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

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

Journal: Canadian Journal of Fisheries and Aquatic Sciences

Abstract

Predicting species abundance is critical for understanding ecological dynamics and guiding conservation and management strategies. Traditional species abundance models (SAMs) 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 variables selection and lake characteristics influences model accuracy 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 were 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 integrating community composition into abundance modelling.

Keywords: Community composition; abundance; co-distribution; co-occurrence; latent; prediction; sport fish

Introduction

Species abundance is a fundamental indicator of population health and viability within ecosystems. It offers crucial information on a species' risk of local extinction, detectability, and ecological influence on their local communities, thereby informing conservation priorities and sustainable management practices. Understanding spatial patterns of species abundance is essential for determining whether populations are declining and require protection, or whether they can be harvested sustainably without compromising long-term viability (Degnbol and Jarre 2004). This knowledge is particularly valuable for policymakers, conservation practitioners, and resource managers striving to balance ecological sustainability with societal needs. Despite its importance, accurately estimating species abundance remains a major challenge. Data collection typically requires intensive fieldwork, make it both costly and time-consuming (Yoccoz et al. 2001; Lindenmayer and Likens 2010; Dickinson et al. 2010). In additional, ethical considerations are increasingly relevant, especially for methods that involve fish capture and handling.

Sampling constraints often limit the frequency and spatial extent of abundance assessments (e.g., across multiple lakes, streams or watersheds), making it challenging to generate comprehensive data over large geographic areas, extended time periods (Jackson and Harvey 1997), and across multiple species. These limitations are especially challenging when timely conservation or management actions are required. To address these challenges, fisheries researchers often 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).

Traditional SAMs typically incorporate local and regional environmental variables 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 frequently overlook complex biotic interactions, such as competition and predation, that also influence abundance distributions (Mack et al. 2000; MacKenzie et al. 2002; Gaston 2003). Consequently, there is a persistent need to improve predictive models by incorporating additional data sources and quantitative frameworks that better account for the diverse factors influencing species abundance.

Stahl et al. (2024) proposed a framework that enhance species abundance predictions by integrating environmental variables with co-occurrence data. While earlier SAMs have included presence-absence data as predictors, they typically focused only on species with well-known interactions with the target species, such as those between a predator and its prey (Boulangeat et al. 2012; Lewis et al. 2017; Olkeba et al. 2020). In contrast, Stahl et al.'s approach incorporated presence-absence data for the entire local community as predictors of local abundance of a target species, offering a more comprehensive basis for predicting the abundance of a focal species. This approach offers at least two key advantages over traditional models. First, it leverages patterns of species co-occurrence as proxies for unmeasured environmental variables. Second, it allows the integration of interaction networks at both local and regional scales, using these networks to predict variation in species abundance for a taget species. The framework employs Gaussian copulas to generate latent variables from species covariation, enabling the identification of complex patterns in multispecies data (Popovic et al. 2018).

Latent variables were initially introduced to reduce dimensionality in community data (e.g., indirect gradient analysis) and have since been adapted to represent unobserved ecological factors and processed inferred from species covariation in ecological models such as hidden environmental drivers or biotic interactions (see Walker & Jackson 2011). These latent factors, estimated from co-occurrence matrices, aim to capture as much variation in community composition as possible. If community structure is primarily shaped by species responses to environmental gradients and local interspecific interactions, these latent variables can effectively stand in for missing predictors in abundance models. Stahl et al. (2024) showed that copula-based latent variables reliably act as proxies for unmeasured environmental gradients when applied to simulated community data. In simulation studies where species abundances were generated as linear functions of environmental conditions and location-specific process error—without including species interactions or nonlinear environmental responses—the latent variables successfully captured the underlying environmental structure driving abundance patterns. This predictive improvement held across a range of scenarios, underscoring the robustness and generality of the framework across diverse ecological contexts.

In real-world ecosystems, species interactions, such as competition, predation, and mutualism, play a fundamental role in shaping community structure and species abundance (Chase and Leibold 2003; Tylianakis et al. 2008). These interactions introduce ecological complexities that can be captured by latent variables, which serve as proxies for unmeasured ecological factors and processes. By capturing both environmental influences and species interactions, latent variables offer a more comprehensive representation of the factors driving species abundances. This dual capacity makes them particularly promising for improving the accuracy and robustness of ecological models when applied to empirical data. Here, we apply a latent abundance-predictive framework to a large empirical dataset of lake fish communities. Lake fish communities, being relatively more isolated systems compared to riverine and terrestrial system, often experience limited dispersal among sites. As a result, local species compositions and abundance are more likely to be shaped by in-lake environmental conditions and species interactions due to limited dispersal between lakes. As a result, local species compositions and abundance distributions are more likely to respond to local-lake influences, raising the possibility that variations between lakes could be effectively captured by latent factors.

Here, we apply the framework developed in Stahl et al. (2024), on 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 of 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 modeling 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 non-sport fish species, and (3) latent variables based on the full community, including both sport and non-sport fishes.

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

Material and method

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 (Figure 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 hours (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 meters of net over an 18-hour period. The number of species per caught per lake ranged from 2 to 25. We assumed that CPUE was an accurate proxy for local density of each species 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 potion of the region and are frequently target species for fisheries management (see Table 1 and Figure S2; 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 applying these filters, we retained 34 non-sport fish species and 14 sport fish species, resulting in a final dataset of 48 species across 594 lakes (Figure 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 S2). 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 S3 and Figure S3).

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 with a binomial family to model species occurrence, 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) non-sport fish species, and (3) all fish species. These latent variable sets were then used as predictors in our single-species abundance models 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. The groups were also structured to reflect management's varying interests. For example, if a group of species is identified as important for predicting the abundance of a target species, it could strengthen the case for incorporating them into management strategies aimed at the target species. To maintain consistency in the numbers of predictors, we limited the number of latent variables to four for each group (Stahl et al. 2024). 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 (Figure S4).

Modelling structure overview

To apply the framework from Stahl et al. (2024) 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) non-sport 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 non-sport 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 the conditional mean abundance for a species in each lake is the sum of (possibly nonlinear) functions of lake-specific covariates (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):

$$LE_{s,m,l} = log_{10}\left(\frac{\hat{Y}_{s,m,l}}{Y_{s,m,l}}\right) \qquad \text{Equation 1}$$

where s, m, 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 (Equation 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, Equation 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).

$$\Delta LE_{s,m} = Med\left(\left|LE_{s,l,m_0}\right|\right) - Med\left(\left|LE_{s,l,m}\right|\right) \qquad \text{Equation 2}$$

where *s*, *m*, *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. 2024), 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, non-sport 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, non-sport 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 calculated (1) the environmental distinctiveness of a lake as the lake pairwise Mahalanobis distance matrix based on environmental variation (i.e., PCA axes), and (2) the ecological distinctiveness of a lake, in terms of species composition, was quantified using its Local Contribution to Beta Diversity (LCBD, Legendre & De Cáceres, 2013). 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.

$$Contribution_{l,s} = \mathrm{Med}_{l\in C_j}\left(\left|LE_{s,j}\right|\right) - \mathrm{Med}_{l\in V_j}\left(\left|LE_{s,j}\right|\right) \qquad \text{Equation 3}$$

where *l, s*, *j* are indices for lakes, and replicates, respectively. The median (referred to as Med in Eq 3) $LE_{j,s}$ 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 Principal Component Analysis (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, 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).

$$\Delta SLE_{s,m} = \left|\frac{1}{M}\sum_{l \in M} LE_{l,s}\right| - \left|\frac{1}{L_s}\sum_{l \in L_s} LE_{l,s}\right| \qquad \text{Equation 4}$$

where s, m, *l, M, 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 (Figure 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, non-sport 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 (Figure 3, Table S3). 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 non-sport fish species. The remaining four species were most accurately predicted by the model that included only sport fish species.

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 (Figure 4). The LE metric showed no correlation with lake rarity, whether defined by environmental characteristics (Mahalanobis distance) or by species composition (LCBD). This suggests that predictive ability is not primarily driven by whether lake types are common or rare, although certain lake characteristics may still influence predictive 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 (TSI) based on phosphorus levels (Figure S5). No clear pattern emerged in relation to key environmental variations. 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 specific environmental factors, reinforcing the general applicability of the predictive framework across diverse lake types.

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, non-sport fish species, and all fish species; Figure S6). 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 (Figure 5). For rainbow smelt, lake trout, 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. 2024, 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. The results indicate that models containing latent variables primarily improved predictions for high-occurrence species, whereas low occurrence species are betted predicted by the environmental model alone, highlighting that the benefits of including latent variables are species-specific rather than uniform across the community.

Our second goal was to assess whether the choice of species subset to generate latent variables impacted predictive performance. We found that no single species subset performs best across all target species. This suggests the framework's effectiveness is relative insensitive to species subsets. The findings are consistent with the original framework assessment using simulated data, which also showed better that incorporating latent variables yielded better predictions for higher-occurrence species. They also align with the broader literature, which suggests that low-occurrence species are generally more vulnerable to stochastic environmental fluctuations and demographic instability (Gaston 1994; Brown et al. 1995), while high-occurrence species tend to engage in more complex biotic interactions (Mouquet et al. 2003; Araújo and Luoto 2007). This could suggest that latent variables are most beneficial for predicting high-occurrence species, which are more influenced by biotic interactions, whereas low-occurrence species are predominantly shaped by stochastic environmental and demographic processes. However, it is possible that these outcomes are system-specific, and the modelling framework could perform better for low-occurrence species in other ecosystems. The framework is flexible enough to be generalized across different taxa and systems. Future applications could explore alternative methods for combining species to generate latent variables that maximize the predictive accuracy for target species, such as using model selection tailored to select species combinations that improve predictions for specific species (see below for other alternative for species selections).

