Received: 29 July 2022 | Revised: 13 October 2023 | Accepted: 29 November 2023

DOI: 10.1111/jbi.14775

Journal of Biogeography

WILEY

SYNTHESIS

Towards causal relationships for modelling species distribution

Daniele Da Re[1] | Enrico Tordoni[2] | Jonathan Lenoir[3] | Sergio Rubin[1] | Sophie O. Vanwambeke[1]

[1]Center for Earth and Climate Research, Earth and Life Institute, UCLouvain, Louvain-la-Neuve, Belgium

[2]Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia

[3]UMR CNRS 7058, Ecologie et Dynamique des Systèmes Anthropisés (EDYSAN), Université de Picardie Jules Verne, Amiens, France

Correspondence
Daniele Da Re, Center Agriculture Food Environment, University of Trento, San Michele all'Adige, Italy.
Email: daniele.dare@unitn.it

Present address
Daniele Da Re, Center Agriculture Food Environment, University of Trento, San Michele all'Adige, Italy

Funding information
Daniele Da Re was supported by a FRS-FNRS ASP Belgian grant No 34766961, Enrico Tordoni is supported by the Estonian Research Council grant (MOBJD1030).

Abstract

Aim: 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 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 the species under investigation. This might result in an oversimplification of the complex niche hypervolume, resulting in a single model formula whose estimates and predictions lack ecological realism.

Location: Not applicable.

Time Period: Not applicable.

Major Taxa Studied: Virtual species.

Materials and Methods: We believe that a causal perspective associated with a finer definition of the modelling target is necessary to develop more ecologically 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.

Results: 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.

Main Conclusions: Including causal processes underlying the spatial distribution of species into an inferential formal system highlights the methodological steps where uncertainty can arise and results 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

Daniele Da Re and Enrico Tordoni equally contributed to the study.

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; GBIF: The Global Biodiversity Information Facility, 2023; Sabatini et al., 2021), environmental predictors (e.g., WorldClim, Fick & 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 (Araújo et al., 2019; Kearney & Porter, 2009; Lee-Yaw et al., 2021), expressing specific criticisms about (i) the conceptual background of correlative SDMs (Austin, 2007; Kearney, 2006); (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 (iv) the ecological realism of SDMs outputs (Lee-Yaw et al., 2021). 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; Sillero et al., 2021; Zurell et al., 2020). 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; Hellegers et al., 2020; Merow et al., 2014).

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., Austin, 2007; Chapman et al., 2019; Kearney & Porter, 2009; Purse & Golding, 2015; Urban et al., 2016). For instance, models based on species life-history traits (i.e., the characteristics influencing individuals' performance or fitness; Dawson et al., 2021; Nock et al., 2016) 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, constitute 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 into models, as when dealing with complex spatio-temporal dynamics or when sampling efforts varied (Mäkinen & 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 to as directed acyclic graphs (DAGs) with a set of rules that can be applied for observational causal inference in ecology (Arif & 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 require 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 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 Robert Rosen's modelling relation framework (Rosen, 1978, 1986, 1993) as a causal scheme to guide the design of SDMs. 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 the 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; Figure 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 an SEM framework and comparing its results with those of a generalized linear model (GLM), a common algorithm used in correlative SDMs.

2 | INCORPORATING HYPOTHESIZED CAUSAL RELATIONSHIPS INTO SDMS

The niche concept is a fundamental notion in ecology and represents the conceptual backbone of SDMs. Different definitions of the niche concept have been proposed (Pocheville, 2015; Sales et al., 2021) but, essentially, the niche concept aims to define the environmental space in which a species could exist, persist and reproduce, allowing us to identify the geographical area where those environmental conditions are met. The design and interpretation of correlative SDMs are usually framed within the niche concept provided by Soberón

