Journal name: Ecological Applications (ESA)

Manuscript type: Article

Manuscript title: Advancing single species abundance models: robust models for predicting abundance using co-occurrence from communities

Author names: Aliénor Stahl[1], Eric J. Pedersen[1] and Pedro R. Peres-Neto[1]

Affiliations: 1: Concordia University, Department of Biology, Montreal, Canada

Corresponding author: Aliénor Stahl, alienor.stahl@gmail.com

Open research statement: Data and code to replicate the presented analyses are available in Zenodo at https://doi.org/10.5281/zenodo.11150229

Key words: abundance; co-distribution; co-occurrence; latent; prediction; simulation

Abstract

Accurate estimates of abundance are crucial for successful conservation and management. However, gathering abundance data is costly. Species Abundance Models (SAMs) are increasingly used to predict variation in abundance for resource management for single species, but collecting enough relevant environmental information to build effective SAMs can often be challenging. Species co-occurrence patterns may provide additional information on missing environmental predictors, and data on presence-absence species co-occurrence are typically easier to collect than abundance or detailed environmental data. However, it is still not clear when supplementing abiotic data with co-occurrence data should improve abundance predictions, as co-occurrence data itself represents a noisy indicator of the local environment. Using simulated data where we manipulated the strength of relevant environmental predictors across multiple species, we assessed the conditions that improve model predictions of a target species by using co-occurrence data on the remaining species as a proxy for missing environmental predictors. Because species often share environmental preferences in nature, an aspect simulated in our data, latent variables are expected to summarize important environmental gradients across co-occurring species. We employed Gaussian copulas to generate presence-absence co-occurrence-based latent variables as proxies. These latent variables, along with various combinations of environmental predictors, were subsequently used as predictors in SAMs. We evaluated the accuracy of these models in predicting the presence and abundance of target species through model validation exercises. Our results showed that incorporating presence-absence latent predictors generally improved model performance when compared to models lacking relevant environmental predictors, although there was considerable variation in performance across simulations. All models tended to have greater error rates when predicting abundant species compared to rare species. The goal of our

proposed framework is to offer a novel and easy to implement method for accurately predicting abundance from both biotic and environmental information.

Introduction

Community ecology has grown increasingly quantitative in response to the demand for a deeper understanding and more accurate predictions regarding how ecological factors and processes influence abundance, biomass, and interactions among both coexisting and non-coexisting species (Flecker and Matthews 1999; Persson 2008). Abundance serves as a critical indicator for individual species, their communities, and/or the state of the environment, enabling us to quantify ecosystem functioning (e.g., predation pressure, densities of preys available, the probability of reproductive encounters) (Degnbol and Jarre 2004). However, abundance data is generally difficult to collect across many different locations in heterogeneous landscapes (e.g., across many lakes in a landscape) whereas data on the presence or absence of communities of species can be easier to collect at landscape scales (Jackson and Harvey 1997). As such, it would be useful for landscape-scale management to be able to predict the local abundance of specific species based on easier-to-sample data such as the presence or absence of other species.

Many conventional models used to predict abundance rely on local (e.g., lake temperature) and regional (e.g., number of growing degree days) environmental variables (Lek et al. 1996; Brosse et al. 1999; VanDerWal et al. 2009; Boyce et al. 2016; Bradley 2016; Sobrino et al. 2020). While environmental variables are relatively easy to gather through sampling or existing datasets, they are unlikely to encompass the multitude of sources of variation necessary for accurately predicting the abundances of target species of interest and other responses related to their communities, such as species composition. This limitation arises because it is not often possible to measure all relevant environmental variables, and many species and community responses depend on factors beyond just environmental ones. Additional factors, such as species interactions and history of introducing exotic species, among many others, also play important roles in shaping species patterns of species distributions, including abundance, and biodiversity (richness and species composition) in local communities and regionally (i.e., large scale variation).

In many cases, however, the environmental data gathered and used for predicting abundance variation in space (e.g., across sites) may stand as the primary source of low predictive accuracy, rather than other additional factors. For instance, relevant environmental variables may be missing or subject to measurement errors, or there could be time lags in environmental fluctuations and related changes in abundances (Myers 1998; Dornelas et al. 2013; Bengtsson, Baillie, and Lawton 1997); and these lags may vary spatially and temporally (i.e., non-stationarity in lag-responses) even for the same species. If an unmeasured driver affects the abundance of at least two species, whether positively, negatively, or even in opposite directions between the species, one can expect that information about the distribution of one of these two species would improve the prediction of the other. This is especially expected when the probability of a species' presence or absence is related to its abundances, and when the presence or absence of other species act as proxies for unmeasured quantitative factors (e.g., low versus high values), or qualitative factors (e.g., presence or absence of the missing factor). Indeed, several studies have shown that, for certain species, the most accurate predictor of abundance was information regarding the presences and absences of other species (González-Salazar, Stephens, and Marquet 2013; Lewis et al. 2017; Öğlü et al. 2019; Olkeba et al. 2020). While pairwise comparisons can be somewhat effective when studying single species, the interactions among multiple species can be complex and may not be adequately captured by pairwise comparisons alone.

It is generally not feasible to include the presence of all species in a regional species pool as predictors in a model targeting even the abundance of a single species. This is because even a moderately sized regional species pool may result in tens or hundreds of additional predictors in any abundance model. As such, incorporating the presence of other species into abundance models requires some form of dimension reduction of the species pool prior to analysis. In addition, many dimension reduction methods can borrow information across species and characterize their patterns of co-occurrence in a much-reduced number of axes, thereby improving predictive power based on these axes rather than considering all species separately (Carreira-Perpinán 1997; Cunningham 2008).

A solution to incorporating complex co-occurrence data while retaining a low dimensionality is to employ latent variable models (Walker and Jackson 2011). Latent variables are unobservable variables or factors that are not directly measured but rather estimated based on the associations (covariation) among species. These latent variables aim to estimate the joint model probability distribution of species presences-absences and represent the underlying structure or patterns in the data by specifying how data points (e.g., species composition across local communities or sites) are likely to be generated. Several methods exist to estimate latent variables from abundance or presence-absence data, including non-model-based (e.g., classic ordination methods such as principal component analysis) and model-based (e.g., mixed-model ordinations) methods (Walker and Jackson 2011; Popovic et al. 2019; Popovic, Hui, and Warton 2022). The power of latent variable methods stems from their ability to capture hidden variation in a dataset in low dimensionality (ter Braak and Prentice 1988; ter Braak 1985). Our contribution here is to demonstrate the robustness of modeling the abundances of single target species as function of latent variables that model the co-occurrence (presence-absence patterns) of the other species. This aspect is particularly important for the management and conservation programs tailored to specific species. We introduce this general modeling framework and evaluate its ability to represent sources of predictive error caused by unmeasured drivers through detailed simulations.

The goal of this study is to assess the robustness of our proposed framework for advancing single species abundance distribution models using species co-occurrence data of other species in their communities. We used detailed simulations to contrast the performance of models containing various levels of information on the environment and community composition. Moreover, because we generate abundance distributions for all species in our simulations, we can contrast our model performance between abundance-based and species-co-occurrence based. Specifically, using comprehensive simulations, we set out to assess the performance of our proposed species-abundance framework by: (1) deriving rules for determining the number of latent variables used in modeling single species abundances, (2) contrasting model performance containing varying levels of information about the true underlying drivers (environment) versus latents (i.e., environmental proxies based on co-occurrence patterns of species sharing variable levels of environmental affinities; Figure 1), and (3) assessing how predictive performance varies as a function of sample size (i.e., number of sites or local communities used as input into the model). In this study, we focused on scenarios in which species and their communities are influenced solely by environmental variation, without considering the impact of species interactions or dispersal, which can either enhance or diminish model performance (i.e., increase or decrease predictive accuracy, respectively).

Material and method

The simulations to test our framework followed the subsequent steps (see Figure 1 for an illustration of how this general workflow for a single simulated landscape):

1. Use stochastic simulations to generate landscape-scale environmental variation for each site in a landscape, and to generate coefficients for each species determining how average species abundance should vary as a function of environmental variables.
2. Simulate the abundance of species in each site, based on the environmental variables and coefficients generated in step 1.
3. Calculate latent variables from the presence-absence data of the previously generated abundance using Gaussian Copulas.
4. Using a subset of the data generated, train a set of statistical models for each species to predict local abundance. Trained models varied in the number of included environmental variables and whether the model included latent variables.
5. Use a suite of metrics to evaluate the ability of each model to predict patterns of presence-absence and abundance for the sites that were not used to estimate the models.

Steps 1 and 2: simulating communities

We used a Poisson model to simulate species abundances across different landscapes representing communities spread across *E* environmental gradients, assuming that the values of the environmental gradients were uncorrelated from one another, and that the log of the mean abundance of each species was equal to the sum of linearly dependent functions of each of the environmental gradients plus a species-specific intercept:

$$A_{s,j,u} \sim Poisson(\mu_{s,j,u}) \quad 1(a)$$

$$\mu_{s,j,u} = \exp\left(b_{0,s,u} + b_{1,s,u}X_{1,j,u} + b_{2,s,u}X_{2,j,u} + \cdots + b_{E,s,u}X_{E,j,u}\right) \quad 1(b)$$

Here $\mu_{s,j,u}$ is the expected number of individuals (abundance) of a species at a site, conditional on the environmental covariates included in the model. The abundance values were drawn from a Poisson distribution with mean $\mu_{s,j,u}$. s denotes species, j sites, and u the landscape. $A_{s,j,u}$ is the abundance of the sth species in site j of landscape u, $X_{1,j,u}$ to $X_{E,j,u}$ are the *E* environmental covariates that vary for each site j of each landscape u, $b_{0,s,u}$ the intercept that vary for each species s and landscape u, and $b_{1,s,u}$ to $b_{E,s,u}$ fixed coefficients relative to environmental variables *1* to *E* for species s in landscape u.

We simulated environmental covariates by drawing *J* independent, normally distributed values for each of the *E* environmental variables for each landscape (step 1). Thus, values for each covariate were statistically independent, with each environmental covariate having a mean of 0 and a variance of 1 across sites. These environmental covariates can be interpreted as environmental gradients given that they were generated independently. The coefficients ($b_{0,s,u}, b_{1,s,u}, \ldots b_{E,s,u}$) for each species were drawn from a uniform distribution with a range of -2.4 to 1.2 for the intercept, and -0.8 to 0.8 for the slopes. The ranges for the coefficients were determined through simulation trials where we identified the minimum and maximum coefficients that allowed for all species to be present in at least 10% of sites and at most in 90% of sites. The selected parameters allowed to generate species with different levels of strength between abundance and environment variables (e.g., narrow versus broad niche breadths; step 2). Table 1 summarizes how each variable in eq. 1 was generated. The distribution across species of spatially averaged species abundance within each landscape was approximately log-normally distributed (Figure 2), resembling common patterns found in natural communities.

Step 3: Latent variables generation and their abilities to represent missing environmental variation

Different methods are available for incorporating presence-absence information into a latent model (Popovic et al. 2019; Zou and Zhang 2009; Blanchet, Cazelles, and Gravel 2020). The copula approach used here is a model-based latent approach to estimate latent variables from multivariate data sets, as implemented in the ecoCopula R package (Popovic et al. 2019). This Gaussian Copula graphical model approach combines a multivariate distribution (e.g., multivariate Gaussian) with a set of marginal distributions (e.g., binomial, Poisson). Due to its high versatility (i.e., allowing for the selection of the multivariate distribution as well as the modeling of the appropriate discrete marginal distributions), it holds significant potential for applications in ecology (Anderson et al. 2019). Additionally, it has been shown to be one of the most accurate latent estimation methods in heterogenous environments (i.e., varying with a binary environmental covariate) (Popovic et al. 2019) and has been identified as the fastest and most robust latent variable quantification method for count and binomial (presence-absence) data (Popovic et al. 2022).

However, the copula model requires specifying the number of latent variables to estimate prior to model fitting. In general, at least E latent variables should be required to capture the variation in E independent environmental gradients, but it may be the case that more latent variables are needed to fully capture environmental variation. One frequently used method for determining the number of latent variables to retain is to compare AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) for models with increasing numbers of latent variables until the chosen matrix reaches a minimum value (i.e., best predictive value of co-occurrence). However, initial testing on landscapes (simulated using the method in step 1) with varying numbers of latent variables consistently showed that, using the BIC method calculated in ecoCopula, the BIC score was always lowest for models with a single latent variable, regardless of the number of environmental predictors used to simulate species abundances. As such, we conducted a preliminary trial to evaluate the number of latent variables needed to best approximate the environmental gradients in our simulated landscapes.

Using eq. 1, we simulated U landscapes of size J (number of sites), containing S species and a varying E number of environmental predictors ($U = 450$, $J \in \{100, 200, 300\}$, $S \in \{10, 20, 30\}$, $E \in [1,5]$; Table 1). To evaluate the optimal number of latent parameters (axes) needed to best approximate the environmental gradients in our simulated landscapes and compare the impact of adding or removing latent variables, we generated several numbers of latent variables for each possible combination of parameter values. Therefore, for each possible combination of parameter values, we fitted the presence-absence data into a stacked species regression model before using a model-based ordination with Gaussian copulas by using the functions *stackedsdm* and *cord* from the package *ecoCopula* (Popovic et al. 2019, version 1.0-2) with L different numbers of latent factors to model them (L $\in$ [1,5]).

