Integrating food webs in species distribution models improves ecological niche estimation and predictions

Giovanni Poggiato[1,2], Jérémy Andréoletti[1] , Laura J. Pollock[3] and Wilfried Thuiller[1]

[1] Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, F-38000 Grenoble

[2] Univ. Grenoble Alpes, CNRS, INRIA, LJK, F-38000 Grenoble, France

[3] Biology department, McGill University, Montréal, Canada

Article type: **method**

Running title: Trophic species distribution models

Statement of authorship: G.P. and W.T. conceived the study, with early advice from L.J.P.. J.A. and G.P. performed all analyses with support from W.T., G.P. wrote the first version of the manuscript with strong input from W.T.. All authors contributed substantially to the writing of the manuscript and its revisions.

Data accessibility statement: The data and codes are available in the GitHub repository of the first author https://github.com/giopogg/webSDM, in the publication folder.

Number of words in the abstract: 145

Number of words in the main text: 4998

Number of words Box 1: 727

Number of boxes: 1

Number of figures: 5

Number of references: 80

Keywords: species distribution models, metanetworks, structural equation models, biotic interactions, biodiversity modeling

Abstract

Biotic interactions drive multitrophic species community assembly. Yet, explicitly incorporating this process in species distribution models (SDMs) is particularly challenging, even when biotic interactions are known. Here, we propose a framework that combines knowledge of trophic interactions with Bayesian structural equation models to model each species as a function of its prey or predators and environmental conditions. We tested and validated our framework on realistic simulated communities spanning different theoretical models and ecological setups. We showed that our framework improves the inference of both species' potential and realized niches compared to single SDMs (mean performances increased by 8% and 6% respectively), especially for species with strong biotic control, thus increasing model predictive performance. Our framework can easily integrate various SDM extensions (e.g., occupancy models) and algorithms, and stands out as a novel solution for modeling multitrophic community distributions when trophic interactions are known or assumed.

Introduction

Environmental changes pose a significant threat to multi-trophic biodiversity, necessitating robust conservation strategies to uphold essential ecosystem functions and services (Rosa *et al.* 2017; Pollock *et al.* 2020). Predicting biodiversity responses to global change has thus emerged as a vibrant research area, fueled by considerable expectations from scientific, conservation, and political communities (Potts *et al.* 2016). Given that accurate biodiversity predictions demand reliable and ecologically sound models, their development is now a critical endeavor (Urban *et al.* 2016).

Species distribution models (SDMs, Guisan & Thuiller 2005) have emerged as the tool of choice for biodiversity modeling, yielding significant advancements in understanding and predicting the impact of environmental conditions on species distributions (Guisan *et al.* 2017). However, despite their widespread applicability, SDMs are constrained by various limitations. These encompass their correlative nature, assumptions of species-environment equilibrium, or the omission of crucial ecological processes like dispersal. Notably, a crucial critique of SDMs is their disregard for biotic interactions. The potential influence of biotic interactions on species distributions has long been acknowledged, highlighted by Gause's Paramecium experiments, showing resource competition leading to exclusion (Gause 1934), or hare-lynx trophic interactions causing temporal cycles in species abundances (MacLulich 1936). This insight has been extensively recognized (Araújo & Luoto 2007; Kissling *et al.* 2012; Wisz *et al.* 2013; Freeman *et al.* 2022), yet the precise nature of biotic interaction effects and their dependence on scale and species characteristics remains an open inquiry (Thuiller *et al.* 2015). Consequently, over the past

two decades, ecological modeling has been significantly focused on the development of models able to elucidate the impact of biotic interactions on species distributions and incorporating them into predictions of species and biodiversity across space and time (Guisan & Rahbek 2011; Warton *et al.* 2015b).

Emerging methodologies like joint species distribution models (JSDMs, Clark *et al.* 2014; Pollock *et al.* 2014; Warton *et al.* 2015a) and Bayesian networks (Larsen *et al.* 2012; Ramazi *et al.* 2021) use co-occurrence data to infer unknown species interactions. Although JSDMs hold statistical merits in modeling community data, they neither explicitly infer interactions (with residual correlations possibly arising from various factors) nor account for these interactions into predictions (Zurell *et al.* 2018; Poggiato *et al.* 2021). Bayesian networks alleviate some challenges (related to identifying signals in residuals), but still infer a species association network from co-occurrence data. This inference approach is indirect at best and has been severely criticized regardless of the method (Blanchet *et al.* 2020).

An alternative involves directly integrating known or assumed biotic interactions into SDMs. When biotic interactions are documented for a focal species, researchers have included the distribution of interacting species (e.g., presence of prey or competing species) as supplementary covariates in SDMs (Araújo & Luoto 2007; Thuiller *et al.* 2018). This practice has generally enhanced model performance, emphasizing the importance of incorporating biotic interactions into SDMs and showing that biotic interactions significantly shape species distributions even at broad spatial scales (Gotelli *et al.* 2010). Not only including biotic interactions can improve

predictions, but allows us to understand how and under which conditions they can constrain species distributions and abundances (i.e., how the realized niche may differ from the potential one, Boulangeat *et al.* 2012). In early works, the interaction direction was typically assumed, rendering the approach particularly suitable for trophic interactions, although it was also employed to model competitive exclusion (Leathwick & Austin 2001). Yet, this promising approach has not been extended to model complex multitrophic communities, primarily due to the scarcity of network data (i.e., the 'Eltonian shortfall', Hortal *et al.* 2015) and technical complexities linked to predicting interdependent species in multitrophic networks (Wisz *et al.* 2013).

Models accommodating biotic interactions gain relevance with advances in interaction measurement (e.g., camera traps, gut content), open databases (e.g., GLOBI), and tools to query the literature (Grenié et al. 2022; Le Guillarme & Thuiller 2022). These growing datasets have accelerated the availability of large species interaction networks, named metawebs (Dunne 2006). These metawebs generalize the regional species pool concept of community ecology by incorporating potential interactions between species of different trophic levels (Albouy *et al.* 2019; Maiorano *et al.* 2020; Calderón-Sanou *et al.* 2021; Potapov 2022). Remaining knowledge gaps can now be filled using models that relate observed (or known) interactions to trait differences (i.e., trait-matching, Pichler *et al.* 2020; Caron *et al.* 2022) or phylogenies (e.g., Strydom *et al.* 2021, 2022). As the Eltonian shortfall is being addressed, we now have the opportunity to directly leverage known trophic interactions to build realistic and ecological sound predictive models (Windsor *et al.* 2022).

106 Here, we introduce a flexible framework based on Bayesian structural equation models to
107 explicitly integrate the known trophic interaction network (metaweb) into SDMs. Our framework
108 models each species as a function of its prey (in case of bottom-up control) or predators (for top-
109 down control) and environmental conditions. It then sequentially predicts the entire species pool
110 as a cascade of predictions. This allows the prediction of unobserved sites or environmental
111 conditions where the distribution of prey is unobserved. Implemented in a fully Bayesian
112 framework, multicollinearity issues and uncertainty propagation are specifically handled. This
113 framework not only potentially improves predictions for species under strong biotic control but
114 can also generate and test ecological hypotheses on the role of biotic interactions. It can show,
115 for example, under which conditions and how biotic interactions modify species distributions and
116 identifies the species that exert strong biotic control.
117
118 In this manuscript, we first describe and present our multitrophic framework, its relationships to
119 the existing literature, and its extensions. Then, to provide a robust and challenging test of the
120 approach we test and validate it on simulated realistic ecological communities. The performances
121 of the framework were studied under a large variety of ecological simulation setups, to highlight
122 when to expect our model to perform better than single (i.e., independent, 'single' hereafter)
123 SDMs, or not. To facilitate users' adoption of our framework and encourage future developments,
124 we implemented it in the R package `webSDM`, available on CRAN.
125

Model framework

We propose to model the distributions of species as a function of the environment and the species with which they interact. In general, this method works by first fitting models for each species independently (Fig. 1a,b) and then by combining these models to sequentially generate predictions on the entire species pool even at unobserved sites where prey distribution is unknown (Fig.1c).

