Canadian Journal of
Fisheries and
Aquatic Sciences

Research Article

Advancing single-species abundance models by leveraging multi-species data to reveal lake-specific patterns for fisheries predictions

Aliénor Stahl, **Eric J. Pedersen**, and **Pedro R. Peres-Neto**

Concordia University, Department of Biology, Montreal, Canada

Corresponding author: **Aliénor Stahl** (email: alienor.stahl@gmail.com)

Abstract

Predicting species abundance is critical for understanding ecological dynamics and guiding conservation and management strategies. Traditional species abundance models rely on environmental variables and the presence or absence of key species, but often overlook community context and unmeasured environmental variation. Community composition can serve as a proxy for both unobserved environmental variables and biotic interactions influencing focal species. Here, we tested whether incorporating community composition via latent variables improves abundance predictions of sport fishing using a large-scale dataset. We assessed how latent variable selection and lake characteristics influence model accuracy in predicting abundance across species. Our results show that low-abundance species were better predicted by models based solely on environment, while high-abundance species benefited from latent variables. Lake contribution to accuracy was correlated among species with similar occurrence, but unrelated to environmental characteristics. Model performance varied by species, with no consistent association with trophic level, occurrence, or abundance. These findings underscore the need to tailor models to species-specific contexts and integrate community composition into abundance modelling.

Key words: community composition, abundance, co-distribution, co-occurrence, latent, prediction

Introduction

Fish species abundance is a fundamental indicator of population health and viability within aquatic ecosystems. It provides crucial information on detectability, local ecological influence, and a species' contribution to community dynamics, thereby informing conservation priorities and sustainable fisheries management. Understanding spatial patterns of species abundance is particularly important: it enables the identification of areas suitable for harvest or protection and supports efficient allocation of management efforts across regions (Degnbol and Jarre 2004). Such knowledge is increasingly critical for policymakers, conservation practitioners, and resource managers seeking to balance ecological sustainability with societal needs. Despite its importance, accurately estimating fish species abundance across broad spatial scales remains a substantial challenge. Field-based surveys are logistically demanding, costly and time-consuming (Yoccoz et al. 2001; Dickinson et al. 2010; Lindenmayer and Likens 2010), often yielding only partial representations of abundance distributions given the vastness and heterogeneity of aquatic ecosystems. These spatial and logistical constraints make it challenging to obtain consistent and representative data across regions, habitats, and species. Moreover, ethical considerations, especially those related to fish capture, handling, and disturbance, are becoming increasingly relevant in contemporary monitoring programs and further constrain sampling intensity in many jurisdictions.

Sampling constraints often limit both the frequency and spatial extent of abundance assessments (e.g., across multiple lakes, streams, or entire watersheds), making it challenging to generate comprehensive data over large geographic areas, long time periods (Jackson and Harvey 1997), and diverse species assemblages. These limitations are especially problematic when rapid conservation or management decisions are needed. To overcome these challenges, fisheries researchers commonly reduce sampling intensity (e.g., number of waterbodies) and rely on predictive models to estimate abundance across broader regions. Species abundance models (SAMs; Waldock et al. 2022) typically incorporate local and regional environmental predictors such as temperature, habitat quality, and substrate to estimate abundance (Lek et al. 1996; Brosse et al. 1999; VanDerWal et al. 2009; Boyce et al. 2016; Sobrino et al. 2020). These variables are generally easy to measure and can capture broad spatial and temporal trends in abundance in space and time. However, while these models can yield useful estimates, they often lack the precision and accuracy needed for fine-scale management and may overlook complex biotic interactions, such as competition and predation, that also shape abundance distributions (Mack et al. 2000; MacKenzie et al. 2002; Gaston

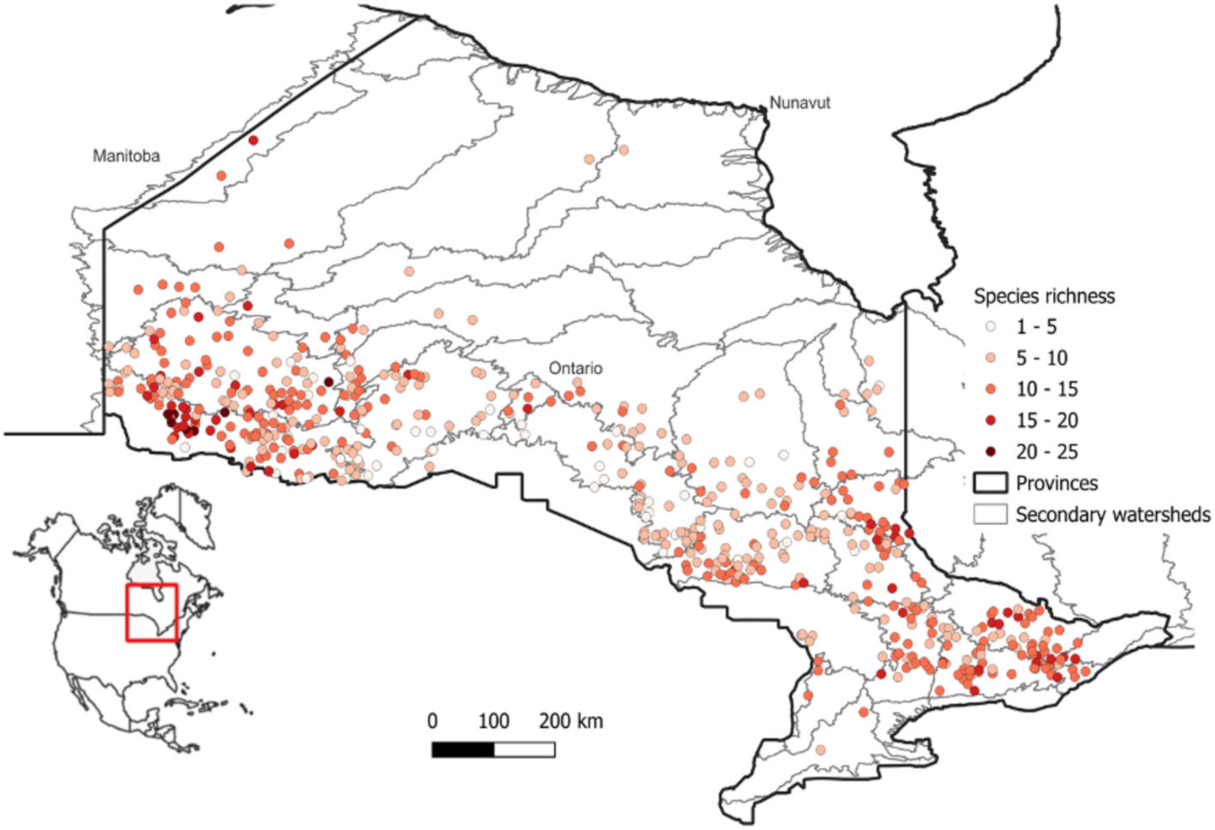
2003). Within community ecology, these interactions represent biotic filters in the broader framework of environmental filtering theory, in which both abiotic conditions and species interactions jointly structure local communities (Kraft et al. 2015). In some contexts, such as biological invasions or shifts in food web dynamics, biotic filters can even explain more variation in abundance than abiotic factors alone (Pagnucco and Ricciardi 2015). This underscores a persistent need for predictive frameworks that integrate environmental data with quantitative representations of biotic context to better capture the multifaceted drivers of species abundance.

Stahl et al. (2026) proposed a framework that enhances species abundance predictions by integrating environmental predictors with co-occurrence data. Although earlier SAMs have included presence–absence information as predictors, they typically focused on species with well-known interactions with the target species, such as predator-prey or strong competitive relationships (Boulangeat et al. 2012; Lewis et al. 2017; Olkeba et al. 2020). In contrast, the approach of Stahl et al. (2026) incorporated presence–absence data for the entire local community as predictors of a target species' abundance, providing a more comprehensive and easily obtainable alternative to intensive abundance sampling. This framework offers at least two key advantages. First, patterns of species co-occurrence can serve as proxies for unmeasured environmental variables, thereby reducing bias arising from incomplete environmental sampling. Second, it allows the integration of co-occurrence-based interaction networks at both local and regional scales, using these networks to predict abundance variation in focal (target) species. The framework uses Gaussian copulas to derive latent variables from species covariation, enabling complex multispecies patterns to be captured efficiently and incorporated into predictive models (Popovic et al. 2018).

Latent variables were originally introduced to reduce dimensionality in community data (e.g., indirect gradient analysis), but they have since been widely adapted to represent unobserved ecological factors, such as hidden environmental gradients or aggregated biotic interactions, derived from species covariation (see Walker and Jackson 2011). Estimated from co-occurrence (presence–absence) matrices, these latent factors aim to capture as much variation in community composition as possible. When community structure is primarily shaped by species' responses to environmental gradients and local interspecific interactions, these latent variables can effectively serve as proxies for missing ecological predictors in abundance models. Stahl et al. (2026) demonstrate that copula-based latent variables reliably recover unmeasured environmental structure in simulated communities. In simulations where species abundances were generated as linear functions of environmental conditions and location-specific process error, without including species interactions or nonlinear environmental responses, latent variables accurately captured the underlying environmental gradients driving abundance patterns. These improvements were consistent across a wide range of scenarios, underscoring the robustness and generality of the framework. In empirical systems, latent variables may also represent the influence of interspecific interactions (e.g., competition, predation, and mutualism), thereby encoding a broader suite of ecological processes that shape abundance. This dual capacity to summarize both environmental and biotic information makes latent-variable approaches particularly promising for improving the accuracy and robustness of SAMs. Here, we apply this latent variables predictive framework to a large empirical dataset of lake fish communities, where limited dispersal among lakes means that local abundance patterns are primarily driven primarily by within-lake environmental conditions and species interactions—factors that can be effectively summarized by latent variables.