We extracted the BIC value of each of these models and subtracted from them the BIC of the best model from any given simulation set (i.e., lowest BIC for the species considered in the current landscape). To evaluate the effectiveness of the latent variables in representing (i.e., serve as a proxy) environmental variation, we conducted a redundancy analysis (RDA) of the original environmental variables used to simulate species abundance regressed against the extracted latents using the function *rda* from the package *vegan* (Oksanen et al. 2022, version 2.6-2). Ability of latents to represent environmental variation was measured via the RDA adjusted R^2 (Peres-Neto et al. 2006). We determined from this trial that, regardless of the number of sites J or species S in the simulation, BIC was always lowest with a single latent variable (Appendix S1: Figure S1), but adjusted R^2 did increase with the number of latent predictors, until the number of latents equalled E, after which the adjusted R^2 did not increase with more latent variables (Appendix S1: Figure S2), so there is no reason to extract more than E latent variables for any given simulation.

Step 4: Contrasting the performance of abundance models

We compared the models containing only the environmental variables used to generate species abundances (eq. 1) against the ones containing selected environmental variables and the latent variables (community composition). This allowed us to compare model performance under ideal conditions because we used the true environmental drivers used to simulate species abundances against models from which we removed various combinations of environmental variables (scenarios) and replaced them with latent variables (proxies) to represent the missing sources of variation. Note, however, that ideal conditions do not imply perfect model performance, as different species were simulated with varying degrees of strength and associated errors relative to environmental variables (e.g., narrow versus broad niche breadths).

For this contrast, we created U landscapes, and for each landscape u, we generated K replicates ($U = 30$, $K = 10$ replicates per landscape). For each replicate k, we simulated abundances for each s species in each site j using eq. 1, using three environmental variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ per landscape containing multiple sites. We simulated 20 species and 1000 sites per landscape. We fixed the number of latent factors to 3 as we had three environmental variables (see RDA results in previous section). Replicates (i.e., landscapes using the same coefficients but had varying values of environmental gradients) were used to allow a reasonable estimate of the metrics used to contrast model performances.

We randomly sampled 100 sites (out of the 1000 simulated) from each landscape u (referred here as to the training set), and for each training set we estimated abundance models with different combinations of environmental and latent predictors (step 4). Each model was estimated using a Generalized Linear Model (GLM), using a Poisson distribution with a log-link function (Kéry and Royle 2015). We used the *manyglm* function from the R package *mvabund* (Wang et al. 2022, version 4.2-1) to fit separate models for each replicate landscape simultaneously for all species separately.

We were interested in comparing models containing different combinations of environmental variables and latent variables. The complete list of model scenarios considered is described in Table 2. As each species had different strengths of relationship with each environmental variable (i.e., different coefficient values in eq. 1 were used to simulate each species), we ordered the models based on the decreasing values of the environmental coefficients used to simulate the species' abundance. For instance, if species *A* had the values of -0.5, 0 and 0.8 as coefficients for the environmental variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$, respectively, $\boldsymbol{X}_3$ had the largest influence on driving abundance values, followed by $\boldsymbol{X}_2$ (i.e., importance is given by decreasing coefficient values) and $\boldsymbol{X}_1$. But if species *B* had values of 0.7, -0.5 and 0.3 as coefficients for the environmental variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$ respectively, its abundance was mostly driven by variations of $\boldsymbol{X}_1$, then $\boldsymbol{X}_3$ and finally $\boldsymbol{X}_2$. When removing $\boldsymbol{X}_1$ from the predictors of a model, species *A* and *B* were not impacted in the same way due to the lesser influence $\boldsymbol{X}_1$ had on the abundance of species *A*. We predicted that including latent variables should increase predictive ability more when added to a model that only included environmental predictors that weakly predicted the abundance of an individual species. To test this, we compared model performance with and without latent variables for models including different combinations of strengths of environmental variables.

For models containing one environmental variable as predictor, we labeled the predictors as "high", "intermediate", and "low", corresponding to the decreasing values of coefficients of the environmental variables. For models incorporating two environmental variables, we designated the model with the two highest coefficients as "high", the model with the highest and lowest coefficient as "intermediate", and the model with the two lowest coefficients as "low".

Step 5: comparison of model performance

For each model estimated for each replicate within the same landscape, we generated predictions for species abundances at the remaining 900 sites in the landscape from which the sites were sampled from (the test set). To establish baselines for optimal model performance, we also calculated predicted abundances in the test set using the oracle model: i.e., the model employing the true coefficients used to simulate each species' abundances to predict the conditional expected abundance for each species in each site. The oracle model represents the best possible model for estimating the simulated abundances in each test set that could be derived using data from the training set. Two other models were singled out: (i) a benchmark model containing all three environmental variables, to identify in which scenarios having access to all environmental variables (drivers of the abundance) did not suffice to properly estimate the environmental coefficients (by comparing the performance of the benchmark model to that of the oracle model), and (ii) a latent model containing only the latent variables, to study how species co-occurrence patterns performed as predictors of their own. We assessed how effectively the different models, including the oracle model, predicted the pattern of presences and absences as well as the true abundances in the test set.

Although our primary focus was on predicting abundance, we evaluated the models for both presence-absence and abundance predictions. This approach was taken because, in many cases, the interest may lie in predicting presence or absence of a particular target species. It is important to note, however, that the latents used as predictors were always derived based on the presence-absence of other species.

Metrics for evaluating presence-absence predictions

The Poisson regression models estimated in step 4 can predict the probability of presence of each species in a given site, but to evaluate the effectiveness of the model for predicting presence, these probabilities need to be translated into concrete predictions for presence or absence (Lawson et al. 2014; Phillips and Elith 2013). If we only treated a model as predicting a species present if the probability of presence was over 50%, models for rare species would only predict absences (and vice versa for common species), so using a fixed probability threshold would lead to all models of rare (common) species having the same predictive performance as a model that just predicts the species always being absent (present).

Therefore, instead of using a fixed probability threshold to convert the probabilities into presence-absence predictions, we used a prevalence-based approach. For each species, we set a threshold equal to the true occurrence (prevalence) rate of the species across a given landscape (e.g., Liu et al. 2005). We used this threshold to generate a predicted presence-absence matrix for each site and each species in each landscape for a given model. This was achieved by determining whether the expected abundance by the model for that site was greater (present) or lower (absent) than the threshold value. We then compared the performance of each model to the oracle model using a range of metrics, the equations for which are provided in Table 3. Using the predicted presence-absence matrices, we calculated the True Skill Statistic (TSS, Peirce 1884; Table 3) for each model, species and landscape replicate. The TSS, which ranges from -1 to +1, measures the difference between the sensitivity and specificity of the model. A score of +1 indicates a perfect agreement between the model's predictions and the true presence-absence, while a score of 0 or lower signifies performance no better than random (Allouche, Tsoar, and Kadmon 2006). We calculated the ratio of the TSS of the model over the TSS of the oracle and computed the mean for each model, species and landscape. Then, we grouped species into bins based on occurrence rates across different landscapes. A TSS ratio of $\geqslant 1$ indicates that the model performed as well or better than the oracle, while a TSS ratio of ≤ 0 or less means that the model predicted presence as badly or worse than random chance.

To compare whether including latent predictors increased model performance relative to just using environmental variables, we also calculated the delta TSS, defined as the TSS of environmental model minus the TSS of corresponding latent model (i.e., models containing the same environmental variables where the only difference in specification was the inclusion of latent variables as predictors). A positive delta TSS indicates the environmental model to have the best performance, whereas a negative value suggests that the model including of latent variables performs best.

Metrics for evaluating abundance predictions

When evaluating how each model predicted species abundance, we limited comparisons to sites where the species was present (i.e., abundance of 1 or higher). To evaluate how well each model predicted species abundance we calculated the following prediction metrics for each model, species and landscape replicate: Mean Absolute Percentage Error (MAPE), Root Mean Squared Percentage Error (RMSPE), Relative Mean Squared Error (RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), and Root Mean Ratio Percentage Error (RMRPE) (see Table 3 for definitions of these metrics). We calculated the ratio of each metric to the corresponding metric calculated for the oracle model (i.e., best possible scenario) and calculated the average ratio for each model, species and landscape (referred to as the ratio metric in the results). We also calculated the delta metric, defined as the metric calculated for a model containing only environmental variables minus the metric calculated for a model with the same environmental variables as well as latent variables. As above, a negative delta metric indicated that the latent model performed better than the same model lacking latent variables.

To illustrate how different metric performances varied with species abundance across simulations, we grouped species in different landscapes into percentile bins, based on the average (true) abundance of the species in its own landscape, and then calculated average ratio metrics and delta metrics for each percentile bin across landscapes and replicates.

Results

Number of latent variables needed to capture environmental variation

We first focus on determining the optimal number of latent dimensions to select when using Gaussian copulas. To assess the goodness of fit of the models, we examined both the RDA adjusted R^2, which represents the proportion of variance explained by the model, and the Bayesian Information Criterion (BIC), which is typically used to determine the optimal number of latent variables to retain. The RDA enabled us to estimate how effectively the latents characterize the original environmental variables (gradients) based on community composition, while the BIC helped us determine whether this criterion indeed allows for selection of an appropriate number of latents to represent community composition.

The adjusted R^2 consistently increased with the number of latent dimensions until it equaled the actual number of environmental variables used to simulate the data, at which point it plateaued (Figure 3, Appendix S1: Figure S2). This indicates that additional latent variables did not improve the model's ability to predict the environmental state of a given location. The maximum fraction of variance explained was not significantly affected by the number of true environmental variables used to generate (simulate) species abundances; capturing variation from one environmental gradient was as feasible as capturing it from three or four environmental gradients (i.e., variables). Note, again, that the interpretation here as gradients is possible because environmental variables were generated independently. The adjusted R^2 was not sensitive to the number of sites in the landscape used to estimate the latent variables, but it was sensitive to the number of species used: models based on 10 species could only explain about 30% of the variation in environmental variables, regardless of the number of latent variables used, whereas models based on 30 species could explain ~60% of variation in the environmental matrix (Appendix S1: Figure S2).

In contrast, the Bayesian Information Criterion (BIC) consistently increased with the number of latent dimensions, without showing any signs of reaching a plateau (Appendix S1: Figure S1). While models with lower BIC are generally expected to have better predictive ability for unobserved data - suggesting that the best model would always retain one latent variable regardless of the environmental dimension - this expectation did not align with our observations for the adjusted R^2. This discrepancy indicates that BIC (as calculated by ecoCopula) is not a good metric of the predictive performance of the latent model, at least when applied to gradients driving abundances while their latents were extracted from presence-absence data. Therefore, we did not report BIC of the estimated latent models for the remainder of our simulations.

Models' performance

Presence-absence predictions

We now focus on the models' performance in predicting presence-absence, including the ratio TSS (representing how well each model performed compared to the oracle model) and delta TSS (represented how well models without latent variables performed relative to models including latent variables). The ratio of the TSS had a mean of 0.7 and ranged from -1.6 to 1.7 (recall that any value below 0 indicates that the model did not perform better than random, while any value above 1 represents better performance compared to the oracle). Initially examining the TSS across species occurrence percentiles, there were no obvious patterns (Figure 4). In this case, the number of occurrences of a target species did not influence model's performance. When comparing models, models containing two environmental variables performed better on average than those with only one, regardless of whether latents are included or not.

When comparing models with and without latent variables, any delta TSS value above 0 indicates that the environmental model performs better, while any negative value indicates a better performance by the latent model. Models containing latent variables generally performed better on average across all (target) species, especially for those with high occurrence and in models containing only one environmental predictor (Figure 4). The differences are less pronounced when comparing models that contain two environmental variables (i.e., where only one environmental predictor is missing from the model). Reducing the number of sites used to fit the model did not affect the performance of the TSS, sensitivity, or specificity (Appendix S1: Figure S3).

When comparing the TSS as performance of the oracle (i.e., a model using the true coefficients of the environmental variables to generate the species' conditional expectations), benchmark (i.e., a model containing all three environmental variables), and latent models (i.e., a model containing only the latent variables), we can notice that they are very correlated across species occurrence percentiles (Figure 5). The benchmark and oracle models have extremely similar performances. Regarding sensitivity, the benchmark and oracle models are also highly correlated, while the latent model demonstrates good correlation for species with low occurrence. For specificity, the benchmark and oracle models are correlated for high occurrence species, while the benchmark and latent models are correlated for low occurrence species.

Abundance predictions

To assess the goodness of fit for abundance-based models (i.e., target species include abundance information while latents are based on presence-absence of the other species), we calculated six metrics to assess the extent to which the models mispredict species abundances. Again, we used the ratio of each metric over the same metric calculated for the oracle model (i.e., representing the best possible predictive scenario), along with the delta metric to compare models that differ in composition due to the inclusion or exclusion of latent variables.