To avoid the potential issue of circularity in the set of equations (i.e., simultaneity bias, (Pearl 2009), the metaweb needs to be a directed acyclic graph (DAG). Directed means that we choose whether we want to model predators according to their prey (i.e., bottom-up control) or reciprocally (i.e., top-down control). Hereafter, we present the framework assuming bottom-up control, although modeling species from top predators to basal species is equally possible, simply by reversing the direction of the metaweb links. 'Acyclic' means that the metaweb does not contain any loops. Our framework is thus built on the following hypothesis: i) the metaweb is fully known and contains no loop, ii) the metaweb is stable in space and time and iii) the model can only account for a single top-down or bottom-up control that needs to be specified up-front (even if the two processes likely occur in reality).

Model fitting

Let be the metaweb a directed acyclic graph, G, and Y the matrix containing the occurrence (e.g., presence-absence, count or biomass) of each species j (where $j = 1, \dots, S$) in each sampling unit

146 i (where $i = 1, \dots, n$). The matrix X contains the environmental covariates (indexed by $k =$
147 $1, \dots, p$).

148 We model the distribution of every species as a function of the environmental covariates and of
149 its prey with a generalized linear model (GLM), so that, for each species j we have:

150 $$y_{ij} \sim f(\mu_{ij})$$

151 $$g(\mu_{ij}) = \beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ik} + \sum_{j':(j',j)\in E} \alpha_{j',j}\, y_{ij'} \quad (1)$$

152 where $f()$ is the probability distribution of the observed species (e.g., binomial or Poisson,
153 depending on the data) with parameter μ_{ij} (we omitted the eventual dispersion parameter in the
154 equation for simplicity) and $g()$ the corresponding link function. β_{jk} denotes the response of
155 species j to covariate k (with $\beta_{j0}$the intercept of species j), $\alpha_{j',j}$ the response of species j to its
156 prey j' and E is the set of links of the metaweb G. Interestingly, given Eq. (1), the probabilistic
157 dependence between species is such that given its prey, a species is conditionally independent
158 from the preys of its preys. In other words, the joint likelihood factorizes, as every node of G is
159 independent from its non-descendants, conditionally to its parents (see also 'the causal Markov
160 condition', Pearl 2009):

161 $$p(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{j=1}^{S} p(\boldsymbol{y}_j \mid \boldsymbol{x}, prey(\boldsymbol{y}_j))$$

162 This is not a model assumption, but a mathematical property of the model arising from Eq. (1).
163 This implies that we can estimate an independent GLM for each species (Grace *et al.* 2012).
164 Interestingly, modeling species as a function of their prey (or their predators) and the
165 environmental covariates is equivalent to a structural equation model (Shipley 2000; Grace 2006)
166 whose structure is given by G, plus a dependence of all species to the environment (see Box 1 and

Fig. 2 for a description of SDM, JSDM, Bayesian networks, and our trophic model through the lens of SEMs). The estimation of the parameters of a SEM (whose structure is a directed acyclic graph) as a set of independent GLMs is a common practice and takes the name of 'local estimation' or piecewise SEM (Shipley 2000; Lefcheck 2016). Therefore, model fitting simply requires fitting S local models, here GLMs, which can be computationally fast and can even be parallelized.

If the model described in (1) can be fitted separately for each species, the ensemble of species distribution models has then to be combined to generate predictions on the whole species pool at unobserved sites, given the metaweb.

Model predictions

While technically possible (Grace *et al.* 2012), using SEMs to predict is not yet a common practice (but see Guerra *et al.* 2021). Indeed, to predict a predator, we need to know the prey distribution, which is unavailable at unobserved sites (or under future conditions). Intuitively, in a simple network of two trophic levels, we would need to predict the prey first and use these predictions to predict the predator. To generalize this idea to complex networks, we predict species following the topological order of the metaweb. This order, that exists for every DAG, guarantees that, for every link (j', j) (i.e., from prey j' to predator j in the bottom-up control scenario), j' comes before j in the ordering (Fig. 1c).

We then can predict the whole set of species in that specific order, thus conditioning predator predictions on the predicted occurrence of their prey (Fig. 1c), which operates as a 'cascade' of

predictions that start from basal species and flows up to top predators. In the case of top-down control, this would simply be the other direction.

By implementing GLMs in a Bayesian framework, we can obtain samples from the posterior predictive distribution of species, which allows a correct uncertainty propagation through this cascade and a proper estimation of the uncertainties on each species' predictions across the metaweb. Technically, we sample from the posterior predictive distribution of basal species and use those values to predict samples of the occurrences of their consumers, and so on throughout the trophic chain (this allows to correctly estimate species' predictive posterior distribution, Supplementary Materials 1). Notice that the width of prediction credibility intervals is likely to increase when moving through the trophic network.

Dimension reduction: sparse modeling and composite variables

Incorporating species as predictors inherently introduces more predictors into the models, potentially leading to multicollinearity issues, particularly in large networks and for generalist species. While multicollinearity may not directly compromise model predictive performance, it can distort coefficient estimates, affecting our understanding of prey-predator effects (Dormann *et al.* 2013; Tredennick *et al.* 2021).

Furthermore, involving insignificant predictors can lead to overfitting. The role of biotic interactions in shaping species' geographic distribution and environmental niche is contentious, especially at coarse resolutions (Pottier *et al.* 2013; Thuiller *et al.* 2015). Consequently, prey may not necessarily influence predator distribution and distort the species-environment relationships in the model.

Both multicollinearity and overfitting can be mitigated by constraining coefficients of unimportant predictors to zero, reducing model complexity—a technique known as sparse modeling or regularization (O'Hara & Sillanpää 2009; Hastie *et al.* 2015). This approach significantly reduces model complexity and should guarantee that in cases where species interactions leave no signals in the data, our model should be equivalent to a single SDM. In an hypothetical case of perfect multicollinearity between a biotic and an abiotic variable, the regularization approach would likely select one of the two variables randomly. Comparing single and trophic SDM would thus avoid misinterpretations.

For some generalist predators, there might simply be an intractable number of prey. In these cases, assuming that every prey has a differential effect on predator distribution is not only a problem but might seem ecologically unjustified. Instead, we could expect the richness or diversity of prey, or whether at least a prey is available, to be important. Hence, we implemented the use of composite variables that summarize the information of a large number of variables from the graph into a few summary variables (Henseler 2021). Implemented examples are prey richness or diversity, or a binary variable set to one if the number of preys is above a certain threshold. These variables assume that all species have the same impact on the predator. An alternative is to group species in the metaweb to represent trophic groups that clump together species that feed on, or are eaten by, the same type of species (Gauzens *et al.* 2015; O'Connor *et al.* 2020). We can then construct composite variables (e.g., their richness) for each of those trophic groups to better represent the variety of resources for species like generalists or top predators.

Implementation

We implemented our model in the R package `webSDM`, available on CRAN (more details on the package in Supplementary Material 2). The independent GLMs are fitted using the R library `rstanarm` (Goodrich *et al.* 2022) that easily incorporates a vast variety of priors and extensions, and exploits the STAN machinery to ensure fast sampling of the MCMC chains. Regularization was implemented with the horseshoe prior (Carvalho *et al.* 2010). We validated the ability of the inference algorithm to retrieve the correct parameters when sampling under the model in Supplementary Material 3.

Validation against realistic simulated data

In a nutshell, to validate our model, we conducted tests on realistic simulated data, which encompass selected ecological processes but exclude excessive stochasticity and sampling biases (Zurell *et al.* 2010; Meynard *et al.* 2019). We used a theoretical model that simulates species distribution data for interacting species in diverse environments, accounting for both bottom-up and top-down control (Fig. 3a,b). We then fitted both single and trophic SDMs on the simulated data and assessed their predictive efficacy for both realized (i.e., observed species distribution along an environmental gradient) and potential niches (i.e., species distribution along an environmental gradient when biotic constraints were released). Manipulating parameters of the theoretical model, such as metaweb complexity or species niche breadth, enabled us to mimic distinct ecological scenarios and deduce important performance drivers for our trophic model (Fig. 3c,d).

