Combining Environmental Factors and Species Co-occurrence Patterns to Predict Species Abundance and Community Biomass: Method Development and Validation in Ontario Lakes

Aliénor Marie Eugénie Stahl

A Thesis in the Department of

Biology

Presented in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy (Biology)

At Concordia University

Montréal, Québec, Canada

October 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis

Prepared by: Aliénor Marie Eugénie Stahl

Entitled: Combining Environmental Factors and Species Co-occurrence Patterns to Predict Species Abundance and Community Biomass: Method Development and Validation in Ontario Lakes

And submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Biology)

Complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____Chair

Dr. Cameron Skinner

_____External Examiner

Dr. Cindy Chu

_____Arm's Length Examiner

Dr. Nicola S. Smith

_____ Examiner

Dr. Dylan J. Fraser

_____ Examiner

Dr James W.A. Grant

_____Thesis Co-supervisor

Dr. Eric J. Pedersen

_____Thesis supervisor

Dr. Pedro R. Peres-Neto

Approved by _____

Dr. Robert B. Weladji, Graduate Program Director

December 18th, 2024 _____

Dr. Pascale Sicotte, Dean for the Faculty of Arts and Science

Abstract

Combining Environmental Factors and Species Co-occurrence Patterns to Predict Species Abundance and Community Biomass: Method Development and Validation in Ontario Lakes

Aliénor M.E. Stahl, Ph.D.

Concordia University, 2024

Predicting species abundance and community biomass is vital for ecosystem management, particularly in freshwater lakes, where this information guides conservation efforts, resource management, and biodiversity assessments. These metrics provide crucial insights into population dynamics, ecosystem productivity, and ecological balance. However, traditional models often rely on abiotic factors or limited species presence-absence data, missing the complex interspecies relationships that shape community structure and ecosystem function. This thesis aims to enhance predictive models of species abundance and biomass by incorporating community-level data, using latent variables derived from species co-occurrence and environmental variables. Latent variables are unobserved or hidden variables that are inferred from observed data. By integrating both biotic and abiotic factors, this approach enhances the accuracy of ecological predictions, offering more reliable tools for ecosystem management and conservation efforts. The three chapters build upon one another, progressively expanding the scope of the models and their applications. Chapter 2 lays the groundwork using simulated data to refine single-species abundance models, exploring how different levels of information (true environmental drivers versus latent variables based on species co-occurrence) affect model accuracy. This simulation framework is essential for understanding the robustness of the models before their application to real-world data. Chapter 3 extends this

work by applying the developed framework to empirical data from lakes, focusing on sport fish species. It examines the role of latent variables and various fish assemblages in improving abundance predictions and explores how lake-specific characteristics influence model performance. This real-world application allows for a deeper understanding of how the framework operates under natural conditions, particularly in aquatic ecosystems. Finally, Chapter 4 uses the abundance predictions from Chapter 3 to develop a stacked model for predicting community biomass. It compares the stacked model's effectiveness to a community model across varying spatial scales and species richness levels. By transitioning from single-species models to multi-species and ultimately biomass prediction, the chapters are sequentially linked, each addressing a broader ecological question while refining and testing the models at different levels of complexity. This cohesive approach enhances both the predictive accuracy of species abundance and the practical applications of these models for ecosystem management.

Résumé

Prédire l'abondance des espèces et la biomasse des communautés est essentiel à la gestion des écosystèmes, particulièrement dans les lacs d'eau douce, où ces informations guident les efforts de conservation, la gestion des ressources, et les évaluations de la biodiversité. Ces métriques fournissent des connaissances cruciales sur la dynamique des populations, la productivité des écosystèmes et l'équilibre écologique. Cependant, les modèles traditionnels s'appuient souvent sur des facteurs abiotiques ou sur des données limitées sur présence-absence d'espèces, omettant les relations interspécifiques complexes qui façonnent la structure de la communauté et le fonctionnement de l'écosystème. Cette thèse vise à améliorer les modèles prédictifs d'abondance des espèces et de biomasse en incorporant des données à l'échelle de la communauté, et en utilisant