To assess across all species the impact on model performance of removing any given environmental predictor, we had to consider the varying strengths in the relationship between each species abundance and each environmental variable to compare the predictive ability of latents. As a reminder, in models containing one environmental variable as predictor, we labeled the predictors as “high”, “intermediate”, and “low”, corresponding to the decreasing coefficients of the environmental variables. For models incorporating two environmental variables, we designated the model with the two highest coefficients as “high”, the model with the highest and lowest coefficient as “intermediate”, and the model with the two lowest coefficients as “low”. Regardless of the metric considered, we observe the following patterns: prediction error increases as species abundance increases, and models containing two environmental variables outperform models containing only one environmental variable (Figure 6, Appendix S1: Figure S4). When comparing models with or without latent variables, highly abundant species were best predicted by models containing latent variables (Figure 6, Appendix S1: Figure S4). For species with low and medium abundances, the inclusion or exclusion of latent did not impact the performance of the models; they exhibited very similar values of error.

When comparing the metrics in relation to the performance of the oracle (i.e., a model using the true coefficients of the environmental variables to generate the species’ conditional expectations), benchmark (i.e., a model containing all three environmental variables) and latent models (i.e., a model containing only the latent variables), we observe identical trends across all metrics. The performance of the three models was very similar for low abundance species; however, the latent model diverged when the abundance percentile was higher than 70%, with an increase in predictive error (Appendix S1: Figure S5). The metrics were not sensitive to the number of sites in the landscape used to fit the models (Appendix S1: Figure S6).

Discussion

Number of latent variables needed to capture environmental variation

Our first goal was to establish rules for determining the number of latent variables used in modeling single species abundances. To achieve this, we examined the behavior of two metrics, the BIC and the adjusted R^2, within a simulated landscape. Our results indicate that the BIC was not a useful metric for deciding the appropriate number of latent variables when employing Gaussians copulas. Instead of plateauing once the latent variables captured as much of the environment as possible, it continued to increase, implying that the best number of latent variables was consistently one even in cases where multiple independent environmental gradients were set to simulate species distributions. It is plausible that current calculation method for BIC is incorrect or does not employ an appropriate penalty measure (number of parameters and sample size). Note that there is a general lack of consensus about the best criteria for assessing latent models (Weller, Bowen, and Faubert 2020). On one hand, the BIC is generally regarded as a reliable metric for latent models (Nylund, Asparouhov, and Muthén 2007); however, it is also criticized for being overly conservative (Mindrila 2023) as it was the case here. Note, however, that the underperformance of BIC to decide the number of latents to use in species abundance models may be due to the fact that, in our simulations, species' responses to environmental gradients were in the form of abundances, whereas latent predictors were extracted from presence-absence data. Consequently, the more liberal AIC might be a preferable option for the Gaussian copulas used in our study. Note that regardless of whether we use AIC or BIC to assess the number of latents to retain, this assessment is intrinsic and solely based on the community data used to estimate the latent variables, which are then used as predictors in abundance distribution models of single species. As we will discuss, an extrinsic selection, in which latents that improve abundance predictive accuracy are chosen, may prove to be a better strategy when using latent models based on co-occurrence data to predict abundance of single (target) species.

Note that the goal of the RDA analysis, based on the R^2 metric, was to assess whether the latent structures used here could serve as a good proxy for the true environmental variables used to simulate species distributions. Given that the adjusted R^2 plateaued when the number of latent variables equalled the true number of environmental dimensions, it instills confidence that these latents serve as robust proxies. However, it is important to note that this analysis cannot generally be performed, as in true empirical cases we do not know whether the measured predictors are important. Further, this plateau of latent predictive ability when the number of latent predictors equals the number of environmental predictors is likely due to the fact that our abundance simulations only used linear environment-abundance relationships; it is likely that if abundance-environment relationships were nonlinear (e.g. uni- or multi-modal), a larger number of latent variables would be needed to capture the same number of environmental dimensions.

Additionally, although the RDA analysis demonstrated that the correct number of latents can represent the true number of environmental gradients structuring co-occurring species, it is important to note that the original simulations generated abundance values that were then transformed into presence-absence for generating latents. Although using presence-absence data allows our models to be applicable across many systems - given that researchers often only have abundance data for a few target species and presence-absence data for multiple other co-occurring species - there is certainly loss of environmental signal by doing so. This explains why the adjusted R^2 is generally not very high.

Model performance

Our second and third objectives were aimed at contrasting model performance that contained varying levels of information (i.e., number of predictors) about the true underlying drivers versus latent predictors and assessing how predictive performance varied as a function of sample size. We first compared model performance based on the presence-absence predictions, with the goal of assessing accuracy and comparing it to current models used by management which in most cases, do not contain all relevant environmental drivers. Although our study was primarily designed to predict abundance, the ability to derive accurate presence-absence predictions would enable researchers to apply an even more general framework for species distribution modeling based on latent predictors.

Presence-absence predictions

As to be expected, adding relevant environmental variables to the models improves predictions. Since the species' abundance - and consequently presence-absence - is linearly related to these variables, any environmental information enables the model to capture more variation and thus predict abundance more accurately. Including all environmental variables leads to a perfect prediction. Although our goal was to develop and assess the performance of a general framework for predicting species distributions of target species based on latents of co-occurring species, different issues could be considered in future studies. For instance, the perfect prediction including all predictors was an outcome to be expected given that we did not include measurement error for environmental predictors or species abundances (i.e., white noise) in our simulations (see McInerny and Purves 2011 for potential approaches for attenuating the potential effects of environmental measurement error species distributional models). It would be interesting to perform a sensitivity analysis after including measurement errors either in the way environment (e.g., spatial variation within sites, temporal lags in species responses to environments) or abundance (e.g., estimates based on mark-recapture) are measured.

The inclusion of species co-occurrence patterns through latent variables also leads to an improvement in predictions, indicating that the latent variables can capture unobserved environmental variation and serve as a proxy for missing (but relevant) environmental drivers. Indeed, models that incorporate two environmental variables and latent variables tended to perform better than models containing only two environmental variables. This is particularly important since empirical datasets are unlikely to capture all relevant environmental drivers. Although presence-absence datasets are common, a model capable of predicting the presence and absence of an invasive species or a rare species based on the rest of the community composition could be useful for conservation efforts, especially with methods such as eDNA surveys that can collect information on presence from relatively few samples (Rees et al. 2014).

The lack of influence of number of sites sampled on model performance may initially seem surprising. However, the training set of sites used to fit the models was sampled independently of the values of the environmental variables and without measurement error. This means that regardless of number of sites used to fit the model, the relationship between abundance and environment would have been accurately captured. It would be interesting to assess how changing the relationship from linear to quadratic would influence the results; as there would be increased complexity in the link, we'd expect to have a greater impact of number of sites sampled on the predictions.

Abundance predictions

The species' average abundance was generally low in our simulations. However, since we were interested in relative abundance error rather than true abundance error, we made a deliberate decision not to adjust the parameters of our simulations, maintaining a low average abundance. The shape of the abundance density curve was, to us, the most salient characteristic we aimed to replicate. Keeping the average abundance low also allowed us to maintain the occurrence of species within an ecologically meaningful range (i.e., between 10% and 90% of occurrence across the landscape).

As expected, adding environmental variables improved the abundance predictions. Since no measurement error was included in the simulations for either environmental variables or species abundances, the inclusion of any environmental variable is likely to improve predictive accuracy. However, it is interesting to note that adding community composition only improved predictions for the high abundance species. One possible explanation for this is that the way we generated species abundances resulted in low-abundance species also being only weakly predictable from environmental variation (and thus only weakly predictable from community composition). In our simulations, a species would have low average abundance if it either had a small intercept (b_0) and values of the environmental slopes (b_1 to b_E values) close to zero (so it would be roughly equally distributed across the landscape), or if it had a very small intercept value (b_0) and one large environmental slope value, so it was well-predicted by a single environmental variable. As such, the low predictive power of latent variables for rare species observed in our results may not generalize to species in natural settings. In fact, one might expect that species with intermediate abundances are likely to be best predicted due to the positive relationship typically observed between occupancy (number of sites occupied) and abundance (Gaston 1996; but see Wright 1991). Species with low abundances may not occupy all suitable habitats, while those with high abundances could be generalists, occupying an excess of environments. Additionally, many other non-environmental factors (e.g., biogeography, dispersal limitation, species interactions, species introductions) may plays an important role in shaping patterns of species distributions and biodiversity in local communities and regionally (Boulangeat, Gravel, and Thuiller 2012; Lewis et al. 2017; Guisan and Thuiller 2005). We suggest that future research could extent these simulations to incorporate nonlinear environmental gradients driving species abundances.

Unlike presence-absence predictions, where no pattern related to species incidence could be identified, we observe a clear trend for the abundance predictions. The more abundant a species is, the higher the model's predictive error. Since we measure the relative error in prediction and not the absolute error, this is not an artefact related to the total abundance of the species but rather it is related to the fact that the high abundance sites are poorly predicted. However, it may be due to the fact that we simulated species abundance from a Poisson distribution, where the variance in outcome increases linearly with the mean abundance, which would lead to higher variability in abundance even between sites with identical environmental variables. This does not make this result an artifact of our simulations, however; positive mean-variance relationships are typical in ecological populations (He and Gaston 2003), so we expect that it should be more difficult in general to predict abundances of common species compared to rare ones. It is important to highlight the fact that using a different statistical family to model species' abundance might allow for a better fit of the model with empirical data and further improve the predictions (see review by Waldock et al. 2022). Note, however, that the main component of our framework - the use of latents based on species co-ocurrence patterns to predict species abundances - can be directly applied to any modeling procedure, whether it is based on maximum likelihood, Bayesian or machine learning models.

One intriguing result was observing the convergence of the models' performance for low-abundance species. Indeed, for species in the 0 to 50 percentiles of abundance, regardless of the metric used, a model containing only community composition can perform as well as one containing all environmental variables. This result may demonstrate the true potential for our framework as a management tool. However, again, this may be due to the Poisson expectation of our simulations as explained earlier. This performance does not apply to high abundance species, where there is a significant divergence in the models' performance, likely caused by a few sites with very high abundances. Applications to empirical datasets may require downweighing the importance of sites containing high abundances to avoid skewing the model's predictive accuracy. The use of more robust models that may account for different types of overdispersion (e.g., very low and high abundances) can be considered within the context of our framework (e.g., Poisson-log normal model, Harrison 2014).

Additionally, increasing the number of sites sampled did not influence predictive performance, a result we anticipated since we sampled uniformly across the landscapes and captured the entire range of variation when fitting the model. However, such uniform sampling across landscapes is unlikely to be realistic when using empirical data, particularly in complex and patchy landscapes in which environmental features are clumped and spatially autocorrelated. This issue extends beyond our study. Various approaches have been proposed to mitigate the impact of complex landscapes on the predictive performance of species distribution models based on environmental features. Different sampling methods (Fortin, Drapeau, and Legendre 1989; Christianson and Kaufman 2016), model validation techniques (Wenger and Olden 2012), and modeling frameworks (e.g., Dormann 2007 for a review, Guélat and Kéry 2018) are among these proposed solutions and could, in principle, be incorporated into our modeling framework given its flexibility.

We did not include any species interactions in our model simulations: as such, our results demonstrate that latent community composition variables can capture similar patterns of environmental interactions even in the absence of species interacting with one another. Although latent variable models can represent species interactions (e.g., competition, trophic interactions) via networks (e.g., Ovaskainen et al. 2016), adjustments to the latent extraction may be necessary in order to incorporate more complex processes underlying pattern of species co-occurrences. It is likely that including direct species interactions (e.g., competition or predation) would increase the power of latent parameters for predicting species abundances as long as strong species interactions were relatively rare, or species interaction networks are relatively sparse; strong species interactions and dense species interaction networks can result in complex feedbacks, such that the net effect of presence or absence of a given species on a focal species may be indeterminant (Tunney, Carpenter, and Vander Zanden 2017).

Finally, it is important to consider that we used all species in any given simulated landscape to generate latents. However, it is likely that certain reduced number of species combinations would better serve as inputs for latent generation. For instance, consider a scenario involving two species and two independent environmental predictors. If one species is highly associated with one environmental predictor but randomly associated with the other; and the second species shows the reverse pattern, then the two species will not effectively predict each other. One possible solution is to cluster species based on their environmental affinities prior to latent generation (see Hui et al. 2013 for a discussion). As such, latents could be tailored to only consider species that increase the model performance of the target species.

Our proposed framework offers considerable promise for several compelling reasons. First, it is highly flexible in terms of parameter estimation, as it can accommodate any regression style approach. This allows to predict both presence-absence and abundance, and it demonstrates very good performance in predicting low-abundance species. Moreover, one can also use other latent modeling procedures and not necessarily Gaussian copulas. The framework could also be used to predict biomass rather than abundance by replacing the family of the GLM used, depending on the variable of highest interest for management. Overall, our proposed framework is incredibly versatile, allowing for significant flexibility and adaptability to accommodate the available data.

Acknowledgments

The data presented in this article were sampled and curated by the Ontario Ministry of Natural Resources and Forestry; we thank Dr. Cindy Chu for graciously giving us access to the Ontario Broadscale Monitoring Program data. AS was funded by the NSERC-CREATE BIOS2 program the Genome Canada GEN-FISH Network. AS and PPN were funded by the Canada Research Chair in Spatial Ecology and Biodiversity.

