


Essential ingredients in Joint Species Distribution Models: how to optimise inference and prediction in species rich communities?

Clément Violet ¹✉, Aurélien Boyé ¹, Mathieu Chevalier ¹, Olivier Gauthier ²,
Jacques Grall ³, Martin P. Marzloff ¹

¹IFREMER, Centre de Bretagne, DYNECO LEBCO, Plouzané, France; ²Laboratoire des Sciences de l'Environnement Marin (LEMAR) UMR 6539 CNRS UBO IRD IFREMER, Institut Universitaire Européen de la Mer, Université de Bretagne Occidentale, Plouzané, France; ³Observatoire des Sciences de l'Univers, UMS 3113, Institut Universitaire Européen de la Mer, Plouzané, France

✉ For correspondence:

clement.violet@ifremer.fr

Present address: IFREMER,
Centre de Bretagne, DYNECO
LEBCO, Plouzané 29280,
France.

Keywords: Community
assembly, Explanatory power,
Joint Species Distribution
Model, jSDM, Model
Performances, Predictive
power, Species Distribution
Model

Competing interests: The
authors declare no competing
interests.

Abstract

1. Joint Species Distribution Models (jSDM) can be powerful and versatile tools to explain and predict the spatio-temporal variability of species communities, which are critical research questions at the forefront of modern ecology. For instance, jSDM can include phylogeny or functional traits and can account for residual co-occurrences patterns between species to capture the processes shaping communities and their evolution in a changing world. However, the effects of including heterogeneous information sources in jSDMs has not been formally assessed while it raises a number of questions related to their influence on model interpretability and performance.
2. Here, we investigated the effects on jSDM of including additional information, either

species-specific information such as phylogeny and/or traits, or community monitoring data related to accompanying species (i.e. sampled at the same time as the target species community). Based on a typical regional case study, we focus on space-time variability in community structure to specifically assess how alternative model formulations (that includes, or not, additional information) affect jSDM interpretability, explanatory and predictive power.

3. Our results show that jSDM ability to predict variability in targeted species assemblage can be improved by including monitoring data related to accompanying species in the model. Moreover, addition of accompanying species clarifies species-environment relationships by filtering out the weak/spurious correlation inferred in the baseline model. While addition of species-specific information for the target species community (e.g. traits) does not improve model predictive performance, it nevertheless provides some insights on how species of interest respond to environmental gradients, and hence improves model interpretability.

4. This work provides important new guidelines for ecological researchers in terms of appropriate strategies to adopt for jSDM fitting as a function of their modelling objective(s) and/or research questions. If the primary goal lies in understanding observed space-time variability in a given species community, then adding species-specific phylogeny or traits appears as an appropriate strategy. Inclusion of accompanying species is however a better strategy if the primary research aim is to predict how the observed species assemblage of interest responds to environmental changes or to alternative scenarios.

Introduction

Community ecology aims at explaining and predicting spatio-temporal variability in species diversity (Whittaker *et al.* 2001) and coexistence (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). Currently, the major challenges faced by ecologists include describing, explaining, and predicting changes in communities (Tredennick *et al.* 2021) in order to inform effective management or restoration measures in a rapidly changing world (Houlahan *et al.* 2017; Dietze *et al.* 2018; Brudvig & Catano 2022). Joint

52 Species Distribution Models (jSDM) are particularly well-suited tools to address these challenges,
53 whether to characterise the processes that shape observed communities (Ovaskainen *et al.* 2017b),
54 or to predict how communities will evolve in the future (Norberg *et al.* 2019).

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

63 Indeed, in jSDMs, the variability in community composition not explained by covariates is
64 captured by a residual covariance matrix representing species co-occurrence patterns potentially
65 representing biotic interactions (Ovaskainen *et al.* 2017b). This feature is highly attractive to
66 ecologists because it provides a way to disentangle the relative influence of abiotic and biotic
67 processes on biodiversity patterns (Godsoe *et al.* 2017) while also improving model's predictive
68 power (Giannini *et al.* 2013; Staniczenko *et al.* 2017). However, in practice, inferring and interpreting
69 residual co-occurrence patterns using jSDMs remains challenging for several reasons (Blanchet *et al.*
70 2020 ; Holt 2020).

71 First, while jSDMs have been applied to a large number of species presence/absence datasets
72 (Norberg *et al.* 2019 ; Wilkinson *et al.* 2019 ; Wilkinson *et al.* 2020), simulation studies showed
73 that co-occurrence networks inferred from such data does not necessarily provide evidence for
74 species interactions (Blanchet *et al.* 2020 ; Dormann *et al.* 2018 ; Sander *et al.* 2017) and only
75 inform about spatial and temporal associations between species (Keil *et al.* 2021). Some authors
76 speculated that jSDMs applied to abundance data - instead of presence/absence data - are likely to
77 provide a better proxy for biotic interactions (Blanchet *et al.* 2020 ; Momal *et al.* 2020). Accordingly,
78 jSDM have progressively been extended and applied to abundance data (Chiquet *et al.* 2021 ; Hui
79 2016 ; Ovaskainen *et al.* 2017b ; Popovic *et al.* 2022). Yet, specific challenges related to modelling
80 abundance data have only been recently explored in the context of species distribution modelling
81 (Waldock *et al.* 2022). To date, the predictive and the explanatory power of jSDM fitted to abundance
82 data remains largely untested compared to presence/absence data (Norberg *et al.* 2019 ; Wilkinson

83 *et al.* 2020).

84 Second, regardless of the type of data considered (i.e. presence/absence or abundance), sev-
85 eral factors may limit or affect the interpretability and predictive ability of jSDM. For instance,
86 co-occurrence patterns estimated in jSDM are affected by unaccounted environmental variables im-
87 plying that jSDMs cannot fully separate the environmental and the biotic niche of species (Blanchet
88 *et al.* 2020 ; Poggiato *et al.* 2021). Beyond missing environmental predictors, one prerequisite
89 for improving biotic inference and thus jSDMs' predictions is to take into account other actors
90 (i.e. species) that could have an influence on the target community (e.g. competitors; Levine *et*
91 *al.* (2017)). However, because many ecological studies only focus on particular taxonomic groups
92 (Pollock *et al.* 2014 ; Häkkinen *et al.* 2018), hence disregarding non-target taxa, co-occurrence patterns
93 estimated from jSDMs are almost always skewed by missing ecological actors (Momal *et al.* 2021).
94 How this bias affects the predictive ability of jSDM remains untested.

95 Finally, similarly to SDMs, jSDMs can theoretically be extended to include additional sources
96 of information about modelled species (Niku *et al.* 2019 ; Ovaskainen *et al.* 2017b). For instance,
97 accounting for phylogenetic relationships between species (Ives & Helmus 2011) or for the link
98 between functional traits and environmental responses (Pollock *et al.* 2012) have been shown
99 to improve both the explanatory and the predictive powers of SDMs (Morales-Castilla *et al.* 2017
100 ; Vesik *et al.* 2021), which supports the hypothesis that similar species in terms of traits and/or
101 recent evolutionary history share similar environmental preferences. While similar effects related
102 to inclusion of species-specific information are expected in jSDMs (Ovaskainen *et al.* 2017b), the
103 relative influence of additional sources of information on their interpretability and predictive power
104 remains untested (Norberg *et al.* 2019 ; Wilkinson *et al.* 2019).

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

114 modelling strategy would lead to greater improvements in model performances.