Simulation settings

To simulate species communities, we used a generalized Lotka-Volterra model that mimics realistic ecological interactions. We also extended simulations using variations of this model and the Ricker model.

The Lotka-Volterra model describes species abundance dynamic over time by a set of ordinary differential equations (Lotka 1920; Volterra 1926), where prey have a positive influence on predators, and vice versa for predators. We have implemented here an abiotic control on species growth rates along an environmental axis. Based on niche theory, we assumed that intrinsic

growth rates decrease following a Gaussian distribution as the species moves away from its niche optimum along the environmental gradient (Hirzel & Le Lay 2008). Species interactions were conserved along the environmental gradient and were specified by a metaweb, where prey and predators have an antisymmetric effect on each other. To be able to compare the true occurrence probabilities with those predicted by the single SDMs/trophic SDMs, we have made the Lotka-Volterra model stochastic, by introducing a stochastic term on the growth rate (see Supplementary materials 4.1-3 for equation and simulation details). For a given directed acyclic graph and given niches, we can then simulate from the stochastic generalized Lotka-Volterra model several communities (i.e., several sites) along different points of the environmental gradient. Simulated species abundances were then transformed to presence-absence (a species was set as presence if its continuous abundance was greater than zero) to obtain a species distribution dataset.

To make sure our results were not dependent of the simulation setting, we manipulated simulation parameters to assess how our trophic model's performances respond to different ecological factors. We played with the size of the species pool, the probability of links in the metaweb, the strength of interspecific interactions, and the species' niche breadths. Other more technical parameters, such as the number of points along the environmental gradient and the number of communities simulated for each environment, were also varied. Using Latin hypercube sampling, we selected 50 combinations of these simulation parameters. Then, for a given set of parameters, we sampled one interaction graph and species niches and used the above-described procedure to obtain one dataset of species distribution. This procedure was repeated 100 times,

each of them with a randomly sampled metaweb and species niches, for each of the 50 parameter combinations. We thus obtained a total of 100 x 50 = 5000 species distribution datasets.

Statistical modeling and data analysis

The statistical analysis of the simulated data was carried out independently for each of the 5000 species distribution datasets (Supplementary Materials 4.4). Our trophic SDM was fitted using the true interaction network as the metaweb, and a two-degree polynomial term to model the effect of the environment.

To compare the performance of our model to single SDMs, we also fitted for each species a generalized linear model as a function of the environment with the same two-degree polynomial term. Therefore, SDMs are equivalent to our trophic models, but without the biotic terms. Since the goal was to test whether including trophic interactions improves predictive performance, we used the same algorithm to avoid differences in the algorithm to determine differences in model performances.

We evaluated model performances through cross-validation, by randomly separating the 51 points along the environmental axis into 5 folds (i.e., training on 4 folds and predicting on the remaining fold, repeated 5 times). We compared our trophic model to SDMs in predicting species realized and potential niches. We estimated the species' realized niche as its probability of presence along the environmental gradient (sensu Hutchinson 1957). This corresponds to the predictions of the model, that were compared against the observed probability of presence.

We defined the potential niche as the environmental conditions where a species could potentially be present if the constraints imposed by biotic interactions were released. This is different from the fundamental niche defined by niche theory (e.g., Soberón 2007), as it does not account for dispersal limitations. To release the biotic constraints, we expressed the species' potential niche as the probability of presence along the environmental gradient when its prey is assumed to be present and its predators absent everywhere. With our trophic model, we can predict species' potential niches by setting the prey as present everywhere across the environmental gradient (i.e., a conditional prediction). Single SDMs cannot condition on the value of other species, so we considered that the predictions of the potential and realized niche coincide. We then compared the predicted potential niches to the true ones of theoretical models, computed as the probability of species having a positive growth rate given the presence of prey, and the absence of predators (see Supplementary materials 4.3).

We evaluated how well the models predicted the realized niche by comparing the predicted probability of presence to observed presence-absences using the area under the ROC curve (AUC, (Fawcett 2006). We also used the Wasserstein distance with p=1 (to compare the distance between the predicted and observed probability of presence), an approximation of the leave-one-out cross-validation likelihood (Vehtari *et al.* 2017), and calibration (i.e., the number of times the 95% credible interval correctly covers the true value, (Norberg *et al.* 2019). To measure model performances in retrieving the potential niche, we only used the Wasserstein distance and calibration (see Supplementary Materials 4.4).

Finally, we checked the width of the 95% credible interval of the predicted realized niche, to understand if our uncertainty propagation technique could lead to overly large credible intervals, especially for top predator species.

The codes to fully reproduce the results are available on the GitHub repository (https://github.com/giopogg/webSDM/tree/main/publication/VirtualEcoSim).

Results

Overall, trophic SDM correctly retrieved species realized niches (mean AUC = 0.85). On average, across all the simulation parameters, our trophic model significantly improved AUC by 6% with respect to SDM (one-side paired t-test p-value < 10^{-16} for non-basal species, Fig. 4a). Our trophic model improved AUC for 66% of species, while for other species predictions were only a little worsened (e.g., only 0.3% of species were worsened by more than 20%). Interestingly, the strongest improvements corresponded to cases where SDMs failed to predict species realized niches (Table S1). Regression of AUC relative improvement as a function of simulation parameters showed that only the size of the species pool and the normalized niche breadth (a measure of species niche breadth that is independent of the size of the species pool) were the main factors explaining a departure from the average improvement (Table S2). The size of the species pool bolstered model performances, while the normalized niche breadth decreased model performances, indicating greater improvement for environmentally specialized species. To understand the reason for these patterns with an example, we identified a typical case of two species corresponding to two simulation scenarios with different normalized niche breadth, all other parameters being equal (Fig. 5). A narrow niche breadth restricted prey to a small

environmental range, leading to a pronounced change in the realized niche shape and thus poor SDM predictions, that were outperformed by our trophic model (Fig. 5a, AUC = 0.68 for SDM and 0.86 for trophic SDM). In contrast, when prey were present in a broader environmental range, predators' distribution was less distorted by biotic interactions, SDMs more accurately predicted the realized niche, and the two models' predictions aligned closely (Fig. 5b, AUC = 0.91 for SDM and 0.90 for trophic SDM).

Our trophic model improved the predictions of the potential niche slightly more (8% mean reduction of the Wasserstein distance to the true potential niche, one-side pair t-test p-value < 10^{-16} for non-basal species, Fig. 4b, table S3), and this mean improvement did not depend on any of the simulation parameters (Table S4). Moreover, the strongest improvements corresponded to cases where SDMs failed to predict species potential niches (Table S3).

These results were consistent across the other evaluation metrics (Fig. S3, S4, Table S1-3) and for the two other theoretical models (Table S5, S6). A complete description of simulation results is available in Supplementary Materials 4.4 and 5.4.

As expected, the width of credible intervals for predicted probability of presence was larger for non-basal species, increasing with the number of prey (Supplementary Materials 4.5). However, this enlargement stabilized around 0.25 regardless of the species' trophic level and prey count, indicating that despite an increase along the trophic network, interval widths remained manageable under all conditions (Fig. S5).

Discussion

In this study, we introduced a versatile SDM framework to account for known trophic interactions. Our framework models species as a function of the environment and their prey (or predators), handles multicollinearity and error propagation, and can predict species at unobserved sites or future conditions where prey (or predators) distribution remains unknown. We conducted a hard and rigorous validation using data simulated under simultaneous top-down and bottom-up control, as this is typically the case in real ecosystems. Despite this difficulty and the fundamentally dynamic simulated ecological properties, our model improved single SDMs in predicting species realized niches when these were strongly controlled by biotic interactions and allowed us to better capture the elusive species' potential niches.