Our third goal was to identify the types of lakes that strongly influenced predictive ability, either positively or negatively, by examining the relationship between log error contribution and both environmental and community composition distinctiveness (LCBD). We found no correlation between log error contribution and the rarity or commonality of lake environmental features, community compositions, or specific environmental features. Essentially, this suggests that large lakes are just as likely to improve predictions as small lakes, and models' predictive ability is not influenced by specific environmental attributes or species compositions. On one hand, this finding is significant as it challenges the common assumption that certain environmental and biotic characteristics inherently enhance predictive power in ecological models. For instance, one might expect larger lakes, being more stable (May 1972) and supporting more diverse habitats, to provide more reliable predictions (Magnuson et al. 2005).Alternatively, larger lakes may be less predictable because their greater abundance of microhabitats and, as a result, local environmental variation (Strayer & Findlay 2010) is often not fully captured by standard environmental measurements. On the other hand, the results suggest that predictive accuracy is not inherently tied to these environmental complexities, increasing the generality of our predictive framework across various and diverse lakes. This implies that our models are robust across different environmental contexts, a valuable attribute 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 Dormann et al. 2013; Elith et al. 2010; Thuiller et al. 2005 among others), highlighting the importance of incorporating species-specific ecological dynamics 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 too 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 (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 has limitations. A key limitation is that it relies on data from lake ecosystems, where dispersal is relatively restricted and species are likely more strongly shaped by local environmental conditions. While our modelling framework is applicable to any system, the empirical findings derived from our studied lake system may limit the generalizability to other ecosystems, particularly those where species dispersal plays a more dominant force in shaping community structure and species distributions (Leibold et al. 2004; Peres-Neto et al. 2012; Thompson & Gonzalez 2017; Urban et al. 2012). Additionally, generating latent variables from presence-absence data may oversimplify the ecological processes influencing species abundance, especially in communities with complex, non-linear, or context-dependent interactions. For example, mutualistic or competitive interactions that vary in strength across different environmental conditions may not be adequately captured by latent variables derived from binary data (Ovaskainen et al. 2017 but see Clark et al. 2018 for a method that does) but see Clark et al. 2018 for a method that does). This simplification can introduce biases in model predictions, particularly when addressing intricate species interactions or generalizing results across different ecosystems.

Another limitation is our use of random sampling to split calibration and validation sets for simplicity and efficiency. DiRenzo et al. (2023) and Roberts et al. (2017) recommend more robust methods such as spatial cross-validation or blocking, especially in cases where data are autocorrelated or where the covariance structure of predictors shifts between datasets. As Wenger & Olden (2012) point out, failing to account for these factors can reduce the transferability and accuracy of ecological models. Incorporating techniques such as stratified sampling may yield more reliable predictions. In summary, 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 data via latent variable into predictive models for species abundance. Our findings highlight the importance of considering species occurrence patterns and environmental affinities when developing predictive models, as clustering species based on these factors can enhance model accuracy. This reinforces the notion that tailored modelling approaches are essential for understanding and managing 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, and received a QCBS travel grant to present this work at a conference (SFE2 2024 in Lyon, France). AS and PPN were funded by the Canada Research Chair in Spatial Ecology and Biodiversity. EP and PPN were funded by the NSERC Discovery Grant program.

Author Contributions

Conceptualization: AS, EP, PPN. Data curation: AS. Formal analysis: AS. Coding: AS, EP, PPN. Methodology: AS, EP, PPN. Visualization: AS. Writing – original draft: AS. Writing – review and editing: AS, EP, PPN. All authors read and approved the final manuscript.

Competing interests statement

The authors declare there are no competing interests.

Data availability statement

Data are currently provided within the manuscript and/or supplemental files, but will be uploaded to an appropriate public repository and a DOI will be provided at the time of acceptance.

References

Appelberg, M. 2000. Swedish Standard Methods for Sampling Freshwater Fish with Multi-mesh Gillnets. Fiskeriverket Information 2000 **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 Ecology and Biogeography **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. Ecological Society of America. doi:10.1002/ecy.3608.

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? Journal of Animal Ecology **85**(1): 11–20. Blackwell Publishing Ltd. 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. *In* Source: Ecology.

Burnham, K.P., and Anderson, D.R. 2004. Model Selection and Multimodel Inference. *Edited By*K.P. Burnham and D.R. Anderson. Springer New York, New York, NY. doi:10.1007/b97636.

Chase, J.M., and Leibold, M.A. 2003. Ecological Niches. University of Chicago Press. doi:10.7208/chicago/9780226101811.001.0001.

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.

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. British Ecological Society. doi:10.1111/2041-210X.14030.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., and Lautenbach, S. 2013. Collinearity: A

Review of Methods to Deal with it and a Simulation Study Evaluating their Performance. Ecography **36**(1): 27–46. Blackwell Publishing Ltd. doi:10.1111/j.1600-0587.2012.07348.x.

Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guénard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., and Wagner, H. 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. Wiley. 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. Oxfort University Press.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. The Elements of Statistical Learning. Springer New York, New York, NY. doi:10.1007/978-0-387-84858-7.

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 vs. 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. Canadian Journal of Fisheries and Aquatic Sciences **54**(12): 2807–2813. doi:10.1139/f97-182.

Legendre, P. 1993. Spatial Autocorrelation: Trouble or New Paradigm? Ecology **74**(6): 1659–1673. doi:10.2307/1939924.

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 Fortin, M.-J. 1989. Spatial Pattern and Ecological Analysis. Vegetatio **80**: 107–138.

Legendre, P., and Legendre, L. 2012. Numerical Ecology. *In* 3rd edition. Elsevier.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D., Loreau, M., and Gonzalez, A. 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 Resour **9**(1): 23–29. ESME - Gauthier-Villars. doi:10.1051/alr:1996004.

Lester, N.P., Sandstrom, S., de Kerckhove, D.T., Armstrong, K., Ball, H., Amos, J., Dunkley, T., Rawson, M., Addison, P., Dextrase, A., Taillon, D., Wasylenko, B., Lennox, P., Giacomini, H.C., and Chu, C. 2021. Standardized Broad-Scale Management and Monitoring of Inland Lake Recreational Fisheries: An Overview of the Ontario Experience. Fisheries (Bethesda) **46**(3): 107–118. John Wiley and Sons Inc. 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**. Nature Publishing Group. doi:10.1038/srep44152.

Lindenmayer, D.B., and Likens, G.E. 2010. Effective Ecological Monitoring. CSIRO Publishing.

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. Ecological Society of America. doi:10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2.

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

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. Oxford University Press.

Mahalanobis, P.C. 1936. On the Generalized Distance in Statistics. Proceedings of the National Institute of Sciences of India.

Marra, G., and Wood, S.N. 2011. Practical Variable Selection for Generalized Additive Models. Comput Stat Data Anal **55**(7): 2372–2387. doi:10.1016/j.csda.2011.02.004.

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. Parasit Vectors **13**(1). BioMed Central Ltd. doi:10.1186/s13071-020-04163-6.

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].

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

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., Roslin, T., and Abrego, N. 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. Blackwell Publishing Ltd. doi:10.1111/ele.12757.

Popovic, G.C., Hui, F.K.C., and Warton, D.I. 2018. A General Algorithm for Covariance Modeling of Discrete Data. J Multivar Anal **165**: 86–100. Academic Press Inc. 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. British Ecological Society. 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. British Ecological Society. 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., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., and Dormann, C.F. 2017, August 1. 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. Elsevier B.V. doi:10.1016/j.fishres.2019.105382.

Stahl, A., Pedersen, E.J., and Peres-Neto, P.R. 2024. Advancing Single Species Abundance Models: Robust Models for Predicting Abundance Using Co-occurrence from Communities. EcoEvoRxiv [preprint]. doi:https://doi.org/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). Nature Publishing Group. 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. Proceedings of the National Academy of Sciences **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. Journal of the Operational Research Society **66**(8): 1352–1362. Palgrave Macmillan Ltd. 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.

Tylianakis, J.M., Didham, R.K., Bascompte, J., and Wardle, D.A. 2008, December. Global Change and Species Interactions in Terrestrial Ecosystems. doi:10.1111/j.1461-0248.2008.01250.x.

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. American Naturalist **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., Tjiputra, J., and Pellissier, L. 2022. A Quantitative Review of Abundance-Based Species

Distribution Models. Ecography **2022**(1). John Wiley and Sons Inc. 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.

Whittaker, R.H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecol Monogr **30**(3): 279–338. doi:10.2307/1943563.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., Elith, J., Dudík, M., Ferrier, S., Huettmann, F., Leathwick, J.R., Lehmann, A., Lohmann, L., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.C., Phillips, S.J., Richardson, K.S., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S.E., and Zimmermann, N.E. 2008. Effects of Sample Size on the Performance of Species Distribution Models. Divers Distrib **14**(5): 763–773. doi:10.1111/j.1472-4642.2008.00482.x.

Wood, S.N. 2003. Thin-Plate Regression Splines. Journal of the Royal Statistical Society (B) **65**(1): 95–114.

Wood, S.N. 2004. Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. J Am Stat 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 Series 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 edition. Chapman and Hall/CRC. doi:https://doi.org/10.1201/9781315370279.

Wood, S.N., Pya, N., and Säfken, B. 2016. Smoothing Parameter and Model Selection for General Smooth Models. J Am Stat Assoc **111**(516): 1548–1563. American Statistical Association. doi:10.1080/01621459.2016.1180986.

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. Journal of Computational and Graphical Statistics **15**(2): 265–286. doi:10.1198/106186006X113430.

Tables

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

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
Non-sport 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

Table 2. Mean and standard deviation of correlation between species groups across models. We calculated the correlation between lake contributions for each species and model, revealing distinct grouping patterns (see Figure S6). 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.

	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

Figure captions

Figure 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).

Figure 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) non-sport 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.