Here, we apply the framework developed in Stahl et al. (2026), to a large landscape-scale fish abundance data set, encompassing nearly 600 lakes and a wide range of environmental gradients. Our primary objective is to evaluate whether the inclusion of latent variables improves the prediction of species abundance in real-world ecosystems, where species interactions and habitat specificity are key drivers. We focus on predicting sport fish abundances due to their ecological importance (e.g., large biomass), cultural and economic value, and heightened sensitivity to fishing pressure. As central targets of fisheries management, sport fishes also provide a practical context in which to assess the utility of our modelling approach for informing conservations and resource management strategies. To evaluate how different components of the fish community contribute to predicting sport fish abundances, we developed three modelling scenarios: (1) latent variables derived solely from other sport fishes, (2) latent variables derived from nonsport fish species, and (3) latent variables based on the full community, including both sport and nonsport fishes. The groups were structured to reflect management's varying interests. For instance, if a particular group of species is identified to be important for predicting the abundance of a target species, this would reinforce the rationale for incorporating those species into management strategies focused on that target. To evaluate model performance, we developed a suite of novel assessment tools that examine both species-level predictions and community-level patterns, providing insights that will also benefit future users of our species abundance modelling framework. First, we assessed which lake types most strongly influenced predictive performance and whether these lakes represent rare or common environmental conditions and species compositions, thereby informing the generalizability of our models across diverse ecological contexts. Second, we analysed shared patterns in species-specific predictive errors, as correlated errors may indicate that species respond to similar interactions and habitat conditions, a consideration for developing conservation strategies that account for community dynamics. Finally, we compared the predictive ability of models trained on all lakes versus only those where each target species for prediction occurs, addressing the trade-off between model generality and specificity. This comprehensive evaluation approach not only tests the robustness of our framework in capturing the complexity of lake ecosystems but also highlights opportunities for refining predictive modelling. Moreover, our modelling and assessment frameworks are flexible and can be readily adapted to other ecological

Fig. 1. Map of the 594 lakes in Ontario, Canada, included in our models. Each point is color-coded to represent the number of species present in the lake (i.e., species richness). Black lines delineate the provincial political boundaries, while grey lines delineate the secondary watersheds (Ontario Ministry of Natural Resources and Forestry—Provincial Mapping Unit 2024, using WGS 84).



modelling applications, offering a roadmap for future use by researchers and fisheries managers.

Material and methods

Dataset

Fish abundance was collected in 707 lakes by the Ontario Broadscale Monitoring Program (Sandstrom et al. 2011; Lester et al. 2021) of the Ontario Ministry of Natural Resources and Forestry (OMNRF 2012), Canada. The lakes spanned from a latitude of 43° to 54° and a longitude of −95° to −76°, with areas of 0.21 to 905 km^2 and maximum depth of 1.2 to 213 m. The lakes were sampled during the summers (June to September) from 2008 to 2012. The lake selection process used a stratified random sampling design, with strata defined by geographic zone and lake surface area. The lakes spanned three primary watersheds and 21 secondary watersheds (Fig. 1). Watershed delimitations were obtained through Ontario Ministry of Natural Resources and Forestry—Provincial Mapping Unit (2024).

A depth-stratified design was employed to sample and estimate fish abundance (see Lester et al. 2021 and Sandstrom et al. 2011 for more details on methods). The number of nets set per stratum was scaled with the surface area and depth strata within each lake to standardize sampling effort. Within each depth stratum, small and large mesh gillnets (small—mesh size between 13 and 38 mm; and large between 38 and 127 mm) were deployed overnight for 18 h (Appelberg 2000; Arranz et al. 2022). All fish captured were identified to the species level. Counts of fish from each lake were converted to catch per unit effort (CPUE) by dividing the number of fish caught by the total length of net deployed. It reflects the expected catch per 100 m of net over an 18-h period. The number of species caught per lake ranged from 2 to 25, reflecting natural variation in community composition along Ontario's latitudinal gradient of post-glacial colonization, with northern lakes typically hosting fewer species. We assumed that CPUE provides a reliable proxy for local species density in each lake (Olin et al. 2009).

The original dataset contained 87 species in total. We modelled the abundance of 14 species considered to be "sport fish", as these species are present across a large portion of the region and are frequently target species for fisheries management (see Table 1 and Fig. S1; selection of sport fish species was made following personal correspondence with Dr. Dylan Fraser, Concordia University, Montreal, Canada). We excluded 39 species that occurred in fewer than 10 lakes (i.e., <2% of all lakes) from the dataset, both to reduce computational time when calculating community latent variables, and because extremely rare species are generally not useful predictors of more widespread species (McGarigal et al. 2000). After apply-

Table 1. List of species included in the dataset, with both common and Latin names.

Category	Common name	Scientific name	Incidence (%)
Sport fish	Yellow perch	*Perca flavescens*	84
	Northern pike	*Esox lucius*	71
	Walleye	*Sander vitreus*	68
	Cisco	*Coregonus artedi*	58
	Lake whitefish	*Coregonus clupeaformis*	53
	Smallmouth bass	*Micropterus dolomieu*	48
	Lake trout	*Salvelinus namaycush*	45
	Burbot	*Lota lota*	38
	Largemouth bass	*Micropterus nigricans*	16
	Brook trout	*Salvelinus fontinalis*	11
	Black crappie	*Pomoxis nigromaculatus*	10
	Rainbow smelt	*Osmerus mordax*	9
	Muskellunge	*Esox masquinongy*	6
	Sauger	*Sander canadensis*	5
Nonsport fish	White sucker	*Castotomus commersonii*	93
	Spottail shiner	*Notropis hudsonius*	48
	Rock bass	*Ambloplites rupestris*	43
	Trout perch	*Percopsis omiscomaycus*	42
	Pumpkinseed	*Lepomis gibbosus*	29
	Logperch	*Percina caprodes*	26
	Common shiner	*Luxilus cornutus*	23
	Golden shiner	*Notemigonus crysoleucas*	23
	Emerald shiner	*Notropis bifrenatus*	21
	Brown bullhead	*Ameiurus nebulosus*	20
	Blacknose shiner	*Notropis heterolepis*	18
	Bluntnose minnow	*Pimephales notatus*	17
	Lake chub	*Couesius plumbeus*	14
	Longnose sucker	*Castotomus castotomus*	12
	Shorthead redhorse	*Moxostoma macrolepidotum*	12
	Bluegill	*Lepomis macrochirus*	9
	Ninespine stickleback	*Pungitius pungitius*	9
	Blackchin shiner	*Notropis heterodon*	7
	Mimic shiner	*Notropis volucellus*	7
	Mottled sculpin	*Cottus bairdii*	7
	Pearl dace	*Margariscus margarita*	7
	Slimy sculpin	*Cottus cognatus*	7
	Brook stickleback	*Culaea inconstans*	6
	Creek chub	*Semotilus atromaculatus*	6
	Fathead minnow	*Pimephales promelas*	6
	Johnny darter	*Etheostoma nigrum*	6
	Northern redbelly dace	*Chrosomus eos*	6
	Spoonhead sculpin	*Cottus ricei*	3
	Yellow bullhead	*Ameiurus natalis*	3
	Common carp	*Cyprinus carpio*	2
	Fallfish	*Semotilus corporalis*	2
	Iowa darter	*Etheostoma exile*	2
	Longnose dace	*Rhinichthys cataractae*	2
	Silver redhorse	*Moxostoma anisurum*	2

Note: The “category” column indicates whether the species is classified as a sport fish, based on guidance from Dr. Dylan Fraser, Concordia University, Montreal, Canada. The study primarily focused on predicting the abundance of sport fish. Within each category, species are ordered by incidence in the dataset (i.e., percentage of lakes in which the species occur), from highest at the top to lowest at the bottom.

ing these filters, we retained 34 nonsport fish species and 14 sport fish species, resulting in a final dataset of 48 species across 594 lakes (Fig. 1).

Environmental predictors

Multiple environmental variables were measured for each lake at the same time they were sampled for fish abundances (see Sandstrom et al. (2011) on the choice of variables to measure, and the sampling methods used for each variable). A total of 64 environmental variables were recorded per lake (Table S1). These variables included measurements of local climate conditions (16 variables), hydro morphology (13 variables), lake chemistry (11 variables), lake productivity (10 variables), human activity on the lake (seven variables), watershed characteristics (five variables), as well as latitude and longitude.

To streamline the analysis and reduce redundancy, we first standardized all variables to mean zero and unit variance, so they had a common scale and then applied principal component analysis (PCA) followed by a sparsification step via a varimax rotation to derive a smaller number of composite environmental variables (Zou et al. 2006). Varimax aims to produce axes where many of the environmental loadings are close to zero, simplifying interpretation by emphasizing the most important relationships (correlations) between environmental variables and PCA axes. We used the *prcomp* and *varimax* from the R package *stats* (R Core Team 2017) for this analysis. Since the dataset was split into calibration and validation sets (see *Modelling structure overview* for more details on the split), we first ran the PCA on the calibration set data and then projected the validation set onto the newly generated multivariate (PCA) environmental axes. This approach reduced dimensionality while maintaining consistent predictive structures between the calibration and validation sets, and it was applied to each validation replicate during the modelling procedure.

To identify the optimal number of PCA environmental axes, we conducted an analysis where the number of latent variables was fixed while the number of environmental PCA axes varied (see Supp. Information for details). The combination yielding the lowest out-of-sample error was selected, leading to the use of 10 composite environmental PCA axes for all subsequent analysis (Table S2 and Fig. S2).

Latent variable generation

We generated latent variables representing species covariation patterns based on presence–absence data for groups of species of interest (see following section *Modelling structure overview*). Latent variables were generated in two steps: (1) we first fitted a stacked species distribution model (SSDM), which integrates individual species-level distribution models to estimate occurrence probabilities across sites (lakes), using a binomial family to model species presence–absence, and (2) we applied a model-based copula ordination using Gaussian copulas to the stacked predictions. These steps were implemented with the functions *stackedsdm* and *cord* from the *ecoCopula* R package (Popovic et al. 2019, version 1.0-2).

The copula method was selected for its robustness with binomial data and computational efficiency (Popovic et al. 2022). We first fit a stacked species regression model without any predictors, used as a null model, to generate Dunn–Smyth residuals (Dunn and Smyth 1996). These residuals, which approximate standard normal residuals, are particularly advantageous for models with non-Gaussian responses, including binary, count, and Poisson-distributed data. The Gaussian copula model was then fitted on these residuals to capture latent dependence structures among species. To mitigate potential bias associated with lake size, the log-transformed lake area was included as a covariate in the stacked species regression model.