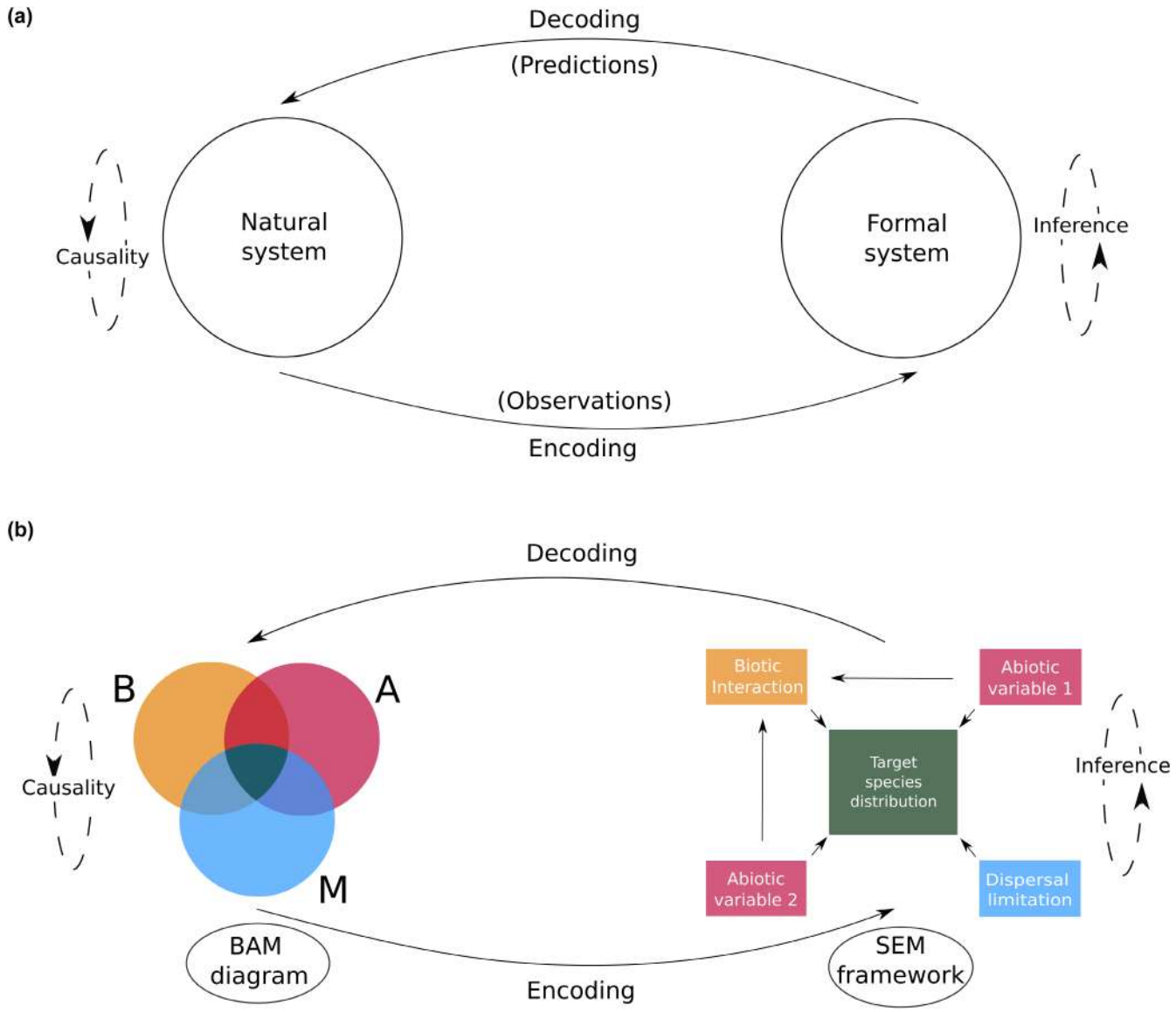

FIGURE 1 (a) Robert Rosen's modelling relation. (b) Example of application of the modelling relation to model the distribution of a species, depicted in green within the biotic abiotic movement diagram (BAM; natural system), by means of a structural equation model (SEM; formal system).

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 & Peterson, 2005). 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., Bible & Peterson, 2018; Escobar & Craft, 2016; Franklin, 2023).

Correlative SDMs' outputs depict (and synthesize) 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 model's predictive accuracy) rather than in terms of their ecological realism. Many efforts have been devoted to solving 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., Aiello-Lammens et al., 2015; Bazzichetto et al., 2023; Brun et al., 2020; Fourcade et al., 2014; Hallgren et al., 2019; Muscarella et al., 2014; Qiao et al., 2015, 2019; Simmonds et al., 2020; Varela et al., 2014; see Sillero & 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 & 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 (Arif & MacNeil, 2023; Briscoe et al., 2019; Feng & Papeş, 2017; Hartemink et al., 2011; Kearney & Porter, 2009; Kraemer et al., 2019; Staniczenko et al., 2017; Urban et al., 2016). Indeed, a modelling perspective based on the biology of the target organism and associated 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 the response variable and predictors, especially when the predictors have no direct causal links with the response variable (Fourcade et al., 2018; Journé et al., 2020; Lozier et al., 2009), 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 occurring in a predictor 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 & 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 the model's biological realism (e.g. Araújo et al., 2019; Chapman et al., 2019; Srivastava et al., 2021; Zurell et al., 2020). Conceptually speaking, we believe the so-called modelling relation framework developed by Robert Rosen in the 1980s (Rosen, 1986) 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 the 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 (Figure 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 Figure 1 represents the observations and measurements of the *natural systems* aiming to capture its causality, while the arrow from the *formal system* towards the *natural system* represents the decoding operation of the prediction into the *natural system* made by the mathematical *formal system*.

