Should we exploit opportunistic databases with joint species distribution models? Artificial and real data suggests it depends on the sampling completeness.

Authors

Daniel Romera-Romera, Diego Nieto-Lugilde

Abstract

Anticipating the effects of global change on biodiversity has become a global challenge requiring new methods. Approaches like species distribution models have limitations which have fueled the development of joint species distribution models (JSDMs). However, JSDMs rely on systematic surveys community data, and no assessment has been made of their suitability with unstructured opportunistic databases data.

We used Hierarchical Modeling of Species Communities (HMSC) to test JSDMs performance when using opportunistic databases. Using artificial data that mimic the limitations of such databases by subsampling complete cooccurrence matrices (i.e., original data), we analysed how the completeness of opportunistic databases affects JSDMs regarding (a) the role of independent variables on species occurrence, (b) residual species cooccurrence (as a proxy of biotic interactions), and (c) species distributions. Moreover, we illustrate how to evaluate completeness at the pixel level of real data with a study case of forest tree species in Europe, and evaluate the role of data completeness in model estimation.

Our results with artificial data demonstrate that decreasing the completion percentage (the rate of original data presences represented in the subsampled matrices) increases false negatives and negative cooccurrence probabilities, resulting in a loss of ecological information. However, HMSC tolerates different levels of degradation depending on the model aspect being considered. Models with 50% of missing data are valid for estimating species niches and

distribution, but interaction matrices require databases with at least 75% of completion data. Furthermore, HMSC's predictions often resemble the original community data (without false negatives) even more than the subsampled data (with false negatives) in the training dataset. These findings were confirmed with the real study case.

We conclude that opportunistic databases are a valuable resource for JSDMs, but require an analysis of data completeness for the target taxa in the study area at the spatial resolution of interest.

Key-words

Artificial data, Community ecology, joint species distribution models, Opportunistic databases, Interacting species, Pinus, Quercus.

Introduction

Anticipating and mitigating the effects of global change on biodiversity has become a global challenge that requires, among others, new methods and procedures to advance our knowledge and understanding of biological processes (Malhi et al., 2020). During the last decades, species distribution models (SDMs) have been one of the most important and popular methodological developments in ecology and biogeography, among other reasons because they are suitable tools to mine opportunistic biodiversity databases and repositories (i.e., databases set with observational data from unstructured and unconnected surveys), especially in assessments of climate change effects (Soroye et al., 2018). However, important limitations in SDMs have fueled the development of joint species distribution models (JSDMs; also known as community level models, multispecies models, or multivariate models; Nieto-Lugilde et al., 2018). However, JSDMs require community data (e.g. sites by species community matrices), more typically coming from systematic surveys (where both presences and absences of the target species are recorded) potentially hindering their use with observational data from opportunistic databases

(where only presences are recorded). Recent studies have discussed the use of opportunistic and presence-only data in JSDM regarding the sampling effort (Escamilla Molgora et al., 2022) and imperfect detection (Hogg et al., 2021). Nonetheless, to our knowledge no assessment of the suitability of JSDMs to work with opportunistic databases and unstructured surveys has been performed. Given the methodological improvements of JSDMs and the unmatched amount of information in opportunistic databases, it seems relevant to evaluate the suitability of JSDMs to mine such resources, which could improve our prediction of biodiversity responses to global changes.

The development of SDMs has been paired with the development of databases and data repositories (Feng et al., 2022), both for biodiversity (e.g., Global Biodiversity Information Facility –GBIF) and for environmental variables (e.g., climate - Worldclim or Chelsa; Fick & Hijmans, 2017; Karger et al., 2016), as well as an increase in computational capacity (Cook et al., 2019; Lin, 2020). SDMs have already been used, tested (Newbold et al., 2010), and validated for decades. During all this time, however, several limitations have been identified that hinder the usefulness of their projections and have boosted the development of new approaches that overcome such limitations (Doser et al., 2022; Wilkinson et al., 2019). The main limitations of SDMs rely on the incapacities/difficulties to include alternative but still important drivers of species distribution, like biotic interactions (Dutra Silva et al., 2019), spatial autocorrelation (Guélat & Kéry, 2018), dispersal abilities (Zurell et al., 2009), species functional traits (Wittmann et al., 2016), or phylogenetic relatedness among species (Morales-Castilla et al., 2017).

JSDMs emerged as an alternative to SDMs that analyze and model a community matrix with multiple species, which allows them to share information among species which in turn leads to more robust parameters estimates (Maguire et al., 2016; Nieto-Lugilde et al., 2018). Further developments in JSDMs allow the estimation of a residual cooccurrence matrix in which coefficients represent the cooccurrence pattern among the species once the effect of

environmental variables (i.e., independent variables) is removed. This matrix has been interpreted as the correlations caused by biotic interactions among the species, which led to the general claim that JSDMs can include biotic interactions (Pollock et al., 2014) in model fitting and, hence, also in their projections. Although it is known that other factors like i) missing important environmental variables, ii) species prevalence, and iii) spatial autocorrelation or unaccounted dispersal limitations can also affect estimates of the residual cooccurrence matrix (Zurell et al., 2018), several studies have used them to correctly capture the effect of biotic interactions (Marjakangas et al., 2021; Ovaskainen et al., 2016; Pollock et al., 2014), or postulate new hypotheses about interactions (Ovaskainen et al., 2010). The internal conformation of JSDMs, whether because they identify biotic interactions or shared response to environmental gradients, represents an advantage because it allows the use of information provided by the most abundant species, to improve parameters and predictions for rare species (Tikhonov et al., 2020), which are usually the most difficult to study with SDMs. Additionally, because JSDMs analyze whole communities, they can include more naturally some innovations which overcome notable weaknesses of SDMs, like phylogenetic relatedness among analyzed species and their functional traits (Kaldhusdal et al., 2015; Pollock et al., 2012; Tikhonov et al., 2020). This progress allows studying the extent to which the environmental filtering is structured by the species traits or their phylogenetic signal. However, due to their community data requirement and their recent development, these models have not been as used and exposed to different validations as SDMs (Nieto-Lugilde et al., 2018; Viana et al., 2022; Zhang et al., 2020a, 2020b).

By giving access to information on species occurrences at a global scale, opportunistic databases (e.g., GBIF, 2023) are a great source for modelling species distributions. However, data from independent observations could be limited and not well suited for use in topics that exceed the species level (e.g., community level) as different species may have been recorded in different sites although they usually cooccur (e.g., depending on taxonomic interest and the usual study area of data collectors, or on other aspects such as species' detectability). Despite

100 this, it has been suggested that multispecies occurrence data from opportunistic databases
101 could be aggregated at the community level (Nieto-Lugilde et al., 2018; Ovaskainen & Abrego,
102 2020). Despite the possibility of taking advantage of the massive amount of data, the resulting
103 community matrix could be an unrealistic representation of the real data. This is a key issue
104 because, potentially at least, the unbalanced representation of presence data in opportunistic
105 databases could cause species cooccurrence patterns in the derived matrices to not be
106 equivalent to actual patterns (Figure 1) by increasing false negatives in the resulting community
107 matrix. This artefact might potentially have several and incremental undesired effects on model
108 calibration and performance. Although these effects have been studied individually for SDMs
109 (Fernandes et al., 2019), the effects remain unclear for JSDMs, thus, the advantages of JSDMs
110 could be missed when using these data sources.

111 The main goal of this work is to test the performance of JSDMs when using opportunistic
112 databases. Specifically, we ask how the completeness of opportunistic databases, due to
113 imperfect sampling, affects JSDMs in terms of estimating (1) the role of independent variables
114 on species occurrence, (2) species cooccurrence residuals (as a proxy of biotic interactions), and
115 (3) species distributions.