des variables latentes dérivées de la cooccurrence des espèces et des variables environnementales. Les variables latentes sont des variables non observées ou cachées qui sont déduites à partir de données observées. En intégrant à la fois les facteurs biotiques et abiotiques, cette approche améliore la précision des prévisions écologiques, offrant des outils plus fiables pour la gestion des écosystèmes et les efforts de conservation. Les trois chapitres sont construits de manière séquentielle, élargissant progressivement la portée des modèles et leurs applications. Le Chapitre 2 pose les bases en utilisant des données simulées pour affiner les modèles d'abondance mono-espèce, en explorant comment différents niveaux d'information (véritables moteurs environnementaux par rapport aux variables latentes basées sur la cooccurrence des espèces) affectent la précision du modèle. Ce cadre de simulation est essentiel pour comprendre la robustesse des modèles avant leur application à des données réelles. Le Chapitre 3 étend ce travail en appliquant le cadre développé à des données empiriques de lacs, en se concentrant sur les espèces de poissons de sport. Il examine le rôle des variables latentes et de divers assemblages de poissons dans l'amélioration des prévisions d'abondance et explore comment les caractéristiques spécifiques des lacs influencent les performances du modèle. Cette application concrète permet de mieux comprendre le fonctionnement du cadre de modélisation dans des conditions naturelles, notamment dans les écosystèmes aquatiques. Enfin, le Chapitre 4 utilise les prévisions d'abondance du chapitre 3 pour développer un modèle empilé permettant de prédire la biomasse de la communauté. Il compare l'efficacité du modèle empilé à un modèle communautaire à travers différentes échelles spatiales et niveaux de richesse en espèces. En passant des modèles mono-espèces aux modèles multi-espèces et finalement à la prédiction de la biomasse, les chapitres sont liés séquentiellement, chacun abordant une question écologique plus large tout en affinant et en testant les modèles à différents niveaux de complexité. Cette approche cohésive améliore à la fois

la précision prédictive de l'abondance des espèces et les applications pratiques de ces modèles pour la gestion des écosystèmes.

Acknowledgments

I would like to express my deepest gratitude to those who have supported and guided me throughout the completion of this thesis.

Firstly, my heartfelt thanks go to my supervisors, Dr. Pedro Peres Neto and Dr. Eric Pedersen. Your expertise, patience, and unwavering support have been invaluable during this journey. Your insightful feedback has greatly enhanced the quality of this work, and for that, I am profoundly grateful.

I am also thankful to my committee members, Dr. Dylan Fraser and Dr James Grant, whose guidance, feedback, and encouragement have been instrumental in shaping this thesis. Your expertise has been crucial in refining my research, and I sincerely appreciate your contributions. My sincere thanks extend to my examiners, Dr. Nicola Smith and Dr. Cindy Chu.

I am also grateful to BIOS2, whose support provided me with the opportunity to undertake an internship focused on studying the American eel, with the guidance and support of Dr. Jonathan Midwood, Jesse Gardner Costa, and Sarah Larocque.

A special thank you to the Ontario Ministry of Natural Resources and Forestry for providing the essential data that were used throughout this thesis. Your contribution has been vital to the success of this research.

To my lab members, thank you for your collaboration, insightful discussions, and camaraderie, which have been integral to the successful completion of this research. Your contributions have created a stimulating and supportive environment that I deeply appreciate. I would particularly like to acknowledge Alex Engler, Dr. Gabriel Khattar, Timothy Law, Dr. Pedro Braga, as well as Fonya Irvine, Allegra Spensieri, Natalie Dupont, and Dr. Paul Savary for their exceptional support.

Lastly, I am grateful to everyone who contributed in one way or another to the successful completion of this thesis. Your support has meant the world to me.

Thank you all.

Dedication

To my family, thank you for your constant love and encouragement. To my parents, Nathalie and Robert, thank you for being attentive listeners, for always freeing yourself when I came back home so we could travel together. I am deeply grateful for all the times you came to visit Spoutnik and I, for giving me reasons to visit Québec, and for that lovely week in Mexico. Your thoughtful care packages, especially during the COVID lockdown, brought comfort and joy when it was most needed. The moments we spent together, your endless support and belief in me have been my driving force during this journey.