Although 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 & MacNeil, 2022), the modelling relation framework represents a valid epistemological tool to guide (and refine) the incorporation of ecological knowledge into biologically more 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; Figure 1a), and to define them as mathematical equations and relations (e.g., translating the BAM diagram into a causal and mathematical diagram; Figure 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 are compared and interpreted in the light of the ecological assumptions they, respectively, made on the *natural system* (Fudge & 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 (Metcalf, 2019; Siekmann, 2018).

2.2 | Applying Rosen's modelling relation

To date, few attempts have been made to include the modelling relations in 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 of the statistical method used (Metcalf, 2019; Siekmann, 2018), but we suggest that the rationale behind the SEM approach (Grace, 2006) fits well within the modelling relation of the *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). All relies on 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 data set), thanks to the piecewiseSEM R package (Lefcheck, 2016; Lefcheck et al., 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 data set (Equation 1):

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

where Σ is the observed variance–covariance matrix, and $\Sigma(\Phi)$ is the model-estimated variance-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 to propose the application of SEM to SDMs, advocating the importance of including and evaluating a causal structure in 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 (Grace, 2022; Lefcheck, 2016). Recent technical developments overcome some of these limitations (e.g., Carvalho-Rocha et al., 2021; Cerqueira et al., 2021; Chu et al., 2019; Quiroga et al., 2021; Walentinowitz et al., 2023), 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 employing 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 precisely 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 & 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 (Figure 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; Figure S1.1a,b). This results in a tree species mostly distributed in the mountainous area of Europe (Figure 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; Figure S1.1a,b), 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 GLMs, or logistic regressions, assuming sigmoid (i.e., non-quadratic) response curves between the occurrence of the species and the chosen predictors (Equation 2), and following the approach described 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 $p_i/(1\text{-}p_i)$, α is the model intercept, β_{pr} is the regression parameter for the linear term (i.e., sigmoid shape) of precipitation and 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}), while 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 (Figure S1.1c).