116 We hypothesize that JSDMs might be robust to a certain degree of degradation in the
117 cooccurrence matrices, while providing relatively good parameters estimation and
118 performance. Thus, this study would serve as a methodological base for new community studies
119 and uncover the potential of JSDMs to be used with massive amounts of data from opportunistic
120 databases, providing new venues to anticipate the effects of climate change.

121 Material and methods

122 Since no opportunistic database can provide an exact representation of the real world, virtual
123 species (i.e., artificially generated data) are a great option to test our hypothesis, because they
124 provide a controlled environment in which the generated data are free of biases that usually

accompany information from opportunistic databases (Zurell et al., 2010). We used virtual species to simulate a whole landscape with 10 cooccurring species. Then, we simulated the representation of this landscape in opportunistic databases by randomly subsampling pixels at which each species was present with different retention percentages and, then, combining the species-specific selected pixels into single community matrices (one for retention percentage). Further, we calibrated JSDMs for each of these datasets and studied their parameters, performance, spatial prediction, and what factors could affect them (e.g., retention percentage and artificial false negatives). Furthermore, we completed the analysis by studying a real case of forest tree species in Europe (i.e., *Pinus* and *Quercus* species). To do so, we downloaded their whole data for Europe from GBIF, analyzed the completeness of the dataset for the study area, calibrated JSDMs with different data subsets (defined by different levels of completeness), and compared their results.

Artificial data

To achieve the proposed objectives, we generated two original datasets, CM_{O1} with species responding to two independent environmental variables (x1 and x2), and CM_{O2} considering the species responses to the same two environmental variables plus known interactions between the species. Models fit with CM_{O1} were used to evaluate model calibration and predictive performance. Models fit with CM_{O2} were used to study the residual cooccurrence matrix among species estimated by the model when the effect of the environmental variables is removed, and how it varied with the retention percentages.

To generate the original datasets (CM_{O1} and CM_{O2}), we defined a squared geographical space of 50 by 50 grid cells. Then, to make the data as simple as possible, we generated two independent, uniformly distributed variables that varied along the geographical space: one along the longitudinal axis (x1) and another one along the latitudinal axis (x2). These two variables were scaled, ensuring that they had a mean of zero and a standard deviation of one.

Next, we used the "communitySimul" function from the HMSC package (https://github.com/guiblanchet/HMSC, note that this is different from the Hmsc package on CRAN) for R (R Core Team, 2022) to create our biological dataset. To simulate the most typical sort of data in real opportunistic databases (i.e., occurrence rather than abundances), we defined species to respond linearly to each of the two environmental variables (species coefficients were assigned randomly) but using a probit link family to generate a dataset with species following a Bernoulli distribution (i.e., presences/absences). To generate CM_{O2}, we added another argument that allows to define interactions between species (i.e., omega parameters; Ω), in the form of an identity matrix with non-diagonal zeros randomly replaced by values drawn from a Wishart distribution with 22 degrees of freedom.

When simulating opportunistic databases, we subsampled the original datasets (i.e., $CM_{O1\ \&\ 2}$) by randomly and independently sampling presences of each species with predefined retention percentages (i.e., 10, 25, 50, 75, and 90% of the presences; $CM_{S10,\ 25,\ 50,\ 75,\ \&\ 90}$). In a specific realization of the subsampling, all species were applied the same percentage (e.g., 10% to all species), generating 5 independent realizations (i.e., one for each retention percentage). Then, the non-selected pixels (true absences plus non sampled presences) per species were transformed into absences, eliminating the localities in which no species was subsampled. This process and its randomizations were repeated 10 times to quantify uncertainty due to stochasticity. Finally, we resampled the resulting communities to get the same sample size (i.e., 500 sites on each of the 50 datasets (5 retention percentages per 10 random iterations).

To model communities, we used the Hierarchical Modelling of Species Communities (HMSC) framework, as implemented in the Hmsc package v 3.0-14 (Tikhonov et al., 2022) for R (R Core Team, 2022). More specifically, we fit one HMSC model for each of the datasets (CM_{O1}, CM_{O2}, and 10 randomizations of $CM_{S10,\ 25,\ 50,\ 75,\ \&\ 90}$). All the models were calibrated with the two environmental variables that determined the occurrence of the artificial species (x1 and x2).

Although we did not define our artificial dataset with a spatial dependency and to simulate the situation in which the modeller does not know the true factors driving species distribution and community structure, we included in the model a spatial random effect given by the spatial coordinates of each of the occurrences, which were considered independent (i.e., no random effect nested on classes or categories). Model parameters were estimated by four Monte Carlo Markov Chains of 18000 samples each (from which 8000 samples were for burn in), and a thinning of 100 to remove autocorrelation in the chains (see Supporting information for details of chain convergence).

To quantify the ability of the model for capturing the "environmental" effects on the species, we built 150 reduced major axis regressions (RMA; three regressions per model replicate, one for each of the model parameters – i.e., intercept, $beta_{x1}$ and $beta_{x2}$) relating the beta parameters estimated for each species by the HMSC models fit with the opportunistic database (i.e., $CM_{S10, 25, 50, 75, 90}$) to those random beta used to create our artificial dataset (CM_{O1}). Here, we were interested in the symmetric relationship between the two groups of variables, as opposed to the assumption that the coefficients estimated by the models are dependent on the actual betas utilized to generate the data. Similarly, to evaluate the capacity of the models for capturing species interactions, we compared the omega parameters used to build the artificial data (i.e., CM_{O2}) with the ones estimated by the HMSC models (i.e., residual cooccurrence matrices) of the full community (i.e., CM_{O2}) and the subsampled counterparts (i.e., $CM_{S10, 25, 50, 75, \& 90}$). The omega comparisons were performed graphically and by means of Procrustes analysis (Peres-Neto & Jackson, 2001).

To validate the model minimizing the possible bias due to spatial autocorrelation in our artificial dataset, we performed a cross validation by splitting input datasets into sixteen spatial blocks, each one belonging to one of five groups. Then, we used the data from four of the groups to calibrate a model and calculate explanatory power (EXP), and data from the fifth to calculate

200 the predictive power (PRED), repeating the process for each group, retention percentage, and
201 randomization. To further control for the effect of artificial false negatives in the evaluation, we
202 also generated predictions for a matrix with the same sites as the subsampled one and
203 calculated the evaluation metrics against the occurrence values for those sites without artificial
204 false negatives (CM_{O1}) getting the explanatory power without false negatives (EXPwfn). For each
205 one of these cases, we calculated the following metrics: area under the receiver operating
206 characteristic curve (AUC), coefficient of discrimination (Tjur R2; Tjur, 2009), Kappa, and true
207 skill statistic (TSS; Allouche et al., 2006).

208 ## Real data study case

209 To illustrate how this could fit in real case studies, we fit HMSC models for a set of species from
210 two of the main European forestry genus: *Pinus* and *Quercus*. We downloaded occurrence data
211 from GBIF (GBIF.org, 2023; GBIF Occurrence Download https://doi.org/10.15468/dl.9sb7mz; n
212 = 4743628), and filtered occurrences without high precision coordinates, date of record, or out
213 of Europe, keeping 1323088 records of *Pinus pinaster* Aiton (n = 76955), *Pinus nigra* J. F. Arnold
214 (n = 29783), *Pinus sylvestris* L. (n = 355610), *Pinus halepensis* Mill. (n = 26214), *Pinus pinea* L. (n
215 = 17481), *Pinus mugo* Turra (n = 25983), *Quercus petraea* (Matt.) Liebl. (n = 92543), *Quercus ilex*
216 L. (n = 71594), *Quercus suber* L. (n = 53814), *Quercus robur* L. (n = 489323), *Quercus Pyrenaica*
217 Willd. (n = 8643), *Quercus pubescens* Willd. (n = 68780), and *Quercus faginea* Lam (n = 6365). As
218 for the environmental variables, we downloaded the 19 bioclim variables from the WorldClim
219 project version 2.1 (Fick & Hijmans, 2017) at 5 km of spatial resolution, for a study area ranging
220 from -11 to 47° longitude and from 34 to 72° latitude. Then, we performed a Principal
221 Component Analysis, keeping the three main axes as our climatic variables. Additionally, we
222 downloaded soil data from the SoilGrids project (Poggio et al., 2021), specifically we considered
223 pH and sand quantity from surface to 60 cm of depth (averaging values from the first three
224 strata; 0-5cm, 5-10cm, 10-60cm). The soil data was downloaded at 1 km of spatial resolution

(https://files.isric.org/soilgrids/former/2017-03-10/aggregated/1km/), and then upscaled and reprojected to match spatial resolution and extent of the climate data.