115 **Methods & Materials**

116 We used the HMSC (Hierarchical Modeling of Species Communities) framework applied to the
117 long-term REBENT coastal monitoring dataset (rebent.ifremer.fr). In the following subsections, we
118 sequentially describe: (1) the HMSC framework; (2) the data used in this study; (3) data splitting
119 between training and testing sets to assess the explanatory and predictive powers of models, re-
120 spectively; (4) the rationales for the suite of alternative models considered; and, (5) the performance
121 metrics used to compare models.

122 **Hierarchical Modelling of Species Community (HMSC)**

123 “HMSC is a multivariate hierarchical generalised linear mixed model adjusted with Bayesian inference
124 rooted in assembly theory” (Ovaskainen & Abrego 2020). A HMSC model is composed of two parts:
125 one taking into account fixed effects and the other taking into account random effects. The fixed
126 part models the realised niche (i.e., the set of environmental conditions that is biotically suitable
127 and accessible to the species; Ovaskainen & Abrego (2020)) of each species (B matrix), where
128 each dimension of the niche is a covariate (e.g. temperature) included in the model (Ovaskainen &
129 Abrego 2020). Including trait data enables estimating of species-specific expected niche value by
130 accounting for trait-environment relationships, where species with similar traits are expected to
131 respond similarly along environmental gradients (Ovaskainen *et al.* 2017b ; Ovaskainen & Abrego
132 2020). It is well-established that phylogenetically-close species tend to share similar trait values or
133 niches (Wiens *et al.* 2010). Adding phylogenetic data to a HMSC model already including traits is
134 not necessarily redundant because it could capture residual ecological information not included in
135 the available trait data. This can theoretically enhance estimating of species niches (Ovaskainen &
136 Abrego 2020). Inclusion of such additional pieces of information can moreover improve model fit
137 for rare species by borrowing information on traits- (or phylogenetic-) environment relationships
138 estimated for common species that are similar in terms of traits (or phylogenetic; Ovaskainen &
139 Abrego (2020)). This property is a main advantage of hierarchical models (Gelman *et al.* 2020).

140 The random part of HMSC relies on latent variables. Specifically, for each random effect, two ma-
141 trices of latent variables are estimated (Ovaskainen *et al.* 2017b ; Tikhonov *et al.* 2019 ; Ovaskainen
142 & Abrego 2020): the *H* matrix (called *site loadings*) contains the values of missing covariates not in-

cluded in the model (Ovaskainen *et al.* 2017b ; Ovaskainen & Abrego 2020); while the Λ matrix (called *species loadings*) corresponds to the response of the species to missing covariates (Ovaskainen *et al.* 2017b ; Ovaskainen & Abrego 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.* 2017b ; Ovaskainen *et al.* 2017a ; Ovaskainen & Abrego 2020).

Datasets

Faunistic data

Faunistic data come from the REBENT program (rebent.ifremer.fr), which is a station-based monitoring network initiated following the dramatic oil spill of the Erika tanker in December 1999 off Brittany's southern coastline (Western France). The goal of the monitoring network is to detect, characterise and explain changes in French coastal benthic ecosystems through space and time. Between 2003 and 2017, this ongoing program has been monitoring four distinct habitats across 49 sites. 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 since the beginning of the program. Here, we focused on benthic communities found in two soft-bottom habitats: intertidal bare sediments and intertidal seagrass meadows (*Zostera marina*). These habitats were sampled following the same protocol across 23 sites along Brittany's coastline (Fig S1). At each site, sampling consists in collection of 3 sediment cores of 0.03m² 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). A detailed description of the sampling methodology is provided in Boyé *et al.* (2017).

Functional traits and phylogeny data

In this study, we collated species-specific information such as functional traits and phylogeny for inclusion in different models. We chose to focus on a particular class, the polychaeta. Polychaeta, which encompasses numerous species that exhibit diverse lifestyles (Jumars *et al.* 2015), are valuable indicators of the health of benthic habitats (Giangrande *et al.* 2005). The polychaeta traits data, which was available for the 99 polychaeta species in the training set, includes 11 fuzzy-coded traits for a total of 41 modalities (Boyé *et al.* 2019). 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

172 *et al.* 2018). The first three axes, which account for 59% of the total variance of the trait matrix,
173 were included in the model as synthetic traits data (Fig S5). The first axis distinguishes mobile
174 predatory species from sessile microphages; the second axis differentiates semelparous species
175 from iteroparous species; and, the third axis separates burrowers from tube-dwellers (Fig S5).

176 In complement to the traits information available for the 99 polychaeta species of interest, we
177 retrieved their taxonomic classification through the WoRMS database (www.marinespecies.org)
178 and used this information as a proxy for phylogenetic relationships (Ovaskainen & Abrego 2020 ;
179 Ricotta *et al.* 2012). Phylogenetic distances between Polychaeta species were then estimated using
180 the *ape* R package (Paradis & Schliep 2019).

181 Environmental data

182 Based on Boyé (2019), we selected seven environmental variables to characterise the ecological niche
183 of each species within the focal communities. These seven variables quantify different components
184 of coastal environmental variability including hydrology (sea water temperature, salinity and current
185 velocity), sedimentology (mud and organic matter content), granulometry (Trask index) and local
186 wave exposure (fetch). For each variable, the first and second degree polynomials have been
187 computed to account for non-linear responses to the environmental predictors.

188 Comparison of alternative model structures

189 The first model (benchmark model abbreviated as “Bench”) only relies on polychaeta community
190 data and environmental covariates. The second model (phylogenetic model abbreviated as “Ph”)
191 adds phylogenetic data to the Bench model, which implies that rare species can thus benefit from
192 phylogenetic-environment relationships estimated for closely related species (Ives & Helmus 2011).
193 The third model (traits-phylogeny model abbreviated as “TrPh”) adds traits data to the Ph model,
194 which means that rare species can benefit from traits-environment relationships estimated for
195 species presenting similar functional traits (whereas phylogeny can capture ecological similarities
196 between species, which are not captured by trait similarity; (Pollock *et al.* 2012)). Finally, the
197 fourth model (whole community model abbreviated as “Whc”), adds information about the whole
198 community (i.e. non-polychaeta species that represents 278 taxa) to the Bench model (only 99
199 polychaeta). This model does not include trait or phylogenetic data for the sake of computation
200 time. Each of these four models were fitted twice, either using occurrence or abundance data. All

models include the same random effects: 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).

Assessing model performance and interpretability

In order to independently assess models' predictive performance, the REBENT dataset was split into a train and a test dataset. The *training dataset* includes 180 sampling units defined as unique combinations of years (varies between 6 and 9 depending on sites), sites (21) and habitats (2). From this dataset, we removed the species that occurred less than 4 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 polychaeta species that made up the target community and the 142 accompanying species that were included in the *Whc* model. The *test dataset* was composed of 35 sampling units resulting from surveys conducted across two specific sites (9 years for both), which included the two habitats. 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). The performance of all models was assessed globally but also separately for each species using AUC for occurrence-based models and root mean squared errors (RMSE) for abundance-based models. We investigated whether models including additional sources of information had a higher performance than the *Bench* model using Dunn's multiple comparison test (Dunn 1964). For the model that demonstrated the best improvement, we examined whether the improvement correlated with individual species

230 occurrence or abundance (e.g., is the improvement higher for abundant species relative to than for
231 rare species?) using the Kendall rank correlation coefficient.