Yet, our framework is not a universal solution that will consistently outperform SDMs. The debated impact of biotic interactions on species distributions and potential errors in the data (e.g., metaweb, environmental covariates) are such that including prey as additional predictors may not always enhance predictions or may even slightly degrade them in specific cases. While regularization may alleviate these issues, making predictions from trophic SDM equivalent to single SDMs at worst, practical results might differ. Hence, we recommend using both SDMs and trophic SDMs and selecting the best model. Our sensitivity analysis has highlighted possible conditions that bolster the relative improvement of trophic SDM performance, yet real-world applications are needed to further refine these conditions.

Model assumptions and differences with (J)SDM

The effect of biotic interactions on species distributions is particularly challenging to model (Thuiller *et al.* 2013). We have proposed to study this effect with a statistical model that relies on co-occurrence data and a directed acyclic interaction network. While the true causal effect of biotic interactions can hardly be unraveled with a static approach (Blanchet *et al.* 2020), we have demonstrated here how specifically accounting for known interactions can still boost the model predictive performance (i.e., the realized niche) and even the inference of ecological processes (i.e., the potential niche). It is useful to understand that in the conditional prediction stage, is it ultimately the environment that determines species' presence, through the interacting species. In other words, species respond to the environment through the cumulative effect of their prey's response to it (see Supplementary Material 6 for a more advanced). Therefore, if a single SDM with a very flexible and complex algorithm and an extremely large set of environmental predictors (e.g., XGBoost, RandomForest) could achieve an equally good representation of the realized niche (i.e., since it implicitly takes interactions into account), there would be a risk of overfitting, leading to erroneous projections in unobserved sites. Our trophic model can instead model a complex response to the environment with a more parsimonious and ecologically meaningful model, despite comparable predictive performances, especially on the training dataset (Fig. S6).

Not only a predictive model

Our framework can help exploring the role of known trophic interactions on species distributions and community composition. We can investigate how the effect of trophic interactions varies with spatial resolution (Thuiller *et al.* 2015), trophic level and species degree of generalism (Fraser

412 *et al.* 2021), functional traits (Boet *et al.* 2020; van der Merwe *et al.* 2021), and environmental
413 conditions (by including an interaction between abiotic and biotic terms, Paquette & Hargreaves
414 2021). We have presented a bottom-up perspective, but it could be interesting to contrast results
415 with a top-down perspective to investigate the stronger direction of the control across different
416 systems. The role and shape of biotic interactions can also be investigated by comparing different
417 specifications of the biotic term of our model. Composite variables can here become particularly
418 interesting to test whether this is the diversity, the richness, or the occurrence of certain types of
419 prey that control the distribution of specific species.
420 Our framework is also valuable for studying network robustness under global change. Indeed, it
421 can spatially predict the potential effects of species extinction on the rest of the trophic chain (by
422 predicting species conditionally on the absence/decline of prey). In other words, this could be a
423 useful framework to investigate the potential impacts of climate and land use change on species
424 distribution, community composition, and trophic diversity.
425
426 **Model extensions**
427 The proposed trophic species distribution model is implemented within a Bayesian framework
428 that enables error propagation and can integrate informative priors that reflect our knowledge of
429 the ecosystem (e.g., uncertainties on the interaction network). Alternatively, a frequentist
430 approach would offer computational speed and generalization to various statistical algorithms.
431 Indeed, thanks to the local estimation of the model, we can extend GLMs to any other type of
432 algorithms like machine learning tools, or even an ensemble of them. Local estimation can also
433 allow our framework to integrate any extension of SDMs, taking advantage of the important

methodological developments in this field such as modeling presence-only data (e.g., Renner *et al.* 2015), integrating different data sources (Isaac *et al.* 2020), or considering imperfect species detection (MacKenzie *et al.* 2006). However, a drawback of this local estimation is that species-environment coefficients cannot be modeled hierarchically as in multi-species models (Pollock *et al.* 2014), which can help to better model rare species. This could be eventually included by switching to a global estimation of our framework. Yet, this will come with the issue of assuming that all species respond to the same set of environmental variables, which might be particularly wrong in large ecological networks or when modeling multiple species with very different life history traits and ecology (e.g., parasites, autotrophs, herbivores).

Conclusions and perspectives

We believe our framework stands out as an exciting solution to integrate known trophic networks into species distribution models. Although it is now only applicable to directed acyclic trophic networks, this limitation could be overcome in several ways. First, symmetrical networks can be made asymmetric by choosing a dominant direction, thus allowing our model to be extended to non-trophic interactions such as competition, in which one species often dominates the other (Hardin 1960; Leathwick & Austin 2001). To extend the model to cyclical and/or symmetrical networks, a possible path of development would be to compute the set of spanning trees in the network (see the matrix tree theorem, Chaiken & Kleitman 1978), fit a model for each of them, and then predict by averaging each tree's predictions. This kind of ensemble approach is already used in the machine learning field (Read *et al.* 2021) and could also integrate link uncertainties in

the metaweb. In other words, we hope this new modeling paradigm will motivate exciting and novel research, and challenge others to improve on our proposed framework.

Acknowledgments. This work is in memory of Marc Ohlmann, who left us too soon in a mountain accident. His deep knowledge of network properties and statistical models helped the development of this model. We thank Stéphane Robin for the insightful discussions around graphical model, and Julyan Arbel for his help with the estimation of the posterior predictive distribution. We also thank Rémi Patin for the help in package development. This study has received funding from the ERA-Net BiodivERsA—Belmont Forum, with the national funder Agence Nationale de la Recherche (ANR-18- EBI4-0009), part of the 2018 Joint call BiodivERsA-Belmont Forum call (project 'FutureWeb'). WT and GP also acknowledge support from the European Union's Horizon Europe under grant agreement number 101060429 (project NaturaConnect), and the French Agence Nationale de la Recherche (ANR) through the Gambas (ANR-18-CE02-0025), EcoNet (ANR-18-CE02-0010) and FORBIC (ANR-18-MPGA-0004) projects.

Bibliography

Albouy, C., Archambault, P., Appeltans, W., Araújo, M.B., Beauchesne, D., Cazelles, K., *et al.* (2019). The marine fish food web is globally connected. *Nat. Ecol. Evol.*, 3, 1153–1161.

Araújo, M.B. & Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Glob. Ecol. Biogeogr.*, 16, 743–753.

Blanchet, F.G., Cazelles, K. & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecol. Lett.*, 23, 1050–1063.

Boet, O., Arnan, X. & Retana, J. (2020). The role of environmental vs. biotic filtering in the structure of European ant communities: A matter of trait type and spatial scale. *PLoS ONE*, 2.

Boulangeat, I., Gravel, D. & Thuiller, W. (2012). Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecol. Lett.*, 15, 584–593.

Calderón-Sanou, I., Münkemüller, T., Zinger, L., Schimann, H., Yoccoz, N.G., Gielly, L., *et al.* (2021). Cascading effects of moth outbreaks on subarctic soil food webs. *Sci. Rep.*, 11, 15054.

Caron, D., Maiorano, L., Thuiller, W. & Pollock, L.J. (2022). Addressing the Eltonian shortfall with trait-based interaction models. *Ecol. Lett.*, 25, 889–899.

Carvalho, C.M., Polson, N.G. & Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97, 465–480.

Chaiken, S. & Kleitman, D.J. (1978). Matrix tree theorems. *J. Comb. Theory Ser. A.*, 24, 377–381.

Chib, S. (1998). Analysis of multivariate probit models. *Biometrika*, 85, 347–361.

Clark, J.S., Gelfand, A.E., Woodall, C.W. & Zhu, K. (2014). More than the sum of the parts: forest climate response from joint species distribution models. *Ecol. Appl.*, 24, 990–999.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., *et al.* (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography (Cop.)*, 36, 27–46.

Dunne, J.A. (2006). The network structure of food webs. In: *Ecological networks: linking structure and dynamics* (eds. M., P. & Dunne, J.A.). Oxford University Press, Oxford, England, pp. 27–86.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27, 861–874.