Author Contributions

Conceptualisation: AS, EP, PPN. Coding: AS, EP, PPN. Methodology: AS, EP, PPN. Visualisation: AS. Writing: AS, EP, PPN. All authors read and approved the final manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

References

Allouche, Omri, Asaf Tsoar, and Ronen Kadmon. 2006. “Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS).” *Journal of Applied Ecology* 43 (6): 1223–32. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

Anderson, Marti J., Perry de Valpine, Andrew Punnett, and Arden E. Miller. 2019. “A Pathway for Multivariate Analysis of Ecological Communities Using Copulas.” *Ecology and Evolution* 9 (6): 3276–94. https://doi.org/10.1002/ece3.4948.

Bengtsson, Jan, Stephen R. Baillie, and John Lawton. 1997. “Community Variability Increases with Time.” Edited by David Warton. *Oikos* 78 (2): 249–56.

https://doi.org/10.2307/3546291.

Blanchet, F. Guillaume, Kevin Cazelles, and Dominique Gravel. 2020. “Co-occurrence Is Not Evidence of Ecological Interactions.” Edited by Elizabeth Jeffers. *Ecology Letters* 23 (7): 1050–63. https://doi.org/10.1111/ele.13525.

Boulangeat, Isabelle, Dominique Gravel, and Wilfried Thuiller. 2012. “Accounting for Dispersal and Biotic Interactions to Disentangle the Drivers of Species Distributions and Their Abundances.” *Ecology Letters* 15 (6): 584–93. https://doi.org/10.1111/j.1461-0248.2012.01772.x.

Boyce, Mark S., Chris J. Johnson, Evelyn H. Merrill, Scott E. Nielsen, Erling J. Solberg, and Bram van Moorter. 2016. “REVIEW: Can Habitat Selection Predict Abundance?” Edited by Luca Börger. *Journal of Animal Ecology* 85 (1): 11–20. https://doi.org/10.1111/1365-2656.12359.

Braak, Cajo J.F. ter. 1985. “Correspondence Analysis Data: In Terms of a Unimodal Properties Response Model.” *Biometrics* 41 (4): 859–73.

Braak, Cajo J.F. ter, and I. Colin Prentice. 1988. “A Theory of Gradient Analysis.” In *Advances in Ecological Research*, 18:271–317. https://doi.org/10.1016/S0065-2504(08)60183-X.

Bradley, Bethany A. 2016. “Predicting Abundance with Presence-Only Models.” *Landscape Ecology* 31 (1): 19–30. https://doi.org/10.1007/s10980-015-0303-4.

Brosse, S., Jean-François Guegan, Jean-Nöel Tourenq, and Sovan Lek. 1999. “The Use of Artificial Neural Networks to Assess Fish Abundance and Spatial Occupancy in the Littoral Zone of a Mesotrophic Lake.” *Ecological Modelling* 120 (2–3): 299–311. https://doi.org/10.1016/S0304-3800(99)00110-6.

Carreira-Perpinán, MA. 1997. “A Review of Dimension Reduction Techniques.” *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, no. 9: 1–69. http://www.pca.narod.ru/DimensionReductionBrifReview.pdf.

Christianson, Danielle S., and Cari G. Kaufman. 2016. “Effects of Sample Design and Landscape Features on a Measure of Environmental Heterogeneity.” Edited by Robert Freckleton. *Methods in Ecology and Evolution* 7 (7): 770–82. https://doi.org/10.1111/2041-210X.12539.

Cunningham, Pádraig. 2008. “Dimension Reduction.” In *Machine Learning Techniques for Multimedia*, 91–112. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_4.

Degnbol, P., and A. Jarre. 2004. “Review of Indicators in Fisheries Management – a Development Perspective.” *African Journal of Marine Science* 26 (1): 303–26. https://doi.org/10.2989/18142320409504063.

Dormann, Carsten F. 2007. “Effects of Incorporating Spatial Autocorrelation into the Analysis of Species Distribution Data.” *Global Ecology and Biogeography* 16 (2): 129–38. https://doi.org/10.1111/j.1466-8238.2006.00279.x.

Dornelas, Maria, Anne E Magurran, Stephen T Buckland, Anne Chao, Robin L Chazdon, Robert K Colwell, Tom Curtis, et al. 2013. “Quantifying Temporal Change in Biodiversity: Challenges and Opportunities.” *Proceedings of the Royal Society B: Biological Sciences* 280 (1750): 20121931. https://doi.org/10.1098/rspb.2012.1931.

Flecker, Alexander S., and William J. Matthews. 1999. “Patterns in Freshwater Fish Ecology.” *Copeia* 1999 (1): 229–30. https://doi.org/10.2307/1447409.

Fortin, Marie-jose, Pierre Drapeau, and Pierre Legendre. 1989. “Spatial Autocorrelation and

Sampling Design in Plant Ecology.” *Vegetatio* 83 (1–2): 209–22. https://doi.org/10.1007/BF00031693.

Gaston, Kevin J. 1996. “The Multiple Forms of the Interspecific Abundance-Distribution Relationship.” *Oikos* 76 (2): 211–20. https://doi.org/10.2307/3546192.

González-Salazar, Constantino, Christopher R. Stephens, and Pablo A. Marquet. 2013. “Comparing the Relative Contributions of Biotic and Abiotic Factors as Mediators of Species’ Distributions.” *Ecological Modelling* 248 (January): 57–70. https://doi.org/10.1016/j.ecolmodel.2012.10.007.

Guélat, Jérôme, and Marc Kéry. 2018. “Effects of Spatial Autocorrelation and Imperfect Detection on Species Distribution Models.” Edited by Nick Isaac. *Methods in Ecology and Evolution* 9 (6): 1614–25. https://doi.org/10.1111/2041-210X.12983.

Guisan, Antoine, and Wilfried Thuiller. 2005. “Predicting Species Distribution: Offering More than Simple Habitat Models.” *Ecology Letters* 8 (9): 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x.

Harrison, Xavier A. 2014. “Using Observation-Level Random Effects to Model Overdispersion in Count Data in Ecology and Evolution.” *PeerJ* 2 (1): e616. https://doi.org/10.7717/peerj.616.

He, Fangliang, and Kevin J. Gaston. 2003. “Occupancy, Spatial Variance, and the Abundance of Species.” *The American Naturalist* 162 (3): 366–75. https://doi.org/10.1086/377190.

Hui, Francis K C, David I Warton, Scott D Foster, and Piers K Dunstan. 2013. “To Mix or Not to Mix: Comparing the Predictive Performance of Mixture Models vs. Separate Species Distribution Models.” *Ecology* 94 (9): 1913–19. https://doi.org/10.1890/12-1322.1.

Jackson, Donald A, and Harold H Harvey. 1997. “Qualitative and Quantitative Sampling of Lake Fish Communities.” *Canadian Journal of Fisheries and Aquatic Sciences* 54 (12): 2807–13. https://doi.org/10.1139/f97-182.

Kéry, Marc, and J. Andrew Royle. 2015. *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS*. Vol. 1. Elsevier. https://doi.org/10.1016/C2015-0-04070-9.

Lawson, Callum R., Jenny A. Hodgson, Robert J. Wilson, and Shane A. Richards. 2014. “Prevalence, Thresholds and the Performance of Presence–Absence Models.” Edited by Robert Freckleton. *Methods in Ecology and Evolution* 5 (1): 54–64. https://doi.org/10.1111/2041-210X.12123.

Lek, Sovan, Alain Belaud, Philippe Baran, Ioannis Dimopoulos, and Marc Delacoste. 1996. “Role of Some Environmental Variables in Trout Abundance Models Using Neural Networks.” *Aquatic Living Resources* 9 (1): 23–29. https://doi.org/10.1051/alr:1996004.

Lewis, Jesse S., Matthew L. Farnsworth, Chris L. Burdett, David M. Theobald, Miranda Gray, and Ryan S. Miller. 2017. “Biotic and Abiotic Factors Predicting the Global Distribution and Population Density of an Invasive Large Mammal.” *Scientific Reports* 7 (1): 44152. https://doi.org/10.1038/srep44152.

Liu, Canran, Pam M Berry, Terence P Dawson, and Richard G Pearson. 2005. “Selecting Thresholds of Occurrence in the Prediction of Species Distributions.” *Ecography* 28 (3): 385–93. http://www.jstor.org/stable/3683850.

McInerny, Greg J., and Drew W. Purves. 2011. “Fine-scale Environmental Variation in Species Distribution Modelling: Regression Dilution, Latent Variables and Neighbourly Advice.” *Methods in Ecology and Evolution* 2 (3): 248–57. https://doi.org/10.1111/j.2041-210X.2010.00077.x.

Mindrila, Diana. 2023. “Bayesian Latent Class Analysis: Sample Size, Model Size, and Classification Precision.” *Mathematics* 11 (12): 2753. https://doi.org/10.3390/math11122753.

Myers, Ransom A. 1998. “When Do Environment-Recruitment Correlations Work?” *Reviews in Fish Biology and Fisheries* 8 (3): 285–305. https://doi.org/10.1023/A:1008828730759.

Nylund, Karen L., Tihomir Asparouhov, and Bengt O. Muthén. 2007. “Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study.” *Structural Equation Modeling: A Multidisciplinary Journal* 14 (4): 535–69. https://doi.org/10.1080/10705510701575396.

Öğlü, Burak, Upendra Bhele, Ain Järvalt, Lea Tuvikene, Henn Timm, Siim Seller, Juta Haberman, et al. 2019. “Is Fish Biomass Controlled by Abiotic or Biotic Factors? Results of Long-Term Monitoring in a Large Eutrophic Lake.” *Journal of Great Lakes Research* 46 (4): 881–90. https://doi.org/10.1016/j.jglr.2019.08.004.

Oksanen, Jari, Gavin L Simpson, F Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R Minchin, R B O’Hara, et al. 2022. “Vegan: Community Ecology Package.” https://cran.r-project.org/package=vegan.

Olkeba, Beekam Kebede, Pieter Boets, Seid Tiku Mereta, Mesfin Yeshigeta, Geremew Muleta Akessa, Argaw Ambelu, and Peter L. M. Goethals. 2020. “Environmental and Biotic Factors Affecting Freshwater Snail Intermediate Hosts in the Ethiopian Rift Valley Region.” *Parasites & Vectors* 13 (1): 292. https://doi.org/10.1186/s13071-020-04163-6.

Ovaskainen, Otso, Nerea Abrego, Panu Halme, and David Dunson. 2016. “Using Latent Variable Models to Identify Large Networks of Species-to-species Associations at Different Spatial Scales.” Edited by David Warton. *Methods in Ecology and Evolution* 7

(5): 549–55. https://doi.org/10.1111/2041-210X.12501.

Peirce, C S. 1884. “The Numerical Measure of the Success of Predictions.” *Science (New York, N.Y.)* 4 (93): 453–54. https://doi.org/10.1126/science.ns-4.93.453-a.

Peres-Neto, Pedro R., Pierre Legendre, Stéphane Dray, and Daniel Borcard. 2006. “Variation Partitioning of Species Data Matrices: Estimation and Comparison of Fractions.” *Ecology* 87 (10): 2614–25. https://doi.org/10.1890/0012-9658(2006)87.

Persson, Lennart. 2008. “Community Ecology of Freshwater Fishes.” In *Handbook of Fish Biology and Fisheries, Volume 1*, 321–40. Oxford, UK: Blackwell Publishing Ltd. https://doi.org/10.1002/9780470693803.ch15.

Phillips, Steven J., and Jane Elith. 2013. “On Estimating Probability of Presence from Use–Availability or Presence–Background Data.” *Ecology* 94 (6): 1409–19. https://doi.org/10.1890/12-1520.1.

Popovic, Gordana C., David I. Warton, Fiona J. Thomson, Francis K.C. Hui, and Angela T. Moles. 2019. “Untangling Direct Species Associations from Indirect Mediator Species Effects with Graphical Models.” *Methods in Ecology and Evolution* 10 (9): 1571–83. https://doi.org/10.1111/2041-210X.13247.

Popovic, Gordana C, Francis K C Hui, and David I Warton. 2022. “Fast Model-based Ordination with Copulas.” *Methods in Ecology and Evolution* 13 (1): 194–202. https://doi.org/10.1111/2041-210X.13733.

Rees, Helen C., Ben C. Maddison, David J. Middleditch, James R.M. Patmore, and Kevin C. Gough. 2014. “The Detection of Aquatic Animal Species Using Environmental DNA – a Review of EDNA as a Survey Tool in Ecology.” Edited by Erika Crispo. *Journal of Applied Ecology* 51 (5): 1450–59. https://doi.org/10.1111/1365-2664.12306.

Sobrino, Ignacio, Lucia Rueda, Maria Pilar Tugores, Candelaria Burgos, Miguel Cojan, and Graham J. Pierce. 2020. “Abundance Prediction and Influence of Environmental Parameters in the Abundance of Octopus (Octopus Vulgaris Cuvier, 1797) in the Gulf of Cadiz.” *Fisheries Research* 221 (August 2019): 105382. https://doi.org/10.1016/j.fishres.2019.105382.

Tunney, Tyler D., Stephen R. Carpenter, and M. Jake Vander Zanden. 2017. “The Consistency of a Species’ Response to Press Perturbations with High Food Web Uncertainty.” *Ecology* 98 (7): 1859–68. https://doi.org/10.1002/ecy.1853.