232 While the AUC and the RMSE can be used to explore model performance globally or for each
233 species, these measures provide no information at the community scale. Hence, we also explored
234 qualitatively the differences between observed and predicted community composition (both for
235 the train and test datasets) by decomposing the total beta diversity (using the Sørensen index) into
236 species turnover and nestedness using the *betapart.temp* function from the betapart R package
237 (Baselga 2010 ; Baselga *et al.* 2022). For abundance-based models, predictions were transformed to
238 presence/absence before computing beta diversity (i.e. all predictions with abundance different than
239 zero were considered as presences). Under this framework, a model predicting the exact observed
240 community would have a total beta diversity of zero whereas a model predicting a community
241 completely different from the one observed would have a total beta diversity of one. As outlined
242 above, the Baselga's framework allows decomposition into two components the type of error when
243 predicting community composition: (1) getting the identity of the species wrong (turnover) or (2)
244 predicting the right species but omitting some (nestedness). In the first case, the model will correctly
245 predict specific richness, while in the other case the model will be more conservative in predicting
246 the correct species but present a bias in species richness.

247 To assess model interpretability, we calculated the proportion of explained variance attributed
248 either to environmental covariates (fixed effects) or to random effects. To evaluate the effect of
249 model structure on estimated species-environment relationships, we classified the shapes of the
250 response curves inferred from the different models according to both their direction (decline, null
251 or increase) and their acceleration (decelerated, constant or accelerated), providing nine different
252 categories (Rigal *et al.* 2020). We then looked for differences between models regarding the
253 proportion of response curves attributed to each category.

254 Finally, we checked the extent to which estimated correlation coefficients differed between the
255 Bench model and the best performing model while also looking for evidence of inversion of effects
256 using the following index:

$$\text{Index} = |corr_{\text{best model}} - corr_{\text{benchmark}}| * \text{sign}(corr_{\text{best model}} * corr_{\text{benchmark}})$$

Results

The MCMC convergence and the effective sample size of the different HMSC models were satisfactory (see Appendix B).

Model Fit & Predictive power

Species level

Occurrence-based models presented an excellent explanatory power, with the AUC being on average greater than 0.9 (Figure S4). Their predictive power was significantly lower with the AUC being about 0.65 on average (Figure S4). For abundance-based models, the RMSE computed on the *training set* ranged from 8.92 to 9.34 on average (Figure S4). Their predictive power was heterogeneous with the *whole community (WhC)* model (RMSE = 5.83 on average) performing better (Figure S4) than the three other models (RMSE values ranged from 54.2 to 95.6, on average).

For the sake of interpretability, all models were compared against *Bench* model (fig. 2). Model's explanatory power was not significantly improved for both *TrPh* and *Ph* models and only slightly increased for the *WhC* models (both occurrence- and abundance-based). This increase in explanatory power was modest with the AUC only increasing by 0.0034 ± 0.0114 (mean \pm sd) for occurrence-based models and the RMSE only decreasing by 0.035 ± 0.796 (mean \pm sd) for abundance-based models. This improvement mainly concerned the most common and abundant species, as reflected by the negative correlations between species-specific RMSE and mean species occurrence (Kendall's $\tau = -0.28$, p-value $< 1e-5$) or mean species abundance (Kendall's $\tau = -0.29$, p-value $< 1e-4$). In terms of predictive power, performance only significantly increased for the *Whc* abundance-based model with a decrease in RMSE of 0.27 ± 0.44 (mean \pm sd) relative to the *Bench* model.

Community level

On the training set, the median Sørensen dissimilarity ranged from 0.36 to 0.38 across models (both occurrence- and abundance-based), suggesting that predicted communities are relatively similar to observed communities (Figure S8 and Figure S9). Errors were equally distributed between turnover and nestedness. With the test data set, abundance-based models presented a median Sørensen dissimilarity of 0.65 while dissimilarity reached 0.72 for occurrence-based models (Figure S8 and Figure S9). This increased dissimilarity relative to predictions made on the training dataset is a direct consequence of the degradation of the predictive power of the various models at the

species scale (see above). Note that the *Whc* model makes more nestedness errors than the others, suggesting that this model is more conservative in terms of community composition (Figure S8 and Figure S9).

Variance partitioning

The amount of variance explained by each model can be decomposed between environmental covariates and random effects. For all models, the environmental variables account for most ($> 75 \% \pm 18$, on average) of the explained variance (Figure S7). However, compared to the *Bench* model, a larger part of variance is explained by the random effect in the *WhC* model (Figure S7). For abundance-based models, the median of the relative change in variance explained by random effects relative to the *Bench* model increased by 0.086 for the *Ph* model, 0.199 for the *TrPh* model and 0.354 for the *WhC* model (fig. 3). Similar results were obtained for occurrence-based models (fig. 3).

Species niche estimated

For abundance-based models, most response curves were neither convex nor concave. For the *Bench*, *TrPh* and *Ph* models, more than 60% of the estimated curves were flat, suggesting a lack of ecologically meaningful species-environment relationships (fig. 4). This rate reached more than 80% for the *WhC* model. Other species-environment relationships included constant or accelerated declines with ~10% and ~15% of such shapes for the three models that do not include the whole community (fig. 4). For the *WhC* model, these percentages dropped to 4.62% and 9.24%, respectively (fig. 4). Similar results were obtained for occurrence-based models (Figure S10).

Considering the *TrPh* model, we further investigated the link between the first fuzzy-PCA axis obtained from the trait matrix and the seven environmental predictors to determine whether some traits were favoured (or hindered) under certain environmental conditions (Figure S6). Both abundance and occurrence-based models highlighted potentially meaningful trait-environment relationships. For instance, mobile predatory species were more negatively affected by fetch than sessile suspensivore. We further found that increasing concentration in organic matter and decreasing current velocities were associated with a higher abundance of suspensivore populations.

Exploring the residual correlation

Since all models included the same random effects, we qualitatively compared the residual correlations estimated by the *Bench* model to the *WhC* model both occurrence- and abundance-based. 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 the random effects in this model relative to others (fig. 3).

The residual correlations estimated by the *WhC* model were similar to those estimated by the *Bench* model, whether the models were occurrence- or abundance-based (fig. 5 and Figure S11). Yet, the agreement between models varied depending on the random effect considered. For instance, considering abundance-based models, the correlation was low for random site effects ($R^2 = 0.66$), moderate for random habitat effects ($R^2 = 0.8$) and high for random year effects ($R^2 = 0.91$).

Our index also qualitatively identifies the residual correlations that have changed the most between the *Bench* and *WhC* models. For abundance or occurrence-based models (fig. 5 and Figure S11), the mode of our index is close to zero, confirming the agreement between the residual correlations obtained from the two models (fig. 5 and Figure S11). Still, our index identifies another mode toward negative values, indicating that the sign of some correlation coefficients have switched from positive to negative. For abundance-based models, the random effects Habitat, Site and Year have respectively 13.3%, 17.7% and 6% of their correlation coefficients that changed sign between the *Bench* model and the *WhC* model. Similar results were obtained for occurrence-based models.