We generated sets of latent variables from three species groups: (1) sport fish species, (2) nonsport fish species, and (3) all fish species. These latent variable sets were then used as predictors in our single-species abundance modelss for each sport fish species. By using different groups of species combinations as a basis for latent variable generation, we were able to contrast their effectiveness in improving abundance predictions. This is particularly important because sampling and identifying all fish species in a lake may not be necessary for predicting the abundance of a target species if they do not contribute to improving predictive accuracy. To maintain consistency in the numbers of predictors, we limited the number of latent variables to four for each group (Stahl et al. 2026). Similarly to environmental variables, we conducted an analysis to identify the optimal number of latent variables to generate, where the number of composite environmental variables was fixed while the number of latent variables varied (see Supp. Information for details). The combination that resulted in the lowest out-of-sample error was selected, resulting in using four latent variables for subsequent analysis (Fig. S3).

Modelling structure overview

To apply the framework from Stahl et al. (2026) to our dataset, we modified the original approach and implemented the following steps:

- Using all lakes ($n = 594$), we derived three sets of latent variables from the presence–absence data of: (1) sport fish species, (2) nonsport fish species, and (3) all fish species.
- The dataset was randomly split into a calibration set and a validation set, representing respectively 70% ($n = 416$ lakes) and 30% ($n = 178$ lakes) of the dataset considered. This split was performed multiple times for each target sport fish species to assess uncertainty over model performance.
- Environmental variables of the calibration set were summarized by PCA with a sparsification step (Zou et al. 2006), and the environmental variables of the validation set were subsequently projected onto the same PCA axes (see section *Environmental predictors* for rationale).
- The calibration set was used to fit (train) statistical models for predicting lake abundance of each of the 14 sport fish

species. The trained models varied in their inclusion of different sets of predictors: (1) environmental variables summarized by sparse PCA axes, (2) environmental PCA axes combined with latent variables generated from presence–absence of the 14 sport fish species, (3) environmental PCA axes with latent variables generated from presence–absence of all nonsport fish species, and (4) PCA environmental axes and latent variables from the presence–absence of all fish species. This approach aimed to contrast the effects of different species groups on predictive ability and provide a comparison with models relying only on environmental data, as is commonly done in abundance modelling.

- The validation set was used to evaluate the performance of each model in predicting species abundance, with accuracy measured by the log error.
- The process of cross validation was replicated 1000 times. To determine the contribution of each lake to the dataset, we calculated the difference in error between two scenarios (1) when the lake was included in the calibration dataset, and (2) when the lake was excluded from the calibration dataset. This step allowed us to assess how influential a particular lake is on model performance and to identify whether certain lakes have a disproportionate effect on prediction accuracy.

Model fitting

We compared models containing (1) PCA environmental axes, (2) PCA environmental axes and latent variables generated from presence–absence of sport fish, (3) PCA environmental axes and latent variables generated from presence–absence of non-sport fish, and (4) PCA environmental axes and latent variables generated from presence–absence of all fish species.

We modelled variation in local abundance for each of the 14 sport fish species via a Generalized Additive Model (GAM) with a Tweedie distribution (Tweedie 1984) with a log-link function, using the functions *tw* and *gam* from the *mgcv* R package (Wood 2004, 2017, version 1.9-1). These models assume that the log of a species' conditional mean abundance in each lake is given by the sum of (potentially nonlinear) smooth functions of lake-specific covariates, along with the contributions of latent variables when included in the model (Wood 2017). We modelled the functional relationship between each predictive variable and log-mean abundance with a 2nd order thin-plate regression spline smoother (Wood 2003) with three basis functions using the function *s* from the R package *mgcv*. All models were estimated using restricted maximum likelihood (Wood 2011) using only data from the calibration set. The Tweedie distribution was selected for its flexibility in modelling a wide range of mean-variance relationships, which is particularly advantageous given that the available abundance data are expressed as a density (number of catches per unit effort, CPUE, a commonly used metric in fisheries research). Since CPUE data often include many zeros and continuous positive values, the Poisson and negative binomial distributions are less appropriate for accurately capturing the underlying structure of the data.

Metrics for evaluating model predictive ability

Although our models can be fit to predict both presence–absence and abundance, we focused exclusively on evaluating their performance in abundance predictions. Given our interest in predictive accuracy, all metrics discussed below compare predicted abundance with observed abundance, but only in the cases where the species was present. Note again, though, that our models were fit considering all lakes regardless of whether the species was present or not. This is important as some applications may require models to estimate potential abundance capacity in lakes where the species is absent, particularly for management purposes such as stocking, and our models are well-suited for such use. To assess whether a specific lake improved or reduced predictive ability, we used log error (LE) of predicted abundance as a measure of the bias of model prediction (eq. 1):

$$(1)\qquad \mathrm{LE}_{s,m,l} = \log_{10}\left(\frac{\widehat{Y}_{s,m,l}}{Y_{s,m,l}}\right)$$

where *s*, *m*, and *l* are indices for individual species, model, and lakes, respectively. *Y* denotes to the observed abundance and $\hat{Y}$ represents the predicted abundance.

This LE metric (eq. 1) assesses whether the model overestimated or underestimated the species' abundance in that lake. A positive LE indicates that the model overestimates abundance, whereas a negative LE reflects an underestimation. By examining the direction of the error, we could assess the impact of each lake on the overall predictive performance. The log error is also useful for evaluating the accuracy of predictive models when dealing with skewed data or data spanning several orders of magnitude (Tofallis 2015).

The log error (LE metric, eq. 1) measures the relative magnitude of the difference between predictions and observations, rather than the absolute difference between the two. As noted earlier, LE was only calculated for lakes where the species was present (i.e., abundance greater than 0). For each calibration replicate (i.e., where lakes were selected randomly to be part of the calibration or validation set), the mean error across the validation set was assigned to the corresponding lakes of the validation set. The median was then calculated across replicates for each model specification based on groups of species, target (response) species, and lake. This approach allowed to stabilize the error metric, as some lakes may have, in certain replicates, been part of a set with an extreme error rate.

Target analyses based on key questions

(1) Does the inclusion of latent variables improve prediction accuracy? To determine whether including latent predictors tended to improve model predictions compared to models with only environmental variables, we calculated a metric, ΔLE, for each model and species, equal to the difference

between the median of the absolute log error of out-of-sample predictions of the model containing only environmental variables to the median of the absolute of the log error of out-of-sample predictions (eq. 1) of the model that incorporated latent variables (eq. 2).

$$(2)\qquad \Delta \mathrm{LE}_{s,m} = \mathrm{Med}\left(|\mathrm{LE}_{s,l,m_0}|\right) - \mathrm{Med}\left(|\mathrm{LE}_{s,l,m}|\right)$$

where s, m, and l are indexes for individual species, models, and lakes, respectively. Med refers to the median across lakes for a single fold and m_0 to the model containing only environmental variables.

Our goal was to determine whether the advantages observed in the original framework (Stahl et al. 2026), which was tested on simulated data, could be replicated in an empirical dataset.

(2) Are predictions of sport fish abundances more accurate when using sport fish, nonsport fish, or all fish species as predictors? We visually contrasted the distribution of log error (eq. 1) of models with latent variables derived from three different community subsets (sport fish, nonsport fish, or all fish).

(3) What types of lakes significantly increase or decrease predictive ability, and are these lakes rare or common in terms of environment and/or species composition? We quantified (1) the environmental distinctiveness of each lake using pairwise Mahalanobis distances based on PCA-derived environmental axes, and (2) the ecological distinctiveness of each lake using its Local Contribution to Beta Diversity (LCBD, Legendre and De Cáceres 2013). We used Mahalanobis distance was because the PCA included a varimax rotation step that induces mild correlations among axes. Although Euclidean distance does not require strict orthogonality, it implicitly treats each axis as independent and places disproportionate weight on variation along correlated gradients. Mahalanobis distance adjusts for covariance structure among axes, providing a more accurate measure of multivariate environmental dissimilarity under these conditions. LCBD values measure how much each local community contributes to the overall beta diversity of the study region, with higher values indicating lakes whose species assemblages are more compositionally unique relative to the regional metacommunity.

To assess each lake's predictive contribution, we compared the median log error when the lake was included in the model calibration to the median log error when the lake was excluded (i.e., the lake was in the validation set, eq. 3). To the best of our knowledge, this represents a novel approach for assessing how individual observations (in this case, lakes) contribute to model performance (i.e., leverage), which can be generalized to any modelling framework whereas based on likelihood approaches (as in here) or machine learning techniques.

$$(3)\qquad \mathrm{Contribution}_{l,s} = \mathrm{Med}_{l\in C_j}\left(|\mathrm{LE}_{s,j}|\right) - \mathrm{Med}_{l\in V_j}\left(|\mathrm{LE}_{s,j}|\right)$$

where l, s, j are indices for lakes, species and replicates, respectively. The median (referred to as Med in eq. 3) $\mathrm{LE}_{j,s}$ was calculated for the lakes in the validation set for species s in replicate j. V_j in eq. 3 represents the validation set for replicate j, and C_j represents the calibration for the same replicate. For each species, we used the log error values of the best-performing model, defined as the one with the absolute median log error closest to zero.

Unlike the Euclidean distance, the Mahalanobis accounts for correlations among environmental variables (Mahalanobis 1936; De Maesschalck et al. 2000). This ensures that distances along strongly correlated environmental axes are not overrepresented and that each lake's environmental distinctiveness reflects departures from typical conditions. The pairwise Mahalanobis distance between lakes was calculated over the first 62 axes of a PCA based on the 64 environmental variables. Note that these PCA axes are somewhat correlated (unlike standard PCA axes) given the sparsification step via a varimax rotation, hence the use of the Mahalanobis distance. Additionally, we applied PCA instead of using the original variables because their correlation structure exhibited rank deficiency cause by the fact that the last two eigenvalues were exactly zero. This indicates that some variables were linearly dependent or provided redundant information, reducing the effective dimensionality of the data. The PCA was conducted using the function *princomp* from the R package *stats* (R Core Team 2017). For each lake, we calculated the average Mahalanobis distance between it and all other lakes. A smaller distance indicates that the lake's environmental conditions are uncommon (rare) compared to the others, while a larger distance suggests that the lake shares many common environmental features with other lakes.