To study the completeness of the database in the study area at the chosen resolution, we calculated completeness of each pixel as the ratio between observed and estimated species richness. The latter was calculated with the Chao2 index given the number of records in each pixel and the number of species reported in that pixel by those records (Chao, 1987; Chao & Colwell, 2017), for all pixels with a minimum of 10 records, as implemented in the bdvis R-package (Barve & Otegui, 2016). Using these completeness values (which, not being calculated using a known reference value, are not completely comparable to the ones used for artificial data) we created three different datasets: all available data (full), records only from pixels over the 75% of completeness (over), and records only from pixels under 75% of completeness (under). Each of these subsets were transformed into community matrices (using pixels as locations) and subsampled to retain 500 locations to ease computing process (each subsample was repeated 10 times). Then, we calibrated HMSC models with the same configuration as for artificial data but including quadratic terms for all independent variables (three climatic PCA axis, and the two soil variables). Finally, we validated the 30 models (three datasets and 10 random subsampling) against the same reference data (i.e., the full community matrix), compared the beta and omega parameters, and compared the spatial predictions among models from different datasets (over, full, and under).

The complete code used in this study is available at https://github.com/<ANONYMISED>/<ANONYMISED>.

Results

Artificial data

Subsampled matrices (CM_S) showed increasing artificial false negative ratios as the retention percentage decreased (Table 1), becoming ten times bigger (from 3.8% to 35.8%) between 90% and 10% of retention percentage. In addition, the variability increased as the percentages went lower, because of the greater stochasticity in the subsample. Overall, this had a clear effect on the raw cooccurrence patterns among species (see Supporting information). In fact, in the complete matrix (CM_{O2}) most of species' pairs had a low but positive cooccurrence probability. However, as the matrix was subsampled with reducing retention percentages, the intensity in the cooccurrence probability decreased as well, gradually turning into negative values. Between 90% and 75% of retention, most cooccurrence values showed the same sign and slightly lower values, but at 50% of retention, most of species' pairs turned already into negative probabilities. Below 50%, all species' pairs showed negative cooccurrences.

All species parameters (i.e., intercept and betas for x1 and x2) were correctly and unbiasedly estimated (slopes close to 1 and intercepts close to 0) by the model with the highest retention percentage (90%; Figure 2). However, the RMA's slopes of the three parameters decreased with the decreasing retention percentages (Figure 2), reaching the lowest values of 0.5 at the 10% of retention level. Interestingly, the RMA's intercepts did not follow the same pattern. Instead, estimations of species intercepts became biased as the retention percentage decreased (models estimated lower intercepts than real ones; Figure 2). However, the estimates of the effect of variable x1 remained unbiased regardless of the retention percentage. Finally, the estimates of the effect of variable x2 were again affected by the subsampling but with increasing values as the retention percentage decreased. It is worth noting, that in all cases, the relationship between species coefficients as estimated by the model and the real ones were linearly related (see Supporting information; $R2 > 0.55$ and $p < 0.01$). Hence, although the actual

values can be biased or miscalculated, the relative effects of variables among species were correctly estimated by the models.

The residuals cooccurrence matrix estimated by the model built with the original community (i.e., CM_{O2}) shows that most interactions signs concur with the omega parameters used to create the community (i.e., 38 of 45 interaction pairs; Figure 3). However, this model overestimated the magnitude of the interactions with much higher values than those used to create the original community matrix (Figure 3). Models fit with subsampled matrices estimated lower omega values, with lower values as the retention percentages decreased. Models fit with subsampled matrices with 90% and 75% of retention correctly estimated the sign of the interaction for most species' pair, although lower retention percentages (≤ 50%) failed even to capture the sign of the interactions (Figure 3). These trends are backed up by the Procrustes analysis, which showed two contrasting tendencies depending on the consideration of the actual values or their sign (Table 1). For the actual value, both the mean RMSE and its standard deviation decreased as the retention percentage decreased. For the sign, both the mean RMSE and its standard deviation increased with decreasing retention percentages.

All evaluation metrics gave similar information: all models correctly estimated species distribution, although their performance decreased as the retention percentage decreased (Figure 4). In general, this reduction was bigger in predictive power (PRED) than in explanatory (EXP). When removing the artificial false negatives on the validation matrix, a different trend appeared; on extreme retention percentages (i.e., 90 and 10%) the explanatory power was similar whether the artificial false negatives were kept or corrected; models with intermediate retention percentages (i.e., 25, 50, 75%) showed higher values in the validation method without artificial false negatives. Thus, the model predicted the original data better than the data with artificial false negatives, although those were used in model calibration. This pattern is clear for three of the four evaluation metrics (AUC, TSS and Kappa, but not so evident for $TjurR^2$).

Real data study case

The number of records and the completeness percentage per pixel shared a strong spatial pattern (see Supporting information), being most of the records and, by consequence, the most complete pixels mainly located in western Europe and the Scandinavian Peninsula. Further, completeness analysis showed that from the original 1323088 GBIF records within 147350 pixels, 715595 were on 21568 pixels which surpassed the 75% of completeness.

Beta parameters (Figure 5a) were generally consistent within the three datasets, aside from the fact that the signal was slightly stronger for the model with the over dataset (i.e., records only from pixels over the 75% of completeness) and only in few cases the sign of the parameters was inverted for the models fit with the other datasets. The residuals cooccurrence matrix in the model with the over dataset (Figure 5b) shows positive residual cooccurrences for most species' pairs (e.g., *Q. petraea* with *P. nigra*, *P. sylvestris*, and *P. pinaster*), but for *Q. suber* and all other species and between *Q. ilex* and *Q. robur*. This model estimates higher omega values, while models fit with the full and the under datasets shows weaker residuals. Furthermore, in models with the full and under datasets, some species were estimated to have a different interaction sign with the rest of the species (i.e., *P. mugo* and *Q. robur*), showing these models a greater number of negative interactions.

The performance of the models was generally good (see Supporting information), according to AUC and TSS but not so much to Kappa. While AUC was very similar among models, TSS and Kappa were highest for the over dataset, intermediate for the full dataset, and lowest for the under dataset.

Spatial predictions from models fit with the three datasets generally agreed (Figure 5c & Supporting information). For most species, spatial predictions by the models fit with the over dataset were slightly more restrictive (e.g., *P. pinea*, Figure 5c). One of the species that showed greater restriction in the spatial predictions for the model fit with the over dataset is *Q. suber*,

which also showed negative residual correlation with the rest of the species (Figure 5b). However, for some species like *P. sylvestris*, which have positive residual correlation values with other species (especially *Q. petraea* and *Q. pubescens*) this model predicted a more extensive distribution.