Discussion

It is the norm in ecological case studies to rely on partial data either in terms of spatio-temporal resolution of monitoring data (Pollock *et al.* 2020), or in terms of available information related to target species groups (e.g. traits, phylogeny; Troudet *et al.* (2017)). In this paper we aimed to better understand how the performance of jSDM varies depending on its architecture, and the sources of information considered. jSDMs have two main goals: explaining and predicting species distribution and community composition across space and/or time (Tredennick *et al.* 2021). To date, jSDMs have mostly been tested with regards to their predictive power (Norberg *et al.* 2019), and to some extent in terms of parameter estimates (Wilkinson *et al.* 2020), but only when fitted on presence-absence data (Norberg *et al.* 2019 ; Wilkinson *et al.* 2020). Yet, jSDMs are increasingly fitted on abundance

342 data (e.g. Brimacombe *et al.* (2020)) and used for explanatory purposes (Häkkinen *et al.* 2018 ; Abrego
343 *et al.* 2017). Hence, there is currently a mismatch between the knowledge we have regarding the
344 performance of jSDMs and their application by ecologists. Here, we bridged both worlds using
345 complementary metrics and evaluation methods. Overall, we showed that the structure of the
346 model and the sources of information considered do impact models' performance in many regards
347 (e.g. predictive power, parameter estimates, estimated response curves, community composition).
348 These changes can have significant consequences on the interpretability and the conclusions drawn
349 from these models, especially for ecosystem management policies.

350 We found that jSDM's performance increased when adding information on the 179 accompanying
351 species that were sampled together with the 99 polychaete species of interest. Specifically, the
352 predictive power of abundance-based models was considerably improved. This improvement is
353 likely related to the hierarchical structure of HMSC (Poggiato *et al.* 2021) where accompanying
354 species can improve model performance by capturing a combination of "signals" operating at a scale
355 relevant for the target community, and that can be related to the environment, to biotic interactions,
356 or to any other factors (e.g. traits or and evolutionary processes) that can help better describe the
357 realised niche of the species (Ovaskainen *et al.* 2017b). These "signals" are captured by the latent
358 residual correlation matrix that was the feature that made jSDMs so popular at first glance owing
359 to its potential to be used as an indicator for biotic interactions between species. Yet, it is now
360 well-established that the potential biotic signal captured by jSDMs is confounded by other factors
361 including a mismatch between the size of the studied organisms and the environmental variables
362 (Potter *et al.* 2013), a too coarse spatial resolution (Zurell *et al.* 2018 ; König *et al.* 2021), or missing
363 environmental variables (Dormann *et al.* 2018 ; Zurell *et al.* 2018 ; Blanchet *et al.* 2020). Importantly,
364 while including accompanying species improved predictive performance in our case study, this
365 does not mean that accounting for accompanying species is always beneficial. These benefits
366 could indeed vary depending on the quality of the data for accompanying species (e.g. detection
367 issues), their role within the ecosystem (e.g. the case of engineer species having a strong influence
368 on local communities) or the target community considered (e.g. when the target community is
369 mostly under the influence of abiotic factors, then adding other species should not have a large
370 influence on model's performance). In practice, most studies focus on a certain guild or taxonomic
371 group (e.g. fish, birds) for data collection and/or availability reasons rather than for ecological
372 reasons (availability of traits or phylogeny, consistent sampling methodology). Similarly, our choice

373 to focus on polychaetes was primarily guided by data availability (trait were available from Boyé *et al.* (2019)), although this is a group characterised by high diversity in terms of lifestyle and functional
374 role (Jumars *et al.* 2015 ; Giangrande 1997). Hence, most studies typically disregard the role of
375 accompanying species, which may bias model estimation and predictive power according to our
376 results. Further studies, considering a gradual increase in the number of accompanying species,
377 while also potentially accounting for their trophic role or position, would help better understand the
378 effect of non-focal species on jSDM's performance. While communities and assemblages remain
379 largely defined based on arbitrary choices (Stroud *et al.* 2015), such sensitivity analysis could lead
380 to a more objective delineation of the appropriate ecological units to be studied (i.e. which species
381 are necessary to improve model performance? Which group can be studied in isolation?).

383 jSDMs have already been used to model the distribution of a wide variety of species ranging
384 from micro-organisms (Minard *et al.* 2019 ; Pichler & Hartig 2021) to megafauna (Rocha *et al.* 2017 ;
385 Brimacombe *et al.* 2020) inhabiting many different ecosystems. Here, while we studied communities
386 associated with two typical coastal habitats, i.e. seagrass and sand, that have original characteristics
387 as they are located at the land-sea interface (Boyé *et al.* 2019), our case study reflects typical
388 aspects of data in ecological research such as data limitation and data availability but also typical
389 aspects of the functioning and structure of communities (e.g. prevalence of rare and transient
390 species; Magurran & Henderson (n.d.) ; Snell Taylor *et al.* (2018)). Our results provide a few insights
391 on trait-environment relationships but we suspect that trait data quality and availability remains
392 limiting to fully exploit what functional ecology can bring to jSDMs (Tyler *et al.* 2012 ; Juan *et al.*
393 2022). For instance, we found an interaction between trophic modalities (i.e. microphagous versus
394 macrophagous diet) and fetch (Fig. S15), which indicates that organisms that filter on small particles
395 are less likely to occur in wave-exposed sites where high levels of sediment resuspension can block
396 their filtering systems (LM *et al.* 2014); conversely macrophagous organisms are less impacted by
397 fetch. Yet, most trait-environment relationships, and most species-environment relationships were
398 flat; a result that could reflect a symptom of the functioning of the system, where neutral processes
399 are expected to dominate (e.g. in seagrass bed communities, Boyé *et al.* (2019)). However, the lack
400 of contribution of other trait-environment relationships in our model could also be related to a
401 mismatch between trait data, environmental data, and the ecological processes at play. For instance,
402 the physical coastal environment is highly dynamic; a feature that is only partially characterised by
403 our environmental variables that summarise average climatological conditions (but not extreme

404 events or annual/seasonal variability). Likewise, the list of available fuzzy-coded traits only partially
405 captures species capacity to adapt to frequent disturbances or environmental variability (Violle *et al.*
406 2012 ; Juan *et al.* 2022). Most studies will face the same trade-off between the potential benefit of
407 including traits within jSDMs and the effort needed to collect relevant information. In our case, while
408 including traits did not improve the model's predictive power, it provides knowledge that is useful
409 to understand the response of species along environmental gradients, and set-up management
410 strategies. Hence, if the goal is not prediction but inference (Tredennick *et al.* 2021), including traits
411 and proxies of phylogeny might still be useful to interpret the models, provided that explanatory
412 power is not affected (as in our case), and that computation time is not too cumbersome given the
413 added parameters.

414 Understanding the behaviour of increasingly complex models from a mathematical and eco-
415 logical point of view is one of the major challenges currently facing numerical ecology and one
416 active research area (Lucas 2020 ; Ryo *et al.* 2021). jSDMs can be particularly difficult to interpret
417 because the number of parameters drastically increases with the size of the community (the curse
418 of dimensionality; Tikhonov *et al.* (2017)). Combined with the use of latent variables, auditing this
419 type of model is very complex, and requires a diverse set of metrics (Wilkinson *et al.* 2020), notably
420 to understand what is hidden in the signal captured by the residual correlations between species.
421 While guidelines have been developed to characterise the performance of jSDM fitted on occurrence
422 data (Wilkinson *et al.* 2020), it is only recently that the predictive power of abundance-based models
423 has been explored Waldock *et al.* (2022). Here, we used a set of complementary metrics to assess
424 the performance of both occurrence- and abundance-based models at the species and community
425 levels, the latter considering both alpha and beta diversity. We also transposed a method initially
426 developed for time series (Rigal *et al.* 2020) to provide an innovative way of characterising the
427 response curves of each species. Further, we bring together a set of approaches and propose
428 a new index to characterise and compare residual correlations networks. Overall, we provide a
429 comprehensive framework ultimately allowing for an integrative assessment and comparison of
430 jSDM performance.