VanDerWal, Jeremy, Luke P. Shoo, Christopher N. Johnson, and Stephen E. Williams. 2009. “Abundance and the Environmental Niche: Environmental Suitability Estimated from Niche Models Predicts the Upper Limit of Local Abundance.” *The American Naturalist* 174 (2): 282–91. https://doi.org/10.1086/600087.

Waldock, Conor, Rick D. Stuart-Smith, Camille Albouy, William W. L. Cheung, Graham J. Edgar, David Mouillot, Jerry Tjiputra, and Loïc Pellissier. 2022. “A Quantitative Review of Abundance-based Species Distribution Models.” *Ecography* 2022 (1): 1–18. https://doi.org/10.1111/ecog.05694.

Walker, Steven C., and Donald A. Jackson. 2011. “Random-Effects Ordination: Describing and Predicting Multivariate Correlations and Co-Occurrences.” *Ecological Monographs* 81 (4): 635–63. https://doi.org/10.1890/11-0886.1.

Wang, Yi, Ulrike Naumann, Dirk Eddelbuettel, John Wilshire, and David Warton. 2022. “Mvabund: Statistical Methods for Analysing Multivariate Abundance Data.” https://cran.r-project.org/package=mvabund.

Weller, Bridget E., Natasha K. Bowen, and Sarah J. Faubert. 2020. “Latent Class Analysis: A Guide to Best Practice.” *Journal of Black Psychology* 46 (4): 287–311.

https://doi.org/10.1177/0095798420930932.

Wenger, Seth J., and Julian D. Olden. 2012. “Assessing Transferability of Ecological Models: An Underappreciated Aspect of Statistical Validation.” *Methods in Ecology and Evolution* 3 (2): 260–67. https://doi.org/10.1111/j.2041-210X.2011.00170.x.

Wright, David Hamilton. 1991. “Correlations between Incidence and Abundance Are Expected by Chance.” *Journal of Biogeography* 18 (4): 463–66. https://doi.org/10.2307/2845487.

Zou, Hui, and Hao Helen Zhang. 2009. “On the Adaptive Elastic-Net with a Diverging Number of Parameters.” *The Annals of Statistics* 37 (4): 1733–51. https://doi.org/10.1214/08-AOS625.

Tables

Table 1. Variable symbols and indexes, and their associated values and distributions used in the simulation study. Bold letters indicate that the variable is a vector or a matrix.

Variable name	Variable	Values
A	Abundance	0 to ∞
S, s	Number of species, species index	{10, 20, 30}
U, u	Number of landscapes, landscape index	30
J, j	Number of sites, site index	
E	Number of environmental variables	3
$b_{0,s,u}$	Intercept for species s and landscape u	Uniform(-2.4, 1.2)
$b_{1,s,u}$ to $b_{E,s,u}$	Slopes for species s, landscape u and environmental variables 1 to E	Uniform(-0.8, 0.8)
$X_{1,u,j}$ to $X_{E,j,u}$	Environmental variables 1 to E for site j of landscape u	Normal(0,1)
L	Number of latent variables	3
X	Environmental variable	
Z	Latent variable	

Table 2. All models considered in this study based on combinations of environmental variables and community composition (latents). The best model is expected to be the true model considering all environmental variables. **A** refers to the abundance matrix, $\mathbf{X}_1$ to $\mathbf{X}_3$ to the environmental variables, and $\mathbf{Z}_1$ to $\mathbf{Z}_3$ to the community composition (latent variables).

Variables included	Model specification	Regression formula
Environmental variables	3 environmental variables	$\boldsymbol{A} \sim \boldsymbol{X}_1 + \boldsymbol{X}_2 + \boldsymbol{X}_3$
	2 environmental variables	$\boldsymbol{A} \sim \boldsymbol{X}_1 + \boldsymbol{X}_2$
		$\boldsymbol{A} \sim \boldsymbol{X}_1 + \boldsymbol{X}_3$
		$\boldsymbol{A} \sim \boldsymbol{X}_2 + \boldsymbol{X}_3$
	1 environmental variable	$\boldsymbol{A} \sim \boldsymbol{X}_1$
		$\boldsymbol{A} \sim \boldsymbol{X}_2$
		$\boldsymbol{A} \sim \boldsymbol{X}_3$
Environmental variables and community composition	2 environmental variables and community composition	$\boldsymbol{A} \sim \boldsymbol{X}_1 + \boldsymbol{X}_2 + \boldsymbol{Z}_1{:}\boldsymbol{Z}_3$
		$\boldsymbol{A} \sim \boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{Z}_1{:}\boldsymbol{Z}_3$
		$\boldsymbol{A} \sim \boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{Z}_1{:}\boldsymbol{Z}_3$
	1 environmental variable and community composition	$\boldsymbol{A} \sim \boldsymbol{X}_1 + \boldsymbol{Z}_1{:}\boldsymbol{Z}_3$
		$\boldsymbol{A} \sim \boldsymbol{X}_2 + \boldsymbol{Z}_1{:}\boldsymbol{Z}_3$
		$\boldsymbol{A} \sim \boldsymbol{X}_2 + \boldsymbol{Z}_1{:}\boldsymbol{Z}_3$

Table 3. Metrics used for assessing model predictive performance based on presence-absence and abundance of target species. *J* represents the number of sites, $\boldsymbol{A_s}$ the true abundance of the (target) species, $\boldsymbol{P_s}$ the predicted abundance, TP the true positives, FP the false positives, TN the true negatives, and FN the false negatives. Bold letters indicate that the variable is a vector or a matrix. The True Skill Statistic (TSS), sensitivity, and specificity are calculated for all sites of the landscape. Having evaluated the presence-absence predictions of the models and to avoid artificially inflating the error rate of the abundance metrics, the Mean Absolute Percentage Error (MAPE), Root Mean Squared Percentage Error (RMSPE), Relative Mean Squared Error (RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), and Root Mean Ratio Percentage Error (RMRPE) are calculated for sites where the species is truly present (i.e., abundance of 1 or more).

Metric	Equation
TSS	$TSS = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1$
Sensitivity	$Sensitivity = \frac{TP}{TP + FN}$
Specificity	$Specificity = \frac{TN}{TN + FP}$
MAPE	$MAPE = \frac{1}{J}\sum_{s} \frac{\lvert \boldsymbol{A}_s - \boldsymbol{P}_s \rvert}{\boldsymbol{A}_s} \times 100$
RMSPE	$RMSPE = \sqrt{\frac{1}{J}\sum_{s} \left(\frac{\boldsymbol{A}_s - \boldsymbol{P}_s}{\boldsymbol{A}_s}\right)^2} \times 100$
RMSE	$RMSE = \sqrt{\frac{1}{J}\sum_{s} \frac{(\boldsymbol{A}_s - \boldsymbol{P}_s)^2}{\boldsymbol{A}_s^2}} \times 100$
SMAPE	$SMAPE = \frac{1}{J}\sum_{s} \frac{\lvert \boldsymbol{A}_s - \boldsymbol{P}_s \rvert}{\lvert \boldsymbol{A}_s \rvert + \lvert \boldsymbol{P}_s \rvert} \times 100$
RMRPE	$RMRPE = \sqrt{\frac{1}{J}\sum_{s} \log\left(\frac{\boldsymbol{P}_s}{\boldsymbol{A}_s}\right)^2} \times 100$

Figure captions

Figure 1. The rationale underlying our model framework and simulation workflow to assess its performance. First, species abundances were simulated for all species (top left panel) as a function of multiple environmental factors. In this example, two environmental variables were used to simulate species abundances ($\mathbf{X}_1$ and $\mathbf{X}_2$; bottom left panel). Species abundances are then transformed into presence-absence data and used to derive latent variables (bottom left panel). Here, only one latent variable is presented for simplicity. allowing one to more easily it association with the abundances of the original simulated species. Variation in species abundances (target species) across sites is then modeled against latent and environmental variables or reduced combinations (e.g., removing an environmental variable and assess the conditions that affect latent performances), depending on specific simulation scenarios. The model can produce either abundance or presence-absence predictions for each site. The black rectangular outline highlights the target species (species 10) that the model aims at predicting.

Figure 2. The density of average species abundance across sites within each landscape. For each landscape, we calculated the average abundance of each species and plotted the density of abundances in each of the 30 landscapes (grey lines). We also plotted the density of abundances across all landscapes to represent the average landscape (black line). The red line is a reference line indicates the probability density function of a log-normal distribution with the same log-mean and log-standard deviation of the average abundance distribution across replicates.

Figure 3. Variation in adjusted R^2 as a function of the number of latent variables used, as well as the true dimensions of the environment and the number of species in the landscape. Here we used 500 sites, and variations according to other number of sites are presented in Appendix S1: Figure S2. Colors represent the varying number of species in the landscape, and

each panel indicates the true dimension of the environment (i.e., number of environmental variables used to simulate the abundance of a given target species).

Figure 4. Ratio TSS and delta TSS for each model and bin of species occurrence percentiles. The ratio TSS was averaged across all landscapes and replicates per model and species, with species binned by percentile of occurrence (percentage of sites occupied) and divided by the TSS of the oracle model. A value of 1 for the ratio TSS indicates an identical performance between the model and the oracle model, while a value below 0 represents a performance similar to that of a random model. To improve contrast between colors, we confined the color scheme between 0 and 1. Any value below 0 indicates a prediction of presence-absence no better than a random model, and any value above 1 a better prediction than the oracle model. The environment panel represents models containing only environmental variables, while the latent panel is for models containing latent variables (mix of latent and environmental predictors); the models were then ordered from bottom to top as fewest to the greatest number of environmental variables included and sorted by coefficients relative to each environmental variable (see Methods for more information, note that the "mid" model refers to the "intermediate" model). The delta TSS was measured as the TSS of the model with environmental variables minus the TSS of the model with the same combination of environmental variables and latent variables. A negative value indicates that the model with latent predicts the presence-absence of the species better than the model containing only environmental variables.

Figure 5. Correlation between the metrics studied (TSS, sensitivity, and specificity) depending on the model across species occurrence percentiles. The vertical panels indicate the different metrics, with models represented in different colors. The oracle model refers to the model using the true environmental coefficients, while the other models were fitted using all environmental variables (benchmark) or latent variables (latent). The True Skill Statistic

(TSS) measures the difference between sensitivity and specificity of the model and ranges from -1 to +1. A score of +1 indicates a perfect agreement between the predictions of the model and the true presence-absence, while a score of 0 or less represents a performance no better than random. Sensitivity represents the ability to correctly classify a species as "present", while specificity represents the ability to correctly classify a species as "absent". Their values can be interpreted as a percentage, with values of 1 indicating perfect classification of either presence or absence, and values of 0.5 no better than random. Here we used 500 sites, and variations according to other number of sites are presented in Appendix S1: Figure S3.

Figure 6. Ratio Mean Absolute Percentage (MAPE) and delta MAPE are presented for each model and bins of species abundance percentiles. The MAPE is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance and divided by the MAPE of the oracle model to derive the ratio MAPE. The environment panel represents models containing only environmental variables, while the latent panel depicts models containing latent predictors. The models are then ordered from bottom to top, from the fewest to the greatest number of environmental variables included and sorted by coefficients relative to each environmental variable. See Methods for more information, note that the "mid" model refers to the "intermediate" model. Delta MAPE was measured as the MAPE of the model with environmental variables only minus the MAPE of the model with the same combination of environmental and latent predictors. A positive value indicates that the model with latent predicts the abundance of the species better than the model containing only environmental variables.

Figures

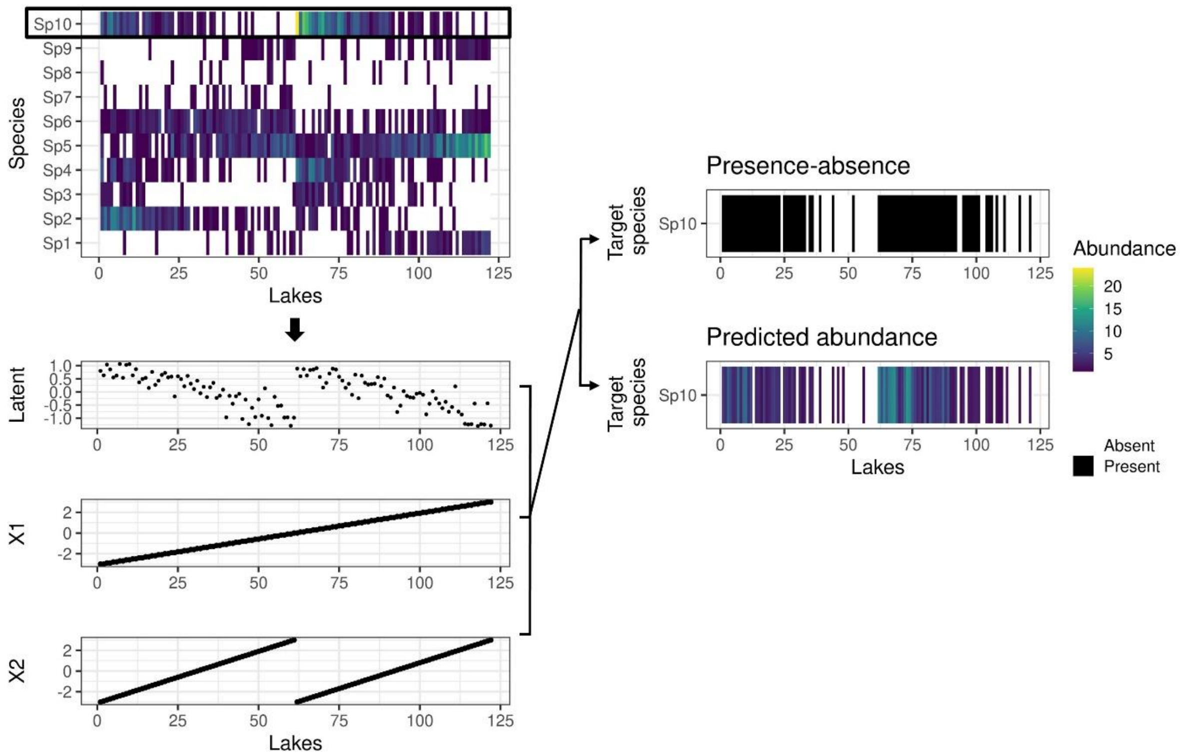


Figure 1.