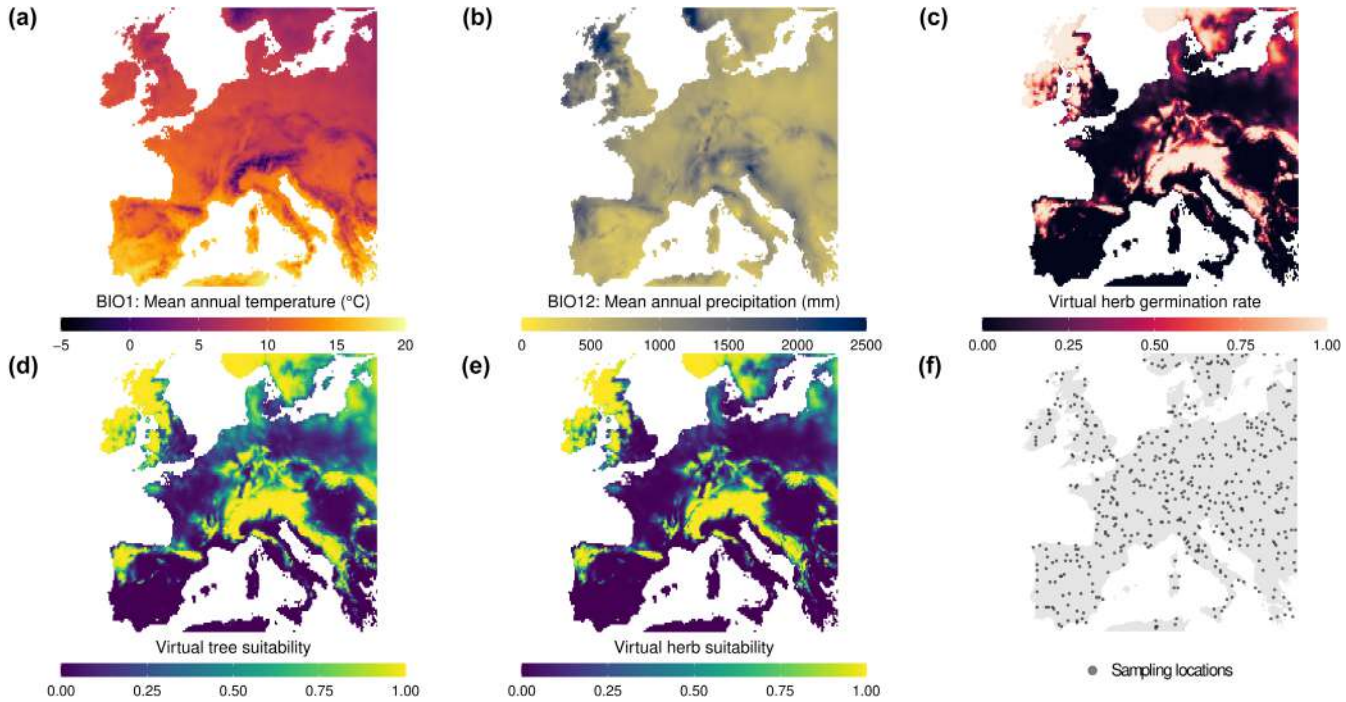

FIGURE 2 (a, b) The set of abiotic variables (BIO1 and BIO1) used to create the two virtual species. (c) The germination rate of the virtual herb species is computed as a function of the habitat suitability of the virtual tree species. (d) The habitat suitability of the virtual tree species. (e) The habitat suitability of the virtual herb species. (f) Sampling locations. The geographic projection used is the WGS84—World Geodetic System 1984, EPSG: 4326.

Eventually, the resulting geographical distribution of the virtual herbaceous species (Figure 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 (Figure 2a–c). The obtained habitat suitability maps of the two virtual species (Figure 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 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 equal 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 (Figure 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 (Figure 2f). We repeated this operation 10 times, and the predictive accuracy of each simulation was estimated using 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 data set to calibrate our SEM models built within 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.

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 the models' 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 (Figure 3), whereby the geographical distribution of the forest herb species represents the *natural system* and the causal diagram from the SEM approach represents the *formal system*. In the causal diagram or DAG (Figure 3):

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

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

2. 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 (Equation 4);

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

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

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

The causal diagram was then converted into a set of candidate models (Equations 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 & 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 initially support the causal diagram proposed in Figure 3 as a valid hypothetical causal structure representing the variance–covariance matrix observed in the training data set ($p<0.05$), suggesting the inclusion of direct effects for both BIO1 and BIO12 on the germination rate of the herb species (Equation 4). Once these two additional direct effects were integrated in Equation 4, 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 data set 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 (Figure S1.3). The 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: while 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 (Figure 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 (Figure 4). As a consequence, 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 a 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 (Table S1.5).

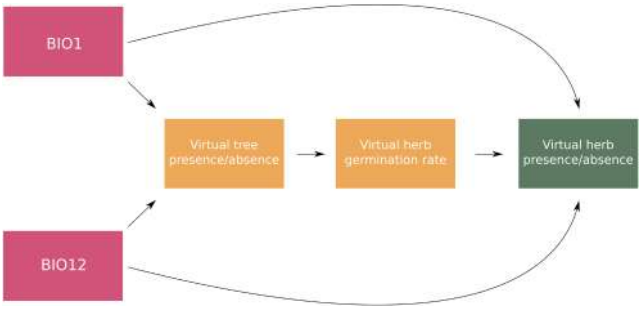

FIGURE 3 Hypothesized causal diagram explaining the distribution of the virtual herb species. Purple boxes indicate abiotic variables, orange boxes indicate biotic variables and green box displays the main response variable.

4 | DISCUSSION

We introduced 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 into correlative SDMs and to account for the hierarchical structure of the links between variables, 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 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