431 As our models performed equally well in terms of explanatory power, it is important to ask
432 whether and to which extent adding extra information is beneficial to jSDM models. Adding
433 accompanying species has two main benefits : increasing predictive power and clarifying species-
434 environment relationships. For the former, while the predictive power was improved by taking

the whole community into account, the gain remained small. Hence, if the objective is prediction, machine learning or neural networks applied to community data may provide an interesting alternative to jSDM (Zhang *et al.* 2020 ; Deneu *et al.* 2021). Still, jSDMs provide an interesting compromise between predictive power and interpretability (Norberg *et al.* 2019 ; Ovaskainen *et al.* 2017b) and adding accompanying species could prove a useful strategy for informing conservation through community models (Pollock *et al.* 2020). For the latter, adding accompanying species helped clarify and remove spurious species-environment relationships (many switching from non-linear responses to flat ones). Nonetheless, it is important to keep in mind that the more species you add, the more weight they have in the likelihood and therefore on the estimated species-environment relationships. Thus if the purpose is inference, adding more species may not be the best strategy. Adding trait information or phylogeny seems more appropriate as it leads to more interpretable models. 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) and the data at hand. Future work could expand on other ecosystems presenting different characteristics (e.g., with stronger environmental filters or competitive processes) or use simulations (Zurell *et al.* 2010) to confirm the generality of our findings.

Author Contributions

All authors conceived these ideas and Violet, Boyé, Chevalier and Marzloff designed the methodology; Boyé, Grall and Gauthier provided the data; Violet analysed the data; Violet, Boyé, Chevalier and Marzloff led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Acknowledgment

The authors acknowledge the Pôle de Calcul et de Données Marines (PCDM) for providing DATARMOR storage and computational resources. <https://pcdm.ifremer.fr>

Conflict of interest

The authors declare no conflict of interest.

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](#). *Oikos*, 126.
- Baselga, A. (2010). [Partitioning the turnover and nestedness components of beta diversity](#). *Global Ecology and Biogeography*, 19, 134–143.
- Baselga, A., Orme, D., Villeger, S., De Bortoli, J., Leprieux, F. & Logez, M. (2022). [betapart: Partitioning Beta Diversity into Turnover and Nestedness Components](#).
- Blanchet, F.G., Cazelles, K. & Gravel, D. (2020). [Co-occurrence is not evidence of ecological interactions](#). *Ecology Letters*.
- Boyé, A. (2019). Diversité taxinomique et fonctionnelle des habitats benthiques dans l'espace et dans le temps: une perspective régionale et décennale. thèse de doctorat. Université de Bretagne Occidentale, Université de Montréal.
- Boyé, A., Legendre, P., Grall, J. & Gauthier, O. (2017). [Constancy despite variability: Local and regional macrofaunal diversity in intertidal seagrass beds](#). *Journal of Sea Research*, 130, 107–122.
- Boyé, A., Thiébaud, 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](#). *Diversity and Distributions*, 25, 1879–1896.
- Brimacombe, C., Bodner, K. & Fortin, M.-J. (2020). [Inferred seasonal interaction rewiring of a freshwater stream fish network](#). *Ecography*, n/a.
- Brudvig, L.A. & Catano, C.P. (2022). [Prediction and uncertainty in restoration science](#). *Restoration Ecology*, n/a, e13380.
- Chesson, P. (2000). [Mechanisms of Maintenance of Species Diversity](#). *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](#). *Frontiers in Ecology and Evolution*, 9.
- Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F. & Joly, A. (2021). Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS Computational Biology*.
- Dietze, M.C., Fox, A., Beck-Johnson, L.M., Betancourt, J.L., Hooten, M.B., Jarnevich, C.S., et al. (2018).

491 [Iterative near-term ecological forecasting: Needs, opportunities, and challenges](#). *Proceedings of*
492 *the National Academy of Sciences*, 115, 1424–1432.

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

496 Dunn, O.J. (1964). [Multiple Comparisons Using Rank Sums](#). *Technometrics*, 6, 241–252.

497 Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., *et al.* (2006). [Novel methods](#)
498 [improve prediction of species' distributions from occurrence data](#). *Ecography*, 29, 129–151.

499 Gelman, A., Hill, J. & Vehtari, A. (2020). [Regression and Other Stories](#). Analytical Methods for Social
500 Research. Cambridge University Press.

501 Gelman, A. & Rubin, D.B. (1992). [Inference from Iterative Simulation Using Multiple Sequences](#).
502 *Statistical Science*, 7, 457–472.

503 Giangrande, A. (1997). [Polychaete reproductive patterns, life cycles and life histories: an overview](#).
504 In: *Oceanography And Marine Biology* (ed. A. D. Ansell, M.B., R. N. Gibson). CRC Press, pp. 310–411.

505 Giangrande, A., Licciano, M. & Musco, L. (2005). [Polychaetes as environmental indicators revisited](#).
506 *Marine Pollution Bulletin*, 50, 1153–1162.

507 Giannini, T.C., Chapman, D.S., Saraiva, A.M., Alves-dos-Santos, I. & Biesmeijer, J.C. (2013). [Improving](#)
508 [species distribution models using biotic interactions: a case study of parasites, pollinators and](#)
509 [plants](#). *Ecography*, 36, 649–656.

510 Godsoe, W., Franklin, J. & Blanchet, F.G. (2017). [Effects of biotic interactions on modeled species'](#)
511 [distribution can be masked by environmental gradients](#). *Ecology and Evolution*, 7, 654–664.

512 Häkkilä, M., Abrego, N., Ovaskainen, O. & Mönkkönen, M. (2018). [Habitat quality is more important](#)
513 [than matrix quality for bird communities in protected areas](#). *Ecology and Evolution*, 8, 4019–4030.

514 Holt, R.D. (2020). [Some thoughts about the challenge of inferring ecological interactions from spatial](#)
515 [data](#). *Biodiversity Informatics*, 15, 61–66.

516 Houlahan, J.E., McKinney, S.T., Anderson, T.M. & McGill, B.J. (2017). [The priority of prediction in](#)
517 [ecological understanding](#). *Oikos*, 126, 1–7.

518 Howard, C., Stephens, P.A., Pearce-Higgins, J.W., Gregory, R.D. & Willis, S.G. (2014). [Improving](#)
519 [species distribution models: the value of data on abundance](#). *Methods in Ecology and Evolution*,
520 5, 506–513.

521 Hui, F.K.C. (2016). [boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance](#)

522 [Data in r](#). *Methods in Ecology and Evolution*, 7, 744–750.

523 IPBES. (2019). [Global assessment report on biodiversity and ecosystem services of the Intergovern-](#)
524 [mental Science-Policy Platform on Biodiversity and Ecosystem Services](#).

525 Ives, A.R. & Helmus, M.R. (2011). [Generalized linear mixed models for phylogenetic analyses of](#)
526 [community structure](#). *Ecological Monographs*, 81, 511–525.