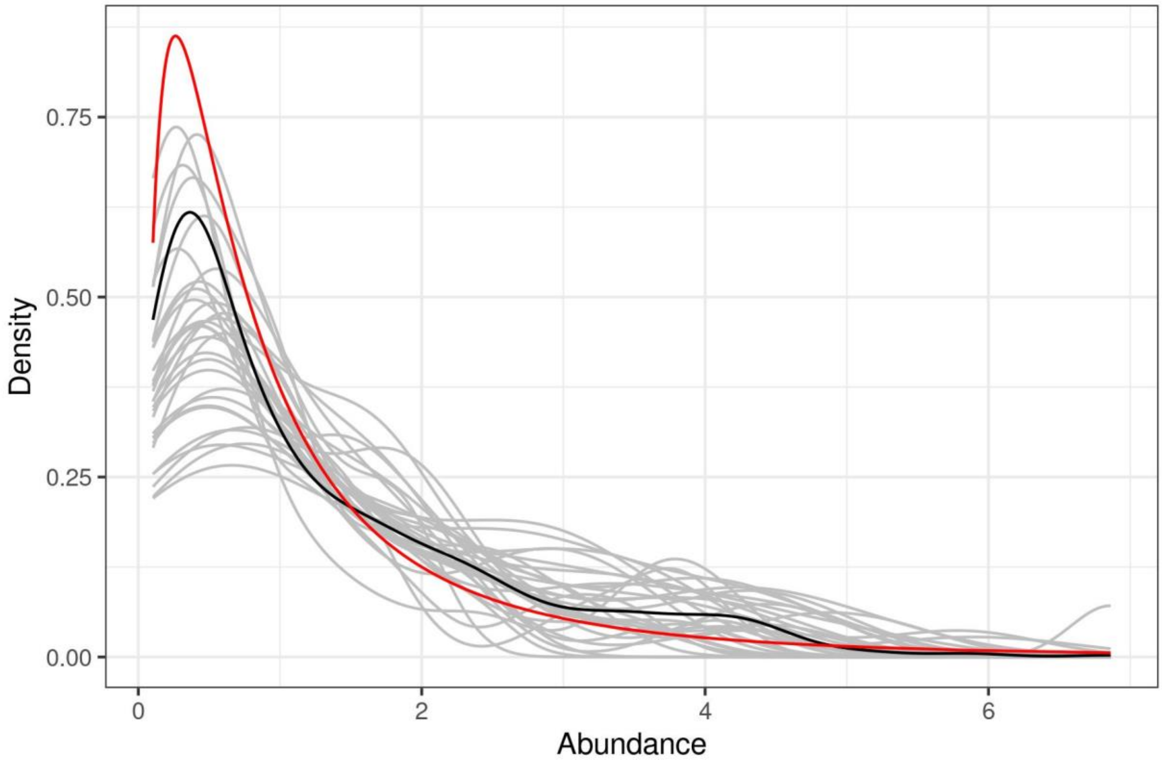
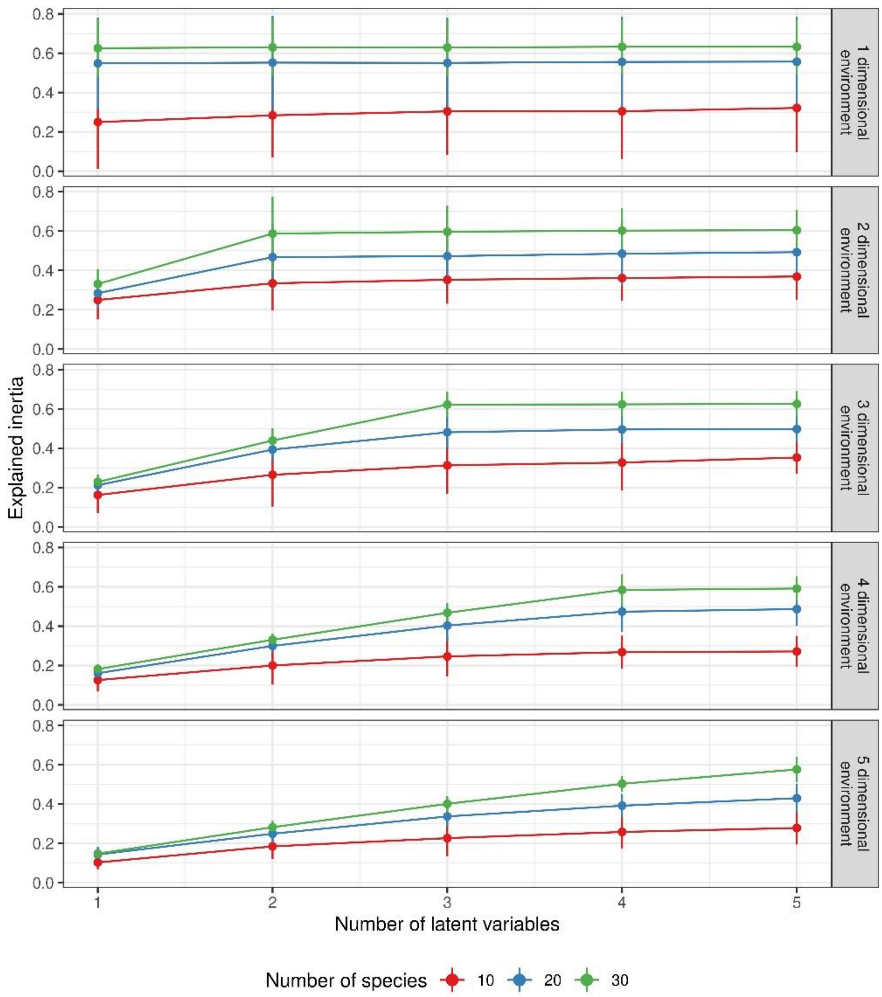
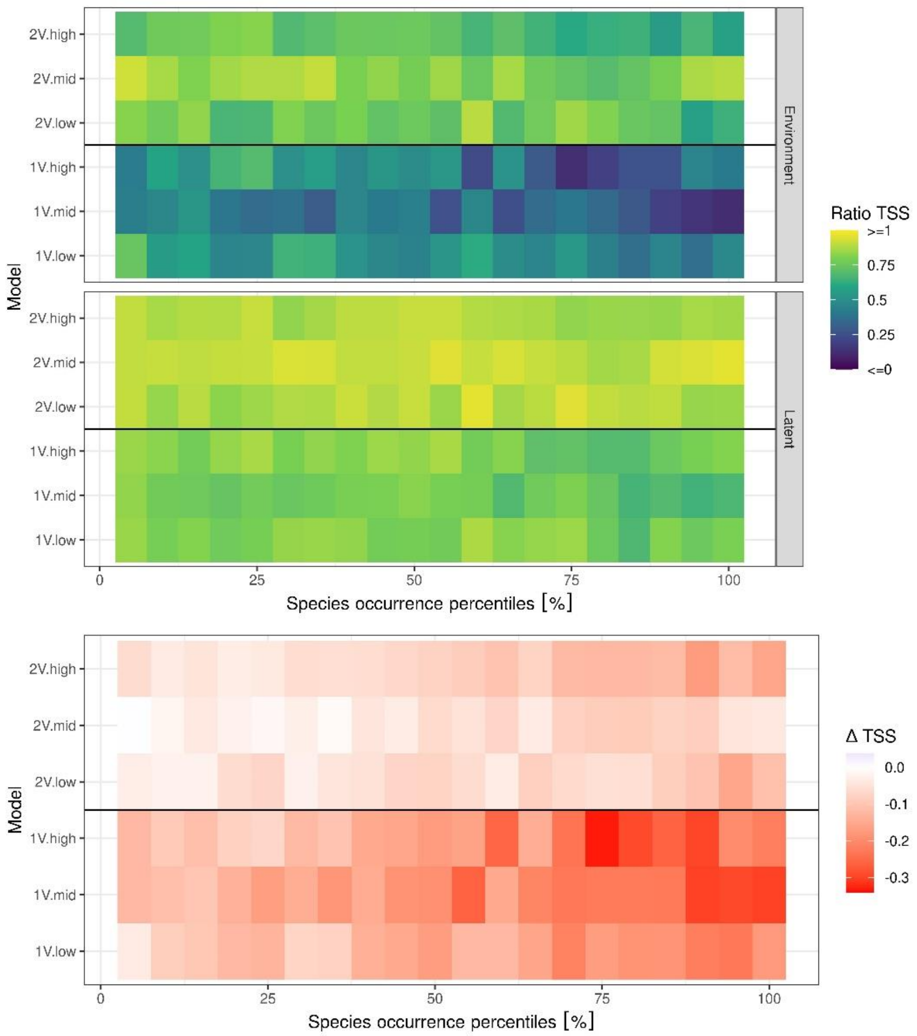


Figure 2.



Explained inertia
Number of latent variables
1 dimensional environment
2 dimensional environment
3 dimensional environment
4 dimensional environment
5 dimensional environment
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
Number of species
10
20
30

Figure 3.



942

943 **Figure 4.**

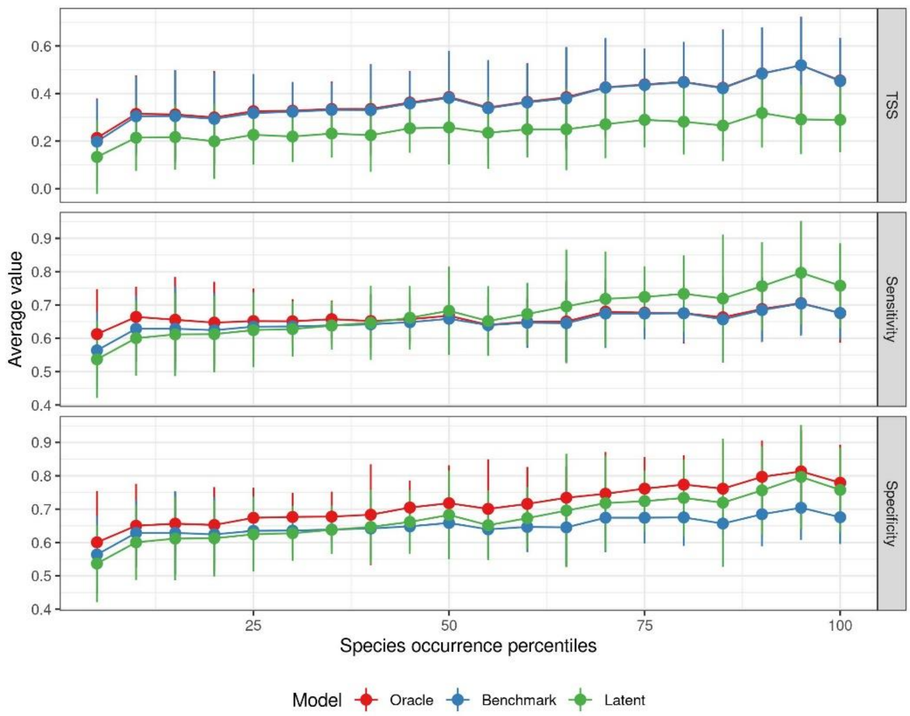

TSS
Sensitivity
Specificity
Average value
Species occurrence percentiles
0.0
0.2
0.4
0.6
0.4
0.5
0.6
0.7
0.8
0.9
25
50
75
100
Model
Oracle
Benchmark
Latent

Figure 5.