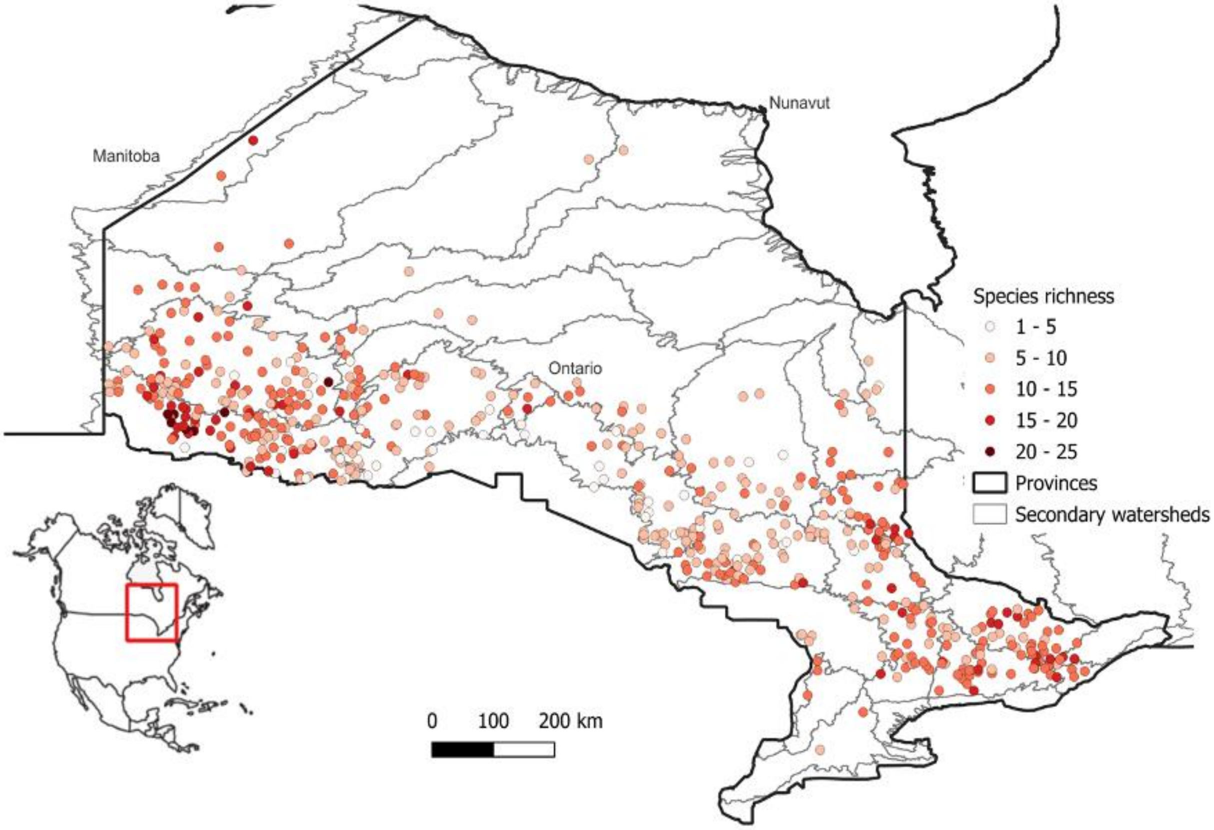
Figure 3. Density plot of the log error as a function of species and model. The log error was calculated following Eq. 1, and for each lake, the median log error was taken across replicates for each species and model. Latent variables were generated using three groups: (1) sport fish species (green), (2) non-sport 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.

Figure 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 Appendix 2 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.

Figure 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.

910 **Figures**



911

912 **Figure 1.**

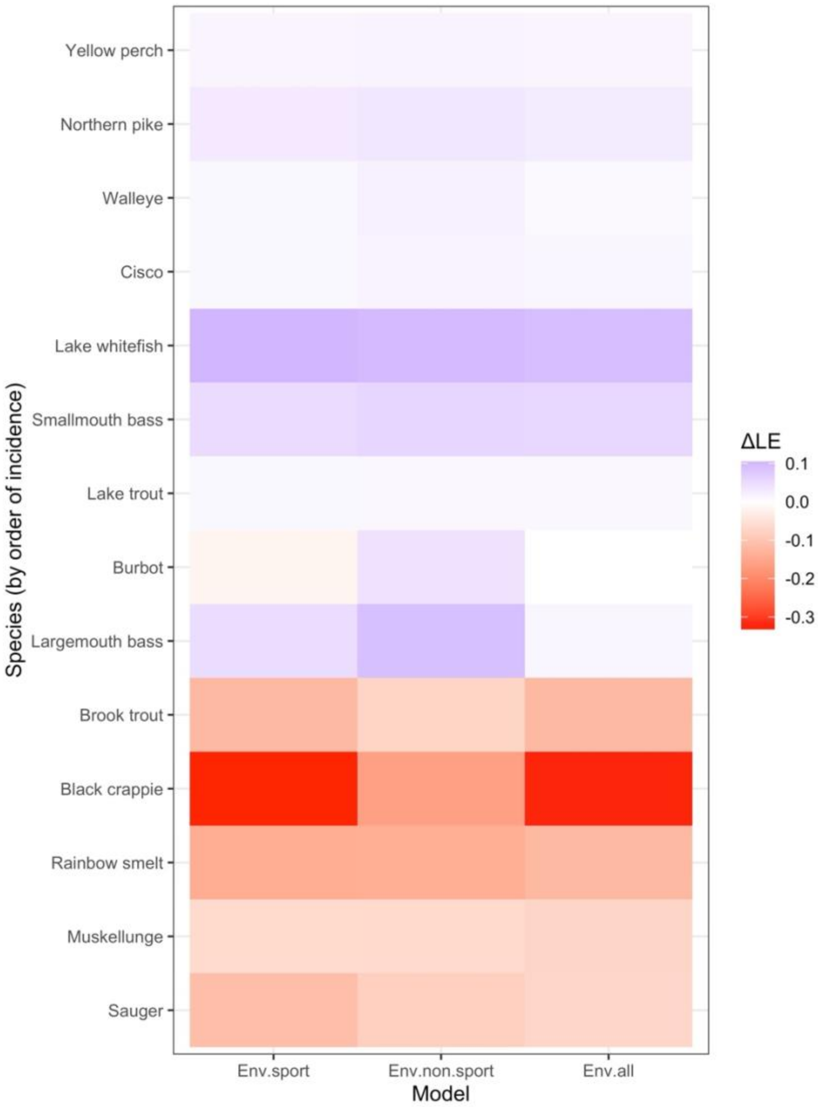
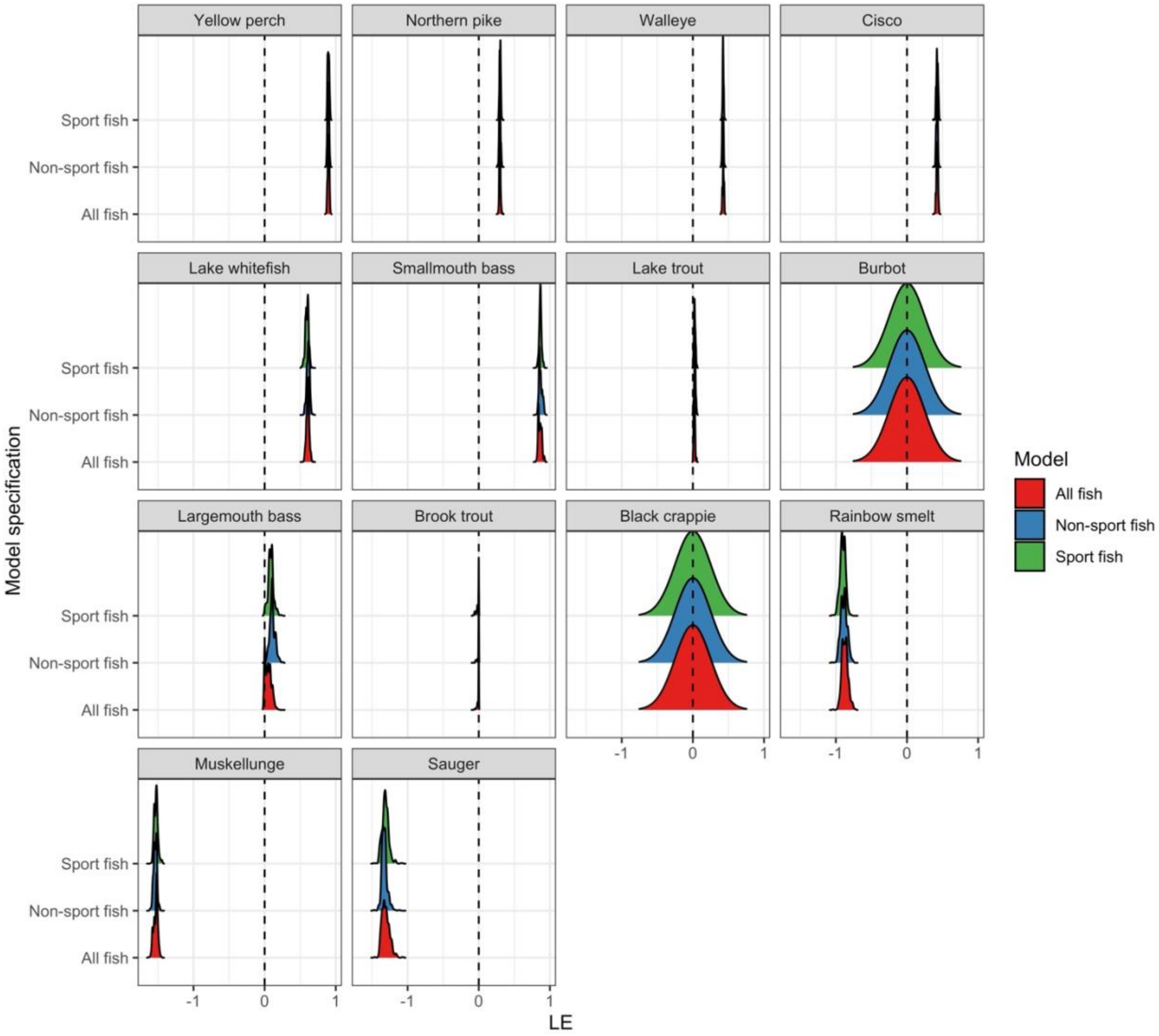

Yellow perch
Northern pike
Walleye
Cisco
Lake whitefish
Smallmouth bass
Lake trout
Burbot
Largemouth bass
Brook trout
Black crappie
Rainbow smelt
Muskellunge
Sauger
Species (by order of incidence)
Env.sport
Env.non.sport
Env.all
Model
ΔLE
0.1
0.0
-0.1
-0.2
-0.3

Figure 2.