To my sister, Adélaïde, thank you for your enthusiasm and boundless encouragement, and for the delightful pictures of Vaettir and Loki that never fail to lift my spirits. Words cannot express how ecstatic I am whenever you visit; those moments breathe life into me, and I cherish every second we spend together. You spoil me beyond measure, and I couldn't have wished for a better sister. Thank you for everything, from the thrilling helicopter ride to our shared pudding adventures and matching tattoos. I am profoundly grateful for your existence.

A special thanks to my bunny, Spoutnik, for your unconditional love and for always, always demanding attention just when inspiration strikes. Your timely interruptions reminded me to take much-needed breaks, and your warm presence has been a constant comfort at the end of my exhausting days.

To my friends, your companionship and understanding have been a source of comfort and motivation. Special thanks to Alex Engler for the countless hot chocolates and lasagnas we shared, for our rants about our pets and shared obsessions, and for the many events we attended together. To Dr. Nicolàs Alessandroni, for the tremendous support you gave me this past year, for the baking (especially your tiramisu), and the concerts we experienced together. To Dr. Gabriel Khattar, for the BBQs and Christmases we spent together. To Dr. Pedro Braga, for our discussions on post-docs and the spinning sessions. To Noémie Samson, for listening to my complains and providing a different point of view. To Oona Versavel, for our deep mutual understanding of each other and our shared escapes into other worlds. To Sun and Lilly Hartmann, for never forgetting me regardless of the time spent apart, and for introducing me to new people and passions. And last but not least, to Melanie Jung, for sharing your passion with me; you've saved me once and I would not be standing here without you. To all my friends who have provided moral support and motivation during this journey, thank you.

> "*I think maybe we die every day. Maybe we're born new each dawn, a little changed, a little further on our own road. When enough days stand between you and the person you were, you're strangers. Maybe that's what growing up is. Maybe I have grown up.*"
>
> Mark Lawrence, *Prince of thorns*

Contribution of authors

As primary author, I led the conception, planning, data extraction, data simulation, data analyses, and writing of all chapters in this thesis. Similarly, Dr. Pedro Peres-Neto and Dr. Eric Pedersen contributed immensely to the conception, planning, interpretation of results, editing and reviewing of all chapters.

For Chapter 3 and Chapter 4, the data were obtained through the Ontario Broad-Scale Monitoring Program for Inland Lakes (BsM) (Sandstrom *et al.* 2011) conducted by the Ontario Ministry of Natural Resources and Forestry (OMNRF, 2012).

Chapter 2 is under review in Ecosphere as: Stahl, A., Pedersen, E. J., & Peres-Neto, P. R. Advancing single species abundance models: robust models for predicting abundance using co-occurrence from communities. A preprint of this chapter has been deposited on EcoEvoRxiv. https://doi.org/10.32942/X2S32J

Table of contents

LIST OF FIGURES ..... **XII**

LIST OF TABLES ..... **XXIII**

CHAPTER 1: GENERAL INTRODUCTION **1**

1.1. Ecology, biodiversity loss and the role of biodiversity monitoring 1
1.2. The significance of species abundance in ecological research 3
1.3. The interplay of species abundance and biomass in ecosystem dynamics and stability 4
1.4. Challenges in sampling species abundance and the need for predictive models 6
1.5. Latent variables: concept, origins, and applications in ecology 10
1.6. The necessity of model testing: from simulations to empirical validation 13
1.7. Chapters overview and novelty of research 16
1.7.1. Chapter 2: Advancing single species abundance models: robust models for predicting abundance using co-occurrence from communities *16*
1.7.1. Chapter 3: Advancing single species abundance models: leveraging multi-species data to uncover lake-specific predictive patterns and improve fisheries predictions *17*
1.7.2. Chapter 4: Predicting biomass: a comparative analysis of community models and stacked abundance models across spatial scales *19*

CHAPTER 2: ADVANCING SINGLE SPECIES ABUNDANCE MODELS: ROBUST MODELS FOR PREDICTING ABUNDANCE USING CO-OCCURRENCE FROM COMMUNITIES **21**

