Essential ingredients in Joint Species Distribution Models: influence on interpretability, explanatory and predictive power

Clément Violet

Aurélien Boyé

Mathieu Chevalier

Olivier Gauthier

Jacques Grall

Martin P. Marzloff

1. Joint Species Distribution Models (jSDM) are increasingly used to explain and predict biodiversity patterns. jSDMs account for species co-occurrence patterns and can include phylogeny or functional traits to better capture the processes shaping communities. Yet, several factors may limit or affect the interpretability and predictive ability of jSDMs : missing abiotic predictors, omitting ecologically-important species, or increasing the number of model parameters by adding phylogeny and/or trait information.
2. We assessed how interpretability, explanatory and predictive power of jSDM varied across four alternative models focusing on 99 coastal benthic marine polychaete species: (1) a baseline jSDM with no additional information sources other than abiotic predictors and residual co-occurrence patterns, (2) a jSDM including phylogeny alone or (3) in combination with traits data and (4) a jSDM including monitoring information related to additional species sampled alongside the target assemblage (i.e. non-target species that are not of direct interest but potentially interact with the target assemblage). The four models fitted on both presence/absence and abundance data from a regional monitoring programme were assessed using complementary metrics. We compared performance at both species- and community-level, considering multiple facets of species responses and assemblage diversity.
3. For both presence/absence and abundance data, all models displayed good and similar explanatory power but varied in their interpretability and predictive power. Considering trait data provides insights on species response along environmental gradients, which is a decisive element for model interpretability. Relative to the baseline model, predictive power increased by 26% when including data on additional species, whereas only marginal changes were detected for the two other models. These patterns are explained by changes in the species-environment relationships and residual co-occurrence patterns inferred by these models.
4. Overall, this study highlights that adequate strategy to fit jSDM depends on data at hand, modelling objective and research question. To understand observed community space-time variability, adding phylogenetic or trait information is most effective. Inclusion of non-target species is however a better strategy to predict how the target species assemblage responds to environmental changes. Importantly, we provide a comprehensive toolbox for the comparative assessment of jSDM performance.

# Introduction

Community ecology aims at explaining and predicting spatio-temporal variability in species diversity ([Whittaker *et al.* 2001](#ref-Whittaker_2001)) and coexistence ([Chesson 2000](#ref-Chesson_2000)). Understanding the processes that determine species distribution around the planet is a prerequisite to characterise and predict community structure and associated ecological dynamics, which is critical to mitigate the effects of global change on biodiversity and prevent the sixth mass extinction ([IPBES 2019](#ref-ipbes_2019)). Currently, the major challenges faced by ecologists include describing, explaining, and predicting changes in communities ([Tredennick *et al.* 2021](#ref-Tredennick_2021)) in order to inform effective management or restoration measures in a rapidly changing world ([Houlahan *et al.* 2017](#ref-Houlahan_2017) ; [Dietze *et al.* 2018](#ref-Dietze_2018) ; [Brudvig & Catano 2022](#ref-Brudvig_2022)). Joint Species Distribution Models (jSDM) are particularly well-suited tools to address these challenges, whether to characterise the processes that shape observed communities ([Warton *et al.* 2015](#ref-Warton_2015) ; [Ovaskainen *et al.* 2017b](#ref-Ovaskainen_2017a)), or to predict how communities will evolve in the future ([Norberg *et al.* 2019](#ref-Norberg_2019) ; [Pollock *et al.* 2020](#ref-Pollock_2020)).

jSDMs are multivariate (i.e. multi-species) extensions of Species Distribution Models (SDMs), which have been broadly applied over the past decades - across all terrestrial and marine realms - to understand and predict both species occurrences ([Elith *et al.* 2006](#ref-Elith_2006) ; [Norberg *et al.* 2019](#ref-Norberg_2019)) and species abundances ([Howard *et al.* 2014](#ref-Howard_2014) ; [Waldock *et al.* 2022](#ref-Waldock_2022)) using a set of covariates (e.g. climatic variables). One advantage of jSDM relies on their explanatory power owing to their tight link with the assembly rule framework ([Ovaskainen *et al.* 2017b](#ref-Ovaskainen_2017a)). In particular, relative to single-species SDMs that only consider the abiotic niche of species (i.e. the Grinellian niche), jSDM can theoretically also account for interspecific interactions (i.e. the Eltonian niche).

Indeed, in jSDMs, the variability in community composition not explained by covariates is captured by a residual covariance matrix representing species co-occurence patterns potentially representing biotic interactions ([Ovaskainen *et al.* 2017b](#ref-Ovaskainen_2017a)). This feature is highly attractive to ecologists because it provides a way to disentangle the relative influence of abiotic and biotic processes on biodiversity patterns ([Godsoe *et al.* 2017](#ref-Godsoe_2017)) while also improving model’s predictive power ([Giannini *et al.* 2013](#ref-Giannini_2013) ; [Staniczenko *et al.* 2017](#ref-Staniczenko_2017)). However, in practice, inferring and interpreting residual co-occurence patterns using jSDMs remains challenging for several reasons ([Blanchet *et al.* 2020](#ref-Blanchet_2020) ; [Holt 2020](#ref-Holt_2020)).

First, while jSDMs have been applied to a large number of species presence/absence datasets ([Norberg *et al.* 2019](#ref-Norberg_2019) ; [Wilkinson *et al.* 2019](#ref-Wilkinson_2019) ; [Wilkinson *et al.* 2020](#ref-Wilkinson_2020)), simulation studies showed that co-occurence networks inferred from such data does not necessarily provide evidence for species interactions ([Sander *et al.* 2017](#ref-Sander_2017) ; [Dormann *et al.* 2018](#ref-Dormann_2018) ; [Blanchet *et al.* 2020](#ref-Blanchet_2020)) and only inform about spatial and temporal associations between species ([Keil *et al.* 2021](#ref-Keil_2021)). Some authors speculated that jSDMs applied to abundance data - instead of presence/absence data - are likely to provide a better proxy for biotic interactions ([Blanchet *et al.* 2020](#ref-Blanchet_2020) ; [Momal *et al.* 2020](#ref-Momal_2020)). Accordingly, jSDM have progressively been extended and applied to abundance data ([Hui 2016](#ref-Hui_2016) ; [Ovaskainen *et al.* 2017b](#ref-Ovaskainen_2017a) ; [Chiquet *et al.* 2021](#ref-Chiquet_2021) ; [Popovic *et al.* 2022](#ref-Popovic_2022)). Yet, specific challenges related to modelling abundance data have only been recently explored in the context of species distribution modelling ([Waldock *et al.* 2022](#ref-Waldock_2022)). To date, the predictive and the explanatory power of jSDM fitted to abundance data remains largely untested compared to presence/absence data ([Norberg *et al.* 2019](#ref-Norberg_2019) ; [Wilkinson *et al.* 2020](#ref-Wilkinson_2020)).

Second, regardless of the type of data considered (i.e. presence/absence or abundance), several factors may limit or affect the interpretability and predictive ability of jSDM. For instance, co-occurence patterns estimated in jSDM are affected by unaccounted environmental variables implying that jSDMs cannot fully separate the environmental and the biotic niche of species ([Blanchet *et al.* 2020](#ref-Blanchet_2020) ; [Poggiato *et al.* 2021](#ref-Poggiato_2021)). Beyond missing environmental predictors, one prerequisite for improving biotic inference and thus jSDMs’ predictions is to take into account other actors (i.e. species) that could have an influence on the target community (e.g. competitors; Levine *et al.* ([2017](#ref-Levine_2017))). However, because many ecological studies only focus on particular taxonomic groups ([Pollock *et al.* 2014](#ref-Pollock_2014) ; [Häkkilä *et al.* 2018](#ref-Hakkila_2018)), hence disregarding non-target taxa, co-occurence patterns estimated from jSDMs are almost always skewed by missing ecological actors ([Momal *et al.* 2021](#ref-Momal_2021)). How this bias affects the predictive ability of jSDM remains untested.

Finally, similarly to SDMs, jSDMs can theoretically be extended to include additional sources of information about modelled species ([Niku *et al.* 2019](#ref-Niku_2019) ; [Ovaskainen *et al.* 2017b](#ref-Ovaskainen_2017a)). For instance, accounting for phylogenetic relationships between species ([Ives & Helmus 2011](#ref-Ives_2011)) or for the link between functional traits and environmental responses ([Pollock *et al.* 2012](#ref-Pollock_2012)) have been shown to improve both the explanatory and the predictive powers of SDMs ([Morales-Castilla *et al.* 2017](#ref-Morales-Castilla_2017) ; [Vesk *et al.* 2021](#ref-Vesk_2021)), which supports the hypothesis that similar species in terms of traits and/or recent evolutionary history share similar environmental preferences. While similar effects related to inclusion of species-specific information are expected in jSDMs ([Ovaskainen *et al.* 2017b](#ref-Ovaskainen_2017a)), the relative influence of additional sources of information on their interpretability and predictive power remains untested ([Norberg *et al.* 2019](#ref-Norberg_2019) ; [Wilkinson *et al.* 2019](#ref-Wilkinson_2019)).

Overall, many practical questions remain concerning the application of jSDMs to ecological community monitoring data in particular related to inclusion of additional sources of information within the models. In this study, we aim to provide a comprehensive assessment of how jSDM predictive and explanatory powers are affected by different sources of information. Specifically, by comparing predictions obtained from a baseline model excluding additional sources of information (i.e. a classical jSDM), we tested the effect of (1) including phylogeny alone and in combination with trait data, (2) incorporating monitoring information related non-target species and (3) considering abundance instead of presence/absence data. We hypothesised that all these sources of information should improve jSDM predictive and explanatory powers, but did not assume a priori that a given modelling strategy would lead to greater improvements in model performances.