915

916 **Figure 3.**

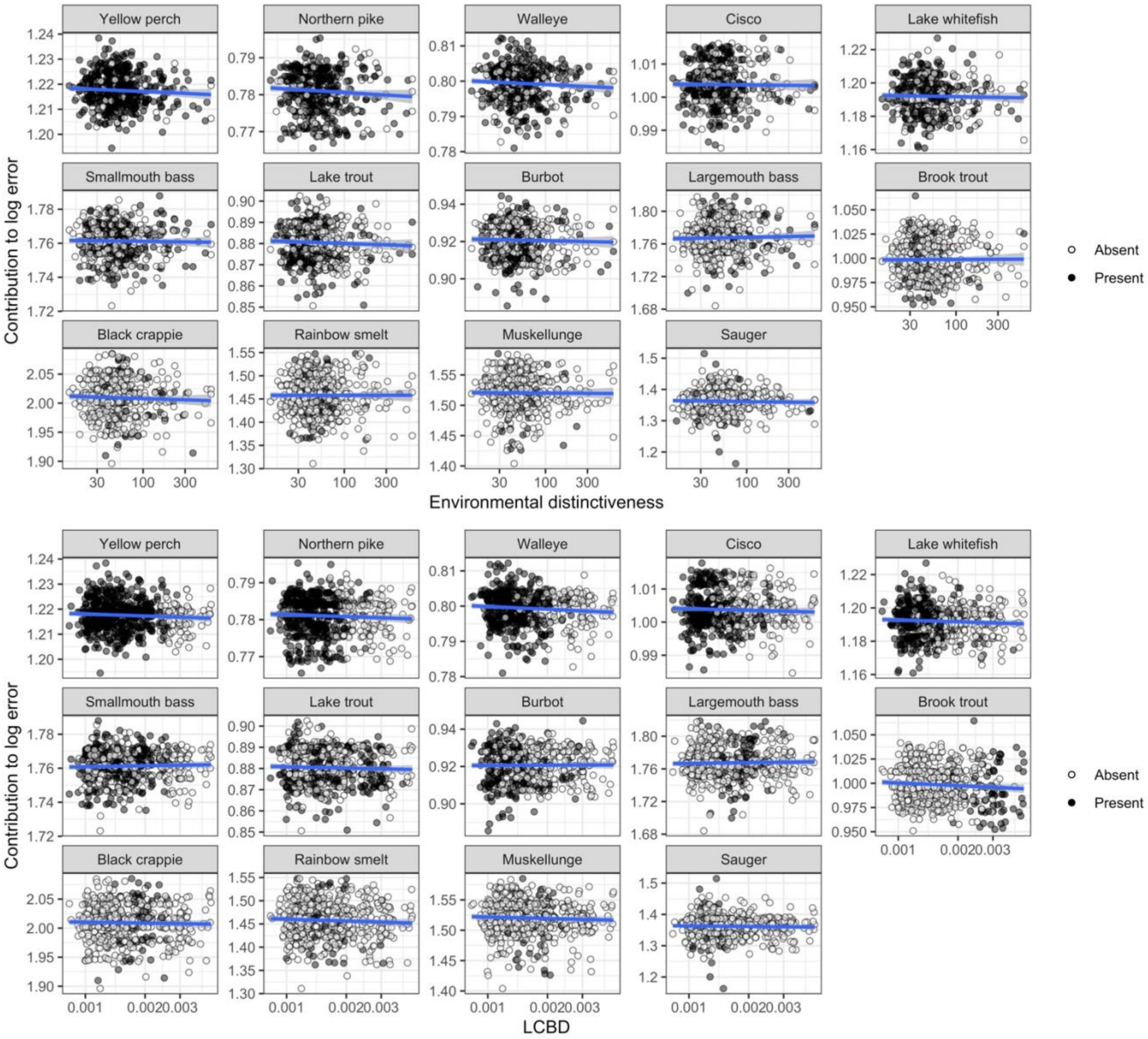
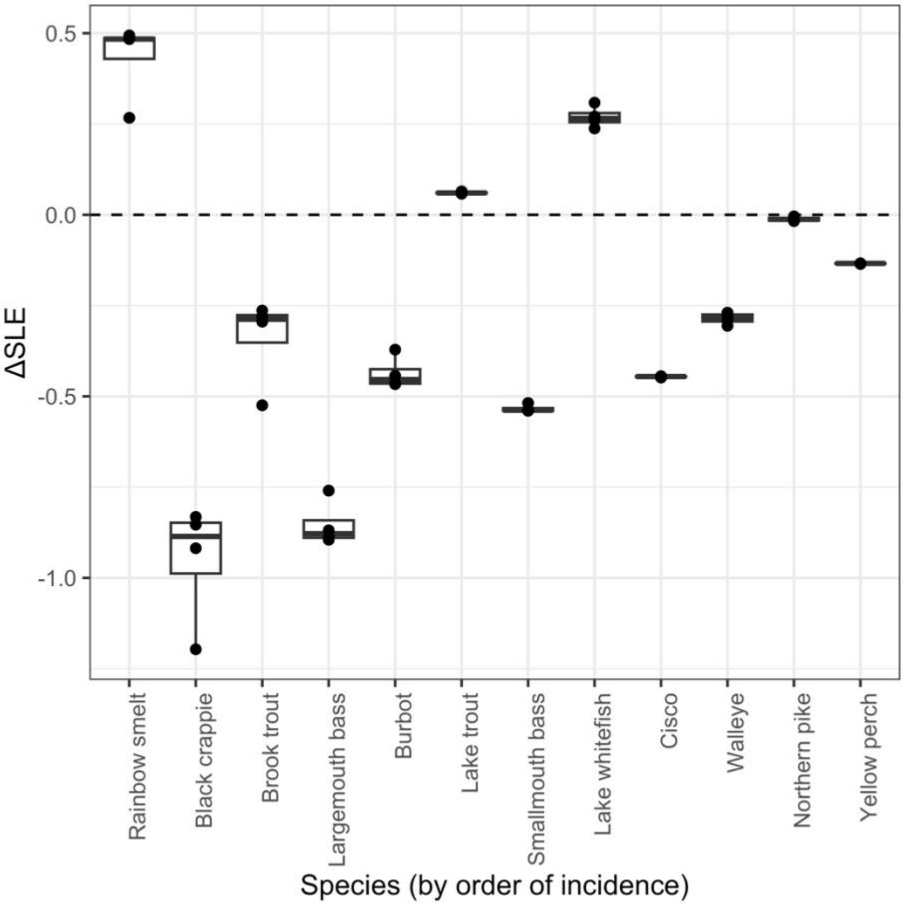

Yellow perch
Northern pike
Walleye
Cisco
Lake whitefish
Smallmouth bass
Lake trout
Burbot
Largemouth bass
Brook trout
Black crappie
Rainbow smelt
Muskellunge
Sauger
Contribution to log error
Environmental distinctiveness
30
100
300
Absent
Present
Yellow perch
Northern pike
Walleye
Cisco
Lake whitefish
Smallmouth bass
Lake trout
Burbot
Largemouth bass
Brook trout
Black crappie
Rainbow smelt
Muskellunge
Sauger
Contribution to log error
LCBD
0.001
0.0020.003
Absent
Present

Figure 4.



919

920 **Figure 5.**

921

Appendix: Identification of optimal number of composite environmental variables and latent variables.

Methods

Given the high dimensionality of our data, we needed to decide how many variables to use in recombining the environmental variables, as well as how many latent variables to generate to best predict species abundance. To optimize these selections, we performed a two-step analysis. First, we fixed the number of one group of variables while varying the other (i.e., environmental variables or latent variables) and then repeated the process in reverse. Specifically, we set the number of variables to five for the fixed group and tested variables ranging from 2 to 15 in increments of 1, as well as 17 and 20 for the varying group. For each tested combination, we randomly split the data into calibration and validation sets (respectively 292 and 291 lakes). We then fitted a Generalized Additive Model (GAM) with a Tweedie distribution, using the functions *tw* and *gam* from the R package *mgcv* (Wood 2004; Wood et al. 2016, version 1.9-1). Each explanatory variable was fitted with a 2nd order thin-plate regression spline smoother (Wood 2003) with 3 bases functions using the function *s* from the R package *mcgv* and linking the smoothing parameters across environmental and latent variables. All models were estimated using restricted maximum likelihood (Wood 2011) using only data from the calibration set and used the double penalty approach for term selection (Marra and Wood 2011). This procedure was repeated 100 times and for six species with different occurrence rates representative of the whole dataset (Table S1). The out-of-sample average prediction was calculated across replicates, and the median across species of the Mean Squared Error (MSE) was derived.

Table S1: List of species considered in the dataset, including both common and scientific name as well as percentage of occurrence in the dataset. Species are organized by occurrence, with high occurrence species at the top of the table and low occurrence species at the bottom of the table.

Common name	**Scientific name**	**Occurrence rate (in %)**
Lake whitefish	*Coregonus clupeaformis*	54
Common shiner	*Luxilus cornutus*	23
Black crappie	*Pomoxis nigromaculatus*	10
Brook stickleback	*Culaea inconstans*	6
Fallfish	*Semotilus corporalis*	2
Channel catfish	*Ictalurus punctatus*	1

Results

When fixing the number of latent variables and varying the number of environmental variables, the lowest Mean Squared Error (MSE) was observed when using 10 environmental variables (Figure S1). Conversely, when fixing the number of environmental variables and varying the number of latent variables, the lowest MSE was achieved with four latent variables. This pattern aligns with expectations, where MSE typically decreases as the number of variables increases until an optimal point is reached, after which overfitting causes the error to rise. Overfitting occurs because the model becomes overly complex, capturing noise in the training data rather than the underlying signal, leading to poorer generalization to new data (Burnham and Anderson 2004; Hastie et al. 2009). Therefore, we selected 10 environmental variables and four latent variables for generating the composite environmental variables and latent variables in the main analysis.