2.1. Abstract 21
2.2. Introduction 22
2.3. Materials and methods 27
2.3.1. Steps 1 and 2: simulating communities *27*
2.3.2. Step 3: Latent variables generation and their abilities to represent missing environmental variation *30*
2.3.3. Step 4: Contrasting the performance of abundance models *32*
2.3.4. Step 5: comparison of model performance *34*
2.4. Results 38
2.4.1. Number of latent variables needed to capture environmental variation *38*
2.4.2. Models' performance *40*
2.5. Discussion 46
2.5.1. Number of latent variables needed to capture environmental variation *46*
2.5.2. Model performance *48*
2.6. Supplementary Information 55

CHAPTER 3: ADVANCING SINGLE SPECIES ABUNDANCE MODELS: LEVERAGING MULTI-SPECIES DATA TO UNCOVER LAKE-SPECIFIC PATTERNS FOR IMPROVED FISHERIES PREDICTIONS **61**

3.1. Abstract 61
3.2. Introduction 62
3.3. Materials and methods 67
3.3.1. Dataset *67*
3.3.2. Environmental predictors *70*
3.3.3. Latent variable generation *71*
3.3.4. Modelling structure overview *72*
3.3.5. Model fitting *74*

3.3.6. Metrics for evaluating model predictive ability *74*
3.3.7. Target analyses based on key questions *76*
3.4. RESULTS 80
3.5. DISCUSSION 86
3.6. SUPPLEMENTARY INFORMATION 93
3.6.1. Identification of optimal number of composite environmental variables and latent variables *93*
3.6.2. Extended results *96*

CHAPTER 4: COMPARATIVE ASSESSMENT OF COMMUNITY AND STACKED ABUNDANCE MODELS FOR PREDICTING BIOMASS ACROSS SPATIAL SCALES **107**

4.1. ABSTRACT 107
4.2. INTRODUCTION 108
4.3. MATERIALS AND METHODS 110
4.3.1. Dataset *110*
4.3.2. Variables and transformation *114*
4.3.3. Model fitting *115*
4.3.4. Analysis *118*
4.4. RESULTS 121
4.5. DISCUSSION 125
4.6. SUPPLEMENTARY INFORMATION 131

CHAPTER 5: CONCLUDING REMARKS, ASSUMPTIONS, AND FUTURE DIRECTIONS **138**

BIBLIOGRAPHY **145**

List of Figures

Figure 1.1: Conceptual diagram of community assembly and the mechanisms assumed to be involved. This diagram illustrates the process of community assembly, starting from the regional species pool, which consists of five species. Species disperse across the landscape through natural movement and chance (depicted by blue arrows). As they disperse, species encounter environmental selection (green) that select based on abiotic conditions, and biotic filters (brown), which involve species interactions such as competition or predation. These selection processes shape the composition of local communities, both in terms of presence-absence (represented by the grey arrows) and relative abundance (illustrated by the pie charts, with species proportions). The outcome reflects how regional dynamics and local selection combine to create unique communities at different locations. This figure has been adapted from HilleRisLambers (2012). 7

Figure 1.2: Conceptual diagram illustrating the commonalities and distinctions across chapters. Chapter 2 focuses on simulated communities where environmental selection (green arrow) is the sole assumed mechanism driving community assembly. In contrast, Chapters 3 and 4 utilize an empirical dataset that incorporates all three assumed selection processes: environmental (green arrow), biotic (brown arrow), and dispersal (blue arrow). Across all chapters, presence-absence data are used to generate latent variables which are then used as predictors. Chapter 2 predicts both presence-absence and abundance, while Chapter 3 focuses solely on predicting abundance. Finally, Chapter 4 expands on this by using abundance predictions to estimate community biomass. 15

Figure 2.1: The rationale underlying our model framework and simulation workflow to assess its performance. First, species abundances were simulated for all species (top left panel) as a function of multiple environmental factors. In this example, two environmental variables were used to simulate species abundances ($\mathbf{X}_1$ and $\mathbf{X}_2$; bottom left panel). Species abundances are then transformed into presence-absence data and used to derive latent variables (bottom left panel). Here, only one latent variable is presented for simplicity, allowing one to more easily associate it with the abundances of the original simulated species. Variation in species abundances (target species) across sites is then modeled against latent and environmental variables or reduced combinations (e.g., removing an environmental variable and assess the conditions that affect latent performances), depending on specific simulation scenarios. The model can produce either

abundance or presence-absence predictions for each site. The black rectangular outline highlights the target species (species 10) that the model aims at predicting.....26