Local contributions to beta diversity (LCBD) is a metric used to quantify the unique contribution of individual communities (here lakes) to the overall beta diversity within a region (Legendre and De Cáceres 2013) and as such can be viewed as a measure of ecological distinctiveness of a lake in the dataset. High LCBD values indicate that a lake has a more distinct (rare) community composition compared to other lake communities, while low values suggest that the species composition is more widespread and common across lakes. LCBD was calculated from the presence–absence dataset of all species using the functions *beta.div.comp* and LCBD*.comp* from the R package *adespatial* (Dray et al. 2023, version 0.3-23).

(4) To what extent do species share lakes that either improve or reduce predictive accuracy? We calculated Pearson correlations between all pairs of species of the lake-specific contributions to model predictive ability for each species (i.e., models containing the same environmental and latent variables, as per eq. 3). By visually examining these correlations, we aimed to identify patterns of shared environmental or biotic factors that might impact multiple species in similar ways. This approach allowed us to determine whether certain lakes consistently played a greater role in predicting abundance for multiple species or if their influence varied by species.

A lack of correlation would indicate that different species respond to distinct, lake-specific factors. This insight is critical for ecological modelling—where it signals that predictive performance for one species may not generalize to others—and for and for conservation and management, as it highlights that protecting or managing a lake for one species may not benefit others with different ecological requirements.

Alternatively, identifying shared drivers across species could streamline management efforts by focusing on key environmental factors that support multiple species simultaneously. Conversely, recognizing species-specific contributions allows for tailored management strategies address the unique needs of individual species.

(5) Are sport fish abundances better predicted using all lakes or only those where the species is present? To address this question, we conducted the same analysis but restricted the pool of lakes to those where species was present (i.e., abundance greater than 0). For this analysis, we excluded two species, muskellunge and sauger, due to their very low occurrences—present in only 38 and 29 lakes, respectively—which resulted in insufficient variation in the community composition of these lakes and made it impossible to fit the various models. As before, we first measured the average log error per lake (eq. 1) across replicates and compared the performance of the two models with the metric ΔSLE (Species Log Error), defined as the difference between absolute mean log error of the model fitted using all lakes and the absolute mean log error of the model fitted using the reduced lake pool (eq. 4).

$$(4)\qquad \Delta\text{SLE}_{s,\,m} = \left|\frac{1}{M}\sum_{l\,\in\,M}\text{LE}_{l,s}\right| - \left|\frac{1}{L_s}\sum_{l\,\in\,L_s}\text{LE}_{l,s}\right|$$

where s, m, l, M, and L_s, are indices for species, models, all lakes of the dataset, and lakes where species s is present, respectively. A positive ΔSLE indicates that the model using only lakes where the species is present performs better, while a negative value suggests that the model fitted with all lakes performs better.

Results

Our first goal was to determine whether incorporating latent variables derived from presence–absence of other species in the lake community improved predictions of target (sport fish) species abundance, which we assessed by comparing ΔLE between the environmental-based model and the latent-based models. Not all target species models benefitted from the inclusion of latent variables (Fig. 2). Importantly, the method used to generate these latent variables did not affect the direction of the ΔLE values and consistently produced the same overall effect on predictive ability, whether as an improvement or a decline relative to the environmental model. A clear trend emerged: species with low occurrences were predicted more accurately by the environmental model, whereas species with higher occurrences were better predicted by models that included latent variables. We then assessed how different species groups influenced predictive performance by comparing models in which the latent variables were derived from sport fish species, nonsport fish species, or all fish species combined. Our analysis showed that the best-performing model varied by species used to build latent variables, but differences in LE densities across models were relatively modest, suggesting that variations in predictive accuracy were not substantial (Fig. 3, Table S2). Cisco, lake whitefish, largemouth bass, northern pike, and smallmouth bass were best predicted by the model using latent variables incorporating all fish species. In contrast, black crappie, lake trout, rainbow smelt, walleye, and yellow perch were better predicted by the model using nonsport fish species. The remaining four species were most accurately predicted by the model that included only sport fish species. Taken together, these results indicate that our models are robust against variations in lake rarity, whether defined by environmental characteristics or community composition, and are not strongly influenced by any single environmental factor. This reinforces the broader applicability of the predictive framework across a wide range of lake types.

Next, we focused on identifying which types of lakes influenced predictive ability by analysing their contributions to LE and evaluating whether these influential lakes were rare or common in terms of their environmental characteristics and/or community composition (Fig. 4).

The LE metric showed no correlation with lake rarity, whether defined by environmental characteristics (Mahalanobis distance) or by species composition (LCBD). This lack of correlation suggests that predictive ability is not primarily driven by whether lake types are common or rare, although certain lake characteristics (e.g., specific environmental variables rather than the overall combination of variables) may still influence predictive ability through their overall characteristics, regardless of their rarity (or commonness). To determine which lake characteristics influenced predictive ability, either positively or negatively, we plotted the contribution to the log error against each environmental variable. These variables included log-transformed area (in km^2), altitude (in meters), maximum water temperature (in °C), and trophic status index based on phosphorus levels (Fig. S4). No clear pattern emerged in relation to key environmental variations.

We evaluated whether the predictive contributions of individual lakes were consistent across species by calculating the correlation of lake-specific contributions between species for each model specification (i.e., sport fish species, nonsport fish species, and all fish species; Fig. S5). Visual analysis revealed three distinct groups with similar correlations across models: (1) rainbow smelt, muskellunge, and sauger; (2) burbot, lake trout, black crappie, brook trout, and largemouth bass; and (3) yellow perch, smallmouth bass, northern pike, walleye, lake whitefish, and cisco.

The first and third groups showed negative correlations with each other but positive correlations within their respective groups (Table 2). In contrast, species in the second group exhibited idiosyncratic responses, with no meaningful correlations either within or between groups. The species groups also appear to be correlated with their occurrence rates (i.e., number of lakes that the species was present): group 1 consisted of low-occurrence species, group 2 included medium-occurrence species, and group 3 represented high-occurrence species.

Finally, we examined whether sport fish abundances (target species) were better predicted by models fitted using data from all lakes or only from lakes where the species was present. The results varied by species but were extremely consistent across models (Fig. 5). For rainbow smelt, lake trout,

Fig. 2. ΔLE as a function of model and species. The ΔLE was calculated as the median absolute log error of the model with only environmental variables, minus the median absolute log error of the model incorporating latent predictors (eq. 2). Positive values (in blue) indicate that the model with latent predictors performed better, while negative values (in red) signify better performance by the environmental model. Latent variables were generated using one of three groups (1) sport fish species, represented ("Env.sport"), (2) nonsport fish species, represented ("Env.non.sport"), or (3) all fish species ("Env.all"). Species are ordered by incidence (number of lakes present) in the dataset, from highest at the top to lowest at the bottom. LE, log error.