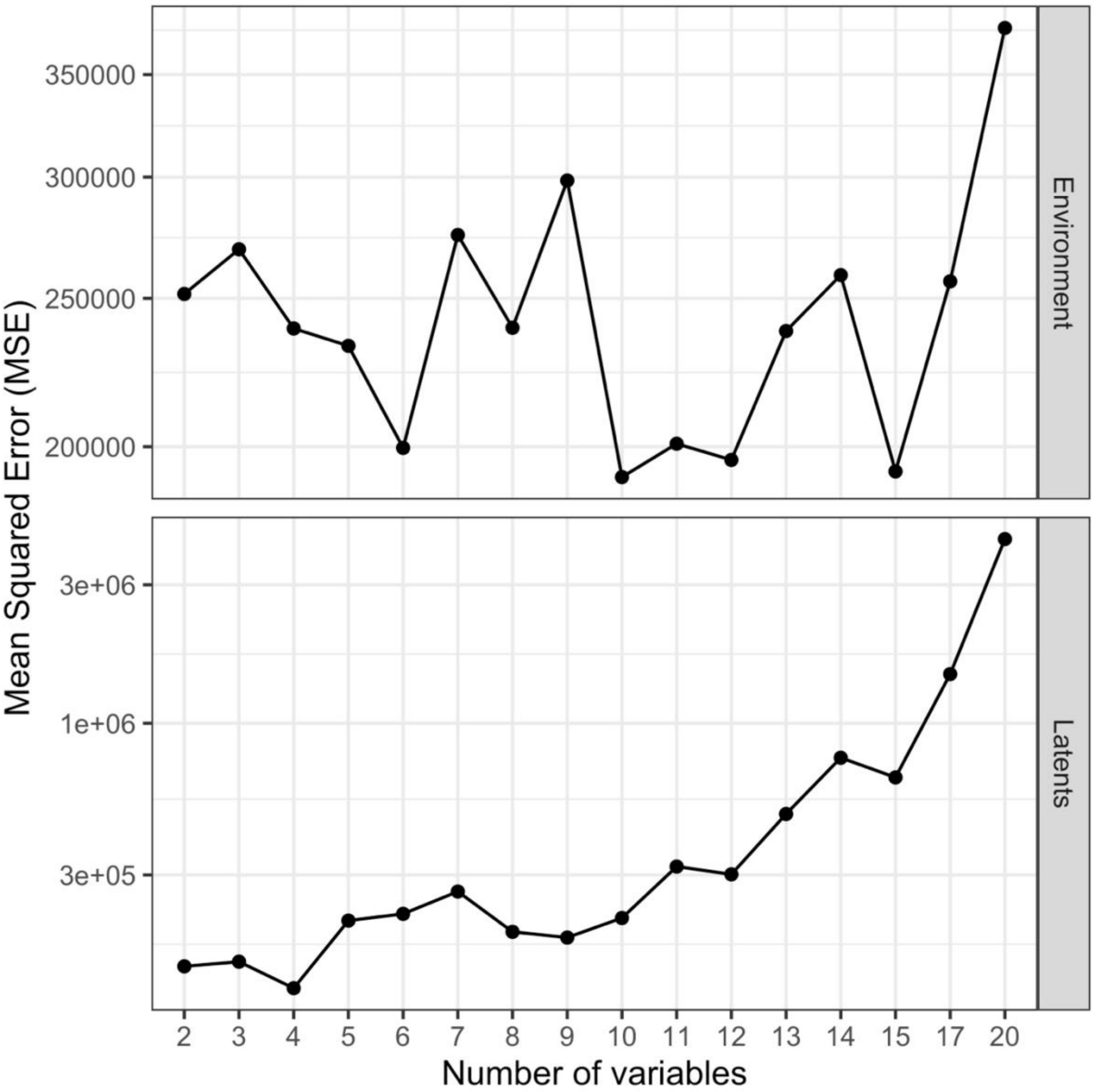


Figure S1: Median Mean Squared Error (MSE) as a function of number of composite environmental and latent variables. The figure shows the median Mean Squared Error (MSE); with the MSE calculated for out-of-sample abundance predictions across replicates and the median calculated across species. The number of variables generated was varied from 2 to 15 in increments of 1, as well as 17 and 20, while the fixed group used 5 variables. Each facet indicates the group being varied. The MSE is represented on a $\log_{10}$ scale, with the expectation of observing a decrease in MSE until an optimal point is reached, after which the error increases due to model overfitting.

Supplementary information

Table S2: Table of environmental variables and their units grouped by categories (e.g., climate, productivity). See Sandstrom et al. (2011) for details on sampling methods.

Category	Environmental variable
Hydro morphology	Area (km^2)
	Maximum lake depth (m)
	Minimum lake depth (m)
	Numeric code indicating lake size
	Observed hypolimnetic area
	Observed hypolimnetic volume
	Observed thermocline depth (m)
	Perimeter lake (no islands, km)
	Proportion of lake area below 20m in depth
	Proportion of littoral (< 4.6m)
	Shoreline development factor
	Total shoreline of lake (perimeter and islands, km)
	Volume (m^3)
Fishing activities	Annual angling pressure based on aerial survey counts (angler-hours/ha-year)
	Conservation status (binary; 1 implies some form of conservation status)
	Fisheries management zone (categorical)
	Mean count of fishing boats in summer
	Mean count of ice huts in winter
	Mean count of open ice fishers in winter
	Mean count of shore fishers in summer
Productivity	Dissolved Inorganic Carbon (mg.L)
	Dissolved Organic Carbon (mg.L)
	Ratio of ammonia over ammonium (mg.L)
	Ratio of nitrate over nitrite (ug.L)
	Secchi depth of lake in spring (m)
	Total dissolved solids (mg.L)
	Total Kjeldahl nitrogen (ug.L)
	Total phosphorus (ug.L)
	Trophic status index based on phosphorous
	True color (TCU) (see Moore et al. 1997 for details)
Climate	Average date of the first day above 0°C (ordinal day)
	Average date of the last day above 0°C (ordinal day)
	Average rainfall from 1981-2010 (mm)
	Cumulative degree days where temperature was above 0°C
	Cumulative degree days where temperature was below 0°C
	Degree days above 5°C from 1981-2010
	Maximum monthly air temperature (°C)
	Maximum surface temperature (°C)

Category	Environmental variable
	Maximum water temperature (°C)
	Mean annual air temperature from 1981-2010 (°C)
	Minimum monthly air temperature (°C)
	Number of days where temperature was above 0°C
	Number of ice-free days
	Proportion of cold days (between 8 and 12°C) during ice free period
	Proportion of cool days (between 22 and 26°C) during ice free period
	Proportion of warm days (between 16 and 20°C) during ice free period
Watershed characteristics	Age of tertiary watershed
	Altitude above sea level (m)
	Elevation within tertiary watershed (max-min, m)
	Tertiary watershed area (km^2)
	Tertiary watershed elevation (meters above sea level)
Water chemistry	Alkalinity (mg.L.CaCO3)
	Calcium concentration (mg.L)
	Chloride concentration (mg.L)
	Conductivity (uS.cm.s)
	Iron
	Magnesium concentration (mg.L)
	pH
	Potassium concentration (mg.L)
	Silicate concentration (mg.L)
	Sodium concentration (mg.L)
	Sulphate concentration (mg.L)

Table S3: Table of the loadings of the PCA conducted on 64 environmental variables. We kept the first 10 axes of the PCA. Environmental variables are grouped by categories (e.g., climate, productivity). See Sandstrom et al. (2011) for details on sampling methods.

Variable	**Axis 1**	**Axis 2**	**Axis 3**	**Axis 4**	**Axis 5**	**Axis 6**	**Axis 7**	**Axis 8**	**Axis 9**	**Axis 10**
Latitude	-0.89	0.11	-0.06	-0.12	-0.01	0.33	0.05	-0.08	0.03	0.15
Longitude	0.63	-0.09	0.04	0.17	0.11	-0.6	-0.28	0.09	-0.08	-0.02
Area (km2)	-0.1	0.2	-0.75	0.04	-0.13	-0.1	-0.08	0.03	-0.11	0
Maximum lake depth (m)	-0.02	-0.22	-0.37	-0.09	-0.76	-0.02	-0.07	-0.06	0.17	-0.14
Minimum lake depth (m)	0	-0.24	-0.13	-0.12	-0.9	0.01	-0.03	-0.03	0.16	-0.09
Numeric code indicating lake size	-0.21	0.02	-0.71	0.16	-0.23	0.08	0.04	-0.05	0.24	0.14
Observed hypolimnetic area	0.07	-0.08	0.06	-0.06	-0.79	0.04	0.07	0.05	-0.39	0.1
Observed hypolimnetic volume	0.03	-0.12	-0.02	-0.08	-0.8	-0.01	0.07	0.04	-0.35	0.06
Observed thermocline depth (m)	-0.15	-0.07	-0.21	0.09	-0.07	0.05	-0.1	-0.01	0.74	-0.01
Perimeter lake (no islands	-0.12	0.02	-0.96	0.01	-0.08	0.04	-0.01	-0.01	0.01	0.01
Proportion of lake area below 20m in depth	0.01	0.24	0.11	0.09	0.87	0.02	0.01	0.01	-0.16	0.05
Proportion of littoral (< 4.6m)	-0.06	0.27	0.06	0.17	0.73	-0.07	-0.09	0.07	-0.06	0
Shoreline development factor	-0.04	-0.12	-0.89	-0.06	0.09	0.14	0.12	-0.07	0.05	0.02
Total shoreline of lake (perimeter and islands	-0.1	0.01	-0.96	0	-0.04	0.03	0.01	0	-0.04	0
Volume (m3)	-0.04	0.15	-0.59	0	-0.33	-0.08	-0.21	0.01	0.01	-0.14
Annual angling pressure based on aerial survey counts (angler-hours/ha-year)	0.46	0.06	0.02	0.31	0.14	0.12	0.03	0.7	-0.02	-0.2
Conservation status (binary; 1 implies some form of conservation status)	0.01	0.03	-0.28	-0.15	-0.2	-0.13	-0.12	-0.08	0.13	-0.08
Fisheries management zone (categorical)	0.85	-0.07	0.04	0.2	0.08	-0.32	-0.21	0.08	-0.05	-0.06
Mean count of fishing boats in summer	0.45	0.07	-0.01	0.36	0.17	0.18	0.02	0.54	-0.03	-0.23
Mean count of ice huts in winter	0.09	0.01	0.05	0.06	-0.1	-0.27	0.03	0.66	0.12	0.22
Mean count of open ice fishers in winter	0.18	-0.11	0.1	-0.01	0.04	0	-0.07	0.71	-0.07	-0.08