Fraser, D., Soul, L.C., Tóth, A.B., Balk, M.A., Eronen, J.T., Pineda-Munoz, S., *et al.* (2021). Investigating biotic interactions in deep time. *Trends Ecol. Evol.*, 36, 61–75.

Freeman, B.G., Strimas-Mackey, M. & Miller, E.T. (2022). Interspecific competition limits bird species' ranges in tropical mountains. *Science*, 377, 416–420.

Gause, G.F. (1934). Experimental analysis of Vito volterra's mathematical theory of the struggle for existence. *Science*, 79, 16–17.

Gauzens, B., Thébault, E., Lacroix, G. & Legendre, S. (2015). Trophic groups and modules: two levels of group detection in food webs. *J. R. Soc. Interface*, 12, 20141176.

Goodrich, B., Gabry, J., Ali, I. & Brilleman, S. (2022). *rstanarm: Bayesian applied regression modeling via Stan*.

Gotelli, N.J., Graves, G.R. & Rahbek, C. (2010). Macroecological signals of species interactions in the Danish avifauna. *Proc. Natl. Acad. Sci. U. S. A.*, 107, 5030–5035.

Grace, J.B. (2006). *Structural Equation Modeling and Natural Systems*. Cambridge University Press, Cambridge, TAS, Australia.

Grace, J.B., Schoolmaster, D.R., Jr, Guntenspergen, G.R., Little, A.M., Mitchell, B.R., Miller, K.M., *et al.* (2012). Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere*, 3, art73.

Guerra, C.A., Delgado-Baquerizo, M., Duarte, E., Marigliano, O., Görgen, C., Maestre, F.T., *et al.* (2021). Global projections of the soil microbiome in the Anthropocene. *Glob. Ecol. Biogeogr.*, 30, 987–999.

Guisan, A. & Rahbek, C. (2011). SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *J. Biogeogr.*, 38, 1433–1444.

Guisan, A. & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.*, 8, 993–1009.

Guisan, A., Thuiller, W. & & Zimmermann, N. (2017). *Habitat Suitability and Distribution Models: With Applications in R (Ecology, Biodiversity and Conservation)*. Cambridge University Press, Cambridge, TAS, Australia.

Hardin, G. (1960). The competitive exclusion principle. *Science*, 131, 1292–1297.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, London, England.

Henseler, J. (2021). *Composite-based structural equation modeling : analyzing latent and emergent variables*. Guilford Press, New York, NY.

Hirzel, A.H. & Le Lay, G. (2008). Habitat suitability modelling and niche theory. *J. Appl. Ecol.*, 45, 1372–1381.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. (2015). Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu. Rev. Ecol. Evol. Syst.*, 46, 523–549.

Hutchinson, G.E. (1957). Concluding remarks. population studies: animal ecology and demography. *Cold Spring Harb. Symp. Quant. Biol.*, 22, 415–427.

Isaac, N.J.B., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., *et al.* (2020). Data integration for large-scale models of species distributions. *Trends Ecol. Evol.*, 35, 56–67.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., *et al.* (2012). Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *J. Biogeogr.*, 39, 2163–2178.

Larsen, P.E., Field, D. & Gilbert, J.A. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nat. Methods*, 9, 621–625.

Leathwick, J.R. & Austin, M.P. (2001). Competitive interactions between tree species in new zealand's old-growth indigenous forests. *Ecology*, 82, 2560.

Lefcheck, J.S. (2016). piecewiseSEM : Piecewise structural equation modelling in r for ecology, evolution, and systematics. *Methods Ecol. Evol.*, 7, 573–579.

Lotka, A.J. (1920). Analytical note on certain rhythmic relations in organic systems. *Proc. Natl. Acad. Sci. U. S. A.*, 6, 410–415.

MacKenzie, D.I., Nichols, J.D., Andrew Royle, J., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006). *Occupancy Estimation and Modeling : Inferring Patterns and Dynamics of Species Occurrence*. Academic Press, San Diego, CA.

MacLulich, D.A. (1936). Fluctuations in numbers of varying hares. *Science*, 83, 162–162.

Maiorano, L., Montemaggiori, A., Ficetola, G.F., O'Connor, L. & Thuiller, W. (2020). TETRA-EU 1.0: A species-level trophic metaweb of European tetrapods. *Glob. Ecol. Biogeogr.*, 29, 1452–1457.

van der Merwe, S., Greve, M., Olivier, B. & le Roux, P.C. (2021). Testing the role of functional trait expression in plant–plant facilitation. *Funct. Ecol.*, 35, 255–265.

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 (Cop.)*, 42, 2021–2036.

Montesinos-Navarro, A., Estrada, A., Font, X., Matias, M.G., Meireles, C., Mendoza, M., *et al.* (2018). Correction: Community structure informs species geographic distributions. *PLoS One*, 13, e0200556.

Norberg, A., Abrego, N., Blanchet, F.G., Adler, F.R., Anderson, B.J., Anttila, J., *et al.* (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecol. Monogr.*, 89, e01370.

O'Connor, L.M.J., Pollock, L.J., Braga, J., Ficetola, G.F., Maiorano, L., Martinez-Almoyna, C., *et al.* (2020). Unveiling the food webs of tetrapods across Europe through the prism of the Eltonian niche. *J. Biogeogr.*, 47, 181–192.

O'Hara, R.B. & Sillanpää, M.J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal.*, 4, 85–117.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., *et al.* (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.*, 20, 561–576.

Paquette, A. & Hargreaves, A.L. (2021). Biotic interactions are more often important at species' warm versus cool range edges. *Ecol. Lett.*, 24, 2427–2438.

Pearl, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.*, 3, 96–146.

Pichler, M., Boreux, V., Klein, A.-M., Schleuning, M. & Hartig, F. (2020). Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods Ecol. Evol.*, 11, 281–293.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J.S. & Thuiller, W. (2021). On the interpretations of joint modeling in community ecology. *Trends Ecol. Evol.*, 36, 391–401.

Pollock, L.J., O'Connor, L.M.J., Mokany, K., Rosauer, D.F., Talluto, M.V. & Thuiller, W. (2020). Protecting biodiversity (in all its complexity): New models and methods. *Trends Ecol. Evol.*, 35, 1119–1128.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., *et al.* (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol. Evol.*, 5, 397–406.

Potapov, A.M. (2022). Multifunctionality of belowground food webs: resource, size and spatial energy channels. *Biol. Rev. Camb. Philos. Soc.*, 97, 1691–1711.

Pottier, J., Dubuis, A., Pellissier, L., Maiorano, L., Rossier, L., Randin, C.F., *et al.* (2013). The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients. *Glob. Ecol. Biogeogr.*, 22, 52–63.

Potts, S.G., Imperatriz-Fonseca, V.L., Ngo, H.T., Biesmeijer, J.C., Breeze, T.D., Dicks, L.V., *et al.* (2016). *IPBES (2016): Summary for policymakers of the assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services on pollinators, pollination and food production*.

Ramazi, P., Kunegel-Lion, M., Greiner, R. & Lewis, M.A. (2021). Exploiting the full potential of Bayesian networks in predictive ecology. *Methods Ecol. Evol.*, 12, 135–149.

Read, J., Pfahringer, B., Holmes, G. & Frank, E. (2021). Classifier Chains: A Review and Perspectives. *J. Artif. Intell. Res.*, 70, 683–718.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., *et al.* (2015). Point process models for presence-only analysis. *Methods Ecol. Evol.*, 6, 366–379.

Rosa, I.M.D., Pereira, H.M., Ferrier, S., Alkemade, R., Acosta, L.A., Akcakaya, H.R., *et al.* (2017). Multiscale scenarios for nature futures. *Nat. Ecol. Evol.*, 1, 1416–1419.

Shipley, B. (2000). *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference with R*. Cambridge University Press, Cambridge, TAS, Australia.

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecol. Lett.*, 10, 1115–1123.

Staniczenko, P.P.A., Sivasubramaniam, P., Suttle, K.B. & Pearson, R.G. (2017). Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. *Ecol. Lett.*, 20, 693–707.

Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M.J., *et al.* (2022). Food web reconstruction through phylogenetic transfer of low-rank network representation. *Methods Ecol. Evol.*

Strydom, T., Catchen, M.D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., *et al.* (2021). A roadmap toward predicting species interaction networks (across space and time). *EcoEvoRxiv*.

Thuiller, W., Guéguen, M., Bison, M., Duparc, A., Garel, M., Loison, A., *et al.* (2018). Combining point-process and landscape vegetation models to predict large herbivore distributions in space and time-A case study of Rupicapra rupicapra. *Divers. Distrib.*, 24, 352–362.

Thuiller, W., Münkemüller, T., Lavergne, S., Mouillot, D., Mouquet, N., Schiffers, K., *et al.* (2013). A road map for integrating eco-evolutionary processes into biodiversity models. *Ecol. Lett.*, 16 Suppl 1, 94–105.

Thuiller, W., Pollock, L.J., Gueguen, M. & Münkemüller, T. (2015). From species distributions to meta-communities. *Ecol. Lett.*, 18, 1321–1328.

Tredennick, A.T., Hooker, G., Ellner, S.P. & Adler, P.B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, 102, e03336.

Urban, M.C., Bocedi, G., Hendry, A.P., Mihoub, J.-B., Pe'er, G., Singer, A., *et al.* (2016). Improving the forecast for biodiversity under climate change. *Science*, 353, aad8466–aad8466.

Vehtari, A., Gelman, A. & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.*, 27, 1413–1432.

Volterra, V. (1926). Fluctuations in the Abundance of a Species considered Mathematically1. *Nature*, 118, 558–560.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., *et al.* (2015a). So many variables: Joint modeling in community ecology. *Trends Ecol. Evol.*, 30, 766–779.

Warton, D.I., Foster, S.D., De'ath, G., Stoklosa, J. & Dunstan, P.K. (2015b). Model-based thinking for community ecology. *Plant Ecol.*, 216, 669–682.

Windsor, F.M., van den Hoogen, J., Crowther, T.W. & Evans, D.M. (2022). Using ecological networks to answer questions in global biogeography and ecology. *Journal of Biogeography*.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., *et al.* (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev. Camb. Philos. Soc.*, 88, 15–30.

Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T., *et al.* (2010). The virtual ecologist approach: simulating data and observers. *Oikos*, 119, 622–635.

Zurell, D., Pollock, L.J. & Thuiller, W. (2018). Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography (Cop.)*, 41, 1812–1819.

Figures

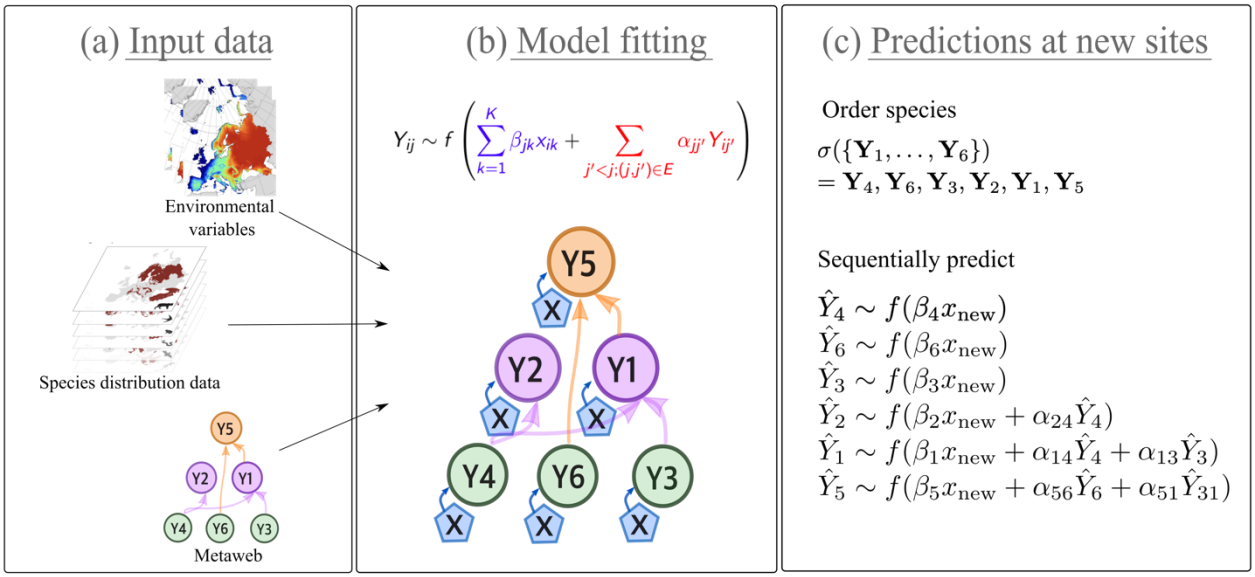
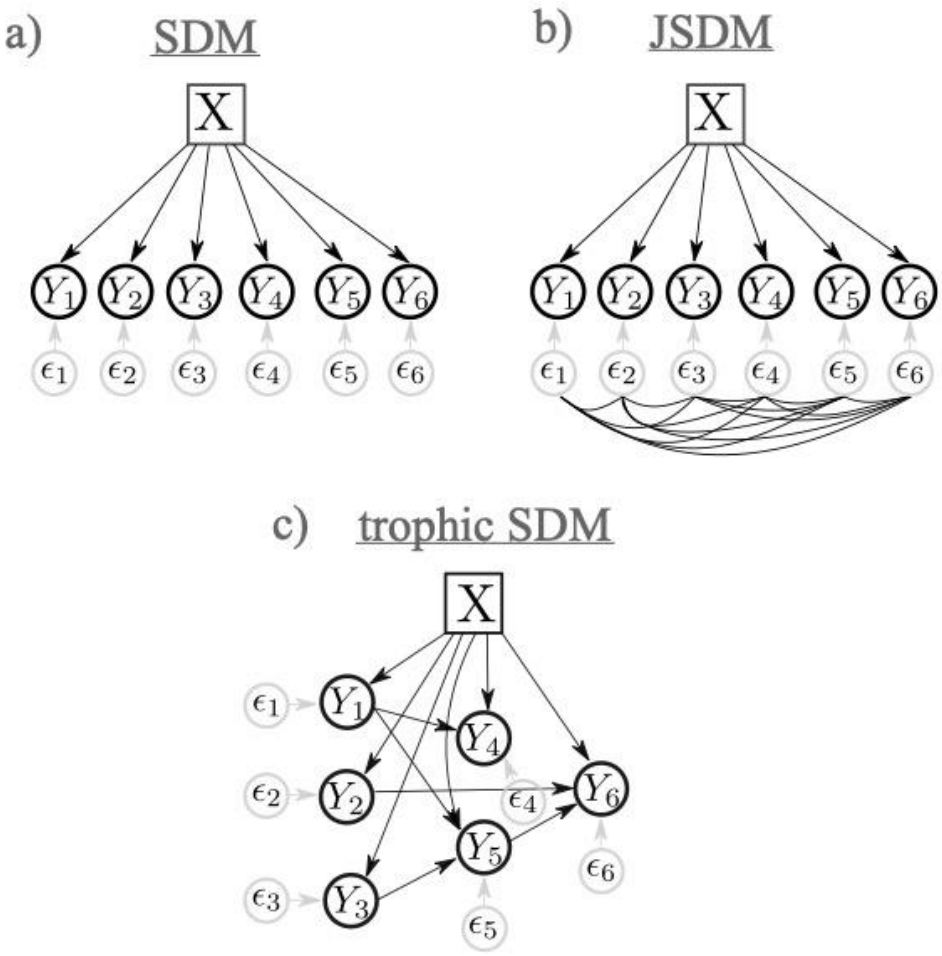


Figure 1. Description of the proposed model. The input data are the environmental variables, species distribution data, and a directed acyclic graph representing the species trophic interactions (a). Assuming six hypothetical species, each species is modeled as a function of the environmental covariates and its prey (or predators). Since the likelihood factorizes, each species can be modelled independently (b). Predictions under new environmental conditions (in space or time) where prey distribution are unknown (c). To predict with our trophic SDM, we need to guarantee that, for each predator, all its prey species have already been predicted. To do so, we order species according to the topological order σ derived from the metaweb and then perform predictions following this order (d). Therefore, predictions of species Y_4 , Y_6 and Y_3 are obtained as a function of the environmental conditions x_{new}. Then, predictions for Y_2 are computed as a function of x_{new} and the predictions of Y_4 and so on for species Y_1 and Y_5.