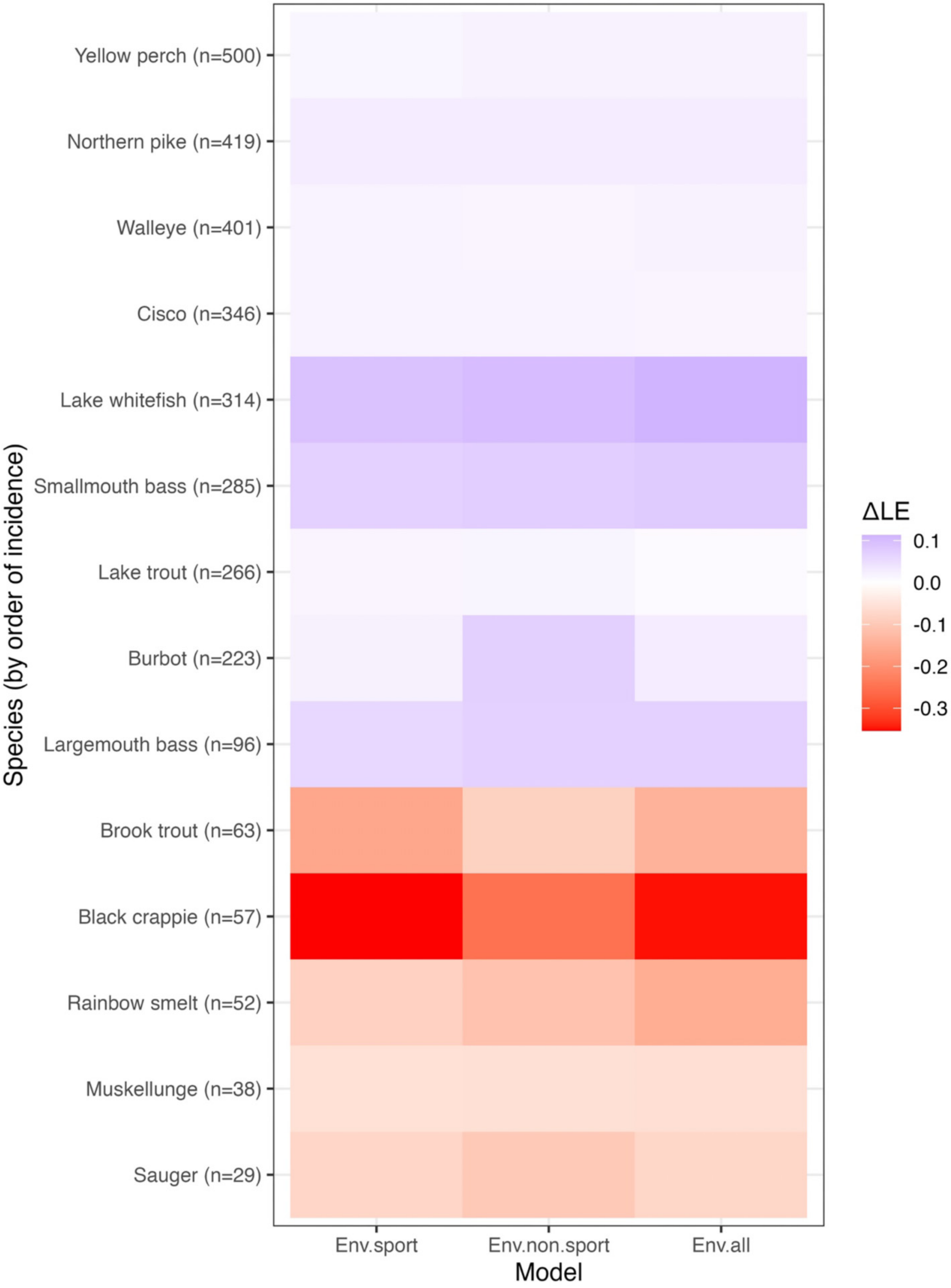
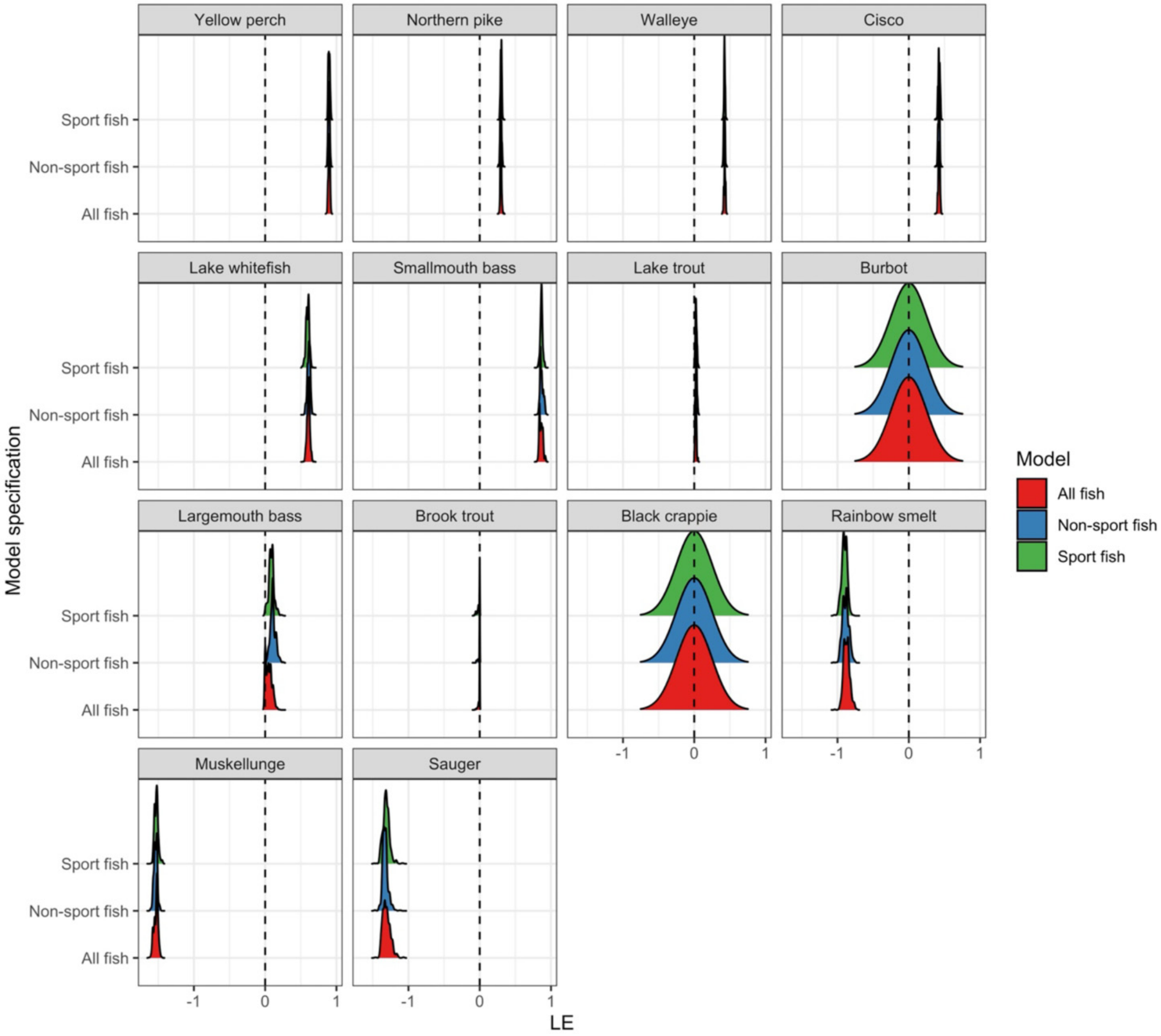


Fig. 3. Density plot of the log error as a function of species and model. The log error was calculated following eq. 1, and for each lake, the median log error was taken across replicates for each species and model. Latent variables were generated using three groups: (1) sport fish species (green), (2) nonsport fish species (blue), and (3) all fish species (red). All models also included environmental variables. The dotted vertical line represents an error of 0, meaning the median prediction equals the median observed values. Species are ordered by their incidence (number of lakes occupied) in the dataset, from highest at the top to lowest at the bottom.

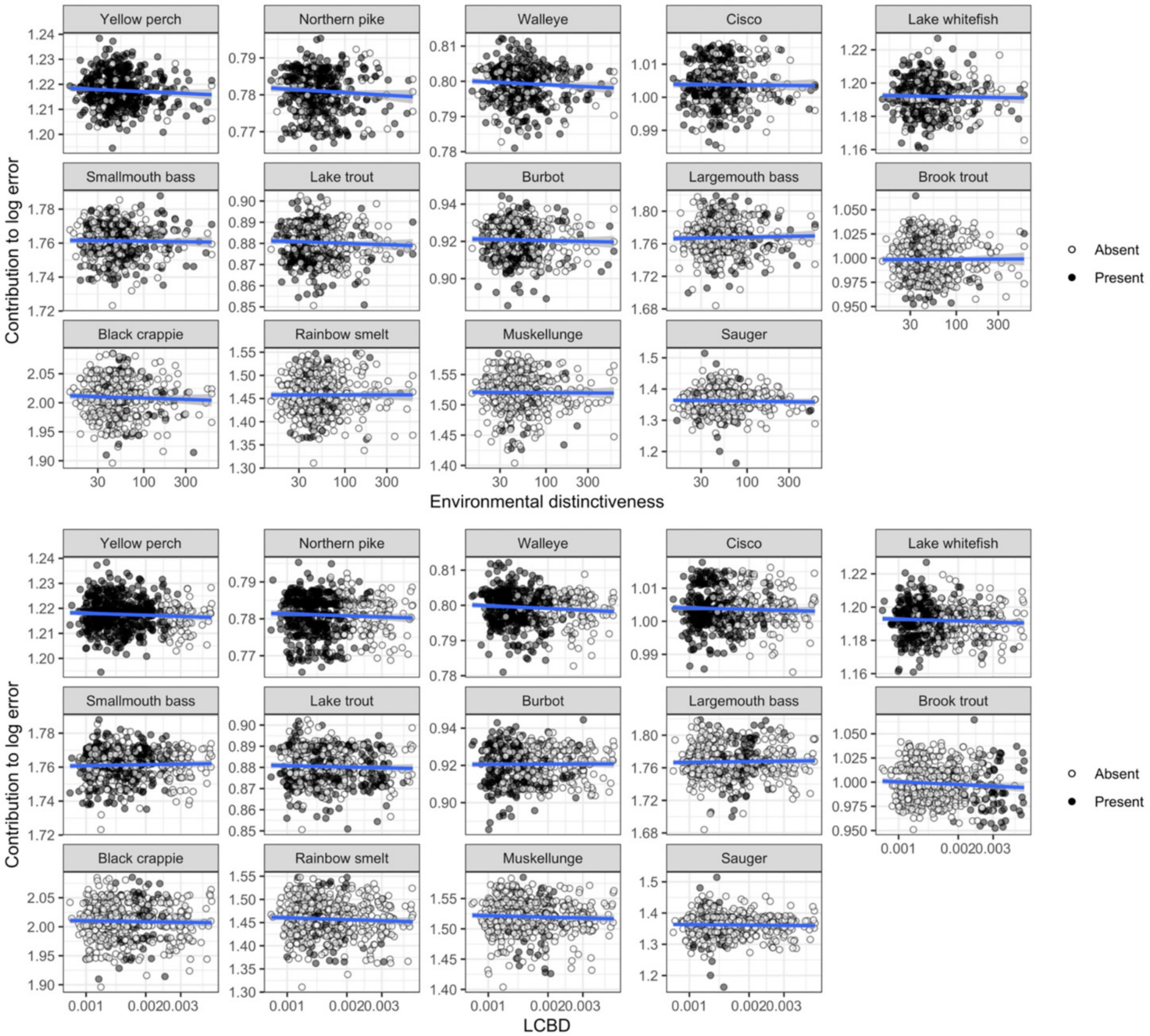


and lake whitefish, models fitted using only the lakes where the species occurred performed better on average. In contrast, for black crappie, brook trout, largemouth bass, burbot, smallmouth bass, cisco, walleye, northern pike, and yellow perch, predictions were more accurate when models included data from all lakes in the dataset. This finding highlights an important aspect of modelling species abundances: a one-size-fits-all approach is not the most effective, as each species may require different model specifications to produce accurate abundance predictions.

