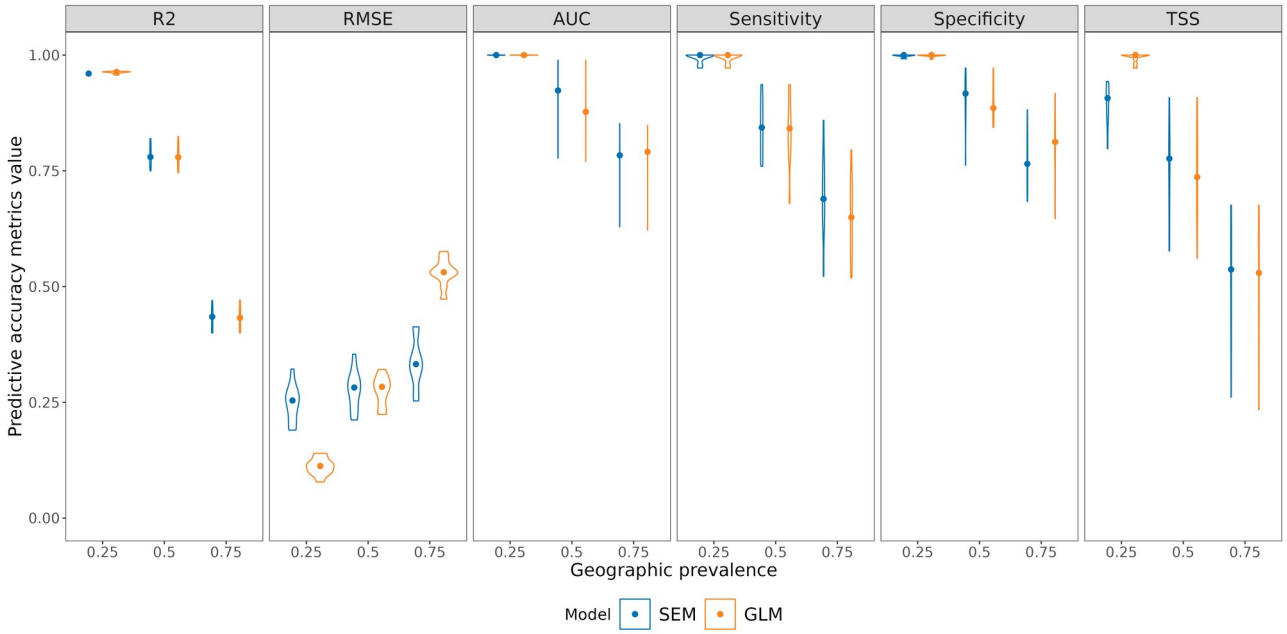


Figure S1.3 Violin plots reporting the distribution of the values of the metrics of predictive performance for the virtual herb species habitat suitability modeled as a function of the tree virtual species presence-absence and virtual herb species germination rate, and varying the geographical prevalence 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).

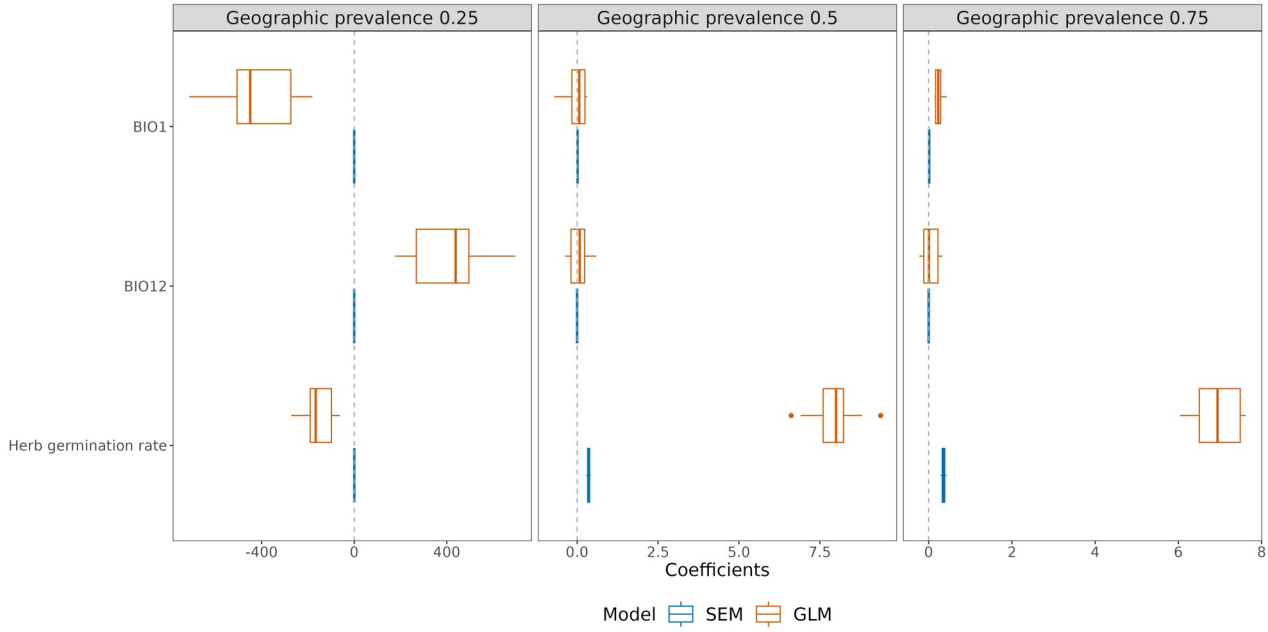


Figure S1.4 Boxplots reporting the distribution of the values of coefficients estimates of the virtual herb species habitat suitability modeled as a function of BIO1, BIO12 and virtual herb species germination rate, and varying the geographical prevalence of the herb species. Colours are associated with 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