Discussion

Our study represents the first analysis on the use of opportunistic databases in ecological research with JSDMs. By subsampling cooccurrence matrices of artificial data, we mimicked the limitations of such databases and found that decreasing the retention percentage can have a significant impact on the estimation of species parameters and interactions, as well as on species distributions by JSDMs. Our results demonstrate that decreasing the retention percentage of subsampled matrices results in an increase in artificial false negatives and negative cooccurrence probabilities, leading to a loss of important ecological information, which ultimately affects model calibration and performance. However, we also found both with artificial data and with a real study case (GBIF data) that JSDMs tolerate different levels of degradation in the original matrices depending on the aspect of the model being considered. The models with at least 50 % of retention data are valid for species niches and species distribution, but interaction matrices would require more complete databases (at least a 75% of data retention). For instance, we found that JSDMs are robust in estimating the relative effects of variables among species, although they may become increasingly biased as the retention percentage decreases in estimating the intercepts of species and the effects of certain variables. In the same vein, our analysis of a real study case, showed no great differences in the estimated niche (beta values) nor in the distribution maps for most of the species between datasets of contrasting quality. Overall, our study provides insight into the challenges and opportunities of using opportunistic databases in ecological research, showing that incomplete opportunistic databases could still be mined with JSDMs to provide valuable information. Considering the increasing amount of data due to new technologies and citizen science projects, like iNaturalist, we believe that JSDMs

could be a suitable tool to mine such databases, and their suitability is expected to improve as the databases get even more complete.

It is usually argued that, by sharing information among species, JSDMs are robust methods to estimate coefficients on the effect of environmental variables on species occurrences (Nieto-Lugilde et al., 2018; Ovaskainen & Soininen, 2011). In fact, they usually estimate less overfitting models than their single-species counterparts (Maguire et al., 2016; Nieto-Lugilde et al., 2018, and references therein), although it might depend on data characteristics, as sample size (Erickson & Smith, 2023). Our results suggest that this ability could also benefit JSDMs when estimating coefficients with incomplete matrices built from opportunistic databases. Subsampling the original matrices (CM_{O1}) while maintaining sample size (n = 500) reduces species' prevalence, which unsurprisingly biases estimation of species' intercepts towards lower values (Figure 2; decreasing the RMAs' intercepts). However, estimates of beta parameters (i.e., the effect of environmental variables on species) are not biased (for x1) or only slightly biased towards higher values (for x2). More interestingly and independently of the retention percentage used to subsample the matrices, JSDMs seem to correctly discriminate the relative roles of environmental variables on the different species (Figure 2; small decrease in RMAs slopes for species intercepts, beta x1 and beta x2), which is also reflected spatially when predicting community composition (see Supporting information). These findings are confirmed also in our models fit with real data from opportunistic databases at different levels of completeness, which estimated very similar species parameters independently of the completeness level. Although no previous work has made a comparison like ours, this result aligns with previous studies using HMSC models, which proved that JSDMs make accurate estimations of the beta parameters (Zhang et al., 2018) and showed that HMSC models can estimate coefficients for species not present in the training dataset (Ovaskainen & Soininen, 2011).

The residual cooccurrence matrix is the result of the cooccurrence pattern captured by the HMSC model after removing the effects of environmental factors driving species (co)occurrence (Ovaskainen et al., 2017; Pollock et al., 2014). The interpretation of this matrix is a controversial issue (i.e., Poggiato et al., 2021), since it could be the result of species interactions, but also several other factors, like unaccounted environmental variables, species prevalence, spatial autocorrelation, or unaccounted dispersal limitations (Ovaskainen et al., 2010; Nieto-Lugilde et al., 2018; Poggiato et al., 2021). In fact, JSDMs usually overestimate the magnitude of real coefficients of species interactions even without other factors driving species cooccurrence (Zurell et al., 2018), which is aligned with our results (CM_{O2}, CM_{S90}, & CM_{S75} in Figure 3). Furthermore, we found that the interpretation of the cooccurrence matrix gets even more complicated when using incomplete matrices from opportunistic databases, since the cooccurrence patterns among species might be altered. Contrary to the effect found for betas, it seems that the matrix degradation makes the model unable to estimate significant interactions, instead of biasing their estimations (Figure 3). Retention percentages above 75% resulted in cooccurrence patterns similar to the original (Figure 3 and Supporting information; most of the cooccurrences sign is kept). In fact, this is the percentage that provides more similar residual cooccurrence matrix as with the original dataset (Figure 3 and Supporting information) both in terms of sign and actual value (Table 1 and Figure 3). Retention percentages below 75% failed to capture both the cooccurrence pattern and the residual cooccurrence matrix (note that at the lowest retention percentages we did not find strong interaction values regardless of the correct sign or the wrong one). This highlights the importance of the completeness of the opportunistic databases and the limitations of JSDMs when it comes to estimation of interactions strength. However, as in previous studies with single species (Soroye et al., 2018; Tiago et al., 2017; van Strien et al., 2013), it also indicates that when the quality of the data is good enough (>75% or retention percentage), the direction of the interactions showed by the model can be reliable. Conversely, it seems also that when model estimates strong interaction

signal is because it has enough data and might not be affected by degraded cooccurrence matrix due to random artificial false negatives.

Again, the real study case with data from GBIF confirms the previous findings, models fit with less complete pixels estimate lower residual cooccurrences, while models with high quality data estimate higher residual cooccurrences for certain species pairs. However, strong positive cooccurrences like *Q. petraea* with *P. nigra*, *P. sylvestris* and *P. pinaster* which positive interaction has been previously reported in regeneration studies (Borderieux et al., 2021), are kept along datasets (though strength decreased).

Previous studies suggest that JSDMs improve predictions of species distributions over SDMs (Nieto-Lugilde et al., 2018; Ovaskainen & Soininen, 2011), but see (Erickson & Smith, 2023). Both of our results (artificial data and real study case) suggest that, in this sense, the models are robust enough to resist missing data in the cooccurrence matrix (at least until 75% of data retention). Even more interesting, in our artificial data analysis we got worse evaluation metrics at intermediate retention percentages on the validation methods that included artificial false negatives in the validation dataset (i.e., EXP, PRED) compared to those based on non-altered data (i.e., EXPwfn). In other words, for some points the model predicted the presence of species (and effectively they were present on the original matrix). However, when they were compared with the evaluation data set in EXP and PRED they were compared with artificial false negatives, being tagged as erroneous predictions, whereas in EXPwfn they were compared with the true positive value, being tagged as correct predictions. We should not forget that this model predicts species presences based on beta parameters and cooccurrence patterns. Considering that up to a 75% of retention percentage the model correctly estimates the residual cooccurrence matrix, and that independently of the retention percentage the model captures well the environmental niche of species, the model is resilient enough to make predictions and reconstruct the original community, except for the most degraded datasets (i.e., $CM_{S10\ \&\ 25}$; blue

boxes in Figure 4). Santini et al. (2021) found that under non-optimal conditions, the true predictive abilities of SDMs dropped considerably. Considering their findings with our results would confirm that JSDMs outperform individual SDMs on this situation. However, as the data treatment was not identical, it would be interesting to repeat their work with JSDMs for a proper comparison. This idea would be in line with Norberg et al. (2019), where they described that the JSDM's predictions are better than SDM's ones, especially when the community includes rare species. Although, this could depend on other factors like sample size or niche similarity between species (Erickson & Smith, 2023).