# Materials & Methods

We used the HMSC (Hierarchical Modeling of Species Communities) framework applied to the long-term REBENT coastal monitoring dataset ([rebent.ifremer.fr](https://rebent.ifremer.fr)). In the following subsections, we sequentially describe Fig. [1](#fig:workflow) : (1) the HMSC framework, (2) the data used in this study, (3) data splitting between training and testing sets to assess the explanatory and predictive powers of models, respectively, (4) the rationales for the suite of alternative models considered and, (5) the performance metrics used to compare models.

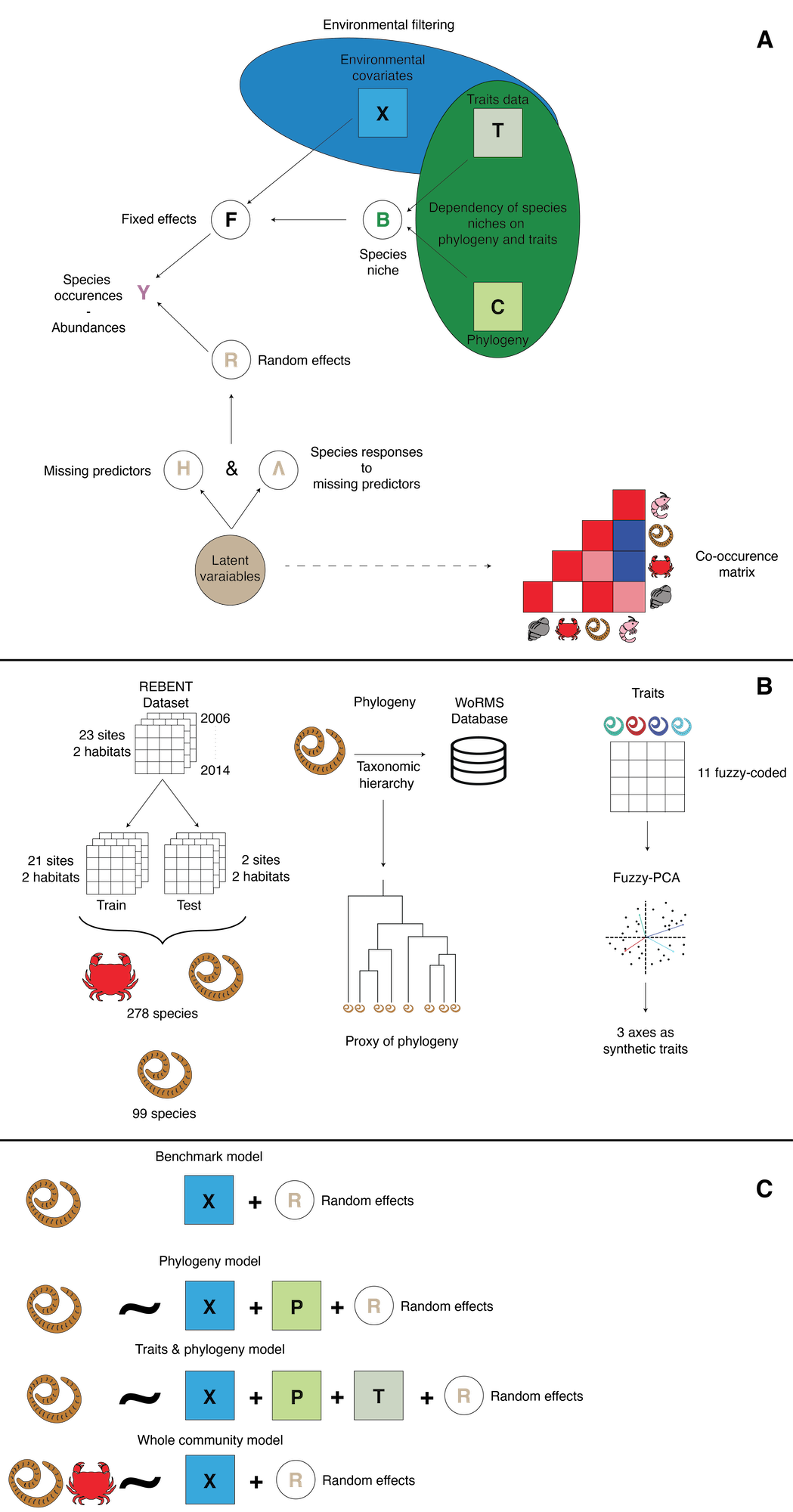


Figure 1: Workflow of the study. A. Structure of a Hierarchical Model of Species Community (HMSC) including environmental variables, phylogeny and species-specific functional traits. B. Data pre-processing: community data partitioning between train and test datasets, estimating of phylogenetic distance between species (using taxonomic classification) and dimension reduction of species-trait matrix using a fuzzy-PCA. C. Summary of the four alternative model structures fitted both on presence/absence and abundance data: the Benchmark, Phylogeny and Traits & Phylogeny models only considered polychaete species assemblage data, while the Whole Community model includes information related to additional species sampled alongside the target assemblage (i.e. non-target species that are not of direct interest but potentially interact with the target assemblage). Random effects accounting for the sampling year, site and habitat were included in all models.

## Hierarchical Modelling of Species Community (HMSC)

“HMSC is a multivariate hierarchical generalised linear mixed model adjusted with Bayesian inference rooted in assembly theory” ([Ovaskainen & Abrego 2020](#ref-Ovaskainen_2020)). A HMSC model is composed of two parts: one taking into account fixed effects and the other taking into account random effects. The fixed part models the realised niche (i.e., the set of environmental conditions that is biotically suitable and accessible to the species; Ovaskainen & Abrego ([2020](#ref-Ovaskainen_2020))) of each species (B matrix), where each dimension of the niche is a covariate (e.g. temperature) included in the model (Fig. [1](#fig:workflow) ; Ovaskainen & Abrego ([2020](#ref-Ovaskainen_2020))). Including trait data enables estimating of species-specific expected niche value by accounting for trait-environment relationships, where species with similar traits are expected to respond similarly along environmental gradients (Fig. [1](#fig:workflow) ; Ovaskainen *et al.* ([2017b](#ref-Ovaskainen_2017a)) ; Ovaskainen & Abrego ([2020](#ref-Ovaskainen_2020))). It is well-established that phylogenetically-close species tend to share similar trait values or niches ([Wiens *et al.* 2010](#ref-Wiens_2010)). Adding phylogenetic data to a HMSC model already including traits is not necessarily redundant because it could capture residual ecological information not included in the available trait data. This can theoretically improve species niche estimates ([Ovaskainen & Abrego 2020](#ref-Ovaskainen_2020)). Inclusion of such additional pieces of information can moreover improve model fit for rare species by borrowing information on traits- (or phylogenetic-) environment relationships estimated for common species that are similar in terms of traits (or phylogenetic; Ovaskainen & Abrego ([2020](#ref-Ovaskainen_2020))). This property is a main advantage of hierarchical models ([Gelman *et al.* 2020](#ref-Gelman_2020)).

The random part of HMSC relies on latent variables. Specifically, for each random effect, two matrices of latent variables are estimated ([Ovaskainen *et al.* 2017b](#ref-Ovaskainen_2017a) ; [Tikhonov *et al.* 2019](#ref-Tikhonov_2019b) ; [Ovaskainen & Abrego 2020](#ref-Ovaskainen_2020)): the $\Eta$ matrix (called site loadings) contains the values of missing covariates not included in the model (Fig. [1](#fig:workflow) ; Ovaskainen *et al.* ([2017b](#ref-Ovaskainen_2017a)) ; Ovaskainen & Abrego ([2020](#ref-Ovaskainen_2020))); while the matrix (called species loadings) corresponds to the response of the species to missing covariates (Fig. [1](#fig:workflow) ; Ovaskainen *et al.* ([2017b](#ref-Ovaskainen_2017a)) ; Ovaskainen & Abrego ([2020](#ref-Ovaskainen_2020))). These covariates thus capture residual variance, which can be due to various factors including missing environmental features or the effect of biotic interactions ([Ovaskainen *et al.* 2017a](#ref-Ovaskainen_2017b) ; [Ovaskainen *et al.* 2017b](#ref-Ovaskainen_2017a) ; [Ovaskainen & Abrego 2020](#ref-Ovaskainen_2020)).

## Datasets

### Faunistic data

Faunistic data come from the REBENT programme ([rebent.ifremer.fr](https://rebent.ifremer.fr)), which is a station-based ongoing monitoring network initiated in 2003 to detect, characterise and explain changes of coastal benthic macrofauna across Brittany’s coastline (Western France). Here, we focused on benthic infaunal communities found in two soft-bottom habitats: intertidal bare sediments and intertidal seagrass meadows (*Zostera marina*). Data from Boyé *et al.* ([2019](#ref-Boye_2019a)), covering 23 sites (Fig. S1) monitored using the same protocol between 2006 and 2014, were used in this study. At each site, sampling consists in the collection of three sediment cores of 0.03m that are pooled together and considered as a single sampling unit at each site. For each sampling event, individuals were identified to the lowest taxonomic level possible (mostly species level; for simplicity we hereafter use the term “species”). A detailed description of the sampling methodology is provided in ([Boyé *et al.* 2017](#ref-Boye_2017) ; [Boyé *et al.* 2019](#ref-Boye_2019a)). Overall, across a total of 375 sampling units (i.e. unique combination of years, sites and habitats), 861,997 individuals belonging to 821 species were collected and identified.