Figure 2.2: The density of average species abundance across sites within each landscape. For each landscape, we calculated the average abundance of each species and plotted the density of abundances in each of the 30 landscapes (grey lines). We also plotted the density of abundances across all landscapes to represent the average landscape (black line). The red line is a reference line indicating the probability density function of a log-normal distribution with the same log-mean and log-standard deviation of the average abundance distribution across replicates.....29

Figure 2.3: Variation in adjusted R^2 as a function of the number of latent variables used, as well as the true dimensions of the environment and the number of species in the landscape. Here we used 500 sites, and variations according to other number of sites are presented in Figure S2.2. Colors represent the varying number of species in the landscape, and each panel indicates the true dimension of the environment (i.e., number of environmental variables used to simulate the abundance of a given target species).....39

Figure 2.4: Ratio TSS and delta TSS for each model and bin of species occurrence percentiles. The ratio TSS was averaged across all landscapes and replicates per model and species, with species binned by percentile of occurrence (percentage of sites occupied) and divided by the TSS of the oracle model. A value of 1 for the ratio TSS indicates an identical performance between the model and the oracle model, while a value below 0 represents a performance similar to that of a random model. To improve contrast between colors, we confined the color scheme between 0 and 1. Any value below 0 indicates a prediction of presence-absence no better than a random model, and any value above 1 indicates a better prediction than the oracle model. The environment panel represents models containing only environmental variables, while the latent panel is for models containing latent variables (mix of latent and environmental predictors); the models were then ordered from bottom to top as fewest to the greatest number of environmental variables included and sorted by coefficients relative to each environmental variable (see Methods for more information, note that the “mid” model refers to the “intermediate” model). The delta TSS was measured as the TSS of the model with environmental variables minus the TSS of the model with the same combination of environmental variables and latent variables. A negative value indicates that the model with latent

predicts the presence-absence of the species better than the model containing only environmental variables. 42

Figure 2.5: Correlation between the metrics studied (TSS, sensitivity, and specificity) depending on the model across species occurrence percentiles. The vertical panels indicate the different metrics, with models represented in different colors. The oracle model refers to the model using the true environmental coefficients, while the other models were fitted using all environmental variables (benchmark) or latent variables (latent). The True Skill Statistic (TSS) measures the difference between sensitivity and specificity of the model and ranges from -1 to +1. A score of +1 indicates a perfect agreement between the predictions of the model and the true presence-absence, while a score of 0 or less represents a performance no better than random. Sensitivity represents the ability to correctly classify a species as "present", while specificity represents the ability to correctly classify a species as "absent". Their values can be interpreted as a percentage, with values of 1 indicating perfect classification of either presence or absence, and values of 0.5 no better than random. Here we used 500 sites, and variations according to other number of sites are presented in Figure S2.3. 43

Figure 2.6: Ratio Mean Absolute Percentage (MAPE) and delta MAPE are presented for each model and bins of species abundance percentiles. The MAPE is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance and divided by the MAPE of the oracle model to derive the ratio MAPE. The environment panel represents models containing only environmental variables, while the latent panel depicts models containing latent predictors. The models are then ordered from bottom to top, from the fewest to the greatest number of environmental variables included and sorted by coefficients relative to each environmental variable. See Methods for more information, note that the "mid" model refers to the "intermediate" model. Delta MAPE was measured as the MAPE of the model with environmental variables only minus the MAPE of the model with the same combination of environmental and latent predictors. A positive value indicates that the model with latent predicts the abundance of the species better than the model containing only environmental variables. 45

Figure 3.1: 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). 67

Figure 3.2: 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. 3.2). 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. 81

Figure 3.3: Density plot of the log error as a function of species and model. The log error was calculated following Eq. 3.1, 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. 82