Discussion

Our first goal was to assess whether abundance models including latent variables, as designed by Stahl et al. (2026), could improve prediction accuracy of species abundances in a large, complex natural system. The original approach was tested only through simulations and did not account for species interactions, such as those found in large scale lake-fish ecosystems. One of the key advantages of this modelling framework is its ability to use presence–absence data, which are easier to generate than abundance data, to extract latent variables that are then used to predict the abundance distributions of target species. However, because presence–absence-based latent variables can partly reflect unmeasured environmental gradients rather than quantitative biotic interactions, their predictive value does not necessarily translate into strong abundance covariation among species. This helps explain why models that included latent variables improved abundance predictions even when pairwise abun-

Fig. 4. Contribution of each lake to the log error as a function of environmental distinctiveness and Local Contribution to Beta Diversity (LCBD) per species (see methods how these values were calculated). The lake's contribution was measured as the median across replicates of the difference between the log error when the lake was included in calibrating the model and the log error when the lake was excluded (i.e., in the validation set, eq. 3). A positive contribution indicates that including the lake in model improved predictions, while a negative contribution indicates that excluding it improved predictions. Point color indicate species presence (black) or absence (white) in the lake. High LCBD values indicate that a lake has a more distinct community composition in relation to other lakes, whereas a low value suggests a common composition. Each sport fish species is shown in a separate panel, and the log error values are from the best model (i.e., the model with a median log error closest to 0; see Table S3 for model details per species). The dotted horizontal line represents an error of 0, indicating that the median prediction equals the observed values). Species were ordered by incidence (number of lakes occupied) in the dataset, from highest at the top to lowest at the bottom.



dance correlations were weak. Our results show that latent-variable models primarily improved predictions for high-occurrence species, whereas low occurrence species are better predicted environmental variables alone. This pattern likely reflects the fact that high-occurrence species, being more widespread and abundant, participate in more consistent and detectable biotic interactions that are captured by the latent variables. In contrast, low-occurrence species may be constrained by rare or highly localized environment conditions that latent variables, derived from broad co-occurrence structure, cannot represent as effectively.

Our second goal was to evaluate whether the choice of species subset to generate latent variables influenced predictive performance. We found that no single species subset that consistently performed best across target species, indicating that the framework is relative insensitive to insensitive to which species are used to construct the latent variables. This pattern is consistent with the original simulated-based

Table 2. Mean and standard deviation of correlation between species groups across models.

	Group 1	Group 2	Group 3
Group 1	0.72 ± 0.04		
Group 2	− 0.09 ± 0.03	− 0.03 ± 0.09	
Group 3	− 0.75 ± 0.06	0.12 ± 0.04	0.80 ± 0.05

Note: We calculated the correlation between lake contributions for each species and model, revealing distinct grouping patterns (see Fig. S5). The species were grouped as follows: (Group 1) rainbow smelt, muskellunge, and sauger; (Group 2) burbot, lake trout, black crappie, brook trout, and largemouth bass; and (Group 3) yellow perch, smallmouth bass, northern pike, walleye, lake whitefish, and cisco.

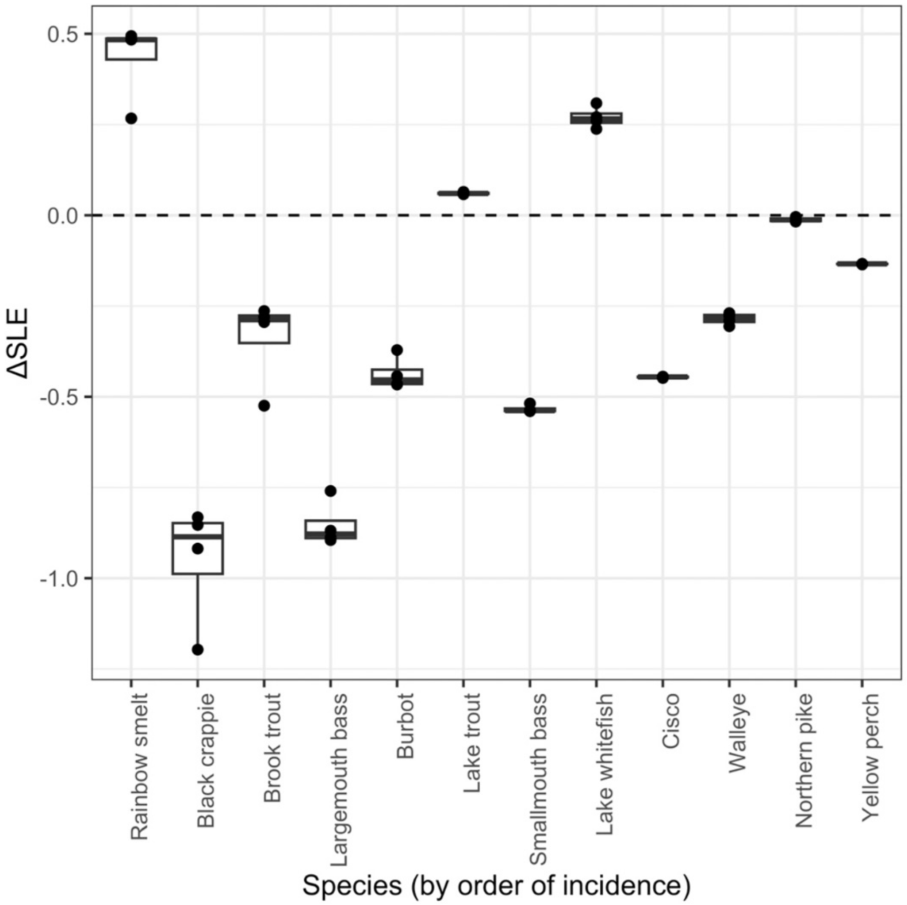
assessment of the framework, which also showed that latent variables mostly strongly improved predictions for higher-occurrence species (Stahl et al. 2026). It also aligns with the broader literature, which suggests that low-occurrence species tend to be more vulnerable to stochastic environmental fluctuations and demographic instability (Gaston 1994; Brown et al. 1995), whereas high-occurrence species more frequently to engage complex biotic interactions (Mouquet et al. 2003; Araújo and Luoto 2007). Consequently, latent variables may be most informative for predicting species that occupy a wide range of habitats and regularly co-occur with many others, making them more influenced by biotic interactions such as competition, predation, or facilitation (Blanchet et al. 2020). In contrast, low-occurrence species appear primarily shaped by localized abiotic constrains and demographic stochasticity. These outcomes may, however, be system-specific, and our framework may perform differently for low-occurrence species in other ecological contexts. Because the approach is flexible and to be generalizable across taxa and systems, future applications could explore alternative strategies for forming species subsets with model selection approaches tailored to maximize predictive accuracy for individual target species (see below for other alternative for species selections). In addition, further work should examine whether constructing latent variables directly from abundance data yields stronger or more consistent predictions, helping to clarify whether the predictive improvements we observed arise largely from shared occurrence patterns (reflecting common environmental drivers) or from true biotic dependencies among species' abundances.

Our third goal was to determine which types of lakes most strongly influenced predictive ability, either positively or negatively, by relating each lake's log error contribution to its environmental distinctiveness and community-composition distinctiveness (LCBD). We found no relationship between prediction error and lake rarity, whether defined by environmental features, species composition, or any individual environmental variable. In practical terms, this means that large lakes were no more likely than small lakes to improve model performance, and that predictive ability was not systematically improved or hindered by particular environmental attributes or community structures (species compositions). This result is noteworthy because it challenges the common assumption that certain environmental and biotic conditions inherently yield better predictions. For example, one could expect larger lakes, being more stable (May 1972) and containing a greater diversity of habitats, to generate stronger and more reliable predictions (Magnuson et al. 2005). Conversely, larger lakes might instead be less predictable because their greater abundance of microhabitat diversity and associated fine-scale environmental heterogeneity (Strayer and Findlay 2010), much of which is not captured by standard broad-scale environmental measurements. Our results suggest that predictive accuracy is not systematically tied to these complexities. Rather, our predictive framework performs comparably across lakes, spanning a wide range of environmental and community profiles, reinforcing the robustness and generality of our predictive approach for broad-scale ecological applications.

The correlation of lake contribution across species allows us to effectively group species by their occurrence rates, revealing underlying ecological patterns that shape species distributions and abundances. This association suggests that species within the same occurrence group (low, medium, or high) likely respond to similar environmental drivers or ecological interactions in lake ecosystems, supporting findings from other studies (Araújo and Guisan 2006; Ovaskainen et al. 2010; Legendre and Legendre 2012). These results underscore the complexity of ecosystem dynamics and the need for sophisticated models that account for diverse species interactions and environmental conditions. Models that incorporate a broad range of variables, including both environmental factors and species interactions, are essential for capturing the intricate nature of ecological communities (Wisz et al. 2008). Given the distinct correlation patterns among the three species groups, generating latent variables specific to each group could be a promising avenue for improving abundance predictions. This strategy leverages ecological similarities within each group, potentially capturing more relevant interactions and environmental gradients that influence species abundance. Moreover, identifying species combinations (groups) that are consistently used across models for multiple target species may be more appropriate for management and conservation practices than identifying different species combinations that maximize abundance predictions for each individual target species as discussed earlier. This is because using a consistent set of species groups simplifies decision-making, enhances the applicability of the models across various contexts, and facilitates the development of broader, ecosystem-wide management strategies rather than focusing on species-specific predictions.

The analysis of whether sport fish abundances were better predicted using data from all lakes or only those where the species was present revealed variations across species, with no clear pattern emerging in relation to occurrence, abundance, or trophic level. This suggests that the predictive success of each approach may be driven by species-specific ecological factors, such as habitat specificity, life history traits, or community interactions—factors that are potentially not fully captured by the diverse and numerous environmental predictors we considered. These findings are consistent with previous studies (see Thuiller et al. 2005; Elith et al. 2010; Dormann et al. 2013 among others), highlighting the importance of incorporating species-specific ecological dynamics

Fig. 5. Boxplot of the ΔSLE per species. The ΔSLE is calculated as the absolute mean log error fitted using all lakes minus the absolute mean log error of the model fitted using only where the species is present (eq. 4). A positive ΔSLE indicates better performance when using the reduced lake pool, while a negative ΔSLE suggests that the model using all lakes performs better. Each point represents a model, and the boxplots group the results of all four models per species. The dotted horizontal line represents an identical performance between models trained on either all lakes or only those where the species is present. Muskellunge and sauger were excluded due to their extremely low occurrences (number of lakes occupied), which rendered the analysis infeasible. Species are ordered by incidence in the dataset, from lowest on the left to highest on the right.



in predictive models. The consistency of our results across models—whether based solely on environmental variables or a combination of environmental variables and community composition factors, highlights the importance of approaches that account for the unique ecological context of each species. This makes it challenging to design broad conservation and management strategies, because any general approach must still account for the specific needs of individual species.