527 Juan, S. de, Bremner, J., Hewitt, J., Törnroos, A., Mangano, M.C., Thrush, S., *et al.* (2022). [Biological](#)
528 [traits approaches in benthic marine ecology: Dead ends and new paths](#). *Ecology and Evolution*,
529 12, e9001.

530 Jumars, P.A., Dorgan, K.M. & Lindsay, S.M. (2015). [Diet of Worms Emended: An Update of Polychaete](#)
531 [Feeding Guilds](#). *Annual Review of Marine Science*, 7, 497–520.

532 Keil, P., Wiegand, T., Tóth, A.B., McGlinn, D.J. & Chase, J.M. (2021). [Measurement and analysis of](#)
533 [interspecific spatial associations as a facet of biodiversity](#). *Ecological Monographs*, n/a.

534 König, C., Wüest, R.O., Graham, C.H., Karger, D.N., Sattler, T., Zimmermann, N.E., *et al.* (2021). [Scale](#)
535 [dependency of joint species distribution models challenges interpretation of biotic interactions](#).
536 *Journal of Biogeography*, 48, 1541–1551.

537 Levine, J.M., Bascompte, J., Adler, P.B. & Allesina, S. (2017). [Beyond pairwise mechanisms of species](#)
538 [coexistence in complex communities](#). *Nature*, 546, 56–64.

539 LM, M., CH, P. & MJ, B. (2014). [Dominant macrobenthic populations experience sustained impacts](#)
540 [from annual disposal of fine sediments on sandy beaches](#). *Marine Ecology Progress Series*, 508,
541 1–15.

542 Lucas, T.C.D. (2020). [A translucent box: interpretable machine learning in ecology](#). *Ecological*
543 *Monographs*, 90, e01422.

544 Magurran, A.E. & Henderson, P.A. (n.d.). [Explaining the excess of rare species in natural species](#)
545 [abundance distributions](#). *Nature*, 422, 714–716.

546 Minard, G., Tikhonov, G., Ovaskainen, O. & Saastamoinen, M. (2019). [The microbiome of the Melitaea](#)
547 [cinxia butterfly shows marked variation but is only little explained by the traits of the butterfly](#)
548 [or its host plant](#). *Environmental Microbiology*, 21, 4253–4269.

549 Momal, R., Robin, S. & Ambroise, C. (2020). [Tree-based inference of species interaction networks](#)
550 [from abundance data](#). *Methods in Ecology and Evolution*, 11, 621–632.

551 Momal, R., Robin, S. & Ambroise, C. (2021). [Accounting for missing actors in interaction network](#)
552 [inference from abundance data](#). *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](#). *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](#). *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](#). *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](#). *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](#). *Ecology Letters*, 20, 561–576.

Paradis, E. & Schliep, K. (2019). [ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R](#). *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](#). *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](#). *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](#). *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](#). *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](#)

584 (JSDM). *Methods in Ecology and Evolution*, 5, 397–406.

585 Popovic, G.C., Hui, F.K.C. & Warton, D.I. (2022). [Fast model-based ordination with copulas](#). *Methods*
586 *in Ecology and Evolution*, 13, 194–202.

587 Potter, K.A., Arthur Woods, H. & Pincebourde, S. (2013). [Microclimatic challenges in global change](#)
588 **biology**. *Global Change Biology*, 19, 2932–2939.

589 Ricotta, C., Bacaro, G., Marignani, M., Godefroid, S. & Mazzoleni, S. (2012). [Computing diversity from](#)
590 [dated phylogenies and taxonomic hierarchies: does it make a difference to the conclusions?](#)
591 *Oecologia*, 170, 501–506.

592 Rigal, S., Devictor, V. & Dakos, V. (2020). [A method for classifying and comparing non-linear trajecto-](#)
593 [ries of ecological variables](#). *Ecological Indicators*, 112, 106113.

594 Rocha, R., Ovaskainen, O., López-Baucells, A., Farneda, F.Z., Ferreira, D.F., Bobrowiec, P.E.D., *et al.*
595 (2017). [Design matters: An evaluation of the impact of small man-made forest clearings on](#)
596 [tropical bats using a before-after-control-impact design](#). *Forest Ecology and Management*, 401,
597 8–16.

598 Ryo, M., Angelov, B., Mammola, S., Kass, J.M., Benito, B.M. & Hartig, F. (2021). [Explainable artificial](#)
599 [intelligence enhances the ecological interpretability of black-box species distribution models](#).
600 *Ecography*, 44, 199–205.

601 Sander, E.L., Wootton, J.T. & Allesina, S. (2017). [Ecological Network Inference From Long-Term](#)
602 [Presence-Absence Data](#). *Scientific Reports*, 7.

603 Snell Taylor, S.J., Evans, B.S., White, E.P. & Hurlbert, A.H. (2018). [The prevalence and impact of](#)
604 [transient species in ecological communities](#). *Ecology*, 99, 1825–1835.

605 Staniczenko, P.P.A., Sivasubramaniam, P., Suttle, K.B. & Pearson, R.G. (2017). [Linking macroecology](#)
606 [and community ecology: refining predictions of species distributions using biotic interaction](#)
607 [networks](#). *Ecology Letters*, 20, 693–707.

608 Stroud, J.T., Bush, M.R., Ladd, M.C., Nowicki, R.J., Shantz, A.A. & Sweatman, J. (2015). [Is a community](#)
609 [still a community? Reviewing definitions of key terms in community ecology](#). *Ecology and*
610 *Evolution*, 5, 4757–4765.

611 Thioulouse, J., Dray, S., Dufour, A., Siberchicot, A., Jombart, T. & Pavoine, S. (2018). [Multivariate](#)
612 [Analysis of Ecological Data with ade4](#). Springer.

613 Tikhonov, G., Abrego, N., Dunson, D. & Ovaskainen, O. (2017). [Using joint species distribution](#)
614 [models for evaluating how species-to-species associations depend on the environmental context](#).

615 *Methods in Ecology and Evolution*, 8, 443–452.

616 Tikhonov, G., Opedal, O., Abrego, N., Lehtikainen, A. & Ovaskainen, O. (2019). [Joint species distribution](#)
617 [modelling with HMSC-R](#). *bioRxiv*.

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

620 Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. (2017). [Taxonomic bias in](#)
621 [biodiversity data and societal preferences](#). *Scientific Reports*, 7.

622 Tyler, E.H.M., Somerfield, P.J., Berghe, E.V., Bremner, J., Jackson, E., Langmead, O., *et al.* (2012).
623 [Extensive gaps and biases in our knowledge of a well-known fauna: implications for integrating](#)
624 [biological traits into macroecology](#). *Global Ecology and Biogeography*, 21, 922–934.

625 Vesk, P.A., Morris, W.K., Neal, W.C., Mokany, K. & Pollock, L.J. (2021). [Transferability of trait-based](#)
626 [species distribution models](#). *Ecography*, 44, 134–147.

627 Violle, C., Enquist, B.J., McGill, B.J., Jiang, L., Albert, C.H., Hulshof, C., *et al.* (2012). [The return of](#)
628 [the variance: intraspecific variability in community ecology](#). *Trends in Ecology & Evolution*, 27,
629 244–252.

630 Waldo, C., Stuart-Smith, R.D., Albouy, C., Cheung, W.W.L., Edgar, G.J., Mouillot, D., *et al.* (2022). [A](#)
631 [quantitative review of abundance-based species distribution models](#). *Ecography*, 2022.