Variable	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Axis 7	Axis 8	Axis 9	Axis 10
Mean count of shore fishers in summer	0.08	0	-0.03	-0.03	-0.04	-0.02	-0.09	-0.02	-0.31	0.05
Dissolved Inorganic Carbon (mg.L)	0.03	0.06	0.01	0.88	0.08	0	-0.04	0	-0.03	-0.05
Dissolved Organic Carbon (mg.L)	-0.43	0.65	-0.01	0.01	0.34	0.12	-0.17	-0.07	0.07	-0.11
Ratio of ammonia over ammonium (mg.L)	0.2	0.31	0.09	0.4	0.25	-0.24	0.26	0	-0.2	-0.05
Ratio of nitrate over nitrite (ug.L)	0.18	0.04	-0.04	0.09	-0.17	-0.07	0.07	0.08	0.05	-0.68
Secchi depth of lake in spring (m)	0.19	-0.69	0.02	0.04	-0.4	0.01	-0.01	0.03	0	0.02
Total dissolved solids (mg.L)	0.25	0.08	0.01	0.94	0.06	-0.05	0.02	0.08	0.05	-0.03
Total Kjeldahl nitrogen (ug.L)	-0.02	0.71	0.05	0.4	0.34	0.07	0.1	-0.03	-0.05	-0.07
Total phosphorous (ug.L)	0.01	0.84	-0.06	0.34	0.1	-0.04	0.15	0.01	-0.1	0.06
Trophic status index based on phosphorous	-0.03	0.81	-0.03	0.33	0.27	0.06	0.06	0	-0.07	0.07
True color (TCU) (see Moore et al. 1997 for details)	-0.31	0.75	-0.04	-0.18	0.24	0.04	-0.18	-0.02	0.08	-0.18
Average date of the first day above 0°C (ordinal day)	-0.96	0.03	-0.03	-0.09	0.09	-0.12	-0.12	-0.03	-0.02	-0.03
Average date of the last day above 0°C (ordinal day)	0.92	-0.09	0.07	0.17	0	-0.24	-0.09	0.11	-0.05	-0.04
Average rainfall from 1981-2010 (mm)	0.71	-0.1	0.03	-0.04	0.02	-0.2	0.13	0.11	-0.02	-0.32
Cumulative degree days where temperature was above 0°C	0.94	-0.01	0.01	0.14	-0.08	0.11	0.06	0.04	0.01	0.16
Cumulative degree days where temperature was below 0°C	0.94	-0.12	0.08	0.09	-0.03	-0.2	0.01	0.08	-0.02	-0.1
Degree days above 5°C from 1981-2010	0.91	0.02	-0.02	0.16	-0.08	0.23	0.05	0.03	0.01	0.16
Maximum monthly air temperature (°C)	0.79	0.06	-0.03	0.1	-0.11	0.3	0.1	-0.01	0.04	0.33
Maximum surface temperature (°C)	0.89	-0.09	0.33	0.01	0.05	-0.07	0	0.07	-0.13	-0.09
Maximum water temperature (°C)	0.75	0.04	0.12	-0.08	0.35	0.02	0.22	-0.01	-0.08	0.16
Mean annual air temperature for 1981 and 2010 (°C)	0.97	-0.07	0.05	0.14	-0.05	-0.04	0	0.07	-0.02	-0.03
Minimum monthly air temperature (°C)	0.93	-0.12	0.08	0.12	-0.02	-0.2	-0.01	0.09	-0.03	-0.12

Variable	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Axis 7	Axis 8	Axis 9	Axis 10
Number of days where temperature was above 0°C	0.96	-0.07	0.05	0.15	-0.04	-0.11	-0.01	0.08	-0.02	-0.01
Number of ice-free days	0.94	-0.04	-0.11	0.21	-0.09	-0.02	-0.01	0.07	0.04	0.01
Proportion of cold days (between 8 and 12°C) during ice free period	-0.46	-0.08	0.08	-0.33	-0.23	-0.08	0.44	-0.12	0.21	0.09
Proportion of cool days (between 22 and 26°C) during ice free period	-0.85	0.08	-0.23	-0.14	-0.11	0.06	0.17	-0.12	0.22	0.06
Proportion of warm days (between 16 and 20°C) during ice free period	0.81	-0.04	0.16	0.21	0.16	-0.02	-0.27	0.13	-0.23	-0.07
Age of tertiary watershed	0.83	0.03	-0.02	0.08	-0.12	0.23	0.08	0.03	-0.04	-0.01
Altitude above sea level (m)	-0.5	-0.11	0.04	-0.43	0.08	0.14	0.36	-0.14	-0.02	-0.36
Elevation within tertiary watershed (max-min	0.24	-0.16	0.14	-0.18	0.02	-0.78	0.02	0.04	-0.08	-0.17
Tertiary watershed area (km2)	-0.56	0.08	-0.11	-0.09	0.02	0.37	0.03	-0.03	0.15	-0.08
Tertiary watershed elevation (meters above sea level)	-0.46	-0.03	-0.05	-0.4	-0.13	0.17	0.52	-0.13	0.09	-0.17
Alkalinity (mg.L.CaCO3)	0.17	0.04	0.03	0.94	0.1	0.03	-0.1	0	0	-0.03
Calcium concentration (mg.L)	0.18	0.05	-0.01	0.94	0.1	-0.03	-0.03	0.05	0.04	-0.04
Chloride concentration (mg.L)	0.39	0.14	-0.02	0.6	0.02	-0.03	0.34	0.23	0.12	-0.02
Conductivity (uS.cm.s)	0.25	0.08	0.01	0.94	0.07	-0.04	0.02	0.08	0.05	-0.03
Iron	-0.07	0.55	-0.01	-0.21	-0.1	0.17	-0.21	0	0.12	-0.06
Magnesium concentration (mg.L)	0.13	0.05	0.06	0.83	0.06	-0.01	-0.15	-0.03	0.03	-0.02
pH	-0.03	0.02	-0.04	0.84	0.1	0.07	-0.1	0.02	-0.04	0.17
Potassium concentration (mg.L)	0.28	0.31	-0.08	0.65	-0.09	0.18	0.3	0.1	0.11	0.01
Silicate concentration (mg.L)	-0.13	0.32	0.1	-0.06	0.12	-0.13	-0.19	-0.04	0.21	-0.42
Sodium concentration (mg.L)	0.33	0.18	-0.02	0.56	-0.01	-0.03	0.37	0.25	0.14	-0.03
Sulphate concentration (mg.L)	0.42	0.04	-0.03	0.32	-0.2	-0.39	0.25	0.22	0.14	0.03

Table S4: Table of the best model of all and the best latent model for each species. The models varied on whether they included (1) recombined environmental variables, (2) recombined environmental variables and latent variables generated from presence-absence of sport fish, (3) recombined environmental variables and latent variables generated from presence-absence of non-sport fish, and (4) recombined environmental variables and latent variables generated from presence-absence of all fish species. When identifying the best model, we selected the model with the median log error closest to 0. For the best model of all, we considered all four models and for the best latent model, we considered models 2, 3, and 4. Species are organised by occurrence, with high occurrence species at the top of the table and low occurrence species at the bottom of the table.

Common name	**Scientific name**	**Best model of all**	**Best latent model**
Yellow perch	*Perca flavescens*	Non sport fish	Non sport fish
Northern pike	*Esox lucius*	All fish	All fish
Walleye	*Sander vitreus*	Non sport fish	Non sport fish
Cisco	*Coregonus artedi*	All fish	All fish
Lake whitefish	*Coregonus clupeaformis*	All fish	All fish
Smallmouth bass	*Micropterus dolomieu*	All fish	All fish
Lake trout	*Salvelinus namaycush*	Non sport fish	Non sport fish
Burbot	*Lota lota*	Environmental	Sport fish
Largemouth bass	*Micropterus nigricans*	All fish	All fish
Brook trout	*Salvelinus fontinalis*	Environmental	Sport fish
Black crappie	*Pomoxis nigromaculatus*	Environmental	Non sport fish
Rainbow smelt	*Osmerus mordax*	Environmental	Non sport fish
Muskellunge	*Esox masquinongy*	Environmental	Sport fish
Sauger	*Sander canadensis*	Environmental	Sport fish