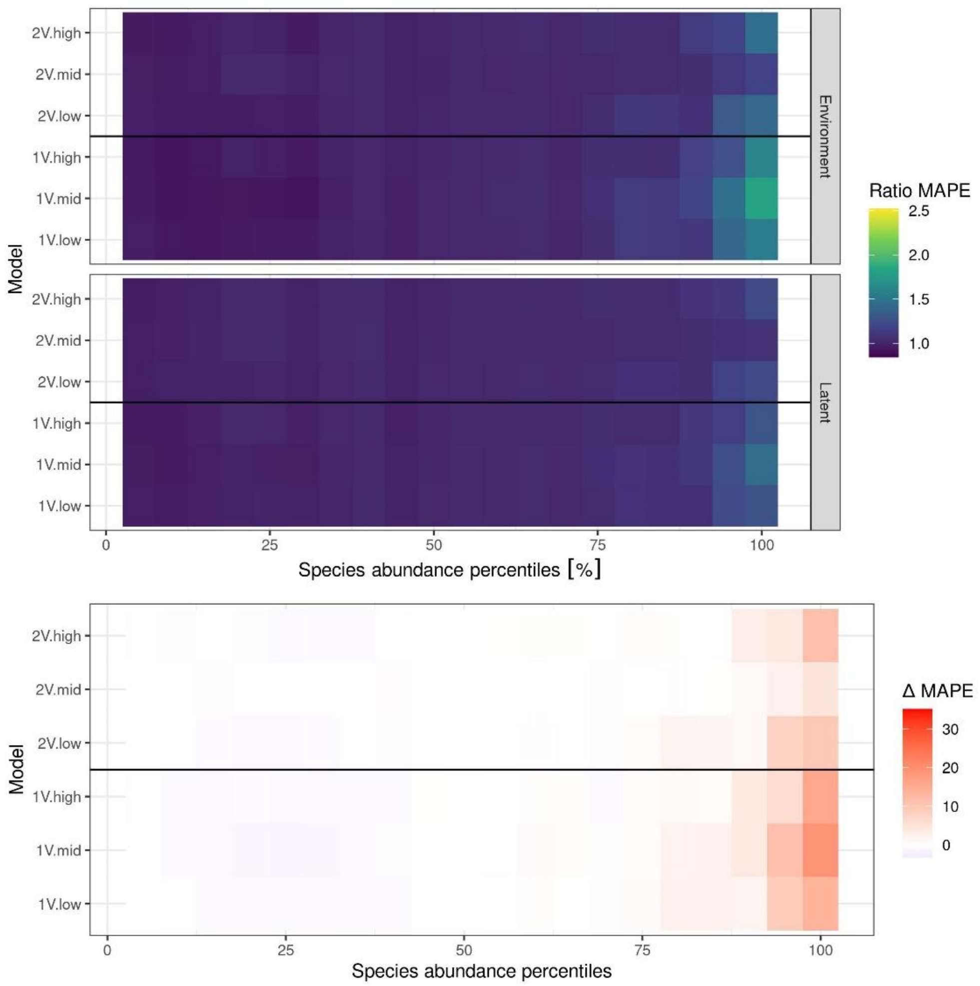

2V.high
2V.mid
2V.low
1V.high
1V.mid
1V.low
Model
Environment
Latent
Ratio MAPE
2.5
2.0
1.5
1.0
0
25
50
75
100
Species abundance percentiles [%]
Δ MAPE
30
20
10
0
Species abundance percentiles

Figure 6.

Supporting Information for review and publication

Journal name: Ecological Applications (ESA)

Manuscript type: Article

Manuscript title: Advancing single species abundance models: robust models for predicting abundance using co-occurrence from communities

Author names: Aliénor Stahl[1], Eric J. Pedersen[1] and Pedro R. Peres-Neto[1]

Affiliations: 1: Concordia University, Department of Biology, Montreal, Canada

Corresponding author: Aliénor Stahl, alienor.stahl@gmail.com

Appendix S1

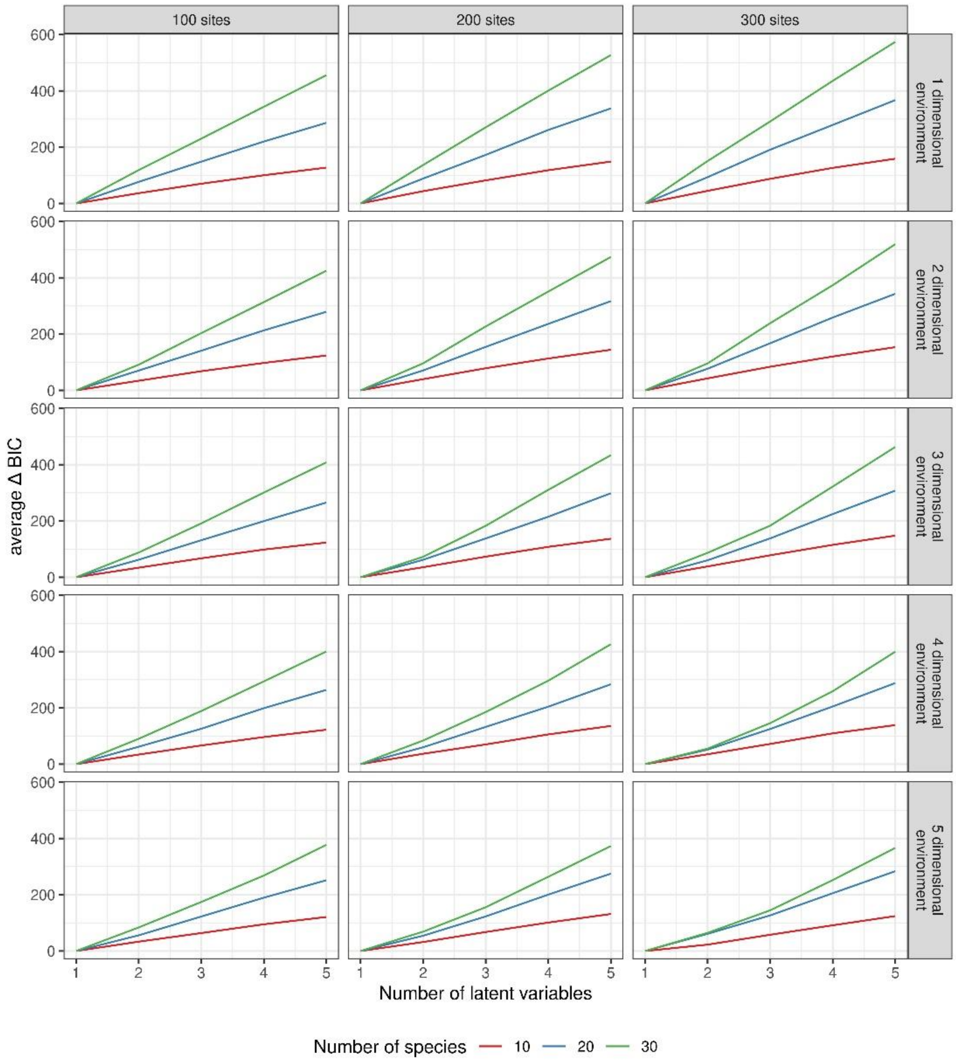


Figure S1. Variation in delta BIC as a function of the number of latent variables used, as well as the true dimensions of the environment, the number of species in the landscape and the number of sites. Horizontal panels represent the number of sites, and each vertical panel indicates the true dimension of the environment (i.e., number of environmental variables used to simulate the abundance of a given target species). Colors represent the varying number of species in the landscape. The delta BIC is calculated as the BIC of the model minus the BIC of the best model for the ongoing simulation.

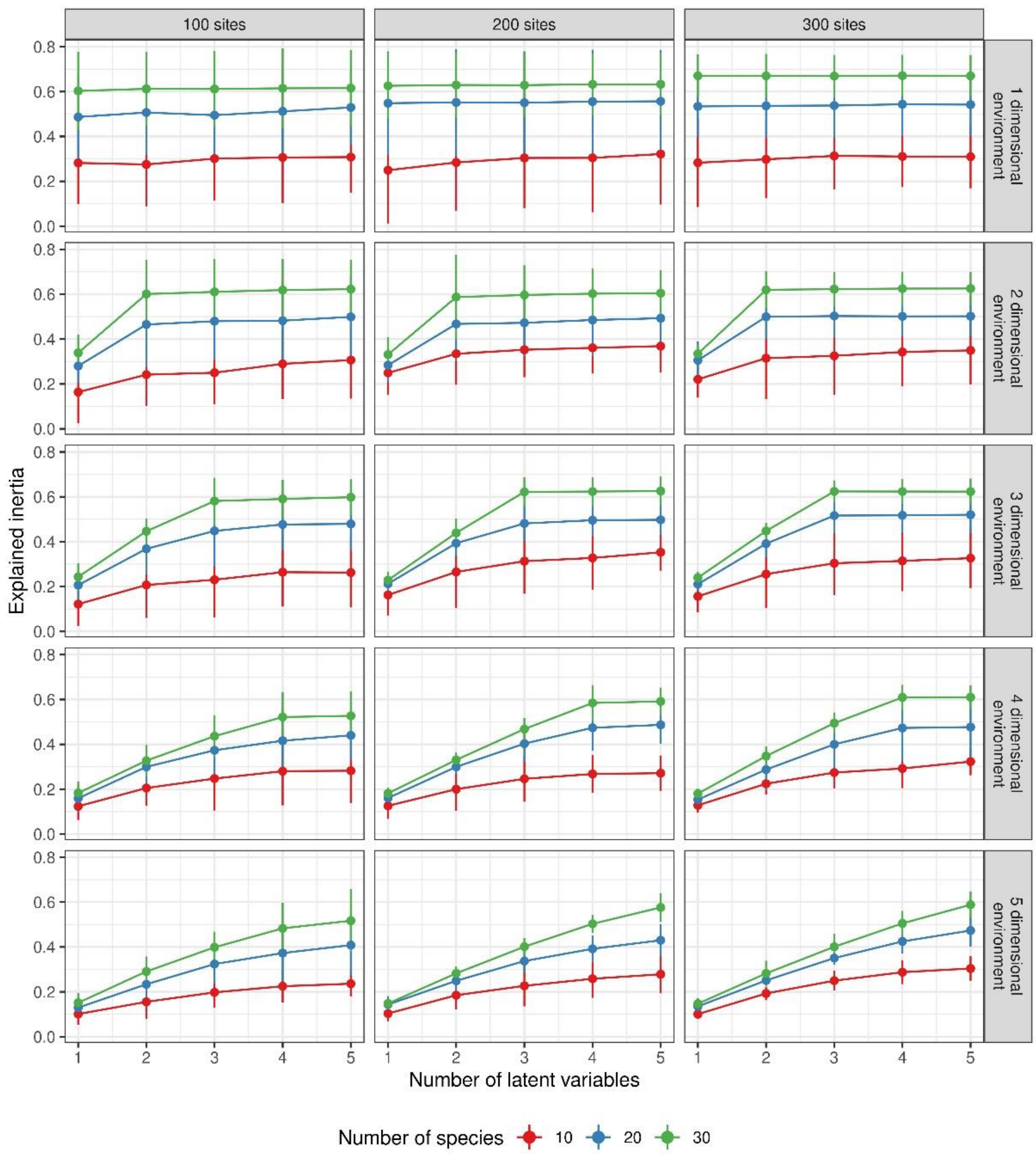


Figure S2. Variation in adjusted R^2 as a function of the number of latent variables used, as well as the true dimensions of the environment, the number of species in the landscape and the number of sites. Horizontal panels represent the varying number of sites, and each vertical panel indicates the true dimension of the environment (i.e., number of environmental variables used to simulate the abundance of a given target species). Colors represent the varying number of species in the landscape.