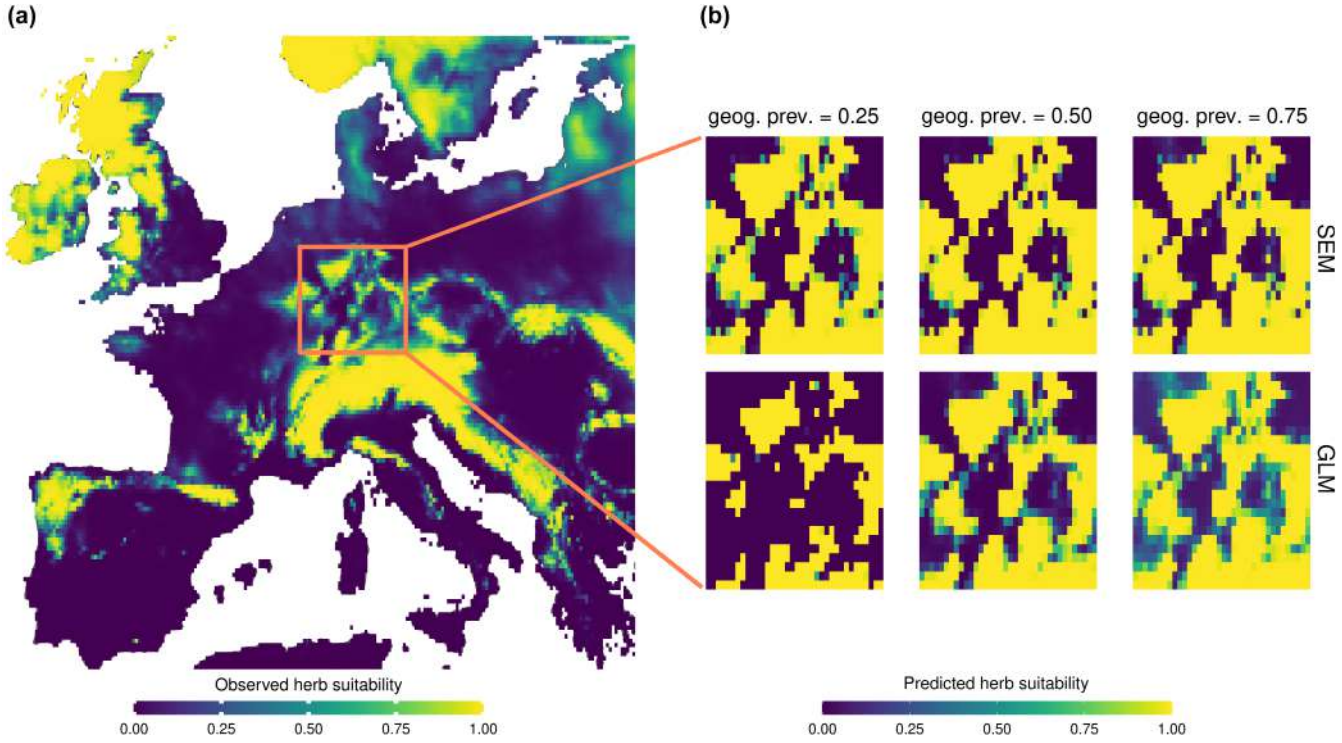

FIGURE 4 The observed (A) and predicted (B) habitat suitability values for the virtual herb species in a subset of the study area under different combinations of geographic prevalences and models (SEM and GLM). The geographic projection used is the WGS84—World Geodetic System 1984, EPSG: 4326.

effective way to assess how appropriate different models are. 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 (Dubos et al., 2022; Fourcade et al., 2018; Jiménez-Valverde, 2021; Rocchini et al., 2023) and the size of the study region (Lobo et al., 2010).

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 & MacNeil, 2022; Pichler and Harting, 2023). 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 aspects of SEM application to SDMs is the capacity to unravel unanticipated mechanisms through conditional independence testing, for example, 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 (Arif & MacNeil, 2022; Lefcheck, 2016; Lefcheck et al., 2020). While the natural-to-formal systems relationships presented in Rosen's modelling relation are 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 Hernan (2008) 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; Meisner et al., 2022; Naimi et al., 2017). 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 offspring) to different factors affecting its performance (e.g., environmental conditions) (Da Re et al., 2022; Kearney et al., 2010). For instance, Larter et al. (2017) showed how a single plant functional trait (xylem resistance to cavitation) displayed a strong statistical relationship with its species distribution in relation to aridity across the climatic range of the species. Process-based SDMs have also been successfully used in invasion ecology to simulate and forecast invasion risk under different global change scenarios (Carboni et al., 2018; Strubbe et al., 2023). Within the family of process-based models, agent-based models (ABMs) aim to predict species population or community dynamics by modelling multiple individuals (agents) that interact with their environment and among each other. For each agent, ABMs require the specification of state variables, which can include age, size and spatial location, as well as physiological and behavioural traits (Zhang & DeAngelis, 2020).