Figure 3.4: 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. 3.3). 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. 83

Figure 3.5: 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. 3.4). 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. 86

Figure 4.1: Map of the 583 lakes surveyed in Ontario, Canada. Each point is color-coded to indicate species richness (i.e., the number of species present in the lake). Black lines denote provincial boundaries within Canada. 111

Figure 4.2: Comparison of predicted versus observed biomass for the two models. The *stacked model* (top panel) predicts species abundance using composite environmental variables and community composition, multiplies the predicted abundance by the average species weight, and fits a second model to estimate community biomass by summing these values. The *community model* (bottom panel) directly predicts community biomass using composite environmental variables and species composition. The dashed line represents the 1:1 line, indicating perfect agreement between predicted and observed biomass. The blue line represents the trend across all lakes (i.e., linear regression between predictive and observed biomass). A version of this figure using $\log_{10}$ scale is available in Figure S4.2..... 121

Figure 4.3: Histogram of predictive error by model type. Predictive error is calculated as the ratio of predicted biomass to observed biomass and displayed on a $\log_{10}$ scale. The stacked model (red) predicts species abundance and then estimates biomass, while the community model (blue) directly predicts biomass using composite environmental variables and community composition (see methods for more details). The dashed line indicates perfect prediction, where predicted biomass matches observed biomass. 122

Figure 4.4: Predictive error of biomass per lake plotted against community diversity. Predictive error is calculated as the ratio of predicted biomass to observed biomass and presented on a $\log_{10}$ scale. Community diversity is measured using both species richness (i.e., number of species per lake; left panels) and Shannon's index (i.e., species diversity weighted by biomass; right panels). The stacked model results are shown in the top panels, while the community model results are in the bottom panels. The dashed line indicates perfect prediction, where predicted biomass matches observed biomass. The blue line represents the trend across all lakes, obtained from linear regression between predictive error and community diversity for each model and diversity metric. 124

Figure 4.5: Histogram and Empirical Cumulative Distribution Function (ECDF) of predictive error across different aggregation levels and models. Predictive error is calculated as the ratio of predicted biomass to observed biomass and presented on a $\log_{10}$ scale. Aggregation levels are shown on the vertical panel, with the lowest aggregated level at the top (lake level) and the most aggregated level (secondary watershed) at the bottom. The stacked model (red) predicts species abundance and then estimates biomass, while the community model (blue) directly predicts biomass using composite environmental variables and species composition. The left panels show the density of predictive error, while the right panels display the ECDF of predictive error. The dashed line indicates a perfect prediction, where predicted biomass matches observed biomass. Alternative aggregation methods are shown in Figure S4.4 and S4.5. 125

Figure SI 2.1: Variation in delta BIC as a function of the number of latent variables used, as well as the true dimensions of the environment, the number of species in the landscape and the number of sites. Horizontal panels represent the number of sites, and each vertical panel indicates the true dimension of the environment (i.e., number of environmental variables used to simulate the abundance of a given target species). Colors represent the varying number of species in the landscape. The delta BIC is calculated as the BIC of the model minus the BIC of the best model for the ongoing simulation. 55

Figure SI 2.2: Variation in adjusted R^2 as a function of the number of latent variables used, as well as the true dimensions of the environment, the number of species in the landscape and the number of sites. Horizontal panels represent the varying number of sites, and each vertical panel indicates

the true dimension of the environment (i.e., number of environmental variables used to simulate the abundance of a given target species). Colors represent the varying number of species in the landscape. 56

Figure SI 2.3: Average value of the studied metrics (Ratio TSS, ratio sensitivity, and ratio specificity) depending on the number of sites used to fit the models, the model used, and the occurance of species. Horizontal panels represent the different occurrence: species with low, medium and high occurrence corresponding respectively to bins of 15, 50, and 80 percentiles of occurrence. Vertical panels indicate the metrics considered, with the models represented in different colors. The ratio metric is calculated as the metric for the predictions of a model for a species of the landscape divided by the same metric calculated for the oracle model. For the ratio TSS, a score of 1 indicates a perfect agreement between the predictions of the considered model and the oracle model, while a score of 0 or less represents a performance no better than random. For the ratio sensitivity, it represents the ability to correctly classify a species as "present", while the ratio specificity represents the ability to correctly classify a species as "absent". For both metrics, values above 1 indicate a better performance than the oracle model and values below 1 indicate a lesser performance. The benchmark model refers to the model containing all environmental variables, 2V.high the model with the two environmental variables with the highest coefficients, 1V.high the model with the environmental variable with the highest coefficient, and Latent the model containing the latent variables. 57