Even though JSDMs could be robust enough for incomplete datasets, our results call for careful consideration when implementing them with opportunistic databases; spatial scale, strength of interactions, number of species, and/or species prevalence are some of the features that might affect their performance with such type of data. For instance, the choice of spatial resolution could be a critical step, since it could affect the number of sites in the analysis, but also the number of cooccurring species at each site, the number of artificial false negatives incorporated into the community matrix, and shifts in cooccurrences patterns (Araújo & Rozenfeld, 2014). This is not trivial, because even species pairs that interact negatively at local scale can show positive cooccurrences at larger scales. On the other hand, the magnitude of biotic interactions (i.e., omega parameters) could affect the models and very importantly the estimation of coefficients and their final projections. Hence, when species distributions are strongly determined by biotic interactions, JSDMs would need a relatively complete dataset to ensure that model performance is not compromised, especially when using opportunistic databases. Although we did not explicitly test the effect of the number of species on JSDMs, previous studies have shown that including rare species can improve JSDMs performance (Zhang et al., 2020b), while species with intermediate prevalence are always the more difficult to predict (Zhang et al., 2018). This suggests that when creating a community matrix from opportunistic databases, one could incorporate as many species as possible. However, our

results suggest that this could depend on the completeness of the sampling for each species. Incorporating a lot of species with less than 50% of completeness in the opportunistic databases could lead to misspecified models. However, many species, especially those that require more information and study (e.g., endemics and rare species), are the ones that might have less complete information in the opportunistic databases, especially when they live in remote and non-accessible places. Further studies should evaluate how JSDMs perform when calibrated with data from opportunistic database in which species have different levels of retention percentage, dispersal capacity, and have non-linear responses to variables which would be more realistic scenarios than ours. Other factors, like artificial clustering in geographic or environmental space and temporal bias (Bowler et al., 2022) can play an important role which has not been considered here and needs further research.

References

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*(6), 1223-1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Araújo, M. B., & Rozenfeld, A. (2014). The geographic scaling of biotic interactions. *Ecography*, *37*(5), 406-415. https://doi.org/10.1111/j.1600-0587.2013.00643.x

Barve, V., & Otegui, J. (2016). bdvis: Visualizing biodiversity data in R. *Bioinformatics*, *32*(19), 3049-3050. https://doi.org/10.1093/bioinformatics/btw333

Borderieux, J., Paillet, Y., Dalmasso, M., Mårell, A., Perot, T., & Vallet, P. (2021). The presence of shade-intolerant conifers facilitates the regeneration of Quercus petraea in mixed stands. *Forest Ecology and Management*, *491*, 119189. https://doi.org/10.1016/j.foreco.2021.119189

Bowler, D. E., Callaghan, C. T., Bhandari, N., Henle, K., Benjamin Barth, M., Koppitz, C., Klenke, R., Winter, M., Jansen, F., Bruelheide, H., & Bonn, A. (2022). Temporal trends in the

spatial bias of species occurrence records. *Ecography*, *2022*(8), e06219. https://doi.org/10.1111/ecog.06219

Chao, A. (1987). Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics*, *43*(4), 783-791. https://doi.org/10.2307/2531532

Chao, A., & Colwell, R. K. (2017). Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling. *SORT-Statistics and Operations Research Transactions*, *41*(1), Article 1.

Cook, C. E., Lopez, R., Stroe, O., Cochrane, G., Brooksbank, C., Birney, E., & Apweiler, R. (2019). The European Bioinformatics Institute in 2018: Tools, infrastructure and training. *Nucleic Acids Research*, *47*(D1), D15-D22. https://doi.org/10.1093/nar/gky1124

Doser, J. W., Finley, A. O., & Banerjee, S. (2022). *Joint species distribution models with imperfect detection for high-dimensional spatial data* (arXiv:2204.02707). arXiv. https://doi.org/10.48550/arXiv.2204.02707

Dutra Silva, L., Brito de Azevedo, E., Vieira Reis, F., Bento Elias, R., & Silva, L. (2019). Limitations of Species Distribution Models Based on Available Climate Change Data: A Case Study in the Azorean Forest. *Forests*, *10*(7), Article 7. https://doi.org/10.3390/f10070575

Erickson, K. D., & Smith, A. B. (2023). Modeling the rarest of the rare: A comparison between multi-species distribution models, ensembles of small models, and single-species models at extremely low sample sizes. *Ecography*, *2023*(6), e06500. https://doi.org/10.1111/ecog.06500

Escamilla Molgora, J. M., Sedda, L., Diggle, P. J., & Atkinson, P. M. (2022). A taxonomic-based joint species distribution model for presence-only data. *Journal of The Royal Society Interface*, *19*(187), 20210681. https://doi.org/10.1098/rsif.2021.0681

Feng, X., Enquist, B. J., Park, D. S., Boyle, B., Breshears, D. D., Gallagher, R. V., Lien, A., Newman, E. A., Burger, J. R., Maitner, B. S., Merow, C., Li, Y., Huynh, K. M., Ernst, K., Baldwin, E., Foden, W., Hannah, L., Jørgensen, P. M., Kraft, N. J. B., … López-Hoffman,

L. (2022). A review of the heterogeneous landscape of biodiversity databases: Opportunities and challenges for a synthesized biodiversity knowledge base. *Global Ecology and Biogeography*, *31*(7), 1242-1260. https://doi.org/10.1111/geb.13497

Fernandes, R. F., Scherrer, D., & Guisan, A. (2019). Effects of simulated observation errors on the performance of species distribution models. *Diversity and Distributions*, *25*(3), 400-413. https://doi.org/10.1111/ddi.12868

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302-4315. https://doi.org/10.1002/joc.5086

GBIF. (2023). *What is GBIF?* GBIF: The Global Biodiversity Information Facility. https://www.gbif.org/what-is-gbif

GBIF.org. (2023, octubre 31). *GBIF Occurrence Download https://doi.org/10.15468/dl.9sb7mz*.

Guélat, J., & Kéry, M. (2018). Effects of spatial autocorrelation and imperfect detection on species distribution models. *Methods in Ecology and Evolution*, *9*(6), 1614-1625. https://doi.org/10.1111/2041-210X.12983

Hogg, S. E., Wang, Y., & Stone, L. (2021). Effectiveness of joint species distribution models in the presence of imperfect detection. *Methods in Ecology and Evolution*, *12*(8), 1458-1474. https://doi.org/10.1111/2041-210X.13614

Kaldhusdal, A., Brandl, R., Müller, J., Möst, L., & Hothorn, T. (2015). Spatio-phylogenetic multispecies distribution models. *Methods in Ecology and Evolution*, *6*(2), 187-197. https://doi.org/10.1111/2041-210X.12318

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2016). *CHELSA climatologies at high resolution for the earth's land surface areas (Version 1.0)* [Dataset]. World Data Center for Climate (WDCC) at DKRZ. https://doi.org/10.1594/WDCC/CHELSA_v1

Lin, C. C. (2020). Ecoinformatics: A Review of Approach and Applications in Ecological Research. *Proceedings of NIE*, *1*(1), 9-21. https://doi.org/10.22920/PNIE.2020.1.1.9

Maguire, K. C., Nieto-Lugilde, D., Blois, J. L., Fitzpatrick, M. C., Williams, J. W., Ferrier, S., & Lorenz, D. J. (2016). Controlled comparison of species- and community-level models across novel climates and communities. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1826), 20152817. https://doi.org/10.1098/rspb.2015.2817

Malhi, Y., Franklin, J., Seddon, N., Solan, M., Turner, M. G., Field, C. B., & Knowlton, N. (2020). Climate change and ecosystems: Threats, opportunities and solutions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1794), 20190104. https://doi.org/10.1098/rstb.2019.0104

Marjakangas, E.-L., Ovaskainen, O., Abrego, N., Grøtan, V., de Oliveira, A. A., Prado, P. I., & de Lima, R. A. F. (2021). Co-occurrences of tropical trees in eastern South America: Disentangling abiotic and biotic forces. *Plant Ecology*, *222*(7), 791-806. https://doi.org/10.1007/s11258-021-01143-3

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*(6), 740-752. https://doi.org/10.1111/geb.12580