### Functional traits and phylogeny data

We collated species-specific information related to functional traits and phylogeny for inclusion in different models. These data were particularly well resolved for the polychaete community which therefore constitutes the main object of inference. Polychaeta is a taxonomic group composed of numerous species exhibiting diverse lifestyles ([Jumars *et al.* 2015](#ref-Jumars_2015)) that can be used to monitor the health of benthic habitats ([Giangrande *et al.* 2005](#ref-Giangrande_2005)). The polychaete traits data, which was available for the 99 polychaete species present in the training set, includes 11 fuzzy-coded traits for a total of 41 modalities ([Boyé *et al.* 2019](#ref-Boye_2019a)). Prior to jSDM fitting, the dimensionality of the trait matrix was reduced using a fuzzy-PCA with the *fpca* function from the *ade4* R package ([Thioulouse *et al.* 2018](#ref-Thioulouse_2018)). The first three axes, which account for 59% of the total variance of the trait matrix, were included in the model as synthetic traits data (Fig. [1](#fig:workflow)). The first axis distinguishes mobile predatory species from sessile microphages; the second axis differentiates semelparous species from iteroparous species; and, the third axis separates burrowers from tube-dwellers (Fig. S5).

In complement to the traits information available for the 99 polychaete species of interest, we retrieved their taxonomic classification through the WoRMS database (([www.marinespecies.org](https://www.marinespecies.org); assessed in january 2020) and used this information as a proxy for phylogenetic relationships (Fig. [1](#fig:workflow) ; Ricotta *et al.* ([2012](#ref-Ricotta_2012)) ; Ovaskainen & Abrego ([2020](#ref-Ovaskainen_2020))). Phylogenetic distances between polychaete species were then estimated using the *ape* R package ([Paradis & Schliep 2019](#ref-Paradis_2019)).

### Environmental data

Following Boyé *et al.* ([2019](#ref-Boye_2019a)), we selected seven environmental variables to characterise the ecological niche of each species within the target community. These seven variables quantify different components of coastal environmental variability including hydrology (sea water temperature, salinity and current velocity), sedimentology (mud and organic matter content), substrate heterogeneity (Trask index) and local wave exposure (fetch). For each of these seven variables, the first and second degree polynomials were computed to account for non-linear responses.

## Comparison of alternative model structures

The first model (benchmark model abbreviated as “Bench”) only relies on polychaete community data and environmental covariates (Fig. [1](#fig:workflow)). The second model (phylogenetic model abbreviated as “Ph”) adds phylogenetic data to the Bench model (Fig. [1](#fig:workflow)), which implies that rare species can thus benefit from phylogenetic-environment relationships estimated for closely related species ([Ives & Helmus 2011](#ref-Ives_2011)). The third model (traits & phylogeny model abbreviated as “TrPh”) adds traits data to the Ph model (Fig. [1](#fig:workflow)), which means that rare species can benefit from traits-environment relationships estimated for species presenting similar functional traits (whereas phylogeny can capture ecological similarities between species, which are not captured by trait similarity; Pollock *et al.* ([2012](#ref-Pollock_2012))). Finally, the fourth model (whole community model abbreviated as “WhC”), adds information about the whole community (i.e. including non-polychaete species for a total of 278 species) to the Bench model (only 99 polychaete; Fig. [1](#fig:workflow)). This model does not include trait or phylogenetic data for the sake of computation time. Each of these four models were fitted twice, either using presence/absence or abundance data. All models include the same random effects (Fig. [1](#fig:workflow)): a temporal random effect to account for variability across years, a spatial random effect to account for variability across sites and another spatial random effect to account for variability across habitats (bare vs seagrass).

## Model fitting and performance

### Model fitting using Markov Chain Monte Carlo

HMSC uses a Bayesian framework for model fitting where the posterior distribution is sampled using a MCMC algorithm. For each model we ran 15 chains, each with 30,000 iterations. The first 10,000 iterations were discarded as burn-in while the remaining were thinned every 20 iterations yielding 1,000 posterior samples per chain. Hence, in total, 15,000 posterior samples were recorded for each parameter. Model convergence for each model parameter was assessed using the potential scale reduction factor ([Gelman & Rubin 1992](#ref-Gelman_1992)).

### Assessing model performance and interpretability

In order to independently assess models’ predictive performance, we splitted the dataset into a train and a test set. The training dataset includes 180 sampling units defined as unique combinations of years (varies between six and nine depending on sites), sites (21) and two habitats (Fig. S1). From this dataset, that originally contained 519 species, we removed the species that occurred less than four times across the 180 observational units to avoid convergence issues and poor model inference, leading to the removal of 241 species. The remaining 278 species encompassed the 99 polychaete species that made up the target community and the 142 non-target species that were included in the WhC model. The test dataset was composed of 35 sampling units related to all surveys over a 9-year period at two specific sites , where both habitats (i.e. seagrass and bare sand) occur. Beyond the presence of both habitats, these two sites were also chosen because they occur in environmental conditions that can be considered average at the scale of the region (thereby limiting extrapolation of the model; Boyé *et al.* ([2017](#ref-Boye_2017)) ; Boyé *et al.* ([2022](#ref-Boye_2022)) ; Toumi Chirine ([n.d.](#ref-Toumi_nd))) while still harbouring different communities, representative of the known diversity gradient across the region ([Boyé *et al.* 2017](#ref-Boye_2017) ; [Toumi Chirine n.d.](#ref-Toumi_nd)).

To investigate jSDM’s performance, models were evaluated using a set of complementary metrics to evaluate both their explanatory (predictions compared to observations of the train dataset) and predictive (predictions compared to observations of the test dataset) powers ([Wilkinson *et al.* 2020](#ref-Wilkinson_2020)). To assess models’ performance, both overall (i.e., across all species) and at individual species level, we used AUC and root mean squared errors (RMSE) for presence/absence and abundance models, respectively. For the “whole community” model that most improved predictive power (see results), we further explored species-specific gain in explanatory power by examining potential correlations between (i) RMSE and the proportion of presences and (ii) RMSE and average abundance using the Kendall rank correlation coefficient.

While the AUC and the RMSE provide estimates of model performance, either overall or for individual species, these measures only partially capture model accuracy (or performance) at the community scale. Hence, we also explored differences between observed and predicted community composition (both for the train and test datasets) by decomposing the total beta diversity (using the Sørensen index) into species turnover and nestedness using the *betapart.temp* function from the *betapart* R package ([Baselga 2010](#ref-Baselga_2010) ; [Baselga *et al.* 2022](#ref-Baselga_2022)). For abundance models, predictions were transformed to presence/absence before computing beta diversity (i.e. all non-zero abundance predictions were considered as presences). Thus, this framework assigns a total beta diversity of zero to a model predicting the exact observed community, whereas a model predicting a completely different community compared to observations is associated with a total beta diversity of one. Moreover, using Baselga *et al.* (2022)’s framework, we decomposed beta diversity (i.e. predicted error in community composition) according to two components: (1) turnover, if model correctly estimates observed species richness but mispredicts species identity and (2) nestedness, if model correctly predicts the identity of observed species but omits some.

To assess model interpretability, we calculated the proportion of explained variance attributed either to environmental covariates (fixed effects) or to random effects. To evaluate the effect of model structure on estimated species-environment relationships, we classified the shapes of estimated response curves inferred from the different models according to nine categories that characterise both their direction (decline, null or increase) and their acceleration (decelerated, constant or accelerated) ([Rigal *et al.* 2020](#ref-Rigal_2020)). We then looked for differences between models in terms of proportion of estimated response curves across each of these nine categories. Finally, to compare changes in random effects across models, we estimated differences between the Bench model and the best performing model in terms of residual co-occurrence patterns. We specifically quantified differences between models in both magnitude and sign of residual species-species correlations using the following index:

# Results

Both MCMC convergence and effective sample size of the different jSDMs were satisfactory (see Appendix B).

## Model Fit & Predictive power

### Species level

Presence/absence models presented an excellent explanatory power as reflected by mean AUC estimates greater than 0.9 (Fig. S4). Conversely, their predictive power was rather low given a mean AUC estimate of ~0.65 (Fig. S4). For abundance models, mean RMSE computed on the training set ranged from 8.94 to 9.35 (Fig. S4). Their predictive power was heterogeneous with the whole community (WhC) model presenting the highest performance () followed by the benchmark model (Bench) (), the phyloheny (Ph) model () and the traits & phylogeny model () (Fig. S4).

Relative to the benchmark model (Fig. [2](#fig:fig2)), model explanatory power only sightly decreased for both TrPh (mean increase in RMSE +0.8% and decrease in AUC of -0.6%) and Ph models (mean increase in RMSE of +0.5% and decrease in AUC -0.6%). Explanatory power only slightly increased for the WhC models (mean decrease in RMSE of -3.6% and increase in AUC +0.3%). In terms of predictive power, performance mostly increased for the WhC abundance model with a mean decrease in RMSE of 26% relative to the benchmark model. This improvement concerned 62 species (mean decrease in RMSE of -49.3%; 10h and 90th deciles [-94.8% ; -9.7%] for these species) whereas 12 species presented a performance decrease (+36% RMSE; 10th and 90th deciles [10% ; 70.1%]). Only 32 and 36 species were improved for TrPH and Ph models, with mean decrease/increase in RMSE of 26.5% and 24.9% across all species, respectively.

Model explanatory performance increased for the most common (correlation between RMSE and proportion of presence: Kendall’s τ = -0.28, p-value < 1e-5) and abundant (correlation between RMSE and average species abundance: Kendall’s τ = -0.29, p-value < 1e-4).