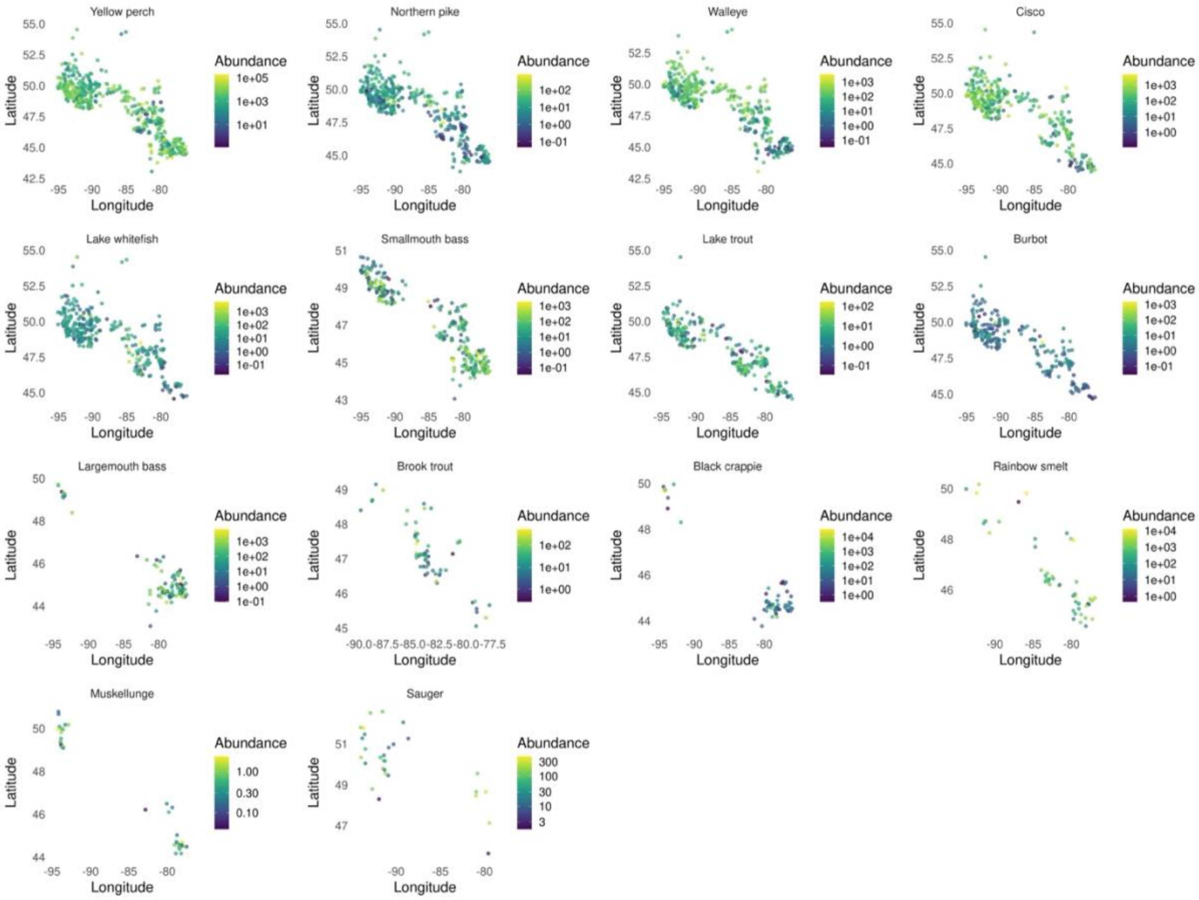


Figure S2: Maps showing the abundance distribution of each sport fish species. Species are organized by incidence within the dataset, with the most common species at the top and the least common at the bottom. Each point represents a lake where the species was observed. Abundance values are represented on a $\log_{10}$ scale, providing a clearer depiction of the wide range of abundance levels across the lakes.

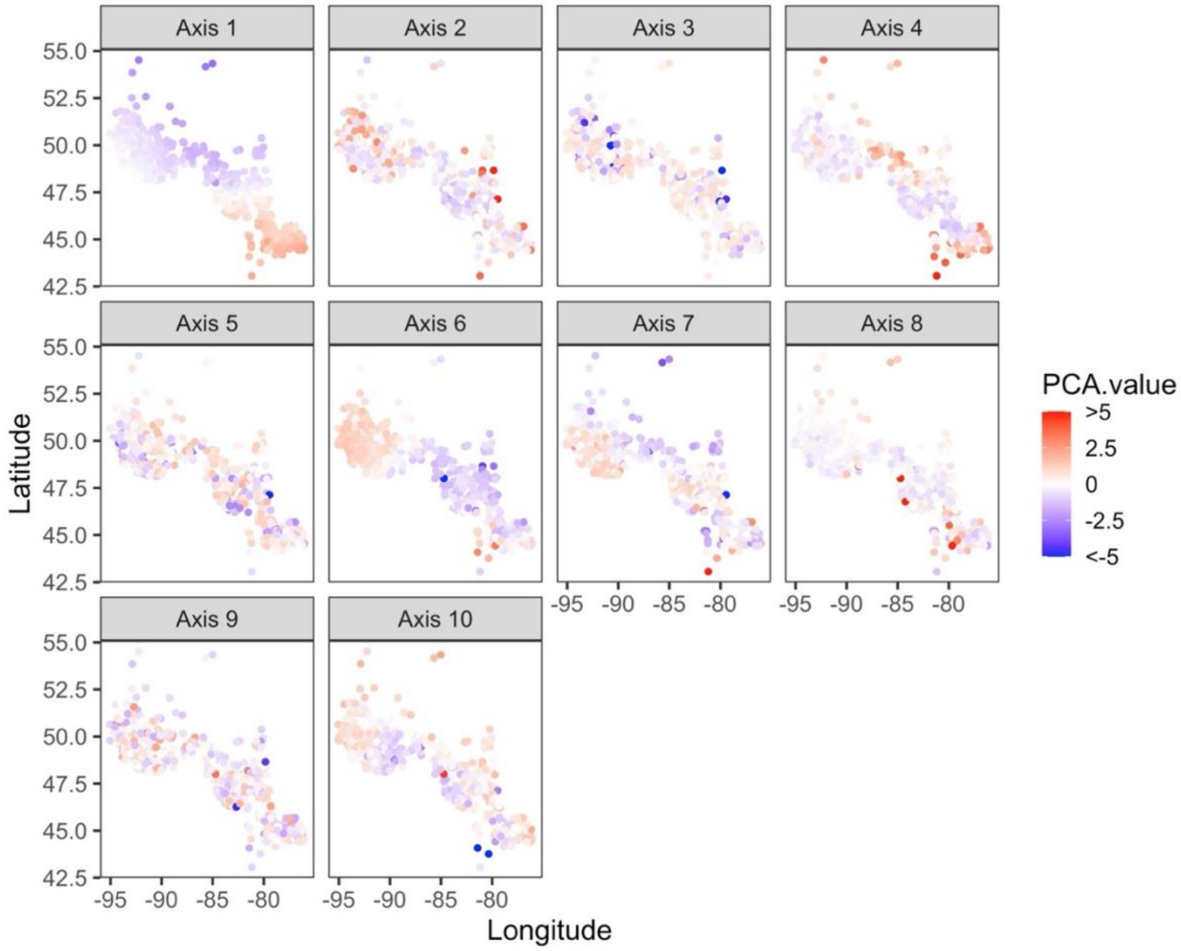


Figure S3: Maps illustrating the spatial patterns for the first 10 axes of the Principal Component Analysis (PCA) conducted on 64 environmental variables. These axes capture the major gradients in environmental variation across the study area, with each map representing one of the top 10 PCA axes.

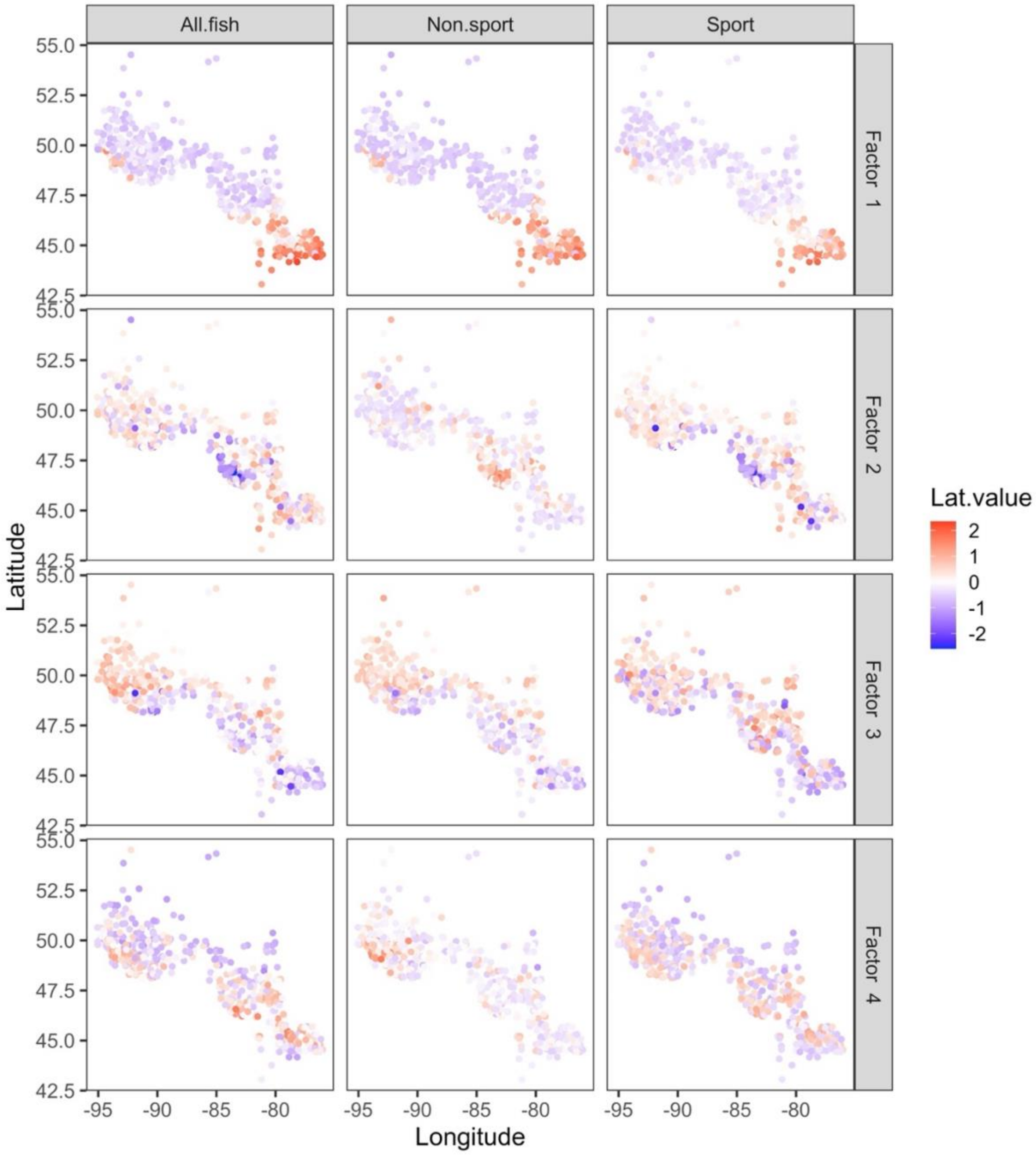


Figure S4: Maps showing the spatial distribution of latent variables derived from three different fish assemblages. We generated the latent variables using (1) sport fish species, labeled as ‘Sport,’ (2) non-sport fish species, labeled as ‘Non.sport,’ and (3) all fish species, labeled as ‘All.fish.’ These latent variables were based on the presence-absence data for the respective fish groups. Each column represents a different model, while each row corresponds to a specific latent variable, visually depicting how these variables vary across the landscape for each fish assemblage.