671

672 Figure 2. Structural dependence of SDM (a), JSDM (b), and trophic SDM (c). Black boxes refer to
673 the environment covariates (X), while black circles represent species (Y). Gray circles represent
674 residuals (ϵ). Bayesian network correspond to the same diagram of trophic SDM (c), except that
675 for Bayesian networks the graph is inferred from the data with obvious uncertainties.

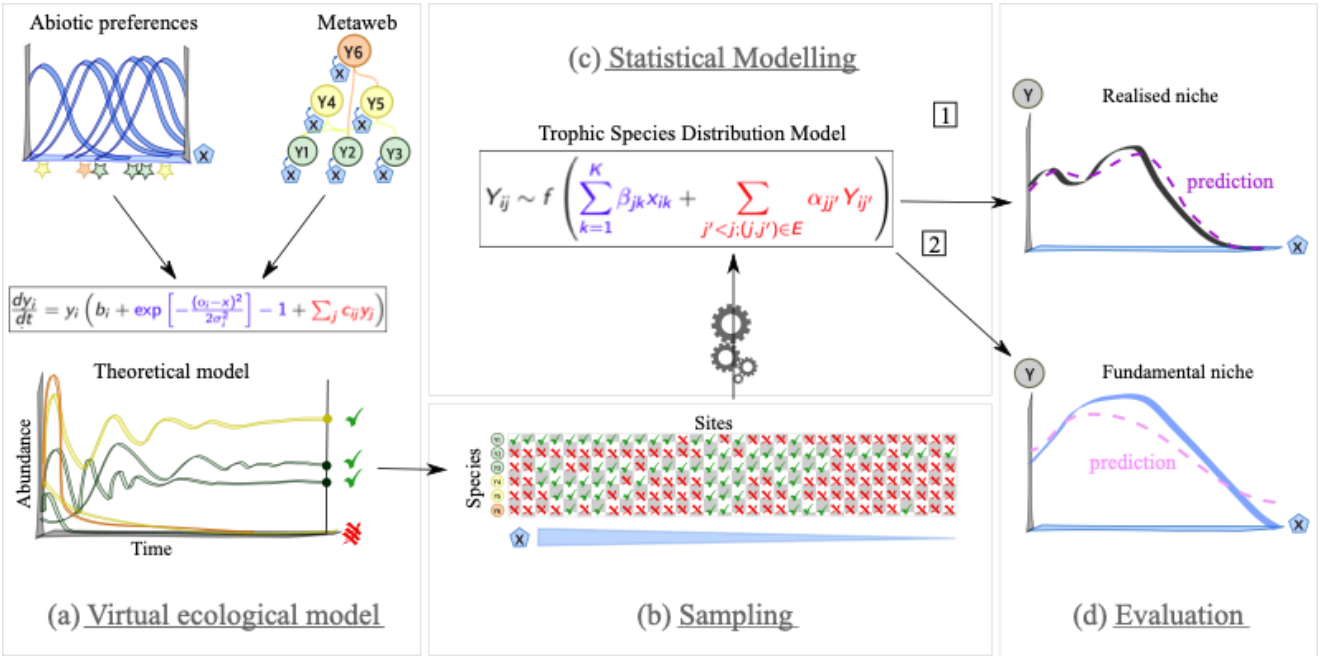
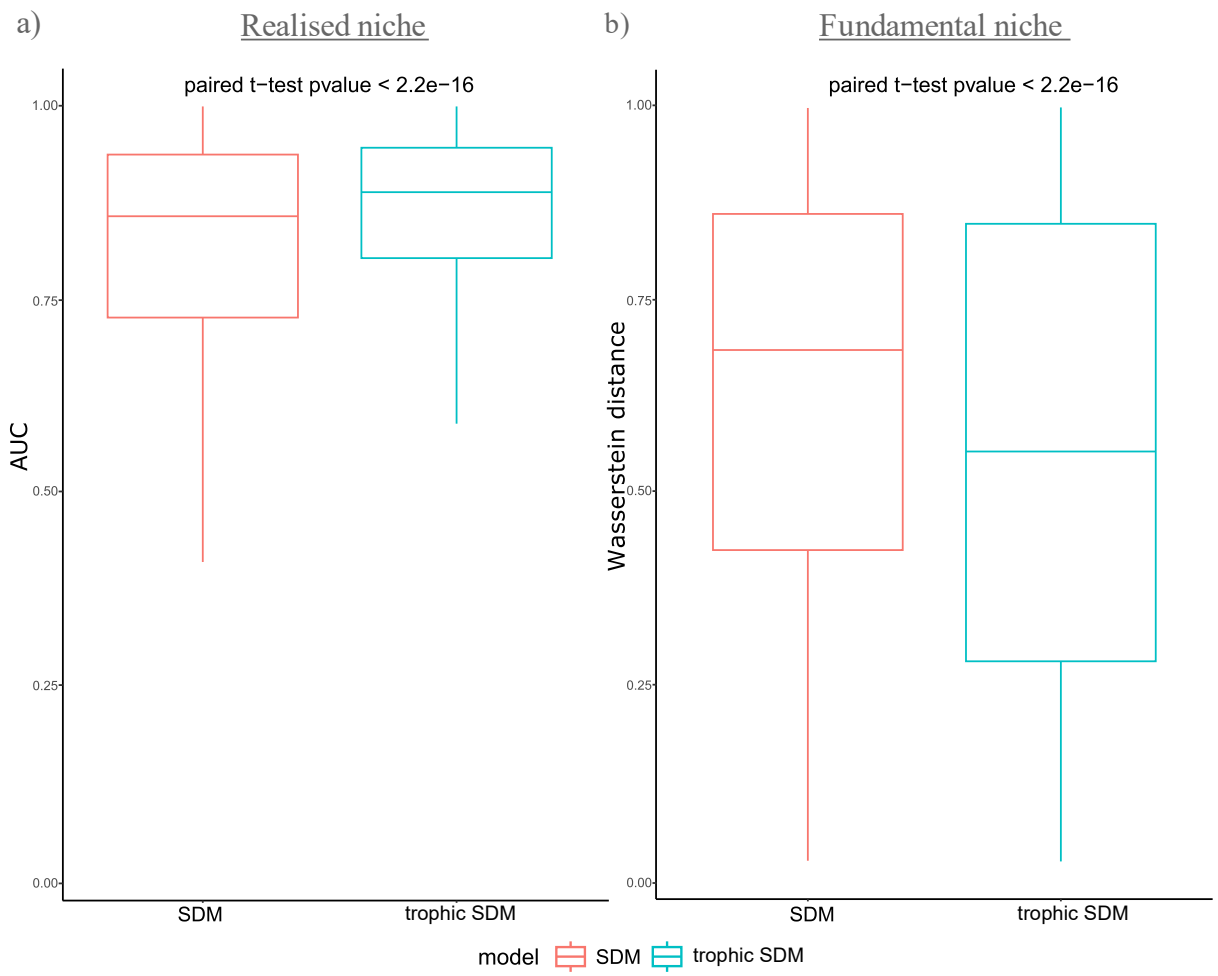
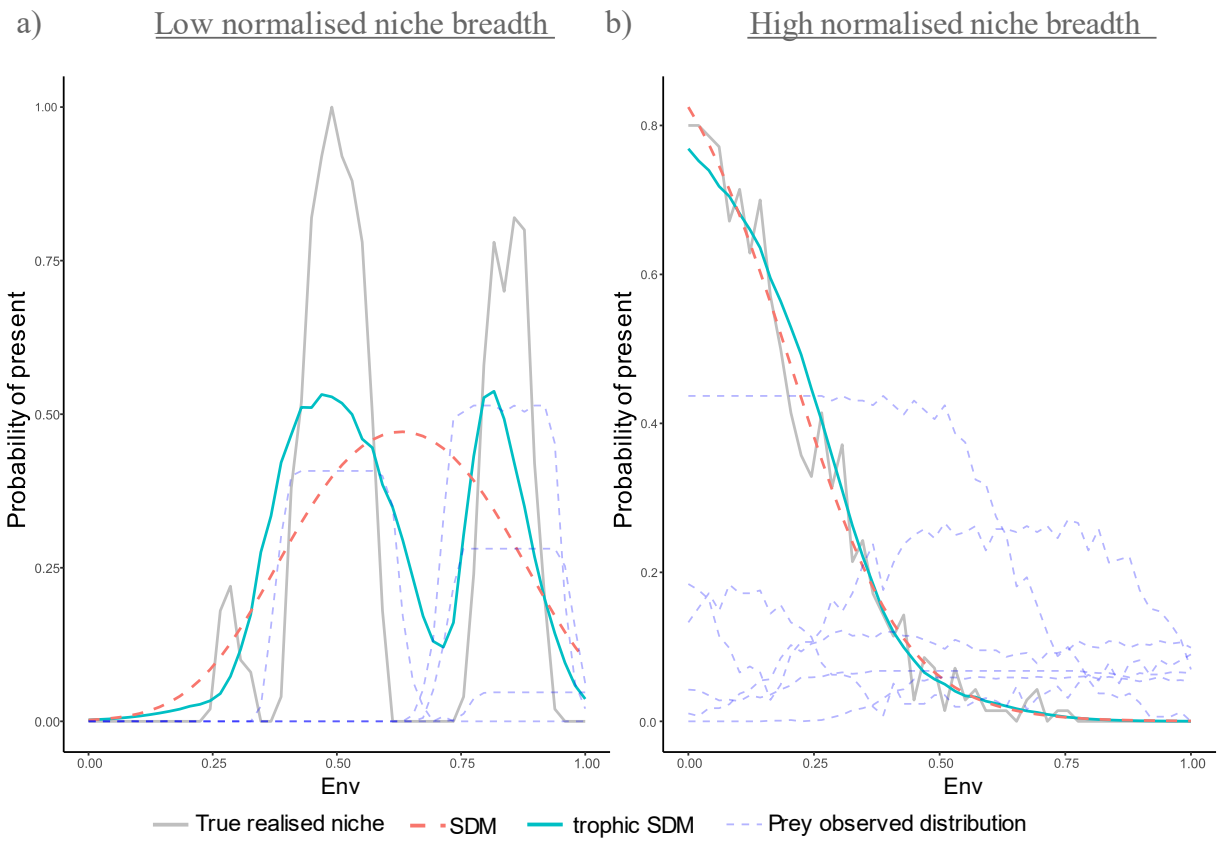