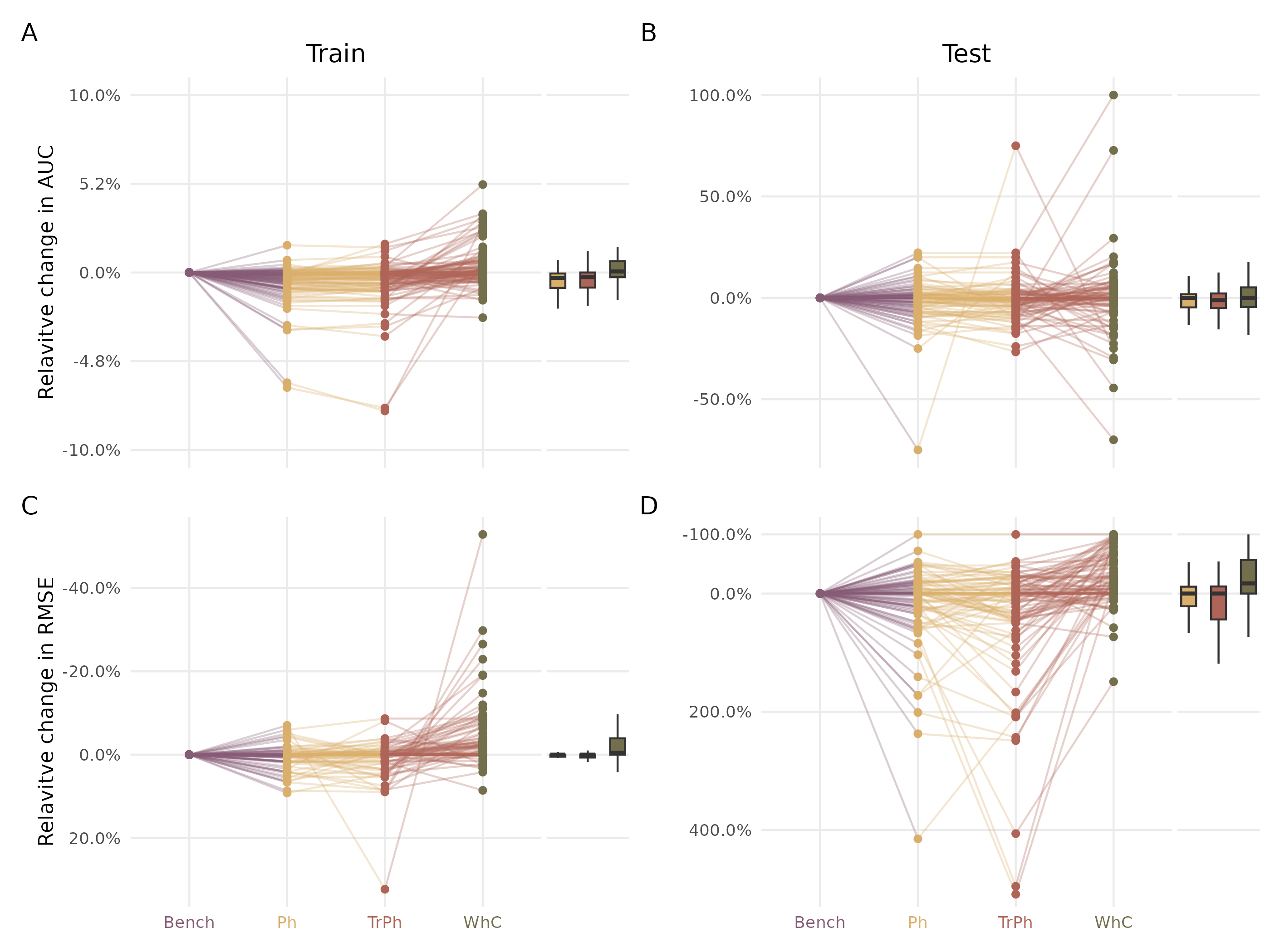


Figure 2: Relative change in explanatory (left column; panels A and C) and predictive (right column; panels B and D) power of different model structures (as colour-coded: purple for benchmark (Bench), yellow for phylogeny (Ph), red for traits & phylogeny (TrPh), and green for whole community (WhC) models). Relative changes (in %) are expressed relative to the benchmark fitted on presence/absence (top row; panels A and B) or abundance (bottom row; panels C and D) data. In all panels, points above the zero line (i.e. increase in AUC for panels A-B but decrease in RMSE for panels C-D) indicate performance gain.

### Community level

On the training set, the median Sørensen dissimilarity index, which ranged from 0.36 to 0.38 across all models (both presence/absence and abundance), suggests that predicted communities are relatively similar to observed communities (Fig. S8 and Fig. S9). Errors were equally distributed between turnover and nestedness (Fig. S8 and Fig. S9). However, relative to observed communities in the test data set, abundance models predictions presented a median Sørensen dissimilarity of 0.65 while dissimilarity reached 0.72 for presence/absence models (Fig. S8 and Fig. S9). Greater dissimilarity between predicted and observed communities in the test dataset relative to the training dataset is a direct consequence of models’ limited predictive power at the species level (see above and Fig. S8 and Fig. S9). Note that proportion of nestedness errors is greater in the WhC model than in other models, suggesting that this model tends to correctly predict the presence of a subset of the observed species assemblage composition (Fig. S8 and Fig. S9).

## Variance partitioning

The amount of variance explained by each model can be partitioned between environmental covariates and random effects. For all models, environmental variables account for most (more than 75% ± 18%, mean ± s.d.) of the variance (Fig. S7). However, a larger part of variance is explained by random effects in the WhC model compared to the Bench model (Fig. S7). Relative to the Bench model, the median relative change in variance explained by random effects increased by 8.6% for the Ph model, 19.9% for the TrPh model and 35.4% for the WhC model (Fig. [3](#fig:fig3)). Similar results were obtained for presence/absence models (Fig. [3](#fig:fig3)).

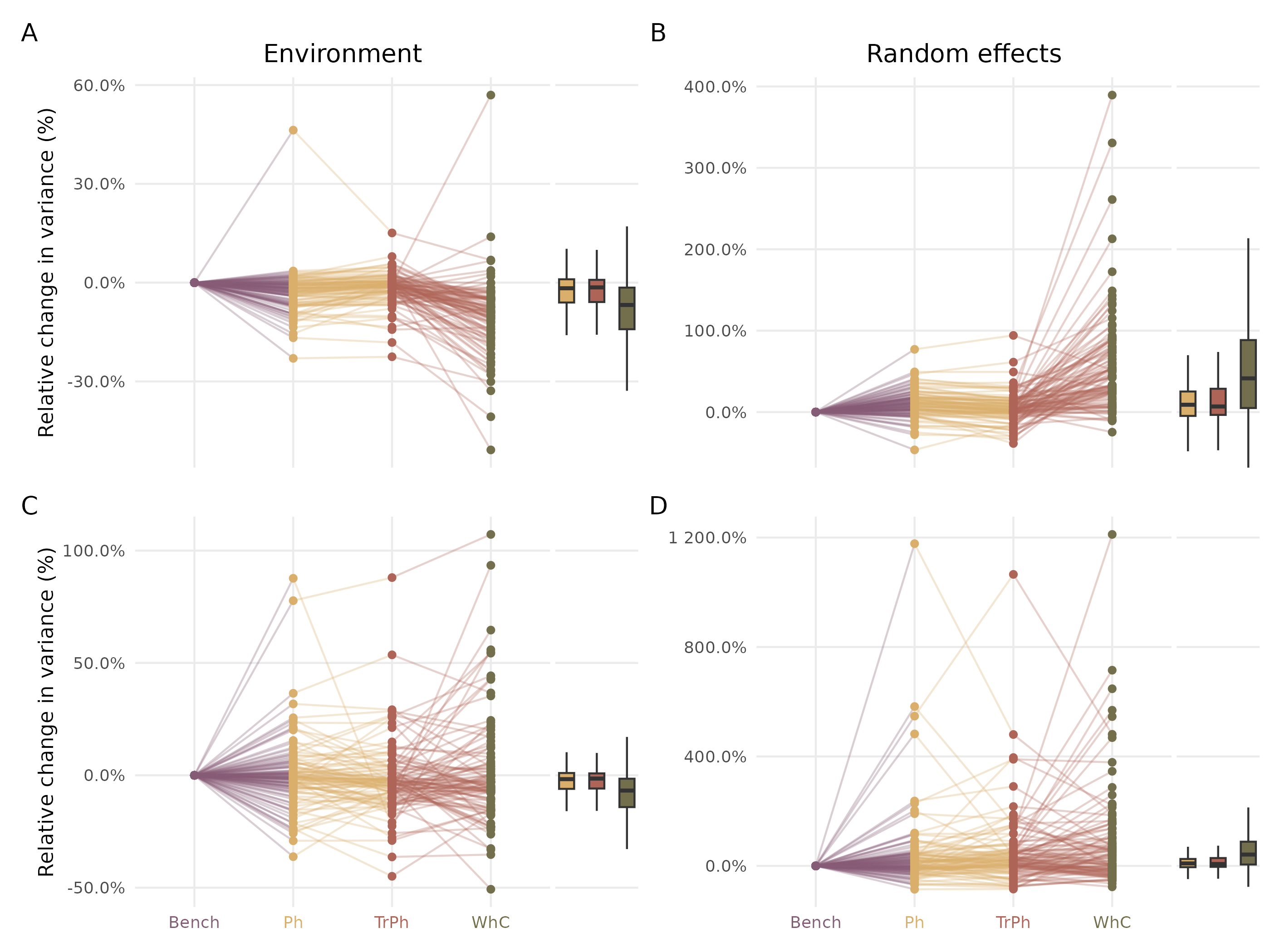


Figure 3: Relative change in variance explained by environmental predictors (left column) and by random effects (right column) according to alternative models as colour-coded (purple for Benchmark (Bench), yellow for phylogeny (Ph), red for traits & phylogeny (TrPh), and green for whole community (WhC) models). Percentage change is expressed relative to the benchmark model fitted with presence/absence (top panels) or abundance (bottom panels) data. In all panels, positive (negative) values point to an increase (decrease) in the proportion of variance explained by the focal model relative to the benchmark model.