Our study provides a series of interconnected insights that link the questions we explored. First, we found that low abundance species are better predicted by environmental models, while high abundance species show improved predictions when latent variables are included (Question 1). This distinction suggests that environmental factors play a more significant role in shaping the distribution of low abundance species, whereas high abundance species may be more influenced by community interactions potentially captured by latent variables. Supporting this, we observed that individual lake contributions to predictive accuracy are correlated within low abundance species as well as within high abundance species (Question 4). However, these correlations do not extend between the two groups, indicating that the factors driving the predictive success of lakes for low abundance species are distinct and inversely related to those influencing high-abundance species. Interestingly, these patterns in lake contributions do not correlate with environmental distinctiveness, species composition distinctiveness, or any of the environmental variables assessed (Question 3). Together, these findings suggest that while environmental variables are key predictors for low abundance species (Gaston 1994; Brown et al. 1995), high abundance species are likely responding to more complex, community-level interactions that are better captured by latent variables (Mouquet et al. 2003; Araújo and Luoto 2007). The distinct and negatively correlated patterns of lake contributions across these species' groups point to underlying ecological processes not linked to traditional environmental or spatial predictors used in species distribution models. These results highlight the need for further investigation into the specific ecological drivers underlying these patterns, particularly species interactions and community dynamics, which may differ fundamentally between low- and high-abundance species.

Our findings echo those of Hui et al. (2013), who demonstrated that clustering species by their environmental affinities, or "archetypes", improved predictive accuracy. In a similar way, we found that clustering species based on their occurrence patterns, particularly low- and high-abundance species, enhanced our ability to predict species distributions. This suggests that identifying and leveraging such clusters, whether based on environmental affinities or other ecological traits such as abundance, is essential for improving ecological predictive models. It underscores that a one-size-fits-all approach may not be optimal when modelling species distributions, especially in complex ecosystems like lakes, where species interactions and community dynamics play a significant role.

While our study provides valuable insights, it also has important limitations. One key limitation is that the empirical evaluation here relies solely on lake ecosystems, where dispersal is relatively restricted and species are likely more strongly shaped by local environmental conditions. Although our modelling framework itself is generalizable, the empirical patterns we report may not extend systems where dispersal plays a more dominant force in shaping community structure and controlling species distributions (Leibold et al. 2004; Urban et al. 2012; Thompson and Gonzalez 2017). A second limitation is that deriving latent variables from presence–absence data may oversimplify the ecological processes influencing species abundances, particularly in communities with complex, nonlinear, or context-dependent interactions. This simplification helps explain the apparent discrepancy between the strong predictive performance of presence–absence-based latent variables and the weak abundance–abundance correlations observed among species. Co-occurrence pattern (i.e., based on presence–absence) often capture broad aspect of community structure and shared environmental filtering or consistent biotic associations (Ovaskainen et al. 2017), whereas abundance data reflect finer-scale demographic processes and interaction strengths that vary from site to site. In this sense, co-occurrence patterns can provide a stable proxy for the potential biotic environment, even when realized abundances fluctuate independently due to demographic variability or local environmental noise. This distinction aligns with the conceptual point raised in the Introduction: that latent variables can reflect both unmeasured environmental gradients and aggregated biotic influences, without necessarily tracking direct abundance covariation. To evaluate this more explicitly, future applications could compare latent variables derived from both presence–absence and abundance data (see Clark et al. 2018), allowing a clearer assessment of their relative ability to represent biotic structure and improve predictive performance. Note again that, in the present study, our primary objective was to improve the prediction of species abundances using lower-resolution presence–absence information to characterize biotic context given that detailed abundance data are often challenging to generate.

Another limitation concerns our reliance on random sampling to split calibration and validation sets for simplicity and computational efficiency. As DiRenzo et al. (2023) and Roberts et al. (2017) note, more robust methods such as spatial cross-validation, blocking, or other spatially structured resampling schemes, are preferable when data exhibit autocorrelation or when the covariance structure of predictors differs between training (calibration) and test (validation) datasets. As emphasized by Wenger and Olden (2012), ignoring these issues can reduce the transferability and reliability of ecological models. Incorporating stratified or spatially sampling strategies could therefore yield more robust predictions. While our study advances the understanding of species abundance prediction, it underscores the need for more comprehensive modelling approaches that better account for the complex interplay of environmental, spatial, and biotic factors.

In conclusion, our study demonstrates the value of integrating co-occurrence information, via latent variables, into predictive models for species abundance, while also clarifying the conceptual and methodological limits of this approach. Presence–absence based latent variables can effectively capture large-scale structure in species distributions and community composition, thereby improving predictions for widespread species whose dynamics reflect consistent biotic and environmental associations. However, because these latent variables do not fully encode quantitative interaction strengths or fine-scale abundance covariation, they may be less informative for rare or locally constrained species. Together, these findings underscore both the promise and the boundaries of latent-variable approaches and highlight the continued need for modelling strategies that are tailored to the ecological processes governing different species. Such approaches will be critical for advancing our capacity to understand, predict, and manage complex ecological systems.

Acknowledgements

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, the QCBS travel grant to present this work at a conference (SFE2 2024 in Lyon, France). AS and PPN were funded by the Canada Research Chairs program (CRC-2023-00139, Spatial Ecology and Biodiversity). EP and PPN (RGPIN-2020-07188, Ecology in Spatially Structured Environments) were funded by the NSERC Discovery Grant program.

Article information

History dates

Received: 5 September 2025
Accepted: 9 December 2025
Accepted manuscript online: 12 January 2026
Version of record online: 18 February 2026

Copyright

Data availability

Data and scripts are available on Dryad: https://doi.org/10.5061/dryad.jh9w0vtq0.

Author information

Author ORCIDs

Aliénor Stahl https://orcid.org/0000-0002-2297-7379
Eric J. Pedersen https://orcid.org/0000-0003-1016-540X
Pedro R. Peres-Neto https://orcid.org/0000-0002-5629-8067

Author contributions

Conceptualization: AS, EJP, PRP
Data curation: AS
Formal analysis: AS, EJP, PRP
Funding acquisition: EJP, PRP
Investigation: AS
Methodology: AS, EJP, PRP
Project administration: EJP, PRP
Resources: EJP, PRP
Software: AS
Supervision: EJP, PRP
Validation: AS, EJP, PRP
Visualization: AS
Writing – original draft: AS
Writing – review & editing: AS, EJP, PRP

Competing interests

The authors declare there are no competing interests.

Supplementary material

Supplementary data are available with the article at https://doi.org/10.1139/cjfas-2025-0290.

References

Appelberg, M. 2000. Swedish Standard Methods for Sampling Freshwater Fish with Multi-mesh Gillnets. Fiskeriverket Inf. **1**(August): 1–32. Available from https://www.fiskeriverket.se/download/18.1e7cbf241100bb6ff0b8000989/provfiskebeskr.pdf.

Araújo, M.B., and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. **33**(10): 1677–1688. doi:10.1111/j.1365-2699.2006.01584.x.

Araújo, M.B., and Luoto, M. 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecol. Biogeogr. **16**(6): 743–753. doi:10.1111/j.1466-8238.2007.00359.x.

Arranz, I., Fournier, B., Lester, N.P., Shuter, B.J., and Peres-Neto, P.R. 2022. Species compositions mediate biomass conservation: the Case of Lake Fish communities. Ecology, **103**(3): e3608. doi:10.1002/ecy.3608.

Blanchet, F.G., Cazelles, K., and Gravel, D. 2020. Co-occurrence is not Evidence of Ecological Interactions. Ecol. Lett. **23**(7): 1050–1063. doi:10.1111/ele.13525.

Boulangeat, I., Gravel, D., and Thuiller, W. 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. Ecol. Lett. **15**(6): 584–593. doi:10.1111/j.1461-0248.2012.01772.x.

Boyce, M.S., Johnson, C.J., Merrill, E.H., Nielsen, S.E., Solberg, E.J., and van Moorter, B. 2016. Can Habitat Selection Predict Abundance? J. Anim. Ecol. **85**(1): 11–20. doi:10.1111/1365-2656.12359.

Brosse, S., Guegan, J.-F., Tourenq, J.-N., and Lek, S. 1999. The Use of Artificial Neural Networks to Assess Fish Abundance and Spatial Occupancy in the Littoral Zone of a Mesotrophic Lake. Ecol. Modell. **120**(2–3): 299–311. doi:10.1016/S0304-3800(99)00110-6.

Brown, J.H., Mehlman, D.W., and Stevens, G.C. 1995. Spatial Variation in Abundance. Source: Ecology.

Clark, A.P., Howard, K.L., Woods, A.T., Penton-Voak, I.S., and Neumann, C. 2018. Why Rate when you Could Compare? Using the "EloChoice" Package to Assess Pairwise Comparisons of Perceived Physical Strength. PLoS One, **13**(1). doi:10.1371/journal.pone.0190393.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D.L. 2000. The Mahalanobis Distance. Chemom. Intell. Lab. Syst. **50**(1): 1–18. doi:10.1016/S0169-7439(99)00047-7.

Degnbol, P., and Jarre, A. 2004. Review of indicators in fisheries management—a development perspective. *In* African Journal of Marine Science. Marine and Coastal Management. pp. 303–326. doi:10.2989/18142320409504063.

Dickinson, J.L., Zuckerberg, B., and Bonter, D.N. 2010. Citizen science as an ecological research tool: challenges and benefits. Annu. Rev. Ecol. Evol. Syst. **41**: 149–172. doi:10.1146/annurev-ecolsys-102209-144636.

DiRenzo, G. V., Hanks, E., and Miller, D.A.W. 2023. A practical guide to understanding and validating complex models using data simulations. Methods Ecol. Evol. **14**(1): 203–217. doi:10.1111/2041-210X.14030.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, **36**(1): 27–46. doi:10.1111/j.1600-0587.2012.07348.x.

Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guénard, G., et al. 2023. adespatial: multivariate multiscale spatial analysis. doi:10.1890/11-1183.1.

Dunn, P.K., and Smyth, G.K. 1996. Randomized quantile residuals. *In* Source: Journal of Computational and Graphical Statistics.

Elith, J., Kearney, M., and Phillips, S. 2010. The art of modelling range-shifting species. Methods Ecol. Evol. **1**(4): 330–342. doi:10.1111/j.2041-210x.2010.00036.x.