Newbold, T., Reader, T., El-Gabbas, A., Berg, W., Shohdi, W. M., Zalat, S., El Din, S. B., & Gilbert, F. (2010). Testing the accuracy of species distribution models using species records from a new field survey. *Oikos*, *119*(8), 1326-1334. https://doi.org/10.1111/j.1600-0706.2009.18295.x

Nieto-Lugilde, D., Maguire, K. C., Blois, J. L., Williams, J. W., & Fitzpatrick, M. C. (2018). Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. *Methods in Ecology and Evolution*, *9*(4), 834-848. https://doi.org/10.1111/2041-210X.12936

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., … Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, *89*(3), e01370. https://doi.org/10.1002/ecm.1370

Ovaskainen, O., & Abrego, N. (2020). *Joint species distribution modelling: With applications in R*. Cambridge university press. https://www.cambridge.org/es/universitypress/subjects/life-sciences/ecology-and-conservation/joint-species-distribution-modelling-applications-r?format=PB&isbn=9781108716789#bookPeople

Ovaskainen, O., Abrego, N., Halme, P., & Dunson, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, *7*(5), 549-555. https://doi.org/10.1111/2041-210X.12501

Ovaskainen, O., Hottola, J., & Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, *91*(9), 2514-2521. https://doi.org/10.1890/10-0173.1

Ovaskainen, O., & Soininen, J. (2011). Making more out of sparse data: Hierarchical modeling of species communities. *Ecology*, *92*(2), 289-295. https://doi.org/10.1890/10-1251.1

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*(5), 561-576. https://doi.org/10.1111/ele.12757

574 Peres-Neto, P. R., & Jackson, D. A. (2001). How well do multivariate data sets match? The
575 advantages of a Procrustean superimposition approach over the Mantel test.
576 *Oecologia*, *129*(2), 169-178. https://doi.org/10.1007/s004420100720
577 Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J. S., & Thuiller, W. (2021). On the
578 Interpretations of Joint Modeling in Community Ecology. *Trends in Ecology &*
579 *Evolution*, *36*(5), 391-401. https://doi.org/10.1016/j.tree.2021.01.002
580 Poggio, L., Sousa, L. M. D., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter,
581 D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial
582 uncertainty. *SOIL*, *7*(1), Article 1. https://doi.org/10.5194/soil-7-217-2021
583 Pollock, L. J., Morris, W. K., & Vesk, P. A. (2012). The role of functional traits in species
584 distributions revealed through a hierarchical model. *Ecography*, *35*(8), 716-725.
585 https://doi.org/10.1111/j.1600-0587.2011.07085.x
586 Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., &
587 McCarthy, M. A. (2014). Understanding co-occurrence by modelling species
588 simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology*
589 *and Evolution*, *5*(5), 397-406. https://doi.org/10.1111/2041-210X.12180
590 R Core Team. (2022). *R: A Language and Environment for Statistical Computing* [Software]. R
591 Foundation for Statistical Computing. https://www.R-project.org/
592 Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., & Huijbregts, M. A. J. (2021). Assessing
593 the reliability of species distribution projections in climate change research. *Diversity*
594 *and Distributions*, *27*(6), 1035-1050. https://doi.org/10.1111/ddi.13252
595 Soroye, P., Ahmed, N., & Kerr, J. T. (2018). Opportunistic citizen science data transform
596 understanding of species distributions, phenology, and diversity gradients for global
597 change research. *Global Change Biology*, *24*(11), 5281-5291.
598 https://doi.org/10.1111/gcb.14358

599 Tiago, P., Pereira, H. M., & Capinha, C. (2017). Using citizen science data to estimate climatic
600 niches and species distributions. *Basic and Applied Ecology*, *20*, 75-85.
601 https://doi.org/10.1016/j.baae.2017.04.001
602 Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M. J., Oksanen, J., &
603 Ovaskainen, O. (2020). Joint species distribution modelling with the r-package Hmsc.
604 *Methods in Ecology and Evolution*, *11*(3), 442-447. https://doi.org/10.1111/2041-
605 210X.13345
606 Tikhonov, G., Ovaskainen, O., Oksanen, J., Jonge, M. de, Opedal, O., & Dallas, T. (2022). *Hmsc:*
607 *Hierarchical Model of Species Communities* (3.0-13) [Software]. https://cran.r-
608 project.org/web/packages/Hmsc/index.html
609 Tjur, T. (2009). Coefficients of Determination in Logistic Regression Models—A New Proposal:
610 The Coefficient of Discrimination. *The American Statistician*, *63*(4), 366-372.
611 https://doi.org/10.1198/tast.2009.08210
612 van Strien, A. J., van Swaay, C. A. M., & Termaat, T. (2013). Opportunistic citizen science data
613 of animal species produce reliable estimates of distribution trends if analysed with
614 occupancy models. *Journal of Applied Ecology*, *50*(6), 1450-1458.
615 https://doi.org/10.1111/1365-2664.12158
616 Viana, D. S., Keil, P., & Jeliazkov, A. (2022). Disentangling spatial and environmental effects:
617 Flexible methods for community ecology and macroecology. *Ecosphere*, *13*(4), e4028.
618 https://doi.org/10.1002/ecs2.4028
619 Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2019). A
620 comparison of joint species distribution models for presence–absence data. *Methods*
621 *in Ecology and Evolution*, *10*(2), 198-211. https://doi.org/10.1111/2041-210X.13106
622 Wittmann, M. E., Barnes, M. A., Jerde, C. L., Jones, L. A., & Lodge, D. M. (2016). Confronting
623 species distribution model predictions with species functional traits. *Ecology and*
624 *Evolution*, *6*(4), 873-879. https://doi.org/10.1002/ece3.1898

625 Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2018). Comparing the prediction of joint species
626 distribution models with respect to characteristics of sampling data. *Ecography*,
627 *41*(11), 1876-1887. https://doi.org/10.1111/ecog.03571
628 Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2020a). Evaluating the influence of spatially
629 varying catchability on multispecies distribution modelling. *ICES Journal of Marine*
630 *Science*, *77*(5), 1841-1853. https://doi.org/10.1093/icesjms/fsaa068
631 Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2020b). Improving prediction of rare species'
632 distribution from community data. *Scientific Reports*, *10*(1), Article 1.
633 https://doi.org/10.1038/s41598-020-69157-x
634 Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., Nehrbass, N.,
635 Pagel, J., Reineking, B., Schröder, B., & Grimm, V. (2010). The virtual ecologist
636 approach: Simulating data and observers. *Oikos*, *119*(4), 622-635.
637 https://doi.org/10.1111/j.1600-0706.2009.18284.x
638 Zurell, D., Jeltsch, F., Dormann, C. F., & Schröder, B. (2009). Static species distribution models
639 in dynamically changing systems: How good can predictions really be? *Ecography*,
640 *32*(5), 733-744. https://doi.org/10.1111/j.1600-0587.2009.05810.x
641 Zurell, D., Pollock, L. J., & Thuiller, W. (2018). Do joint species distribution models reliably
642 detect interspecific interactions from co-occurrence data in homogenous
643 environments? *Ecography*, *41*(11), 1812-1819. https://doi.org/10.1111/ecog.03315
644
645