## Species niche estimated

For all abundance models, more than 60% of estimated response curves were flat, suggesting a lack of ecologically meaningful species-environment relationships (Fig. [4](#fig:fig4)). This proportion even reached more than 80% for the WhC model. Almost no convex or concave responses curves were estimated for abundance models. Meaningful species-environment relationships essentially included constant or accelerated declines, which respectively represented ~10% and ~15% of estimated response curves for the three models that do not include the whole community (i.e. Bench, TrPh and Ph) (Fig. [4](#fig:fig4)). For the WhC model, these percentages dropped to 4.62% and 9.24%, respectively (Fig. [4](#fig:fig4)). Similar results were obtained for presence/absence models (Fig. S10).

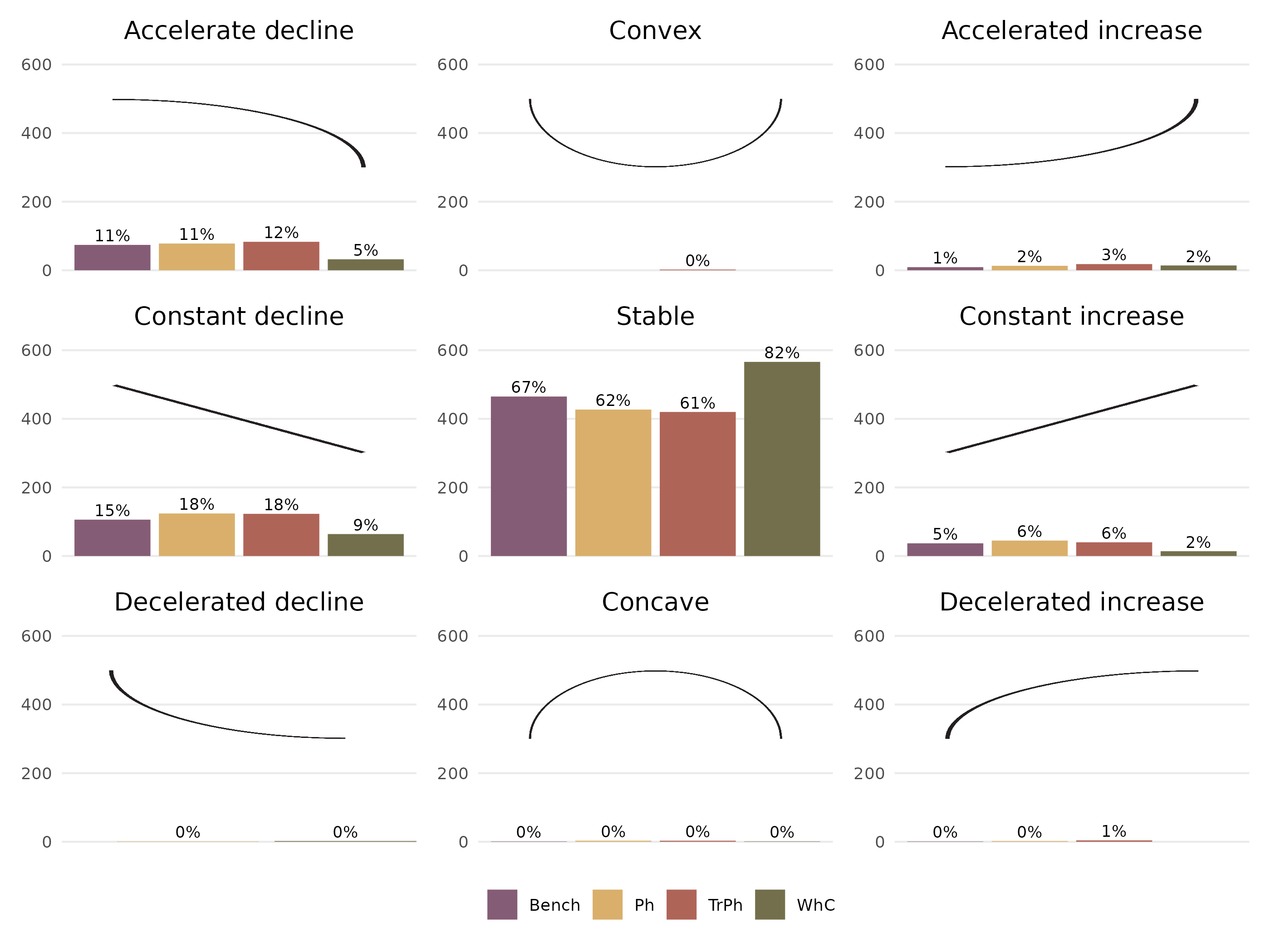


Figure 4: Number (y-axis) and proportion (indicated above individual bars) of response curves (i.e. one for each species-predictor combination) according to the nomenclature (nine shapes highlighted by the black curve in each panel) defined by Rigal *et al.* ([2020](#ref-Rigal_2020)) for different abundance model structures. See Fig. S10 for a similar representation for presence-absence models.

Both abundance- and presence/absence TrPh models (which include species functional traits) reveal some meaningful trait-environment relationships between the first fuzzy-PCA axis and the seven environmental predictors. This suggests that the occurrence of certain traits is likely favoured (or hindered) under certain environmental conditions (Fig. S6). For instance, mobile predatory species were more negatively affected by fetch than sessile suspensivores (Fig. S6). Moreover, increase in organic matter concentration and decrease in current velocities were associated with higher abundances of suspensivore populations.

## Exploring the residual correlation

Since all models included the same random effects, residual correlation matrices are comparable. Here, we compared residual correlations between the Bench model and the WhC model both when fitted to presence/absence or abundance data. We specifically considered the WhC model for this specific comparison, because of (1) its higher performances relative to alternative models and (2) the larger proportion of variance explained by random effects in this model relative to others (Fig. [3](#fig:fig3)).

Residual correlations estimated by the WhC model were similar to those estimated by the Bench model, regardless of whether the models were fitted on presence/absence or abundance data (Fig. [5](#fig:fig5) and Fig. S11). Yet, agreement between models varied across the different random effects. For instance, when comparing residuals between Bench and WhC fitted on abundance data, correlation was low for random site effects (), moderate for random habitat effects () and high for random year effects ().

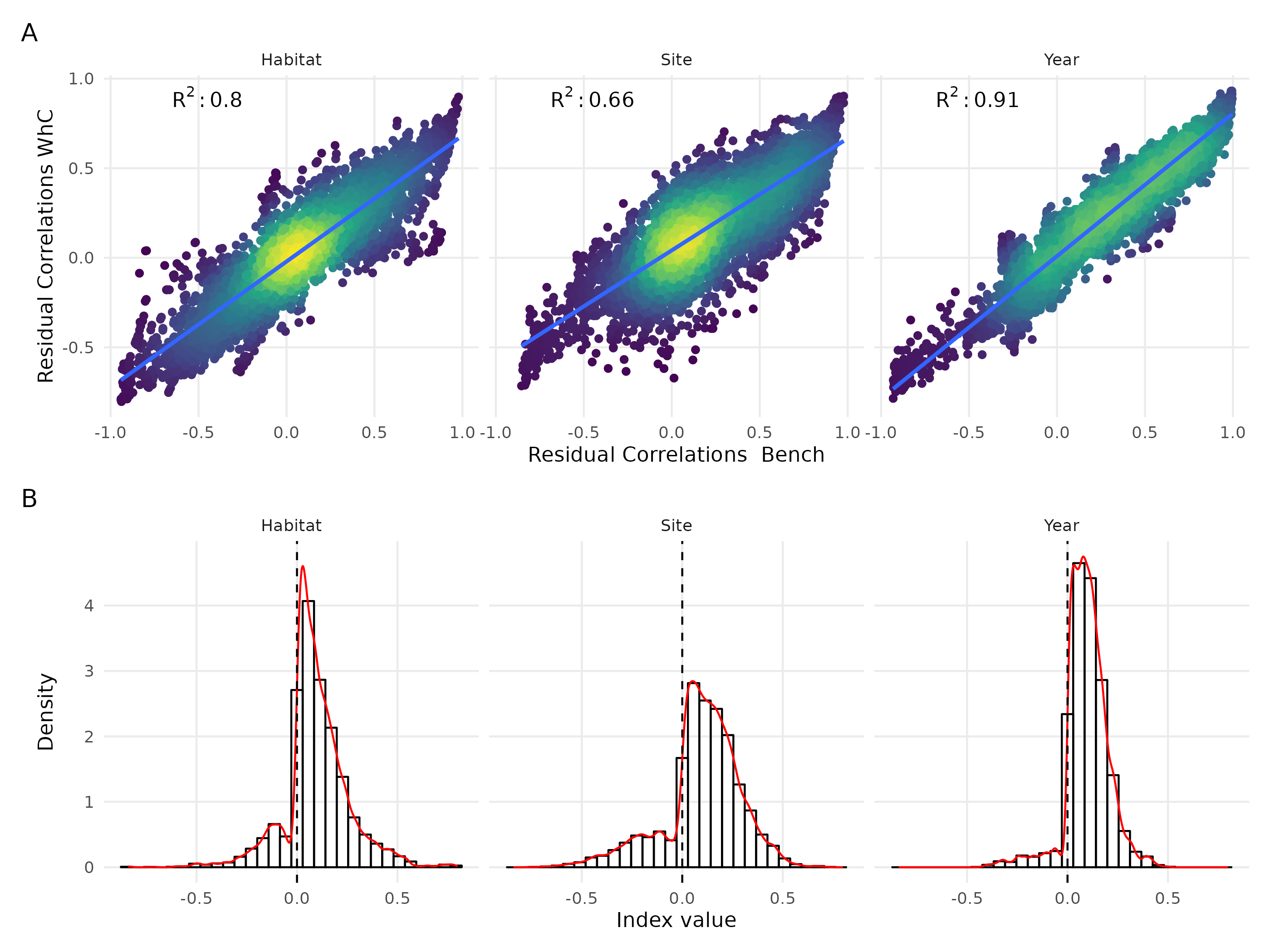


Figure 5: (A) Comparison of residual correlations associated with the three random effects estimated by the Whole Community Model (y-axis) and the Benchmark model (x-axis) fitted on abundance data. The colour scale highlights the density of points in each scatter plot. (B) Distribution of the index measuring change in sign (sign change left to the zero line, no change to the right) and magnitude (higher departure from the zero line indicate higher difference) between residual correlations estimated by the whole community model and the benchmark model adjusted with abundance data for the three random effects (Habitat, Site, Year).