Figure SI 2.4: Abundance metrics and the comparison of performance between environmental models and latent models measured as delta metrics. Each metric is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance, and divided by the metric of the oracle model to give the ratio metric. The environment panel represents models containing only environmental variables, while the latent panel depicts models containing latent predictors. The models are then ordered from bottom to top, from the fewest to the greatest number of environmental variables included and sorted by coefficients relative to each environmental variable. See Methods for more information, note that the "mid" model refers to the "intermediate" model. The delta metric was measured as the metric of the model with environmental variables only minus the metric of the model with the same combination of environmental and latent predictors. A positive value indicates that the model with latent predicts the abundance of the species better than the model containing only environmental variables.....59

Figure SI 2.5: Correlation between the metrics studied (MAPE, RMSPE, RMSE, SMAPE, and RMRPE) depending on the model across species abundance percentiles. The vertical panels indicate the different metrics, with models represented in different colors. Each metric is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance. The oracle model refers to the model using the true environmental coefficients while the other models were fitted using all environmental variables (benchmark) or latent variables (latent). 59

Figure SI 2.6: Average value of the studied metrics (MAPE, RMSPE, RMSE, SMAPE, and RMRPE) depending on the number of sites used to fit the models, the model used, and the abundance of species. Horizontal panels represent the different abundances: species with low, medium and high occurrence corresponding respectively to bins of 15, 50, and 80 percentiles of occurrence. Vertical panels indicate the metrics considered, with the models represented in different colors. Each metric is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance and divided by the metric of the oracle model to give the ratio metric. The benchmark model refers to the model containing all environmental variables, 2V.high the model with the two environmental variables with the highest coefficients, 1V.high the model with the environmental variable with the highest coefficient, and Latent the model containing the latent variables. 60

Figure SI 3.1: 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. 95

Figure SI 3.2: 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. 102

Figure SI 3.3: 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.....103

Figure SI 3.4: 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.....104

Figure SI 3.5: 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 106

Figure SI 3.6: 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. 106

Figure SI 4.1: Map of the 583 lakes surveyed in Ontario, Canada. Each point is color-coded to indicate community biomass (i.e., total weight of fish caught per unit effort (BPUE) for each lake). Black lines denote provincial boundaries within Canada. 133

Figure SI 4.2: Comparison of predicted versus observed biomass for the two models on the $\log_{10}$ scale. The *stacked model* (top panel) predicts species abundance using composite environmental variables and community composition, multiplies the predicted abundance by the average species weight, and fits a second model to estimate community biomass by summing these values. The *community model* (bottom panel) directly predicts community biomass using composite environmental variables and species composition. The dashed line represents the 1:1 line, indicating perfect agreement between predicted and observed biomass. The blue line represents the trend across all lakes (i.e., linear regression between predictive and observed biomass).....134

Figure SI 4.3: Map and estimated spatial smooths of the prediction errors for the two models. The stacked model (left panel) predicts abundance and then estimates biomass, while the community model (right panel) directly predicts biomass from composite environmental variables and community composition. The predictive error is calculated as the $\log_{10}$ of predicted biomass over observed biomass. The maps show underprediction in blue and overprediction in red. The black lines represent the delimitations of the secondary watersheds. The spatial smooths show in blue, areas where lake biomass tend to be more underestimated and in red, areas where lake biomass tends to be more overestimated. 135

Figure SI 4.4: Histogram and Empirical Cumulative Distribution Function (ECDF) of predictive error across different aggregation levels and models using the nearest neighbor method. Predictive error is calculated as the ratio of predicted biomass to observed biomass and presented on a $\log_{10}$ scale. Aggregation levels are shown on the vertical panel, with the lowest aggregated level at the top (lake level) and the most aggregated level (50 closest lakes to the focal lake) at the bottom. The stacked model (red) predicts species abundance and then estimates biomass, while the community model (blue) directly predicts biomass using composite environmental variables and species composition. The left panels show the density of predictive error, while the right panels display the