Figure 3. Application of the Virtual Ecologist approach to trophic SDMs (Zurell et al. 2010). The virtual ecological model (a) consists of defining, for a given number of species, their abiotic preferences as well as a trophic interaction network, and then making their abundances vary according to a selected theoretical model combining abiotic (blue) and biotic (red) controls. After a certain time (i.e., at the stationary state), species abundances are transformed in presence absences (b) and used to fit the SDM and trophic SDM (c). Finally, the models are evaluated (d) by looking at the congruence of predicted realized and potential niches (dashed lines, and continuous lines represent the true niches).



685

686 Figure 4: Model performances for the realized niche (a) and potential niche (b).



687

688 Figure 5: True and predicted realized niches of two species corresponding to simulations where
689 species had low normalized niche breadth (a) and high normalized niche breadth (b). For
690 species(a), trophic SDM improved predictions (AUC = 0.86 for trophic SDM and 0.68 for SDM). For
691 species (b) performances were almost equal (AUC = 0.91 for SDM, 0.90 for trophic SDM).

692

Boxes

Box 1: Structural dependence in species distribution modeling

Like any multivariate model, all kinds of species distribution models introduce probabilistic dependencies between variables (i.e., both species and environmental covariates). These dependencies can be expressed using the graphical representation of structural equation models (SEM), i.e., the 'path diagram', to understand the differences between models and, in particular, the relationships they introduce between species. In SEM, each variable is represented as a box, a direct arrow from one box to another indicates a direct relationship between the two variables (e.g., an arrow from A to B means that A affects B), and an undirected arrow between two variables indicates a correlation. From such a graphical representation, we can depict the directed and indirect relationships between variables, as well as the marginal and conditional dependencies that these relationships imply (see Koller and Friedman 2009 for a thorough introduction to probabilistic graphical models). Here, we express the dependencies introduced by the single SDM, JSDM, Bayesian networks, and trophic SDM when modeling a hypothetical set of six interacting species (Y1 to Y6) and a set of environmental covariates X (Fig. 2).

SDM

Single SDMs model each species independently as a function of the environment (Guisan & Thuiller 2005). This is therefore equivalent to an SEM diagram with arrows pointing from environment X to all species (Fig. 2a). As a consequence, species are marginally dependent, as they are correlated through their response to environment (e.g., species with similar niches will

be predicted to co-occur), but they are conditionally independent given the environment. Since the effect of interacting species is not controlled for, the species-environment relationships capture both the environmental and biotic effect, and, as such, SDMs only infer species' realized niches (i.e., biotic interactions are implicitly taken into account, Araujo and Guisan 2006).

JSDM

JSDMs model each species as a function of the environment assuming residuals are correlated across species (Pollock *et al.* 2014; Ovaskainen *et al.* 2017). This corresponds to an SEM diagram where the environment X points to all species, and species are linked with each other, through their residuals, with undirected arrows representing their correlations (Fig. 2b). These correlations imply a symmetrical, undirected, relationship between species residuals, so that species are conditionally dependent given the environment. Moreover, since conditional independencies are expressed in a latent layer, every pair of species is conditionally dependent given the rest of the network (i.e., the inferred graph at the species level is fully connected). However, due to the residual nature of these correlations, they have little effect on the estimates of the species-environment relationships (Chib 1998; Poggiato *et al.* 2021) and do not modify species marginal predictions (Poggiato *et al.* 2021).

Bayesian network models

Bayesian network models (Larsen *et al.* 2012; Ramazi *et al.* 2021) infer a directed acyclic graph (DAG) from species distributions and environmental covariates. Once this network of co-occurrences is inferred, species are modeled as a function of their parents (similarly to the trophic SDM), depending on the method used (Montesinos-Navarro *et al.* 2018; Ramazi *et al.* 2021). The SEM diagram of a Bayesian network is then simply the inferred DAG. The advantage of Bayesian

736 networks over JSDM is that they model the species together with the environment, so the two
737 effects can be properly separated. However, the inferred DAG does not necessarily correspond
738 to the true interaction networks, which leads to a conceptual difference between the Bayesian
739 network and trophic SDM. Moreover, to our knowledge, Bayesian networks have never been used
740 to generate species predictions at unobserved sites where prey distribution is unknown
741 (Staniczenko *et al.* 2017).
742
743 ***Trophic SDM***
744 Our approach proposes to inject the knowledge of the metaweb by modeling species as a function
745 of their prey (or predators) and of the environment. This introduces a direct arrow from prey to
746 predators on top of the arrows from the environment to species (Fig. 2c). So, the effect of prey is
747 controlled for and the distribution of prey directly determines the prediction of the predators
748 (reciprocally for a top-down control). Each species is therefore conditionally independent, given
749 its prey and the environment, to the preys of its preys. However, for any connected metaweb,
750 species are marginally correlated given the environment only, so that, for example, two predators
751 feeding on the same prey are marginally correlated due to the indirect effect of sharing the same
752 prey (i.e., they tend to co-occur because they both feed on the same prey).