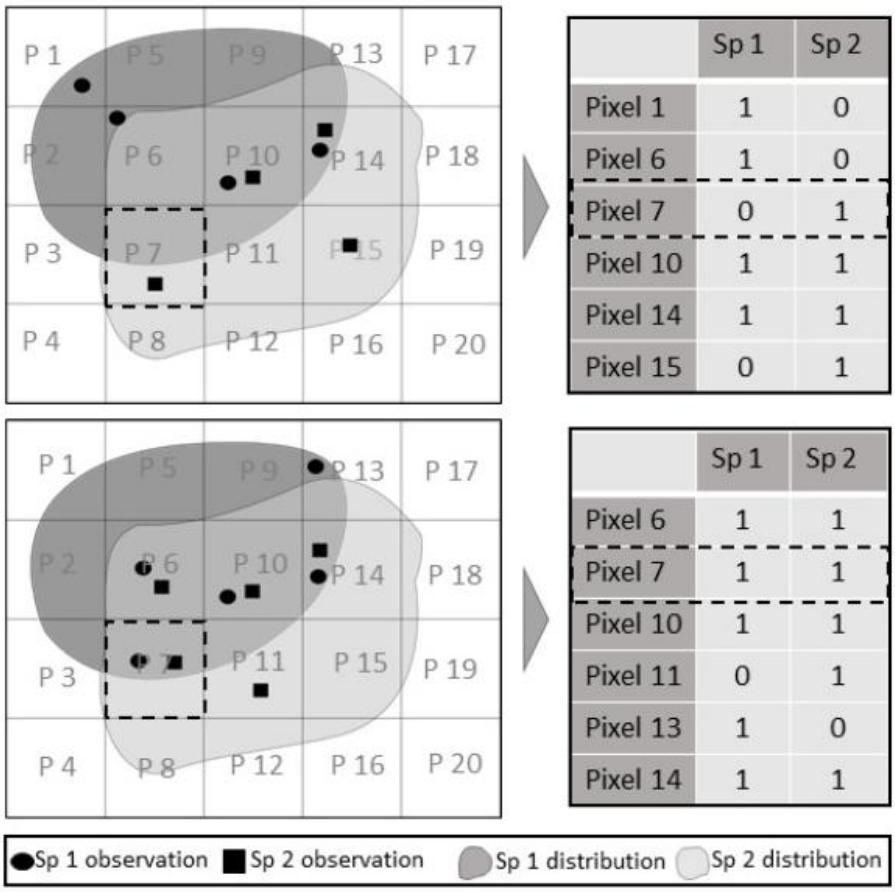
646 TABLES

647 *Table 1. Effect of retention percentages on the False Negative Ratios in the subsampled community matrices and on*
648 *the estimation of the residuals cooccurrence matrix by HMSC models. The False Negative Ratio is the number of false*
649 *negatives in each subsampled matrix divided by the number of occurrences (5000: 500 pixels × 10 species). The*
650 *Procrustes analysis reports the Root Mean Squared Error (RMSE) of the Procrustes transformation of the estimated*
651 *residuals cooccurrence matrix to fit the original matrix of Omega parameters used to create the original community*
652 *matrix (CM_{O2}). That is, higher values indicate greater differences between matrices. Because each retention*
653 *percentage was applied 10 times, all results are summarized as the mean False Negative Ratio or RMSE and their*
654 *standard deviations across the 10 iterations.*

Retention Percentage	False Negatives Ratio (Mean ± SD)	Procrustes RMSE (Mean ± SD)	Procrustes RMSE (sign) (Mean ± SD)
10	0.358 ± 0.099	0.729 ± 0.118	2.618 ± 0.205
25	0.299 ± 0.085	0.754 ± 0.120	2.695 ± 0.489
50	0.193 ± 0.053	0.772 ± 0.165	2.519 ± 0.389
75	0.097 ± 0.027	1.00 ± 0.505	2.285 ± 0.392
90	0.038 ± 0.011	1.10 ± 0.430	2.137 ± 0.271

655

656 FIGURES



657

658 *Figure 1. Illustration of how the distribution of two hypothetical (not related with our data) cooccurring species (light*
659 *and dark grey polygons; Sp1 and Sp2) can be derived into community matrices (right tables) by aggregating*
660 *observations (square and circle symbols) at specific locations (i.e., pixels – P) and assuming unobserved species as*
661 *absences. The figure illustrates two different realizations of observations simulating opportunistic databases (top and*

bottom illustrations). In the example at the top, the random sampling of both species results in a selection of pixels that more or less capture a cooccurrence pattern representative of reality. In the example at the bottom, the sampling of both species is biased towards areas of cooccurrence leading to a community matrix with a higher degree of cooccurrence between the two species. In both examples, there are omission errors for one (e.g., species 1 at pixel 7 and species 2 at pixel 6 in the example at the top, or species 2 at pixel 13 in the example at the bottom) or both species (e.g., both species at pixel 9 in both examples).

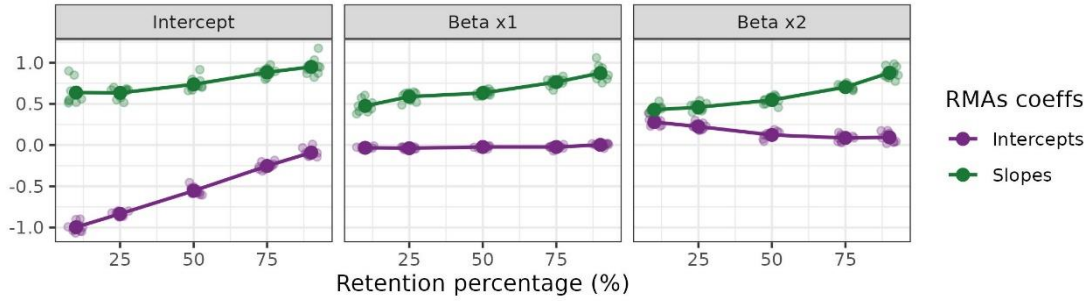


Figure 2. Intercepts and slopes of reduced major axis (RMA) regressions among estimated and real coefficients of virtual species in each HMSC model. Each retention percentage displays 10 values of slopes and 10 values of intercepts for each of the 10 random iterations in the subsampling process with the same retention percentage (these points were jittered to avoid overlapping). Ideally, slopes of RMA regressions should be close to one and intercepts close to zero. Darker points connected with a line represent the mean values across the iterations.

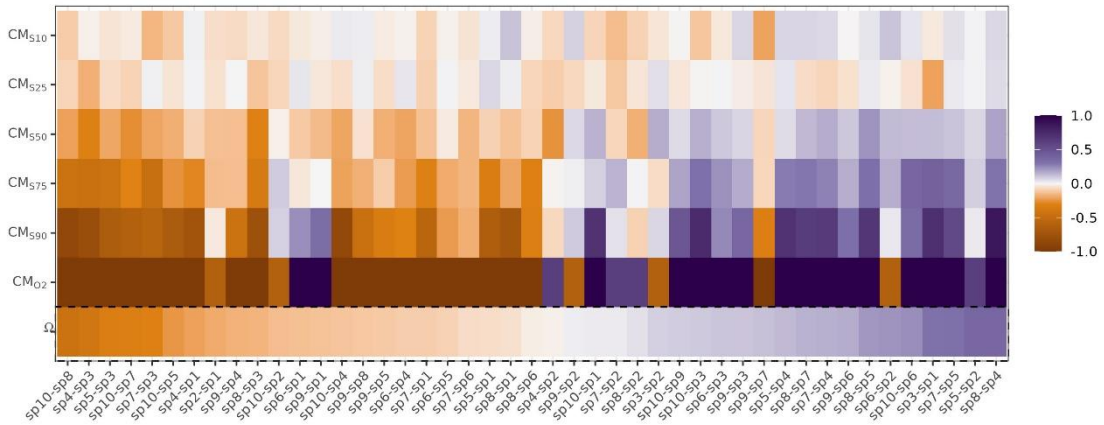


Figure 3. Cooccurrence parameters for all virtual species pairs according to the omega parameters used to build the original community matrix (Ω) and as estimated by HMSC models fit on the original matrix (CM_{O2}) and its subsampled counterparts ($CM_{S90, 75, 50, 25, \& 10}$). Because each retention percentage was applied 10 times, results for the models fit with subsampled matrices are summarized as the mean residual cooccurrence matrix across the 10 iterations.