632 Whittaker, R.J., Willis, K.J. & Field, R. (2001). [Scale and species richness: towards a general, hierarchical](#)
633 [theory of species diversity](#). *Journal of Biogeography*, 28, 453–470.

634 Wiens, J.J., Ackerly, D.D., Allen, A.P., Anacker, B.L., Buckley, L.B., Cornell, H.V., *et al.* (2010). Niche
635 conservatism as an emerging principle in ecology and conservation biology. *Ecology letters*, 13,
636 1310–1324.

637 Wilkinson, D.P., Golding, N., Guillerá-Arroita, G., Tingley, R. & McCarthy, M.A. (2019). [A comparison](#)
638 [of joint species distribution models for presence–absence data](#). *Methods in Ecology and Evolution*,
639 10, 198–211.

640 Wilkinson, D.P., Golding, N., Guillerá-Arroita, G., Tingley, R. & McCarthy, M.A. (2020). [Defining and](#)
641 [evaluating predictions of joint species distribution models](#). *Methods in Ecology and Evolution*, n/a.

642 Zhang, C., Chen, Y., Xu, B., Xue, Y. & Ren, Y. (2020). [Improving prediction of rare species' distribution](#)
643 [from community data](#). *Scientific Reports*, 10, 12230.

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

⁶⁴⁶ Zurell, D., Pollock, L.J. & Thuiller, W. (2018). [Do joint species distribution models reliably detect](#)
⁶⁴⁷ [interspecific interactions from co-occurrence data in homogenous environments?](#) *Ecography*, 41,
⁶⁴⁸ 1812–1819.

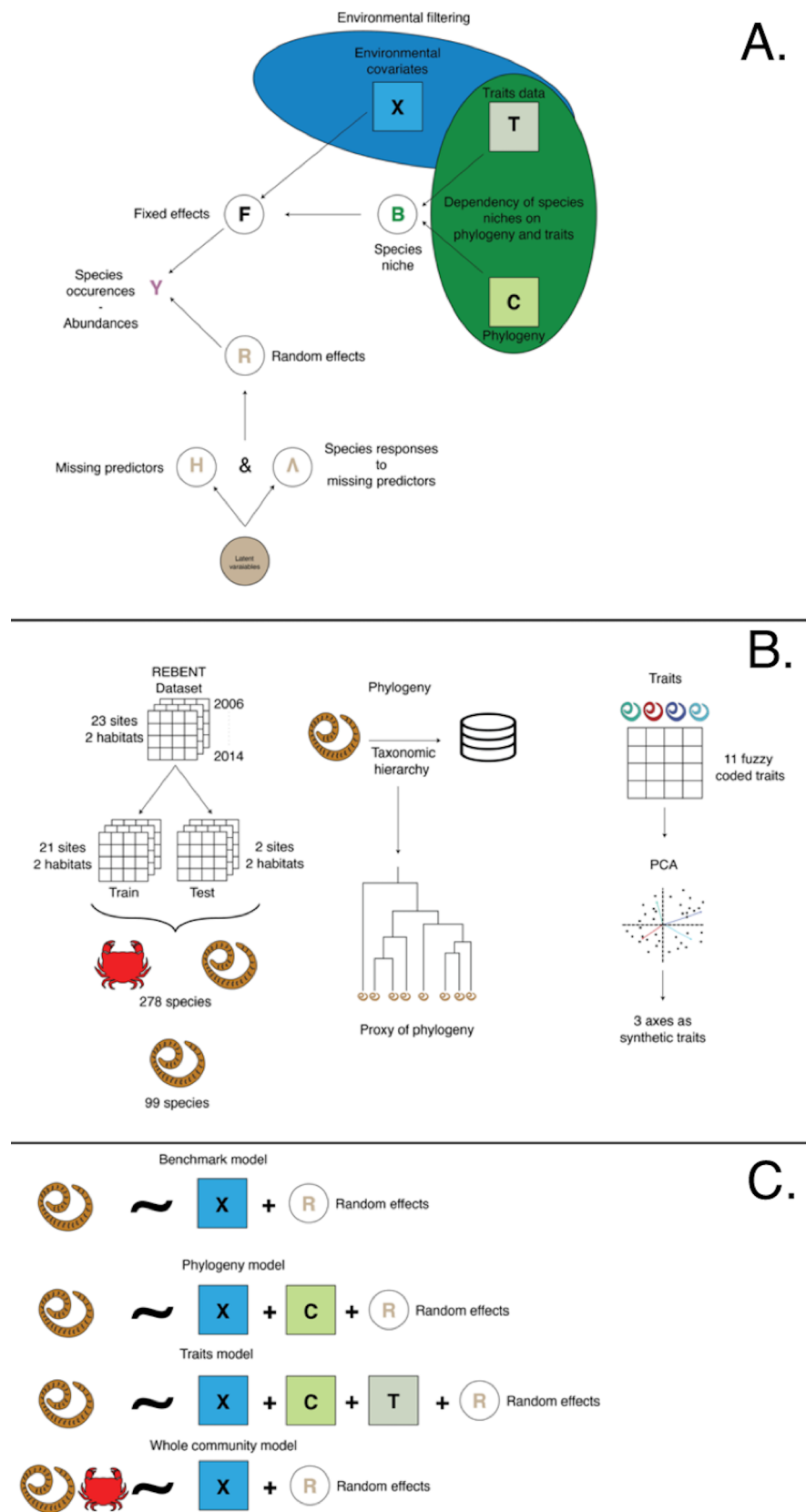


Figure 1. Workflow of the study. A. Hierarchical design of an HMSC model set-up with environmental variables, phylogeny and traits. B. Data acquisition workflow. C. Configuration of the different models.

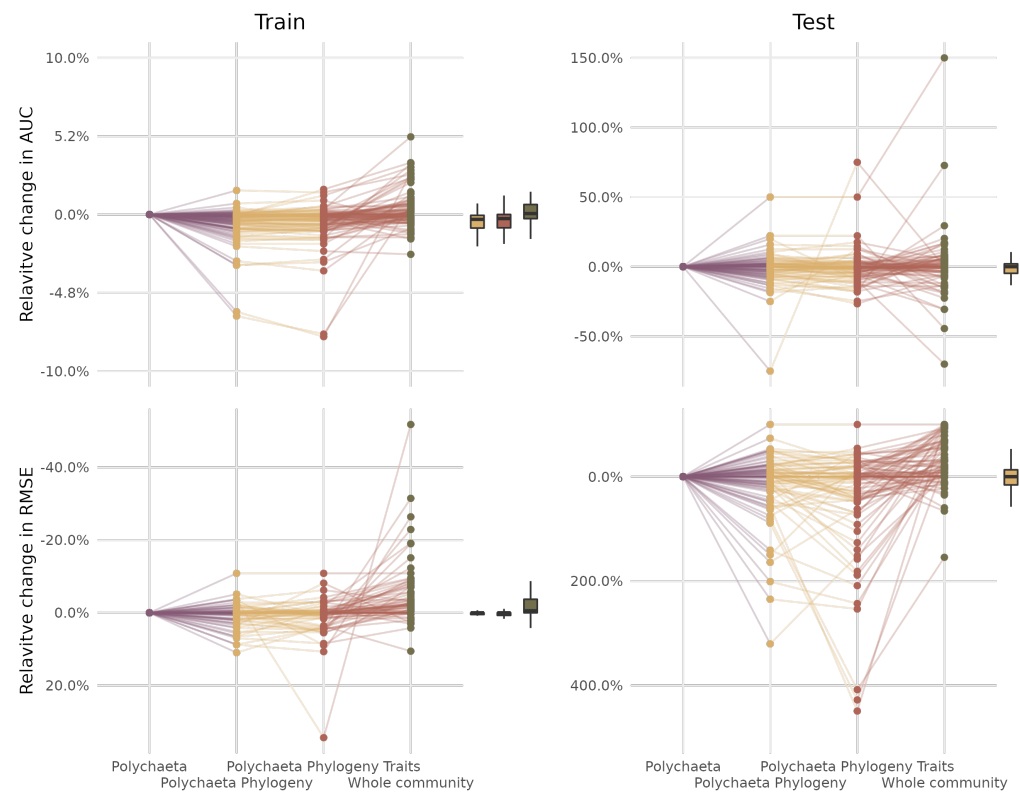


Figure 2. Relative change in explanatory (left column) and predictive (right column) power of different model architectures with respect to the benchmark fitted with occurrence (top line) or abundance (bottom line) data.