The dedicated comparison index helps qualify how pairwise species-species residual correlations change in sign and magnitude between the Bench and the WhC models. For abundance- or presence/absence models (Fig. [5](#fig:fig5) and Fig. S11), the index main modal distribution, which is centred on zero, suggests an overall agreement between residual correlations obtained from both models (Fig. [5](#fig:fig5) and Fig. S11). While the right part of the distribution highlights variation in the estimated magnitude of effect, on the left-hand part of the distribution (negative values) indicates a sign inconsistency in residual correlations between the two models. For abundance models, the Habitat, Site and Year random effects are respectively associated with 13.3%, 17.7% and 6% of sign inconsistencies in residual correlations between the Bench model and the WhC model. Similar results were obtained for presence/absence models.

# Discussion

Case studies in community ecology typically rely on partial and heterogeneous observations ([Pollock *et al.* 2020](#ref-Pollock_2020)) but also on incomplete knowledge of target species ecological features (e.g. traits, phylogeny; Troudet *et al.* ([2017](#ref-Troudet_2017))). In this paper we aimed to better understand how jSDM performance varies depending on the type of information included (i.e. phylogeny, traits or data on non-target species). While jSDMs have two main goals, i.e. explaining and predicting species distribution and community composition across space and/or time ([Tredennick *et al.* 2021](#ref-Tredennick_2021)), to date, they have mostly been tested with regards to their predictive power ([Norberg *et al.* 2019](#ref-Norberg_2019)), and to some extent in terms of parameter estimates ([Wilkinson *et al.* 2020](#ref-Wilkinson_2020)), but only when fitted on presence/absence data ([Norberg *et al.* 2019](#ref-Norberg_2019) ; [Wilkinson *et al.* 2020](#ref-Wilkinson_2020)). Yet, jSDMs are increasingly fitted on abundance data (e.g. Brimacombe *et al.* ([2020](#ref-Brimacombe_2020))) and used for explanatory purposes ([Abrego *et al.* 2017](#ref-Abrego_2016); [Häkkilä *et al.* 2018](#ref-Hakkila_2018)). Hence, there is currently a mismatch between the knowledge we have regarding the performance of jSDMs and their application by ecologists. Here, we consolidate the current understanding of jSDMs performance using complementary metrics and evaluation methods. Overall, our results highlight that changes in model structure depending on the type of information considered do impact many aspects of model performances (e.g. predictive power, parameter estimates, estimated response curves, community composition). These changes can have significant consequences on the interpretability and the conclusions drawn from these models, especially for ecosystem management policies.

We found that jSDM’s performance, in particular predictive power of abundance models, most increased when including information related to the 179 non-target species sampled alongside with the 99 polychaete species of interest. Given the hierarchical structure of HMSC ([Poggiato *et al.* 2021](#ref-Poggiato_2021)), inclusion of additional monitoring data related to other species likely improves model performance by capturing a combination of relevant drivers that influence the target assemblage but that are not explicitly considered in the model. For instance, inclusion of monitoring data for other species can help describe the realised niche of the species of interest by capturing important ecological processes related to environmental conditions (including trait-mediated responses) or biotic interactions that are not well captured when considering only the target assemblage ([Ovaskainen *et al.* 2017a](#ref-Ovaskainen_2017b)). In jSDMs, these unquantified ecological processes can be estimated using latent variables from model residual correlation matrix. While this feature of jSDMs originally yielded the potential to capture biotic interactions, it is now well-established that potential biotic signals captured by jSDMs are largely confounded by other factors. These include missing environmental variables ([Dormann *et al.* 2018](#ref-Dormann_2018) ; [Zurell *et al.* 2018](#ref-Zurell_2018) ; [Blanchet *et al.* 2020](#ref-Blanchet_2020)), scale mismatch between study organism responses and available environmental variables ([Potter *et al.* 2013](#ref-Potter_2013)), coarse spatial resolution of environmental variables ([Zurell *et al.* 2018](#ref-Zurell_2018) ; [König *et al.* 2021](#ref-Konig_2021)).

Importantly, while including non-target species improved predictive performance in our case study, this does not mean that accounting for non-target species is always beneficial. These benefits could indeed vary depending on robustness of non-target species monitoring data (e.g. detection issues), their role within the ecosystem (e.g. engineer species are likely more influential on local communities than rare transient species) or the process shaping the target assemblage (e.g. when the target assemblage is mostly under the influence of abiotic factors, then adding other species will have limited influence on model performance). Furthermore, how many non-target species should be included warrants further investigations. In the future, simulated datasets, overcoming limitations related to real world datasets ([DiRenzo *et al.* 2022](#ref-DiRenzo_2022)), could be used to more systematically assess how and to which extent jSDM performance is affected when increasing the number of non-target species and whether the functional or the trophic position of these species matters in this context. While species communities and assemblages are largely defined arbitrarily ([Stroud *et al.* 2015](#ref-Stroud_2015)), such sensitivity analysis could help delineate which ecological units to include (or not) to improve model performance for the species of management/conservation interests.

In practice, ecological studies often focus on a certain guild or taxonomic group (e.g. fish, birds) given data collection (consistent sampling methodology) or availability constraints (traits and/or phylogeny biased toward some taxonomic groups, Tyler *et al.* ([2012](#ref-Tyler_2012)) and usually centralised in taxonomic-centred repositories [e.g. FishBase; Froese & Pauly ([2022](#ref-Froese_2022))]), rather than for ecological reasons (e.g. all potential interactions well captured by the data at hand). In this study, dedicated focus on polychaetes was primarily guided by data availability (species-traits matrices available from Boyé *et al.* ([2019](#ref-Boye_2019a)) only included polychaetes) rather than for ecological reasons, although the fact that this taxonomic group is numerically dominant and highly diverse in terms of lifestyles and functional roles ([Giangrande 1997](#ref-Giangrande_1997) ; [Jumars *et al.* 2015](#ref-Jumars_2015)) was the reason that originally motivated trait-data collection.

jSDMs have already been used to model the distribution of a wide variety of species ranging from micro-organisms ([Minard *et al.* 2019](#ref-Minard_2019) ; [Pichler & Hartig 2021](#ref-Pichler_2021)) to megafauna ([Rocha *et al.* 2017](#ref-Rocha_2017) ; [Brimacombe *et al.* 2020](#ref-Brimacombe_2020)) inhabiting many different ecosystems. Here, while we studied communities associated with two specific coastal habitats, i.e. seagrass and sand, that have original characteristics as they are located at the land-sea interface ([Boyé *et al.* 2019](#ref-Boye_2019a)), our case study reflects typical aspects of applied ecological research. These include issues related to data limitation and availability but also typical features of ecological communities (e.g. prevalence of rare and transient species; ([Magurran & Henderson 2003](#ref-Magurran_2003) ; [Snell Taylor *et al.* 2018](#ref-SnellTaylor_2018))). Our results provide some insights on trait-environment relationships but these contributions of functional ecology in jSDMs are likely limited by trait data quality and availability ([Tyler *et al.* 2012](#ref-Tyler_2012) ; [Juan *et al.* 2022](#ref-deJuan_2022)). For instance, we found an interaction between trophic modalities (i.e. microphagous versus macrophagous diet) and fetch (Fig. S15), indicating that organisms that filter on small particles are less likely to occur in wave-exposed sites where high levels of sediment resuspension can block their filtering systems ([Manning *et al.* 2014](#ref-Manning_2014)). Conversely macrophagous organisms are less impacted by fetch. Yet, most trait-environment relationships, and most species-environment relationships were flat suggesting that polychaete assemblages are driven by processes other than abiotic ones, including neutral processes ([Boyé *et al.* 2019](#ref-Boye_2019a)). However, the lack of contribution of other trait-environment relationships in our model could also be related to a mismatch between trait data, environmental data, and the ecological processes at play ([Juan *et al.* 2022](#ref-deJuan_2022)). For instance, the physical coastal environment is highly dynamic; a feature that is only partially characterised by our environmental variables that summarise average climatological conditions (but not extreme events or annual/seasonal variability). Likewise, the list of available fuzzy-coded traits only partially captures species capacity to adapt to frequent disturbances or environmental variability ([Violle *et al.* 2012](#ref-Violle_2012) ; [Juan *et al.* 2022](#ref-deJuan_2022)). Most ecological studies are likely to face similar trade-offs where the potential benefit of including traits within jSDMs is balanced out by the effort needed to collect relevant trait information when missing. In our case, while inclu=sding traits does not improve model predictive power, it enhances our understanding of species responses along environmental gradients. Hence, if the goal is not prediction but inference ([Tredennick *et al.* 2021](#ref-Tredennick_2021)), including traits and proxies of phylogeny can facilitate model interpretation, providing that explanatory power does not decrease (as in our case), and that additional model parameters do not make computation time impractical.

While guidelines have been developed to characterise the performance of jSDM fitted on presence-absence data ([Wilkinson *et al.* 2020](#ref-Wilkinson_2020)), it is only recently that the predictive power of abundance-based models has been explored ([Waldock *et al.* 2022](#ref-Waldock_2022)). Here, we used a set of complementary metrics to assess the performance of both presence-absence and abundance models at the species and community levels, the latter considering both alpha and beta diversity. We also transposed a method initially developed for time series ([Rigal *et al.* 2020](#ref-Rigal_2020)) to provide an innovative way of characterising the response curves of each species. Further, we bring together a set of approaches and propose a new index to characterise and compare residual correlation matrices. Overall, we provide a comprehensive framework for integrative assessment and comparison of alternative jSDM performance.