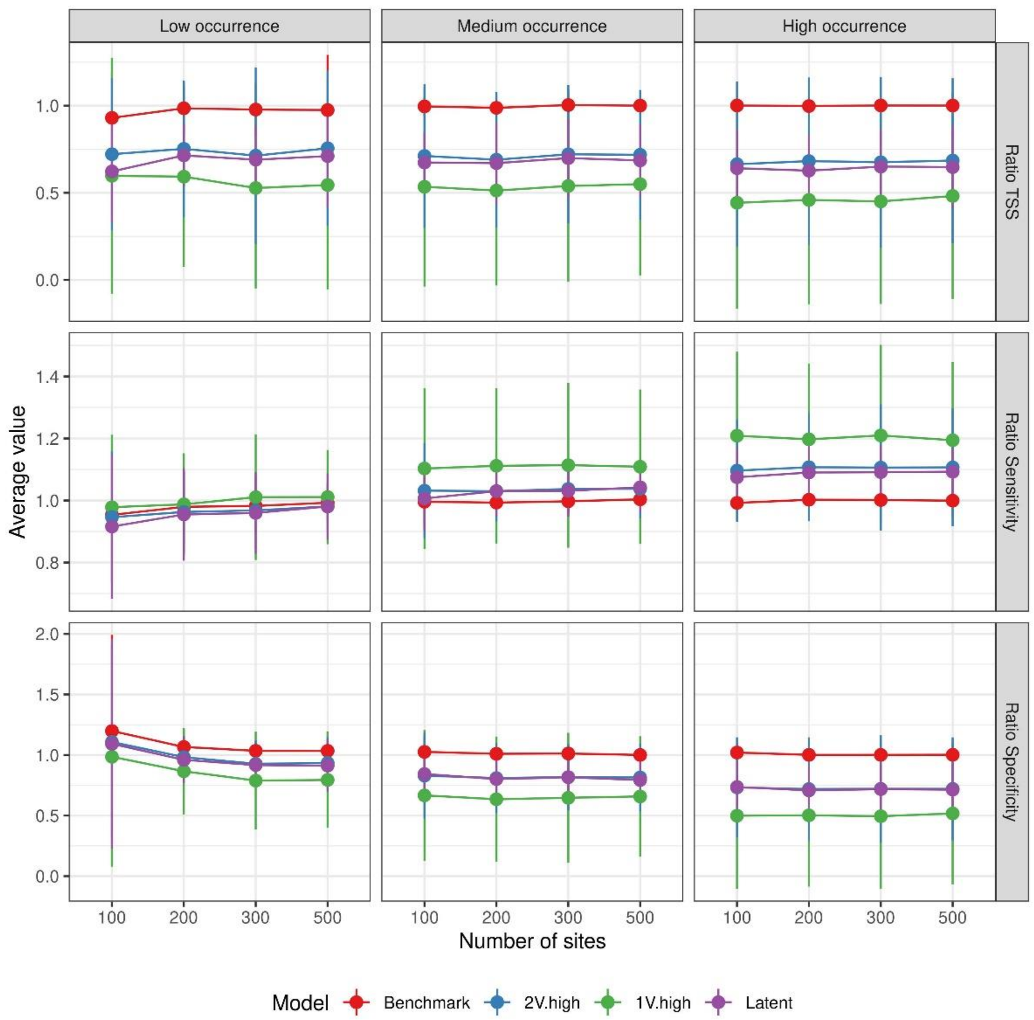


Figure S3. Average value of the studied metrics (Ratio TSS, ratio sensitivity, and ratio specificity) depending on the number of sites used to fit the models, the model used, and the occurance of species. Horizontal panels represent the different occurence: species with low, medium and high occurrence corresponding respectively to bins of 15, 50, and 80 percentiles of occurrence. Vertical panels indicate the metrics considered, with the models represented in different colors. The ratio metric is calculated as the metric for the predictions of a model for a species of the landscape divided by the same metric calculated for the oracle model. For the ratio TSS, a score of 1 indicates a perfect agreement between the predictions of the considered model and the oracle model, while a score of 0 or less represents a performance no better than random. For the ratio sensitivity, it represents the ability to correctly classify a species as “present”, while the ratio specificity represents the ability to correctly classify a species as “absent”. For both metrics, values above 1 indicate a better performance than the oracle model and values below 1 indicate a lesser performance. The benchmark model refers to the model containing all environmental variables, 2V.high the model with the two environmental variables with the highest coefficients, 1V.high the model with the environmental variable with the highest coefficient, and Latent the model containing the latent variables.

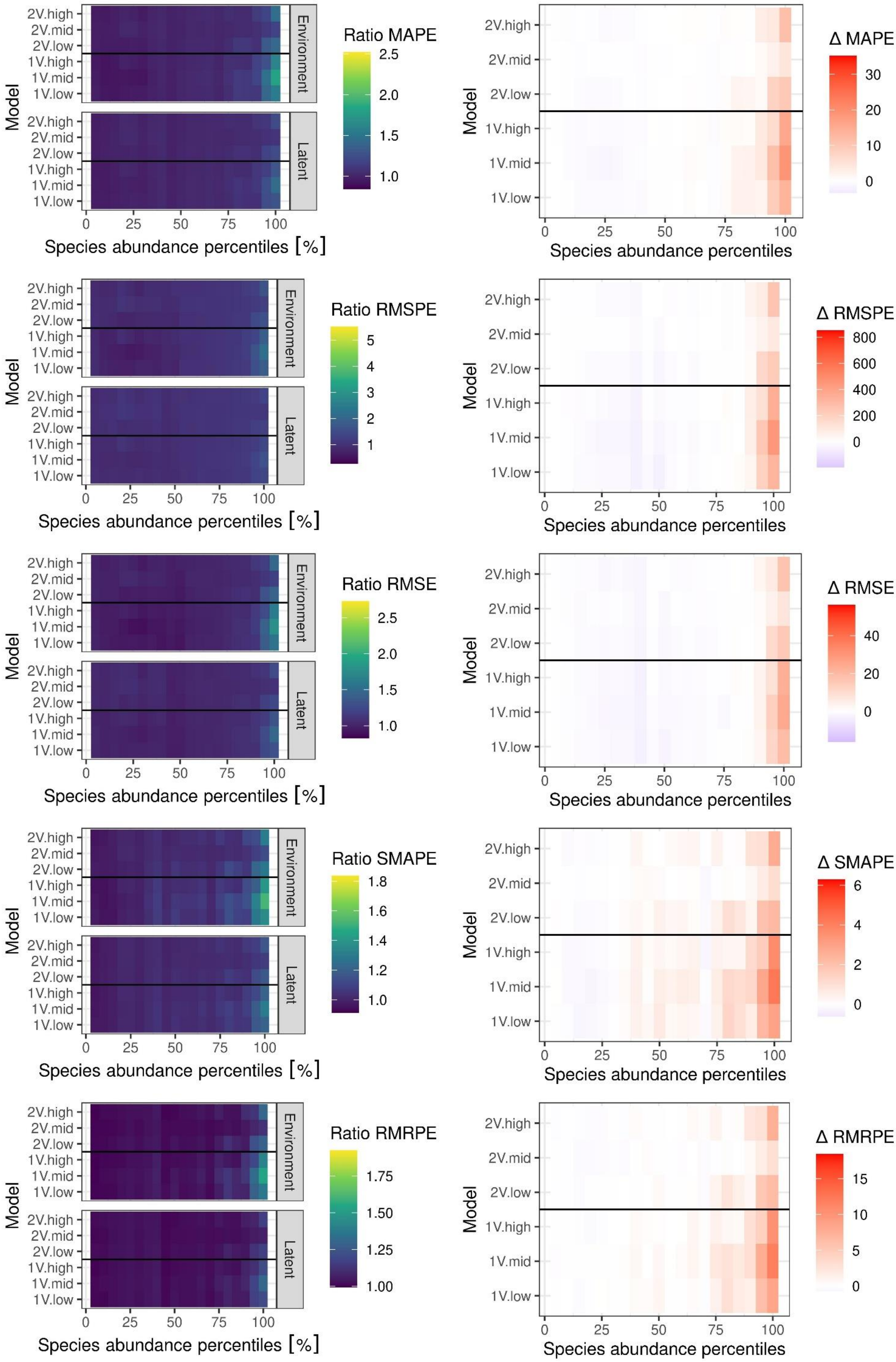

Model
2V.high
2V.mid
2V.low
1V.high
1V.mid
1V.low
Environment
Latent
0
25
50
75
100
Species abundance percentiles [%]
Ratio MAPE
2.5
2.0
1.5
1.0
Δ MAPE
30
20
10
0
Species abundance percentiles
Ratio RMSPE
5
4
3
2
1
Δ RMSPE
800
600
400
200
0
Ratio RMSE
2.5
2.0
1.5
1.0
Δ RMSE
40
20
0
Ratio SMAPE
1.8
1.6
1.4
1.2
1.0
Δ SMAPE
6
4
2
0
Ratio RMRPE
1.75
1.50
1.25
1.00
Δ RMRPE
15
10
5
0

Figure S4. Abundance metrics and the comparison of performance between environmental models and latent models measured as delta metrics. Each metric is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance, and divided by the metric of the oracle model to give the ratio metric. The environment panel represents models containing only environmental variables, while the latent panel depicts models containing latent predictors. The models are then ordered from bottom to top, from the fewest to the greatest number of environmental variables included and sorted by coefficients relative to each environmental variable. See Methods for more information, note that the "mid" model refers to the "intermediate" model. The delta metric was measured as the metric of the model with environmental variables only minus the metric of the model with the same combination of environmental and latent predictors. A positive value indicates that the model with latent predicts the abundance of the species better than the model containing only environmental variables.

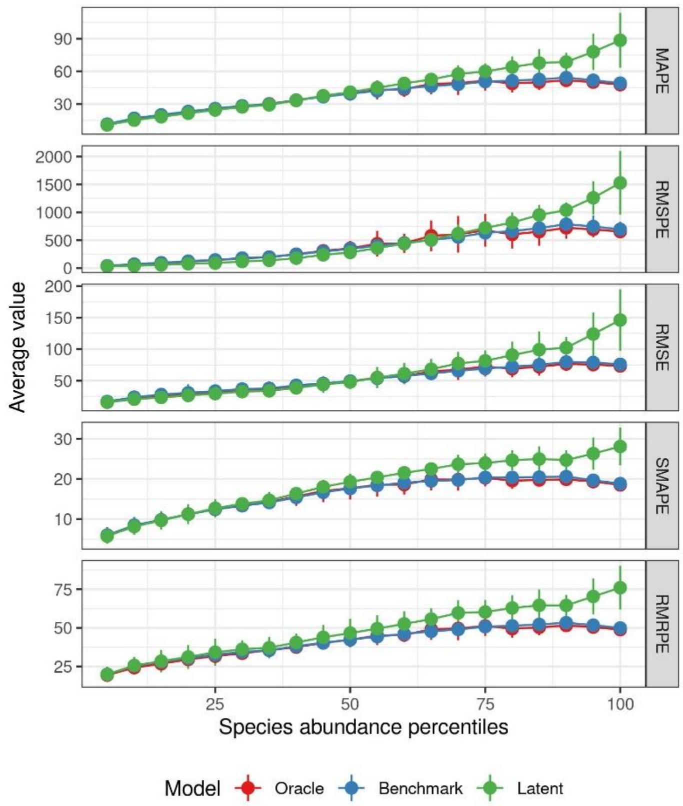


Figure S5. Correlation between the metrics studied (MAPE, RMSPE, RMSE, SMAPE, and RMRPE) depending on the model across species abundance percentiles. The vertical panels indicate the different metrics, with models represented in different colors. Each metric is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance. The oracle model refers to the model using the true environmental coefficients while the other models were fitted using all environmental variables (benchmark) or latent variables (latent).

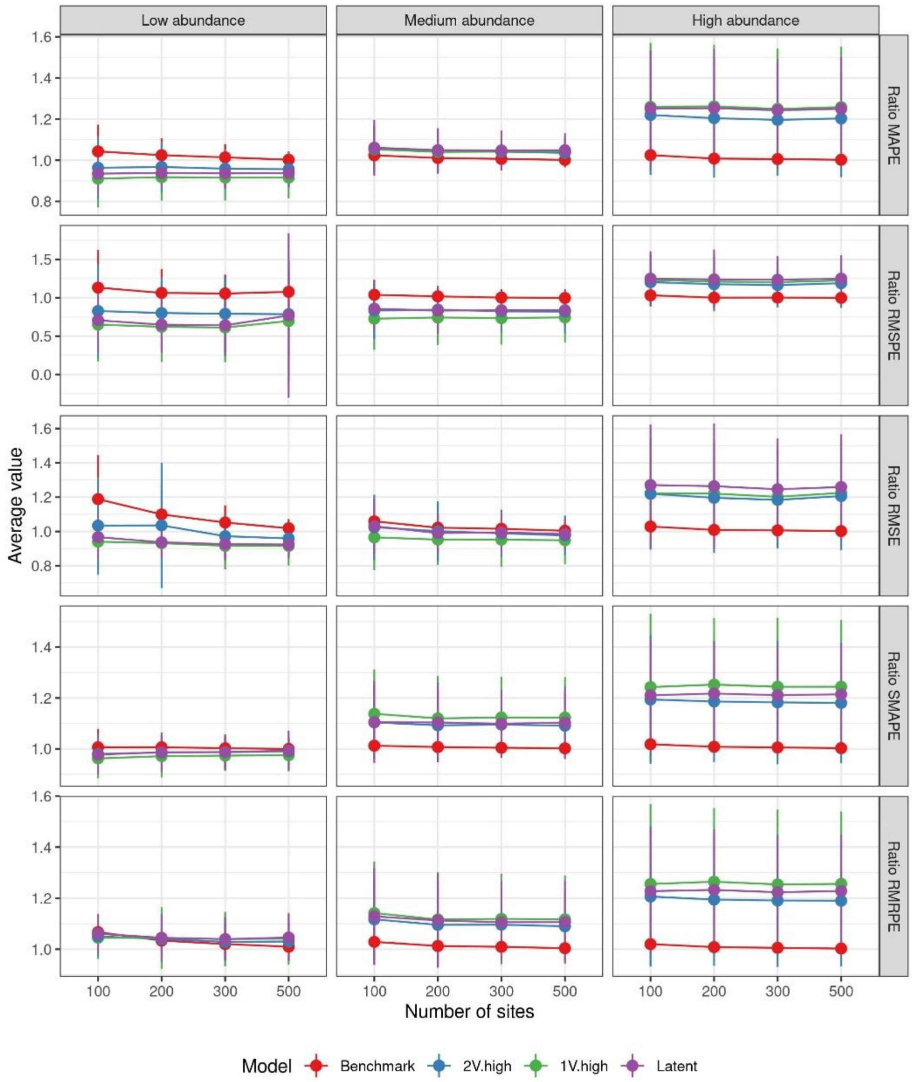


Figure S6. Average value of the studied metrics (MAPE, RMSPE, RMSE, SMAPE, and RMRPE) depending on the number of sites used to fit the models, the model used, and the abundance of species. Horizontal panels represent the different abundances: species with low, medium and high occurrence corresponding respectively to bins of 15, 50, and 80 percentiles of occurrence. Vertical panels indicate the metrics considered, with the models represented in different colors. Each metric is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance and divided by the metric of the oracle model to give the ratio metric. The benchmark model refers to the model containing all environmental variables, 2V.high the model with the two environmental variables with the highest coefficients, 1V.high the model with the environmental variable with the highest coefficient, and Latent the model containing the latent variables.