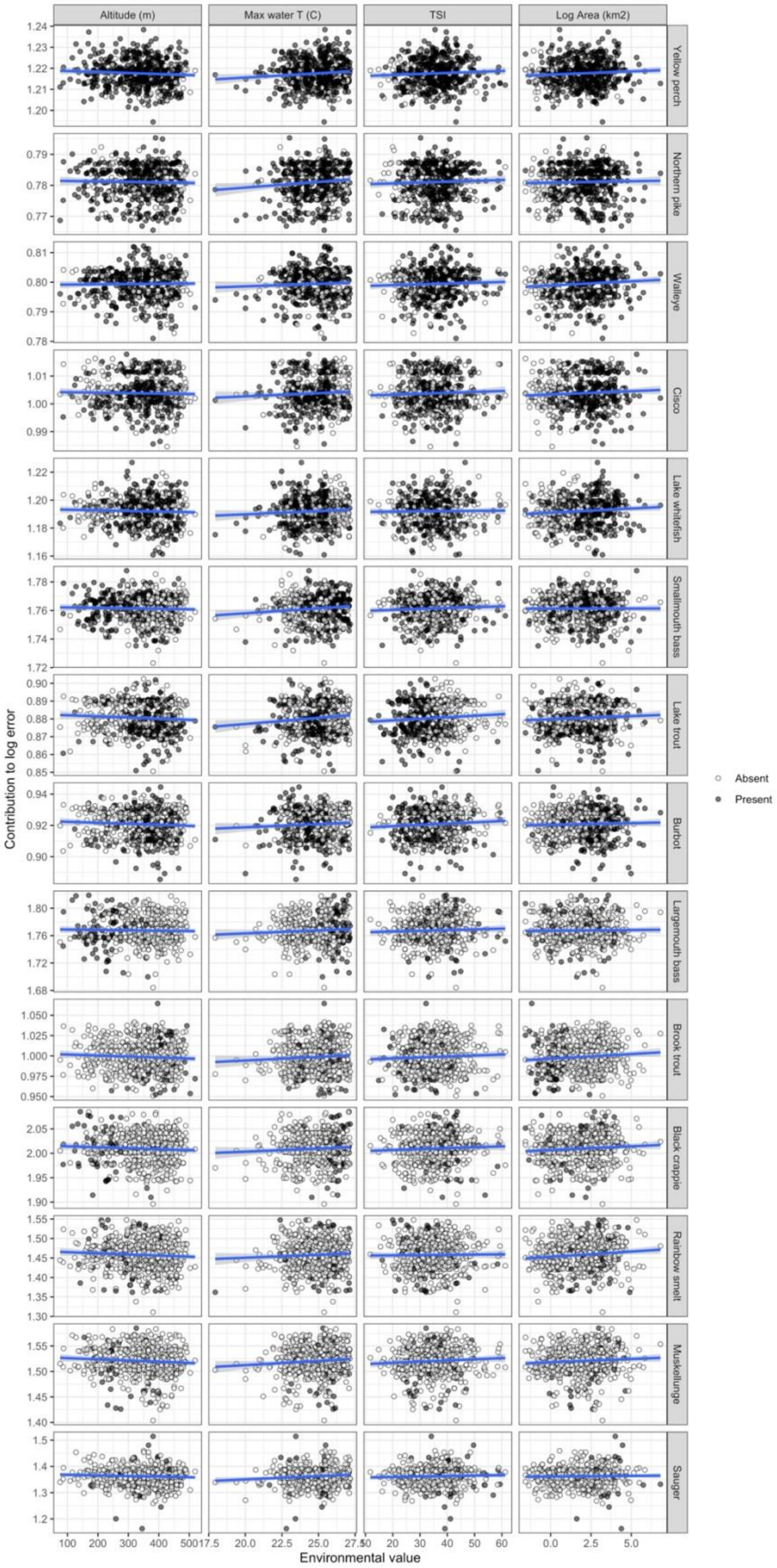

Altitude (m)
Max water T (C)
TSI
Log Area (km2)
Yellow perch
Northern pike
Walleye
Cisco
Lake whitefish
Smallmouth bass
Lake trout
Burbot
Largemouth bass
Brook trout
Black crappie
Rainbow smelt
Muskellunge
Sauger
Contribution to log error
Environmental value
Absent
Present
100
200
300
400
500
17.5
20.0
22.5
25.0
27.5
10
20
30
40
50
60
0.0
2.5
5.0

Figure S5: Contribution of each lake to the log error as a function of environmental variables. The contribution was calculated as the median log error when the lake was part of the calibration set minus the median log error when the lake was part of the validation set. A positive contribution indicates that including the lake in the calibration set improved predictions. Color of the points represents whether the species is present (black) or absent (white) from the considered lake. The blue line represents the linear trend across all lakes. The four environmental variables selected were: log transformed area (in km^2), altitude (in m), maximum water temperature in °C, and Trophic Status Index based on phosphorus levels (TSI). The environmental variables selected are meant to represent different types of lakes in terms of, respectively, hydro-morphology, watershed characteristics, climate, and productivity. Species are organised by occurrence, with high occurrence species at the top of the table and low occurrence species at the bottom of the table.

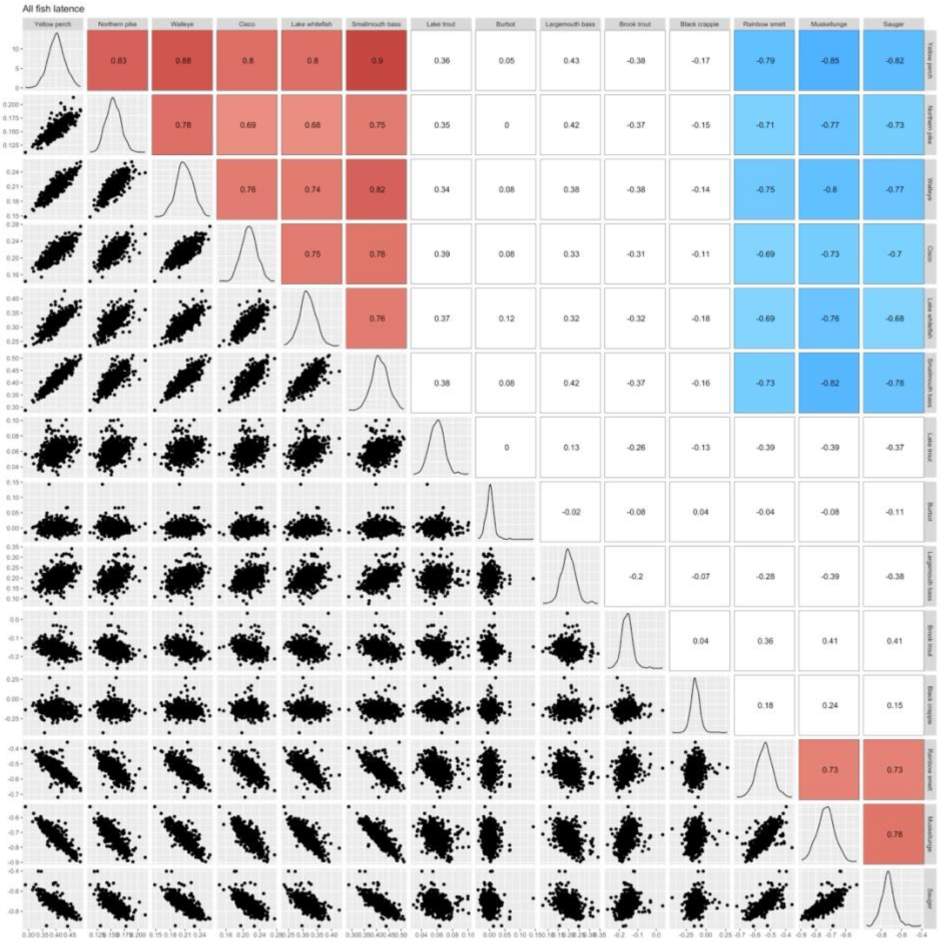


Figure S6: Correlation of lake contributions between species for model containing latent variables generated from all fish species. The patterns observed allowed us to group species in the following manner: (Group 1) rainbow smelt, muskellunge, and sauger; (Group 2) burbot, lake trout, black crappie, brook trout, and largemouth bass; and (Group 3) yellow perch, smallmouth bass, northern pike, walleye, lake whitefish, and cisco. Correlations above 0.5 are highlighted in red and correlations below -0.5 in blue. Species are organised by occurrence, with high occurrence species on the right and low occurrence species on the left.