Overall, our results provide new insights into the most appropriate strategies for jSDM fitting, according to the objective of the modelling exercise ([Troudet *et al.* 2017](#ref-Troudet_2017)) and the data at hand. While the four considered models performed equally well in terms of explanatory power, adding extra information to traditional jSDM (i.e. that only consider abiotic predictors) can still prove useful in some cases. For instance, adding monitoring data for other non-target species can substantially increase model predictive power by modifying inferred species-environment relationships and residual correlation matrices. Similarly, adding traits or phylogeny can lead to better model interpretability. To confirm the generality of our findings, future studies should focus on ecosystems that display different characteristics such as stronger environmental filtering or competitive processes; or by resorting to simulated case studies ([Zurell *et al.* 2010](#ref-Zurell_2010) ; [DiRenzo *et al.* 2022](#ref-DiRenzo_2022)).

## Author Contributions

MPM conveived the original idea with imputs from CV, AB, MC. CV analysed the data and wrote the first draft of the manuscript. All authors provided critical comments on earlier version of the manuscript and ave final approval for publication.

## Acknowledgment

The authors are very grateful to Marion Maguer and Vincet Le Garrec who conducted the fieldwork and laboratory analysis that made this study possible, as well as to all students, technicians, engineers and Ph.D. that have been involved in this long-term effort. This study was supported by the REBENT programme coordinated by Sandrine Derrien (MNHN) and its funding partners (Agence de l’eau Loire-Bretagne, Région Bretagne, DREAL Bretagne). The authors would also like to acknowledge the Pôle de Calcul et de Données Marines (PCDM) for providing DATARMOR storage and computational resources. <https://pcdm.ifremer.fr>. MPM is the recipient of an ANR early career grant ANR-21-CE02-0006.

# References

Abrego, N., Dunson, D., Halme, P., Salcedo, I. & Ovaskainen, O. (2017). [Wood-inhabiting fungi with tight associations with other species have declined as a response to forest management](https://doi.org/10.1111/oik.03674). *Oikos*, 126.

Baselga, A. (2010). [Partitioning the turnover and nestedness components of beta diversity](https://doi.org/10.1111/j.1466-8238.2009.00490.x). *Global Ecology and Biogeography*, 19, 134–143.

Baselga, A., Orme, D., Villeger, S., De Bortoli, J., Leprieur, F. & Logez, M. (2022). [*betapart: Partitioning Beta Diversity into Turnover and Nestedness Components*](https://CRAN.R-project.org/package=betapart).

Blanchet, F.G., Cazelles, K. & Gravel, D. (2020). [Co-occurrence is not evidence of ecological interactions](https://doi.org/10.1111/ele.13525). *Ecology Letters*.

Boyé, A., Gauthier, O., Becheler, R., Le Garrec, V., Hily, C., Maguer, M., *et al.* (2022). [Drivers and limits of phenotypic responses in vulnerable seagrass populations: Zostera marina in the intertidal](https://doi.org/10.1111/1365-2745.13791). *Journal of Ecology*, 110, 144–161.

Boyé, A., Legendre, P., Grall, J. & Gauthier, O. (2017). [Constancy despite variability: Local and regional macrofaunal diversity in intertidal seagrass beds](https://doi.org/10.1016/j.seares.2017.06.004). *Journal of Sea Research*, 130, 107–122.

Boyé, A., Thiébaut, Éric, Grall, J., Legendre, P., Broudin, C., Houbin, C., *et al.* (2019). [Trait-based approach to monitoring marine benthic data along 500 km of coastline](https://doi.org/10.1111/ddi.12987). *Diversity and Distributions*, 25, 1879–1896.

Brimacombe, C., Bodner, K. & Fortin, M.-J. (2020). [Inferred seasonal interaction rewiring of a freshwater stream fish network](https://doi.org/10.1111/ecog.05452). *Ecography*, n/a.

Brudvig, L.A. & Catano, C.P. (2022). [Prediction and uncertainty in restoration science](https://doi.org/10.1111/rec.13380). *Restoration Ecology*, n/a, e13380.

Chesson, P. (2000). [Mechanisms of Maintenance of Species Diversity](https://doi.org/10.1146/annurev.ecolsys.31.1.343). *Annual Review of Ecology and Systematics*, 31, 343–366.

Chiquet, J., Mariadassou, M. & Robin, S. (2021). [The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances](https://doi.org/10.3389/fevo.2021.588292). *Frontiers in Ecology and Evolution*, 9.

Dietze, M.C., Fox, A., Beck-Johnson, L.M., Betancourt, J.L., Hooten, M.B., Jarnevich, C.S., *et al.* (2018). [Iterative near-term ecological forecasting: Needs, opportunities, and challenges](https://doi.org/10.1073/pnas.1710231115). *Proceedings of the National Academy of Sciences*, 115, 1424–1432.

DiRenzo, G.V., Hanks, E. & Miller, D.A.W. (2022). [A practical guide to understanding and validating complex models using data simulations](https://doi.org/10.1111/2041-210X.14030). *Methods in Ecology and Evolution*, n/a.

Dormann, C.F., Bobrowski, M., Dehling, D.M., Harris, D.J., Hartig, F., Lischke, H., *et al.* (2018). [Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions](https://doi.org/10.1111/geb.12759). *Global Ecology and Biogeography*, 27, 1004–1016.

Elith, J., H. Graham, C., P. Anderson, R., Dudı́k, M., Ferrier, S., Guisan, A., *et al.* (2006). [Novel methods improve prediction of species’ distributions from occurrence data](https://doi.org/10.1111/j.2006.0906-7590.04596.x). *Ecography*, 29, 129–151.

Froese, R. & Pauly, D. (2022). [FishBase](https://www.fishbase.org).

Gelman, A., Hill, J. & Vehtari, A. (2020). [*Regression and Other Stories*](https://doi.org/10.1017/9781139161879). Analytical Methods for Social Research. Cambridge University Press.

Gelman, A. & Rubin, D.B. (1992). [Inference from Iterative Simulation Using Multiple Sequences](https://doi.org/10.1214/ss/1177011136). *Statistical Science*, 7, 457–472.

Giangrande, A. (1997). [Polychaete reproductive patterns, life cycles and life histories: an overview](https://doi.org/10.1201/b12590-8). In: *Oceanography And Marine Biology* (ed. A. D.Ansell, M.B., R. N.Gibson). CRC Press, pp. 310–411.

Giangrande, A., Licciano, M. & Musco, L. (2005). [Polychaetes as environmental indicators revisited](https://doi.org/10.1016/j.marpolbul.2005.08.003). *Marine Pollution Bulletin*, 50, 1153–1162.

Giannini, T.C., Chapman, D.S., Saraiva, A.M., Alves-dos-Santos, I. & Biesmeijer, J.C. (2013). [Improving species distribution models using biotic interactions: a case study of parasites, pollinators and plants](https://doi.org/10.1111/j.1600-0587.2012.07191.x). *Ecography*, 36, 649–656.

Godsoe, W., Franklin, J. & Blanchet, F.G. (2017). [Effects of biotic interactions on modeled species’ distribution can be masked by environmental gradients](https://doi.org/10.1002/ece3.2657). *Ecology and Evolution*, 7, 654–664.

Häkkilä, M., Abrego, N., Ovaskainen, O. & Mönkkönen, M. (2018). [Habitat quality is more important than matrix quality for bird communities in protected areas](https://doi.org/10.1002/ece3.3923). *Ecology and Evolution*, 8, 4019–4030.

Holt, R.D. (2020). [Some thoughts about the challenge of inferring ecological interactions from spatial data.](https://doi.org/10.17161/bi.v15i1.13302) *Biodiversity Informatics*, 15, 61–66.

Houlahan, J.E., McKinney, S.T., Anderson, T.M. & McGill, B.J. (2017). [The priority of prediction in ecological understanding](https://doi.org/10.1111/oik.03726). *Oikos*, 126, 1–7.

Howard, C., Stephens, P.A., Pearce-Higgins, J.W., Gregory, R.D. & Willis, S.G. (2014). [Improving species distribution models: the value of data on abundance](https://doi.org/10.1111/2041-210X.12184). *Methods in Ecology and Evolution*, 5, 506–513.

Hui, F.K.C. (2016). [boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r](https://doi.org/10.1111/2041-210X.12514). *Methods in Ecology and Evolution*, 7, 744–750.

IPBES. (2019). [Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services](https://doi.org/10.5281/zenodo.6417333).

Ives, A.R. & Helmus, M.R. (2011). [Generalized linear mixed models for phylogenetic analyses of community structure](https://doi.org/10.1890/10-1264.1). *Ecological Monographs*, 81, 511–525.

Juan, S. de, Bremner, J., Hewitt, J., Törnroos, A., Mangano, M.C., Thrush, S., *et al.* (2022). [Biological traits approaches in benthic marine ecology: Dead ends and new paths](https://doi.org/10.1002/ece3.9001). *Ecology and Evolution*, 12, e9001.

Jumars, P.A., Dorgan, K.M. & Lindsay, S.M. (2015). [Diet of Worms Emended: An Update of Polychaete Feeding Guilds](https://doi.org/10.1146/annurev-marine-010814-020007). *Annual Review of Marine Science*, 7, 497–520.

Keil, P., Wiegand, T., Tóth, A.B., McGlinn, D.J. & Chase, J.M. (2021). [Measurement and analysis of interspecific spatial associations as a facet of biodiversity](https://doi.org/10.1002/ecm.1452). *Ecological Monographs*, n/a.