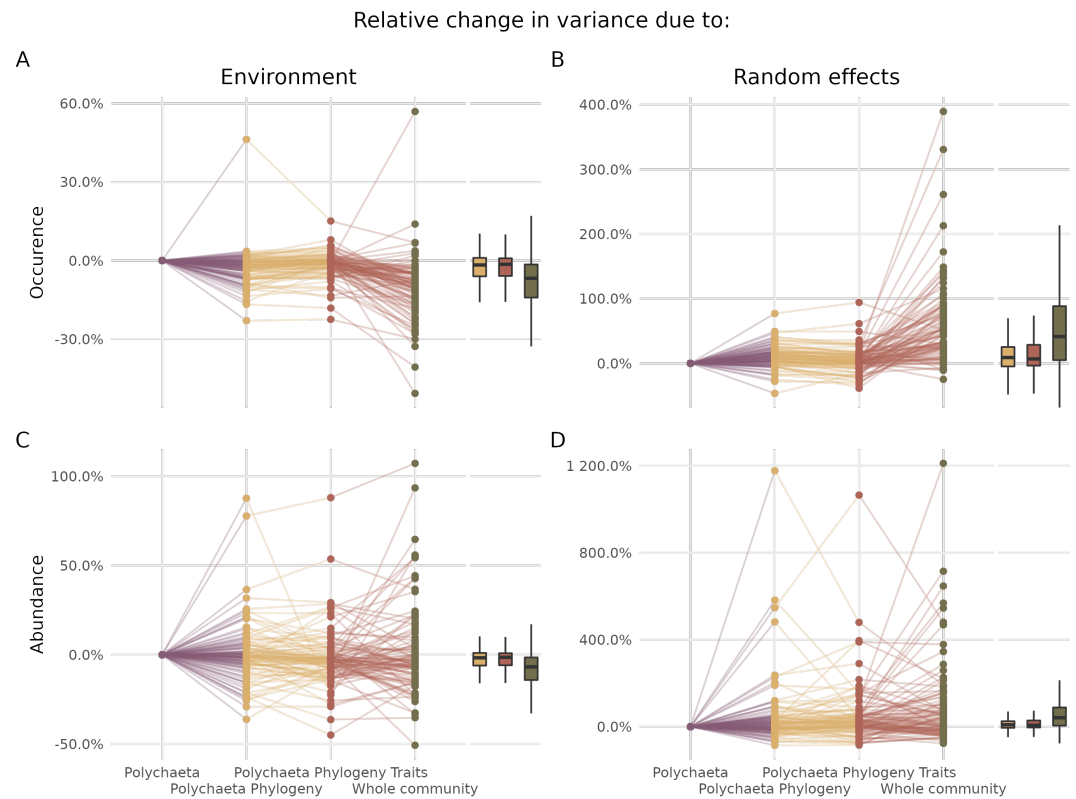


Figure 3. Relative change in variance explained by environmental predictors (left column) and by random effects (right column) power of different model architectures with respect to the benchmark fitted with occurrence (top line) or abundance (bottom line) data.



Figure 4. Proportion of response curves according to the nomenclature defined by Rigal *et al.* (2020) for different model architectures. All models have been fitted with abundance data. Each response is characterised by a shape (column) and an intensity (line).

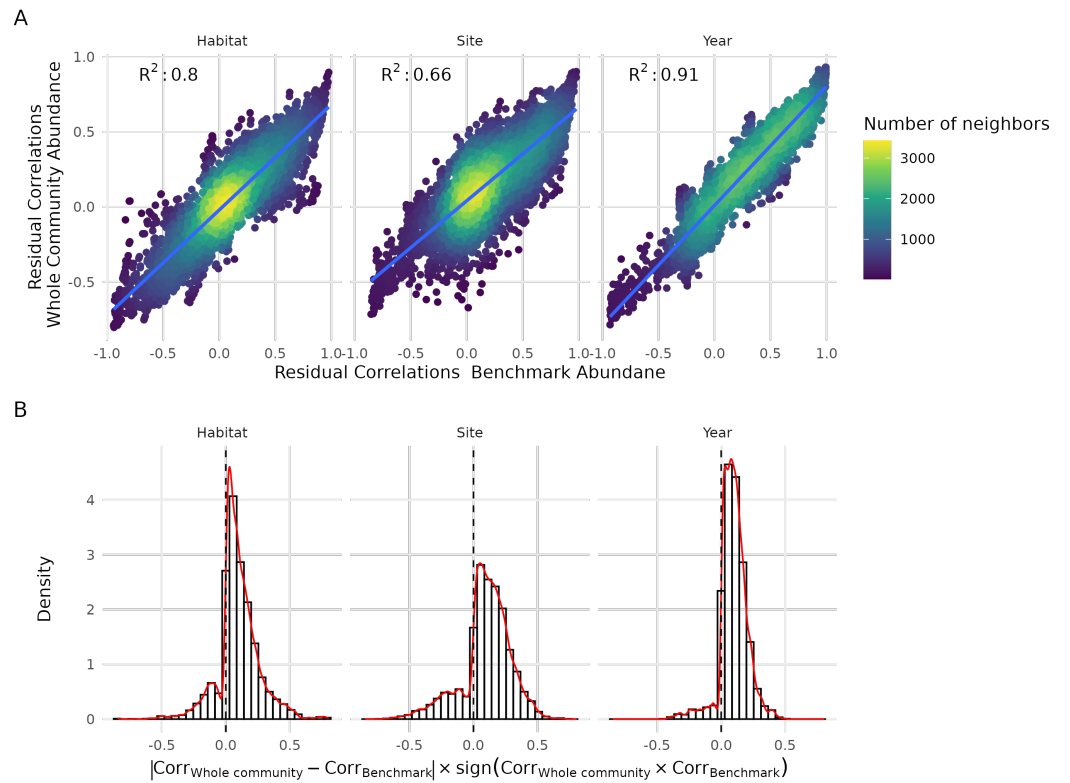


Figure 5. Comparison of residual correlations estimated by Whole Community Model and Benchmark model for the three random effect, the two models were fitted with abundance data. A. Scatter plot of the residual correlations estimated by the whole community model as a function of the residual correlations estimated by the benchmark model. B. Distribution of the index measuring the change of sign and magnitude between residual correlations estimated by the Whole community model and the Benchmark model adjusted with abundance data.