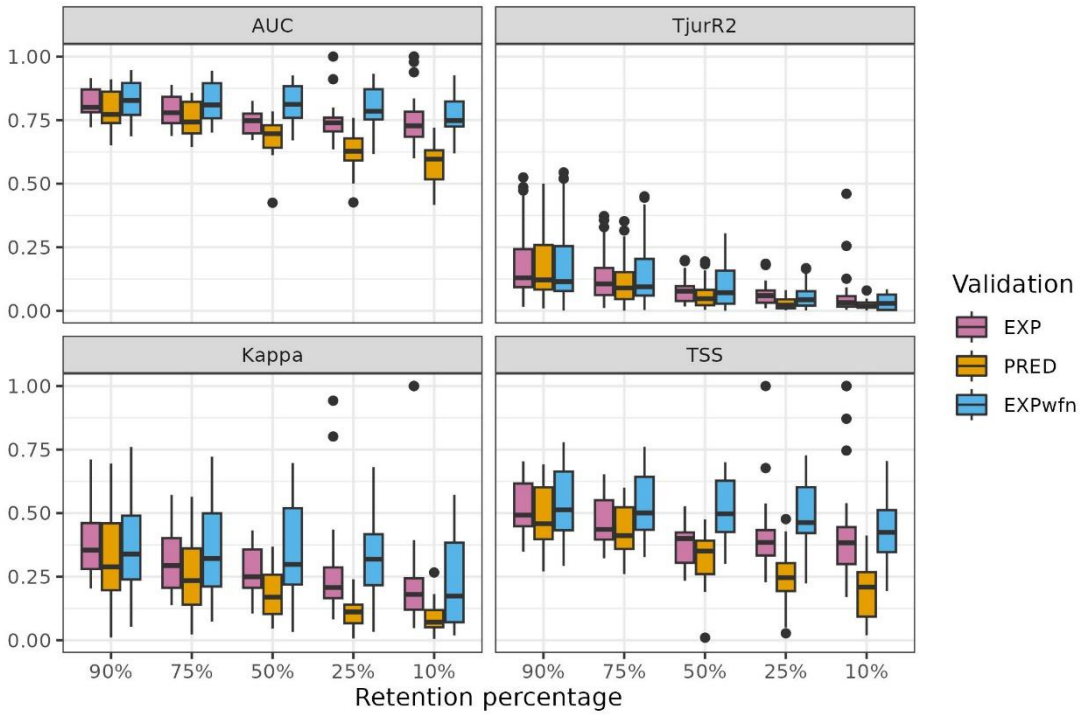


Figure 4. Model ability to estimate virtual species distribution with four different evaluation metrics (Area Under the ROC Curve – AUC –, the Tjur's Coefficient of Discrimination – TjurR2 –, Kappa, and True Skill Statistic – TSS) along the retention percentages for three validation strategies: explanatory power (EXP) evaluates the capacity of the model to predict the calibration data or explain the data; predictive power (PRED) evaluates the ability of the model to predict points outside the calibration data or predict new data; and explanatory power without false negatives (EXPwfn) evaluates the ability of the model to predict the calibration data without false negatives introduced by the subsampling scheme.

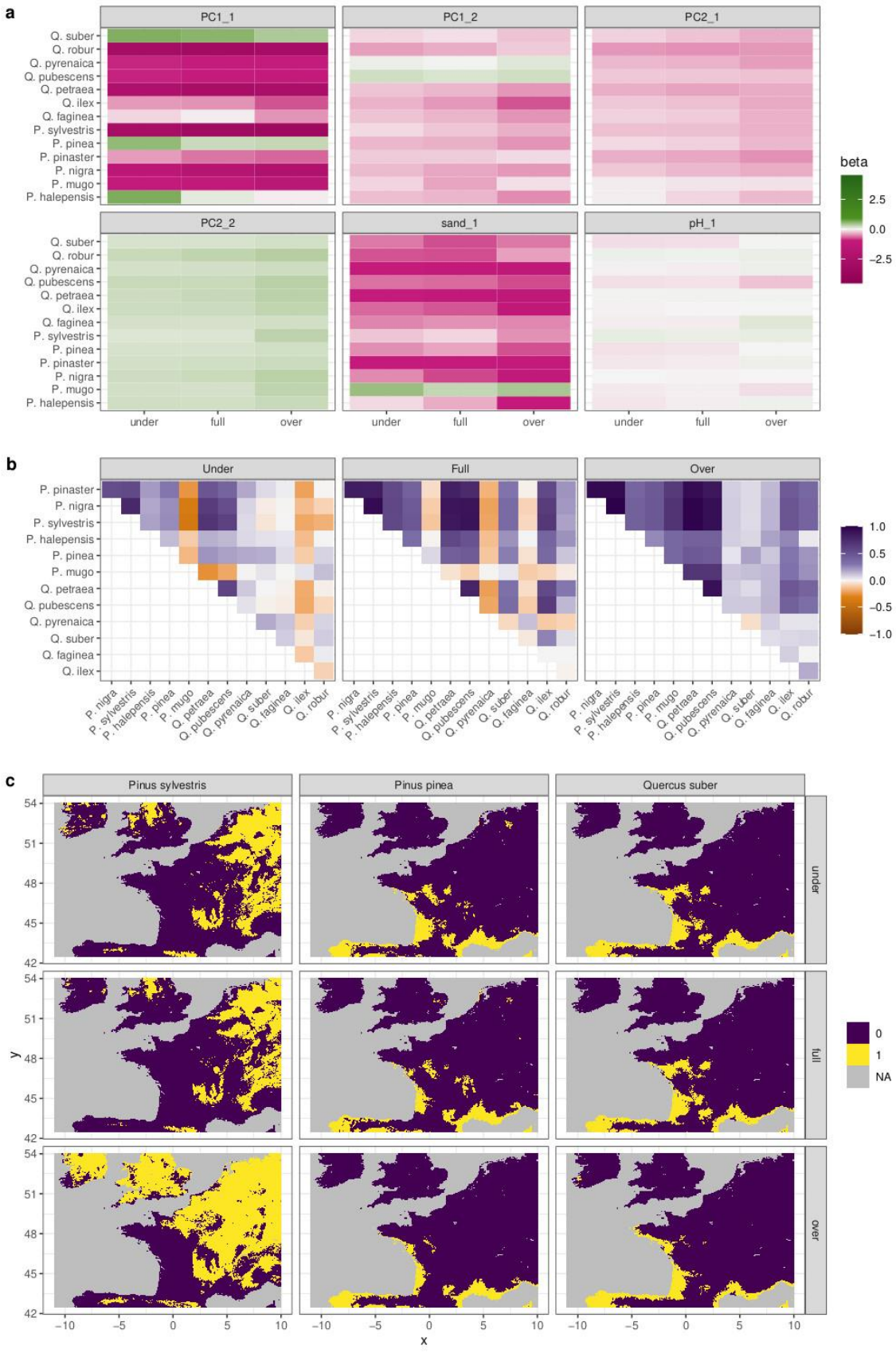


Figure 5. Example of HMSC models coefficients and predictions in the real study case of Pinus and Quercus species: a) Subset of beta parameters of the first two PCA of climate variables and the two soil variables for all species among the three datasets (under 75% threshold, full data, over 75% threshold); b) Cooccurrence parameters for all species pairs as estimated by models fit on the same three datasets (over, full, and under); and c) Predicted

694 *distribution maps (for a subset region) for Pinus sylvestris, Pinus pinea and Quercus suber (binarized using a species-*
695 *specific threshold that maximized the sum of specificity and sensitivity) by models fit with the same three datasets*
696 *(over, full, and under). Because each dataset was sampled 10 times, results are summarized as the mean values*
697 *across 10 randomizations.*

698

699

700