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

Ecologists aspire to foster knowledge of global environmental changes induced by human activities, such as climate change, biological invasions and habitat loss. To efficiently tackle such challenges, clear, robust and well-defined epistemological premises about the main determinants of species distribution and species distribution changes are needed to design realistic experiments (Currie, 2019; Pigliucci, 2002). Epistemological premises are not just philosophical murmuring but allow us to set the boundaries of the modelling exercise, increasing model robustness in depicting natural patterns and resulting in clear practical applications (Currie, 2019; Dawson et al., 2023). Rosen's modelling relation and its implementation by means of the SEM approach requires to clearly define the *natural system* (the key response variable of interest), such as the *niche*, *habitat* or *biome* (see Box 1), which inherently defines different biological

BOX 1 Natural systems glossary.

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., tundra), *sensu* Pennington et al., 2004).

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

Fitness: individual reproductive success.

Functional trait: Those characteristics influencing the 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 immigration.

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.

entities and cannot be used interchangeably. It may also help to identify when model assumptions are causal or not and to develop a suite of model comparisons (hypothesis-driven modelling) that can robustly explain the variation in the data while accounting for ecological observations.

ACKNOWLEDGMENTS

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

FUNDING INFORMATION

This project did not receive specific funding.

CONFLICT OF INTEREST STATEMENT

No conflict of interest has been declared by the authors.

DATA AVAILABILITY STATEMENT

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

ORCID

Daniele Da Re https://orcid.org/0000-0002-3398-9295
Enrico Tordoni https://orcid.org/0000-0002-9722-6692
Jonathan Lenoir https://orcid.org/0000-0003-0638-9582
Sergio Rubin https://orcid.org/0000-0002-3387-7760
Sophie O. Vanwambeke https://orcid.org/0000-0001-6620-6173

REFERENCES

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & 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., & O'Hara, R. B. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, *5*(1), eaat4858.

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

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

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

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., Doherty, M., & 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., Barták, V., & 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., & 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.

Briscoe, N. J., Elith, J., Salguero-Gómez, R., Lahoz-Monfort, J. J., Camac, J. S., Giljo hann, K. M., Holden, M. H., Hradsky, B. A., Kearney, M. R., McMahon, S. M., & Phillips, B. L. (2019). Forecasting species range dynamics with process-explicit models: Matching methods to applications. *Ecology Letters*, *22*(11), 1940–1956.

Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wüest, R. O., Wang, Z., & Zimmer mann, N. E. (2020). Model complexity affects species distribution projections under climate change. *Journal of Biogeography*, *47*(1), 130–142.

Carboni, M., Guéguen, M., Barros, C., Georges, D., Boulangeat, I., Douzet, R., Dullinger, S., Klonner, G., van Kleunen, M., Essl, F., Bossdorf, O., Haeuser, E., Talluto, M. V., Moser, D., Block, S., Conti, L., Dullinger, I., Münkemüller, T., & Thuiller, W. (2018). Simulating plant invasion dy namics in mountain ecosystems under global change scenarios. *Global Change Biology*, *24*(1), e289–e302.

Carvalho-Rocha, V., Peres, C. A., & Neckel-Oliveira, S. (2021). Habitat amount and ambient temperature dictate patterns of anuran diversity along a subtropical eleva tional gradient. *Diversity and Distributions*, *27*(2), 344–359.

Cerqueira, R. C., de Rivera, O. R., Jaeger, J. A., & Grilo, C. (2021). Direct and indirect effects of roads on space use by jaguars in Brazil. *Scientific Reports*, *11*(1), 22617.

Chapman, D., Pescott, O. L., Roy, H. E., & Tanner, R. (2019). Improving species distribution models for invasive non-native species with biologically informed pseudo-absence selection. *Journal of Biogeography*, *46*(5), 1029–1040.

Chu, C., Lutz, J. A., Král, K., Vrška, T., Yin, X., Myers, J. A., Abiem, I., Alonso, A., Bourg, N., Burslem, D. F., & Cao, M. (2019). Direct and indirect effects of climate on richness drive the latitudinal diversity gradient in forest trees. *Ecology Letters*, *22*(2), 245–255.

Currie, D. J. (2019). Where newton might have taken ecology. *Global Ecology and Biogeography*, *28*(1), 18–27.

Da Re, D., Van Bortel, W., Reuss, F., Müller, R., Boyer, S., Montarsi, F., Ciocchetta, S., Arnoldi, D., Marini, G., Rizzoli, A., L'Ambert, G., Lacour, G., Koenraadt, C. J. M., Vanwambeke, S. O., & Marcantonio, M. (2022). dynamAedes: A unified modelling framework for invasive Aedes mosquitoes. *Parasites & Vectors*, *15*(1), 1–18.