Gaston, K.J. 1994. Rarity. Springer, Netherlands, Dordrecht. doi:10.1007/978-94-011-0701-3.

Gaston, K.J. 2003. The Structure and Dynamics of Geographic Ranges. Oxford University Press. doi:10.1093/oso/9780198526407.001.0001.

Hui, F.K.C., Warton, D.I., Foster, S.D., and Dunstan, P.K. 2013. To mix or not to mix: comparing the predictive performance of mixture models versus separate species distribution models. Ecology, **94**(9): 1913–1919. doi:10.1890/12-1322.1.

Jackson, D.A., and Harvey, H.H. 1997. Qualitative and quantitative sampling of lake fish communities. Can. J. Fish. Aquat. Sci. **54**(12): 2807–2813. doi:10.1139/f97-182.

Kraft, N.J.B., Adler, P.B., Godoy, O., James, E.C., Fuller, S., and Levine, J.M. 2015. Community assembly, coexistence and the environmental filtering metaphor. Funct. Ecol. **29**(5): 592–599. doi:10.1111/1365-2435.12345.

Legendre, P., and De Cáceres, M. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. Ecol. Lett. **16**(8): 951–963. doi:10.1111/ele.12141.

Legendre, P., and Legendre, L. 2012. Numerical ecology. *In* 3rd ed. Elsevier.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., et al. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecol. Lett. **7**(7): 601–613. doi:10.1111/j.1461-0248.2004.00608.x.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., and Delacoste, M. 1996. Role of some environmental variables in trout abundance models using neural networks. Aquat. Living Res. **9**(1): 23–29. doi:10.1051/alr:1996004.

Lester, N.P., Sandstrom, S., de Kerckhove, D.T., Armstrong, K., Ball, H., Amos, J., et al. 2021. Standardized broad-scale management and monitoring of inland lake recreational fisheries: an overview of the ontario experience. Fisheries, **46**(3): 107–118. doi:10.1002/fsh.10534.

Lewis, J.S., Farnsworth, M.L., Burdett, C.L., Theobald, D.M., Gray, M., and Miller, R.S. 2017. Biotic and Abiotic Factors Predicting the Global Distribution and Population Density of an Invasive Large Mammal. Sci. Rep. **7**. doi:10.1038/srep44152.

Lindenmayer, D.B., and Likens, G.E. 2010. Effective ecological monitoring. doi:10.1071/9780643100190.

Mack, R.N., Simberloff, D., Lonsdale, W.M., Evans, H., Clout, M., and Bazzaz, F.A. 2000. Biotic invasions: causes, epidemiology, global consequences, and control. *In* Ecological applications. pp. 689–710. doi:10.1890/1051-0761.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A.A., and Langtimm, C.A. 2002. Estimating Site Occupancy Rates when Detection Probabilities are Less Than One. Ecology, **83**(8): 2248–2255. doi:10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2.

Magnuson, J.J., Kratz, T.K., and Benson, B.J. 2005. Long-term dynamics of lakes in the landscape: long-term ecological research on north temperate lakes. doi:10.1093/oso/9780195136906.001.0001.

Mahalanobis, P.C. 1936. On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India.

May, R.M. 1972. Will a Large Complex System be Stable? Nature, **238**(5364): 413–414. doi:10.1038/238413a0.

McGarigal, K., Stafford, S., and Cushman, S. 2000. Multivariate statistics for wildlife and ecology research. Springer, New York, New York, NY. doi:10.1007/978-1-4612-1288-1.

Mouquet, N., Munguia, P., Kneitel, J.M., and Miller, T.E. 2003. Community assembly time and the relationship between local and regional species richness. Oikos, **103**(3): 618–626. doi:10.1034/j.1600-0706.2003.12772.x.

Olin, M., Malinen, T., and Ruuhijärvi, J. 2009. Gillnet catch in estimating the density and structure of fish community—comparison of gillnet and trawl samples in a eutrophic lake. Fish. Res. **96**(1): 88–94. doi:10.1016/j.fishres.2008.09.007.

Olkeba, B.K., Boets, P., Mereta, S.T., Yeshigeta, M., Akessa, G.M., Ambelu, A., and Goethals, P.L.M. 2020. Environmental and Biotic Factors Affecting Freshwater Snail Intermediate Hosts in the Ethiopian Rift Valley Region. Parasites Vectors, **13**(1). doi:10.1186/s13071-020-04163-6.

Ontario Ministry of Natural Resources and Forestry (OMNRF). 2012. Dataset. Ontario Hydro Network—Waterbody.

Ontario Ministry of Natural Resources and Forestry—Provincial Mapping Unit. 2024. Ontario Watershed Information Tool (OWIT). Available from https://lio.maps.arcgis.com/home/item.html?id=67546fd352d24b97b126f181fb650600 [accessed 9 October 2024].

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

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., et al. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecol. Lett. **20**(5): 561–576. doi:10.1111/ele.12757.

Pagnucco, K.S., and Ricciardi, A. 2015. Disentangling the influence of abiotic variables and a non-native predator on freshwater community structure. Ecosphere **6**(12). doi:10.1890/ES15-00371.1.

Popovic, G.C., Hui, F.K.C., and Warton, D.I. 2018. A general algorithm for covariance modeling of discrete data. J. Multivariate Anal. **165**: 86–100. doi:10.1016/j.jmva.2017.12.002.

Popovic, G.C., Hui, F.K.C., and Warton, D.I. 2022. Fast Model-based Ordination with Copulas. Methods Ecol. Evol. **13**(1): 194–202. doi:10.1111/2041-210X.13733.

Popovic, G.C., Warton, D.I., Thomson, F.J., Hui, F.K.C., and Moles, A.T. 2019. Untangling direct species associations from indirect mediator species effects with graphical models. Methods Ecol. Evol. **10**(9): 1571–1583. doi:10.1111/2041-210X.13247.

R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.r-project.org/.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Blackwell Publishing Ltd. doi:10.1111/ecog.02881.

Sandstrom, S., Rawson, M., and Lester, N. 2011. Manual of Instructions for Broadscale Fish Community Monitoring Using North American (NA1) and Ontario Small Mesh (ON2) Gillnets. *In* 2011.1. Ontario Ministry of Natural Resources and Forestry, Peterborough, Ontario.

Sobrino, I., Rueda, L., Tugores, M.P., Burgos, C., Cojan, M., and Pierce, G.J. 2020. Abundance Prediction and Influence of Environmental Parameters in the Abundance of Octopus (Octopus vulgaris Cuvier, 1797) in the Gulf of Cadiz. Fish. Res. **221**: 105382. doi:10.1016/j.fishres.2019.105382.

Stahl, A., Pedersen, E.J., and Peres-Neto, P.R. 2026. Advancing single species abundance models: robust models for predicting abundance using co-occurrence from communities. Ecosphere. doi:10.32942/X2S32J.

Strayer, D.L., and Findlay, S.E.G. 2010. Ecology of freshwater shore zones. Aquat. Sci. **72**(2): 127–163. doi:10.1007/s00027-010-0128-9.

Thompson, P.L., and Gonzalez, A. 2017. Dispersal governs the reorganization of ecological networks under environmental change. Nat. Ecol. Evol. **1**(6). doi:10.1038/s41559-017-0162.

Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T., and Prentice, I.C. 2005. Climate Change Threats to Plant Diversity in Europe. Proc. Natl. Acad. Sci. **102**(23): 8245–8250. doi:10.1073/pnas.0409902102.

Tofallis, C. 2015. A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation. J. Oper. Res. Soc. **66**(8): 1352–1362. doi:10.1057/jors.2014.103.

Tweedie, M.C.K. 1984. An Index which Distinguishes Between some Important Exponential Families. *In* Statistics: applications and new directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Calcutta. pp. 579–604.

Urban, M.C., De Meester, L., Vellend, M., Stoks, R., and Vanoverbeke, J. 2012. A Crucial Step Toward Realism: Responses to Climate Change from an Evolving Metacommunity Perspective. Evol. Appl. **5**(2): 154–167. doi:10.1111/j.1752-4571.2011.00208.x.

VanDerWal, J., Shoo, L.P., Johnson, C.N., and Williams, S.E. 2009. Abundance and the environmental niche: environmental suitability estimated from niche models predicts the upper limit of local abundance. Am. Nat. **174**(2): 282–291. doi:10.1086/600087.

Waldock, C., Stuart-Smith, R.D., Albouy, C., Cheung, W.W.L., Edgar, G.J., Mouillot, D., et al. 2022. A quantitative review of abundance-based species distribution models. Ecography, **2022**(1). doi:10.1111/ecog.05694.

Walker, S.C., and Jackson, D.A. 2011. Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. Ecol. Monogr. **81**(4): 635–663. Ecological Society of America. doi:10.1890/11-0886.1.

Wenger, S.J., and Olden, J.D. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol. Evol. **3**(2): 260–267. doi:10.1111/j.2041-210X.2011.00170.x.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., et al. 2008. Effects of sample size on the performance of species distribution models. Diversity Distributions **14**(5): 763–773. doi:10.1111/j.1472-4642.2008.00482.x.

Wood, S.N. 2003. Thin-plate regression splines. J. R. Stat. Soc. Ser. B: Stat. Methodol. **65**(1): 95–114. doi:10.1111/1467-9868.00374.

Wood, S.N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Am. Statist. Assoc. **99**(467): 673–686. doi:10.1198/016214504000000980.

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Ser. B: Stat. Methodol. **73**(1): 3–36. doi:10.1111/j.1467-9868.2010.00749.x.

Wood, S.N. 2017. Generalized additive models: an introduction with R. *In* 2nd ed. Chapman and Hall/CRC. doi:10.1201/9781315370279.

Yoccoz, N.G., Nichols, J.D., and Boulinier, T. 2001. Monitoring of biological diversity in space and time. Trends Ecol. Evol. **16**(8): 446–453. doi:10.1016/S0169-5347(01)02205-4.

Zou, H., Hastie, T., and Tibshirani, R. 2006. Sparse principal component analysis. J. Comput. Graph. Statist. **15**(2): 265–286. doi:10.1198/106186006X113430.