König, C., Wüest, R.O., Graham, C.H., Karger, D.N., Sattler, T., Zimmermann, N.E., *et al.* (2021). [Scale dependency of joint species distribution models challenges interpretation of biotic interactions](https://doi.org/10.1111/jbi.14106). *Journal of Biogeography*, 48, 1541–1551.

Levine, J.M., Bascompte, J., Adler, P.B. & Allesina, S. (2017). [Beyond pairwise mechanisms of species coexistence in complex communities](https://doi.org/10.1038/nature22898). *Nature*, 546, 56–64.

Magurran, A.E. & Henderson, P.A. (2003). [Explaining the excess of rare species in natural species abundance distributions](https://doi.org/10.1038/nature01547). *Nature*, 422, 714–716.

Manning, L.M., Peterson, C.H. & Bishop, M.J. (2014). [Dominant macrobenthic populations experience sustained impacts from annual disposal of fine sediments on sandy beaches](https://www.int-res.com/abstracts/meps/v508/p1-15/). *Marine Ecology Progress Series*, 508, 1–15.

Minard, G., Tikhonov, G., Ovaskainen, O. & Saastamoinen, M. (2019). [The microbiome of the Melitaea cinxia butterfly shows marked variation but is only little explained by the traits of the butterfly or its host plant](https://doi.org/10.1111/1462-2920.14786). *Environmental Microbiology*, 21, 4253–4269.

Momal, R., Robin, S. & Ambroise, C. (2020). [Tree-based inference of species interaction networks from abundance data](https://doi.org/10.1111/2041-210X.13380). *Methods in Ecology and Evolution*, 11, 621–632.

Momal, R., Robin, S. & Ambroise, C. (2021). [Accounting for missing actors in interaction network inference from abundance data](https://doi.org/10.1111/rssc.12509). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70, 1230–1258.

Morales-Castilla, I., Davies, T.J., Pearse, W.D. & Peres-Neto, P. (2017). [Combining phylogeny and co-occurrence to improve single species distribution models](https://doi.org/10.1111/geb.12580). *Global Ecology and Biogeography*, 26, 740–752.

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2019). [gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r](https://doi.org/10.1111/2041-210X.13303). *Methods in Ecology and Evolution*.

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](https://doi.org/10.1002/ecm.1370). *Ecological Monographs*, e01370.

Ovaskainen, O. & Abrego, N. (2020). *Joint Species Distribution Modelling: With Applications in R*. Ecology, Biodiversity and Conservation. Cambridge University Press.

Ovaskainen, O., Tikhonov, G., Dunson, D., Grøtan, V., Engen, S., Sæther, B.-E., *et al.* (2017a). [How are species interactions structured in species-rich communities? A new method for analysing time-series data](https://doi.org/10.1098/rspb.2017.0768). *Proceedings of the Royal Society B: Biological Sciences*, 284, 20170768.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., *et al.* (2017b). [How to make more out of community data? A conceptual framework and its implementation as models and software](https://doi.org/10.1111/ele.12757). *Ecology Letters*, 20, 561–576.

Paradis, E. & Schliep, K. (2019). [ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R](https://doi.org/10.1093/bioinformatics/bty633). *Bioinformatics*, 35, 526–528.

Pichler, M. & Hartig, F. (2021). [A new joint species distribution model for faster and more accurate inference of species associations from big community data](https://doi.org/10.1111/2041-210X.13687). *Methods in Ecology and Evolution*, 12, 2159–2173.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J.S. & Thuiller, W. (2021). [On the Interpretations of Joint Modeling in Community Ecology](https://doi.org/10.1016/j.tree.2021.01.002). *Trends in Ecology & Evolution*.

Pollock, L.J., Morris, W.K. & Vesk, P.A. (2012). [The role of functional traits in species distributions revealed through a hierarchical model](https://doi.org/10.1111/j.1600-0587.2011.07085.x). *Ecography*, 35, 716–725.

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](https://doi.org/10.1016/j.tree.2020.08.015). *Trends in Ecology & Evolution*, 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)](https://doi.org/10.1111/2041-210X.12180). *Methods in Ecology and Evolution*, 5, 397–406.

Popovic, G.C., Hui, F.K.C. & Warton, D.I. (2022). [Fast model-based ordination with copulas](https://doi.org/10.1111/2041-210X.13733). *Methods in Ecology and Evolution*, 13, 194–202.

Potter, K.A., Arthur Woods, H. & Pincebourde, S. (2013). [Microclimatic challenges in global change biology](https://doi.org/10.1111/gcb.12257). *Global Change Biology*, 19, 2932–2939.

Ricotta, C., Bacaro, G., Marignani, M., Godefroid, S. & Mazzoleni, S. (2012). [Computing diversity from dated phylogenies and taxonomic hierarchies: does it make a difference to the conclusions?](https://doi.org/10.1007/s00442-012-2318-8) *Oecologia*, 170, 501–506.

Rigal, S., Devictor, V. & Dakos, V. (2020). [A method for classifying and comparing non-linear trajectories of ecological variables](https://doi.org/10.1016/j.ecolind.2020.106113). *Ecological Indicators*, 112, 106113.

Rocha, R., Ovaskainen, O., López-Baucells, A., Farneda, F.Z., Ferreira, D.F., Bobrowiec, P.E.D., *et al.* (2017). [Design matters: An evaluation of the impact of small man-made forest clearings on tropical bats using a before-after-control-impact design](https://doi.org/10.1016/j.foreco.2017.06.053). *Forest Ecology and Management*, 401, 8–16.

Sander, E.L., Wootton, J.T. & Allesina, S. (2017). [Ecological Network Inference From Long-Term Presence-Absence Data](https://doi.org/10.1038/s41598-017-07009-x). *Scientific Reports*, 7.

Snell Taylor, S.J., Evans, B.S., White, E.P. & Hurlbert, A.H. (2018). [The prevalence and impact of transient species in ecological communities](https://doi.org/10.1002/ecy.2398). *Ecology*, 99, 1825–1835.

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](https://doi.org/10.1111/ele.12770). *Ecology Letters*, 20, 693–707.

Stroud, J.T., Bush, M.R., Ladd, M.C., Nowicki, R.J., Shantz, A.A. & Sweatman, J. (2015). [Is a community still a community? Reviewing definitions of key terms in community ecology](https://doi.org/10.1002/ece3.1651). *Ecology and Evolution*, 5, 4757–4765.

Thioulouse, J., Dray, S., Dufour, A., Siberchicot, A., Jombart, T. & Pavoine, S. (2018). [*Multivariate Analysis of Ecological Data with ade4*](https://doi.org/10.1007/978-1-4939-8850-1). Springer.

Tikhonov, G., Opedal, O., Abrego, N., Lehikoinen, A. & Ovaskainen, O. (2019). [Joint species distribution modelling with HMSC-R](https://doi.org/10.1101/603217). *bioRxiv*.

Toumi Chirine, G.J., De Cáceres Miquel. (n.d.). Long-term coastal macrobenthic Community Trajectory Analysis reveals habitat-dependent stability patterns. *Ecography*.

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](https://doi.org/10.1002/ecy.3336). *Ecology*, 102, e03336.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. (2017). [Taxonomic bias in biodiversity data and societal preferences](https://doi.org/10.1038/s41598-017-09084-6). *Scientific Reports*, 7.

Tyler, E.H.M., Somerfield, P.J., Berghe, E.V., Bremner, J., Jackson, E., Langmead, O., *et al.* (2012). [Extensive gaps and biases in our knowledge of a well-known fauna: implications for integrating biological traits into macroecology](https://doi.org/10.1111/j.1466-8238.2011.00726.x). *Global Ecology and Biogeography*, 21, 922–934.

Vesk, P.A., Morris, W.K., Neal, W.C., Mokany, K. & Pollock, L.J. (2021). [Transferability of trait-based species distribution models](https://doi.org/10.1111/ecog.05179). *Ecography*, 44, 134–147.

Violle, C., Enquist, B.J., McGill, B.J., Jiang, L., Albert, C.H., Hulshof, C., *et al.* (2012). [The return of the variance: intraspecific variability in community ecology](https://doi.org/10.1016/j.tree.2011.11.014). *Trends in Ecology & Evolution*, 27, 244–252.

Waldock, C., Stuart-Smith, R.D., Albouy, C., Cheung, W.W.L., Edgar, G.J., Mouillot, D., *et al.* (2022). [A quantitative review of abundance-based species distribution models](https://doi.org/10.1111/ecog.05694). *Ecography*, 2022.

Warton, D.I., Blanchet, F.G., O’Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., *et al.* (2015). [So many variables: joint modeling in community ecology](https://doi.org/10.1016/j.tree.2015.09.007). *Trends in Ecology & Evolution*, 30, 766–779.

Whittaker, R.J., Willis, K.J. & Field, R. (2001). [Scale and species richness: towards a general, hierarchical theory of species diversity](https://doi.org/10.1046/j.1365-2699.2001.00563.x). *Journal of Biogeography*, 28, 453–470.

Wiens, J.J., Ackerly, D.D., Allen, A.P., Anacker, B.L., Buckley, L.B., Cornell, H.V., *et al.* (2010). Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology letters*, 13, 1310–1324.

Wilkinson, D.P., Golding, N., Guillera-Arroita, G., Tingley, R. & McCarthy, M.A. (2019). [A comparison of joint species distribution models for presence–absence data](https://doi.org/10.1111/2041-210X.13106). *Methods in Ecology and Evolution*, 10, 198–211.

Wilkinson, D.P., Golding, N., Guillera-Arroita, G., Tingley, R. & McCarthy, M.A. (2020). [Defining and evaluating predictions of joint species distribution models](https://doi.org/10.1111/2041-210X.13518). *Methods in Ecology and Evolution*, n/a.

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](https://doi.org/10.1111/j.1600-0706.2009.18284.x). *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?](https://doi.org/10.1111/ecog.03315) *Ecography*, 41, 1812–1819.