Dawson, M. N., Mainali, K., Meyer, R., Noonan, M., Papeş, M., Parenti, L. R., & Villalobos, F. (2023). Reshaping biogeography: Perspectives on the past, present and future. *Journal of Biogeography*, *50*(8), 1405–1408.

Dawson, S. K., Carmona, C. P., González-Suárez, M., Jönsson, M., Chichorro, F., Mallen Cooper, M., Melero, Y., Moor, H., Simaika, J. P., & Duthie, A. B. (2021). The traits of "trait ecologists": An analysis of the use of trait and functional trait terminology. *Ecology and Evolution*, *11*(23), 16434–16445.

Dubos, N., Préau, C., Lenormand, M., Papuga, G., Monsarrat, S., Denelle, P., Le Louarn, M., Heremans, S., Roel, M., Roche, P., & Luque, S. (2022). Assessing the effect of sample bias correction in species distribution models. *Ecological Indicators*, *145*, 109487.

Enquist, B. J., Condit, R., Peet, R. K., Schildhauer, M., & Thiers, B. M. (2016). Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. Technical report, PeerJ Preprints.

Escobar, L. E., & 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., & 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., & Papeş, M. (2019). Physiology in ecological niche modeling: Using zebra mussel's upper thermal tolerance to refine model predictions through bayesian analysis. *Ecography*, *43*, 270–282.

Feng, X., & 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., & 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., & 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., & 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*, *50*, 1533–1545. https://doi.org/10.1111/jbi.14617

Fudge, D. S., & Turko, A. J. (2020). The best predictions in experimental biology are critical and persuasive. *Journal of Experimental Biology*, *223*.

Garrido, M., Hansen, S. K., Yaari, R., & 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? https://www.gbif.org/what-is-gbif

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., & O'Sullivan, D. (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. (2006). *Structural equation modeling and natural systems*. Cambridge University Press.

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

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*, 37–48.

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: with applications in R*. Cambridge University Press.

Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y., & Mackey, B. (2019). Species distri bution models can be highly sensitive to algorithm configuration. *Ecological Modelling*, *408*, 108719.

Hartemink, N., Vanwambeke, S. O., Heesterbeek, H., Rogers, D., Morley, D., Pesson, B., Davies, C., Mahamdallie, S., & Ready, P. (2011). Integrated mapping of establish ment risk for emerging vector-borne infections: A case study of canine leishmaniasis in Southwest France. *PLoS One*, *6*(8), e20817.

Hellegers, M., Ozinga, W. A., Hinsberg van, A., Huijbregts, M. A., Hennekens, S. M., Schaminée, J. H., Dengler, J., & Schipper, A. M. (2020). Evaluating the ecological realism of plant species distribution models with ecological indicator values. *Ecography*, *43*(1), 161–170.

Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M., & Baselga, A. (2008). Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, *117*(6), 847–858.

Hutchinson, G. (1957). *Concluding remarks cold spring harbor symposia on quantitative biology* (Vol. *22*, pp. 415–427). GS SEARCH.

Jiménez-Valverde, A. (2021). Prevalence affects the evaluation of discrimination capacity in presence-absence species distribution models. *Biodiversity and Conservation*, *30*(5), 1331–1340.

Journé, V., Barnagaud, J. Y., Bernard, C., Crochet, P. A., & Morin, X. (2020). Correl ative climatic niche models predict real and virtual species distributions equally well. *Ecology*, *101*(1), e02912.

Kearney, M. (2006). Habitat, environment and niche: What are we modelling? *Oikos*, *115*(1), 186–191.

Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. *Ecology Letters*, *12*(4), 334–350.

Kearney, M., Simpson, S. J., Raubenheimer, D., & Helmuth, B. (2010). Modelling the ecological niche from functional traits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1557), 3469–3483.

Keil, A. P., Edwards, J. K., Richardson, D. R., Naimi, A. I., & Cole, S. R. (2014). The parametric G-formula for time-to-event data: Towards intuition with a worked example. *Epidemiology (Cambridge, Mass.)*, *25*(6), 889.

Kineman, J. J. (2007). Relational complexity in natural science and the design of ecological informatics (PhD thesis). Citeseer.

Kineman, J. J. (2009). Relational theory and ecological niche modelling. In *Proceedings of the 53rd Annual Meeting of the ISSS-2009*.

Kineman, J. J., & Wessman, C. A. (2021). Relational Systems Ecology: The Anticipatory Niche and Complex Model Coupling. In G. S. Metcalf, K. Kijima, & H. Deguchi (Eds.), *Handbook of Systems Sciences*. Springer. https://doi.org/10.1007/978-981-15-0720-5_79

Kraemer, M. U., Reiner, R. C., Jr., & 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., & 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., McCune, L. J., Pironon, S. N., & Sheth, S. (2022). Species distribution models rarely predict the biology of real populations. *Ecography*, *2022*(6), e05877.

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., & 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., & García, R. A. (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., & 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., & 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.

Meisner, J., Kato, A., Lemerani, M. M., Mwamba Miaka, E., Ismail Taban, A., Wakefield, J., Rowhani-Rahbar, A., Pigott, D. M., Mayer, J. D., & 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.

Merow, C., Smith, M. J., Edwards, T. C., Jr., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., & 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.

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.

Meynard, C. N., Leroy, B., & 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.

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., & 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., & Beisner, B. E. (2016). *Functional traits* (pp. 1–8). American Cancer Society.

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

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, *14*, 994–1016.

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

Pocheville, A. (2015). The ecological niche: History and recent controversies. In *Handbook of evolutionary thinking in the sciences* (pp. 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., & Papeş, M. (2019). An evaluation of transferability of ecological niche models. *Ecography*, *42*(3), 521–534.

Qiao, H., Soberón, J., & 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., & 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.

Regos, A., Gagne, L., Alcaraz-Segura, D., Honrado, J. P., & 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. In *Handbooks of modern statistical methods* (pp. 553–599). Chapman & Hall/CRC.

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.

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

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., & Jiménez-Alfaro, B. (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., & 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., & 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., & Barbosa, A. M. (2021). Common mistakes in ecological niche models. *International Journal of Geographical Information Science*, *35*(2), 213–226.
Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., & O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*, 1413–1422.
Soberon, J., & Townsend Peterson, A. (2005). Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas. *Biodiversity Informatics*, *2*. https://doi.org/10.17161/bi.v2i0.4
Srivastava, V., Roe, A. D., Keena, M. A., Hamelin, R. C., & Griess, V. C. (2021). Oh the places they'll go: improving species distribution modelling for invasive forest pests in an uncertain world. *Biological Invasions*, *23*, 297–349.
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., & 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., & Gonzalez, A. (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., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1–26.
Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*(11), 1084–1091.
Walentowitz, A., Ferreira-Arruda, T., Irl, S. D., Kreft, H., & Beierkuhnlein, C. (2023). Disentangling natural and anthropogenic drivers of native and non-native plant diversity on North Sea islands. *Journal of Biogeography*.
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., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*, 1261–1277.

BIOSKETCH

Daniele Da Re is a postdoctoral researcher at the University of Trento. His current research focuses on developing and applying correlative and mechanistic species distribution models to unravel the complexities of invasive arthropods to address the challenges posed by vector-borne diseases.

Enrico Tordoni is a Research Fellow at the University of Tartu. His current research focuses on disentangling global diversity spatiotemporal patterns of plants and vertebrates also accounting for functional and evolutionary aspects under the current scenario of global change.

Jonathan Lenoir is a CNRS researcher in Ecology and biostatistics. He is broadly interested in the ecological dynamics associated with spatial and temporal global changes, with particular emphasis on the biotic responses to contemporary climate change. His current research focuses on biodiversity redistribution and forest microclimates.

Sergio Rubin is Topical Advisory Panel Editor of Mathematical Biology Journal and belongs to the Editorial Board of the Second order Cybernetics, Autopoiesis & Cybersemiotics Journal. His current research focuses on the Earth System Functioning and the Gaia hypothesis using advanced theories of cognitive sciences such as the free energy principle and the autonomous biochemical organization of cellular metabolism.

Sophie Vanwambeke is a professor of Geography at UCLouvain. Her research in medical geography and spatial epidemiology focuses on health as a feature of human-environment interactions.

Author contributions: Daniele Da Re and Sergio Rubin 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 Sophie O. Vanwambeke and Jonathan Lenoir on the use of structural equation modelling. Daniele Da Re and Enrico Tordoni performed the data analysis. All the authors critically commented the results and their interpretation; Daniele Da Re and Enrico Tordoni led the writing of the manuscript and produced a first draft, which was further improved by all other authors.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Da Re, D., Tordoni, E., Lenoir, J., Rubin, S., & Vanwambeke, S. O. (2024). Towards causal relationships for modelling species distribution. *Journal of Biogeography*, *51*, 840–852. https://doi.org/10.1111/jbi.14775