ECDF of predictive error. The dashed line indicates a perfect prediction, where predicted biomass matches observed biomass. 136

Figure SI 4.5: Histogram and Empirical Cumulative Distribution Function (ECDF) of predictive error across different aggregation levels and models using the distance-based method. Predictive error is calculated as the ratio of predicted biomass to observed biomass and presented on a $\log_{10}$ scale. Aggregation levels are shown on the vertical panel, with the lowest aggregated level at the top (lake level) and the most aggregated level (all lakes within a 100 km radius around the focal lake) at the bottom. The stacked model (red) predicts species abundance and then estimates biomass, while the community model (blue) directly predicts biomass using composite environmental variables and species composition. The left panels show the density of predictive error, while the right panels display the ECDF of predictive error. The dashed line indicates a perfect prediction, where predicted biomass matches observed biomass. 137

List of Tables

Table 2.1: Variable symbols and indexes, and their associated values and distributions used in the simulation study. Bold letters indicate that the variable is a vector or a matrix.....28

Table 2.2: All models considered in this study based on combinations of environmental variables and community composition (latents). The best model is expected to be the "true" model considering all three environmental variables. **A** refers to the abundance matrix, $\mathbf{X}_1$ to $\mathbf{X}_3$ to the environmental variables, and $\mathbf{Z}_1$ to $\mathbf{Z}_3$ to the community composition (latent variables).....34

Table 2.3: Metrics used for assessing model predictive performance based on presence-absence and abundance of target species. *J* represents the number of sites, $\boldsymbol{A}_s$ the true abundance of the (target) species, $\boldsymbol{P}_s$ the predicted abundance, TP the true positives, FP the false positives, TN the true negatives, and FN the false negatives. Bold letters indicate that the variable is a vector or a matrix. The True Skill Statistic (TSS), sensitivity, and specificity are calculated for all sites of the landscape. Having evaluated the presence-absence predictions of the models and to avoid artificially inflating the error rate of the abundance metrics, the Mean Absolute Percentage Error (MAPE), Root Mean Squared Percentage Error (RMSPE), Relative Mean Squared Error (RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), and Root Mean Ratio Percentage Error (RMRPE) are calculated for sites where the species is truly present (i.e., abundance of 1 or more). 37

Table 3.1: List of species included in the dataset, with both common and scientific 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.....69

Table 3.2: 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 S3.6). 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. 85

Table 4.1: List of species included in the dataset, with common and scientific names, along with their average weight (in kg). The average adult weight was calculated as the mean of the minimum and maximum weights reported in the Ontario Freshwater Fishes Life History Database by Eakins (version 5.31, 2024). 112

Table SI 3.1: 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. 94

Table SI 3.2: Table of environmental variables and their units grouped by categories (e.g., climate, productivity). See Sandstrom et al. (2011) for details on sampling methods..... 96

Table SI 3.3: 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..... 98

Table SI 3.4: 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. 101

Table SI 4.1: Table of environmental variables and their units grouped by categories (e.g., climate, productivity). See Sandstrom et al. (2011) for details on sampling methods..... 131

Chapter 1: General introduction

> "*The thing the ecologically illiterate don't realize about an ecosystem is that it's a system. A system! A system maintains a certain fluid stability that can be destroyed by a misstep in just one niche. A system has order, a flowing from point to point. If something dams that flow, order collapses. The untrained might miss that collapse until it was too late. That's why the highest function of ecology is the understanding of consequences.*"
>
> Frank Herbert, *Dune*

1.1. Ecology, biodiversity loss and the role of biodiversity monitoring

The importance of ecology (i.e., the study of interactions among organisms and their environment) has surged in response to escalating biodiversity loss and ecosystem degradation. Human activities such as habitat destruction, pollution, overexploitation of natural resources, and climate change have accelerated extinction rates to 100–1,000 times the natural background rate (Barnosky *et al.* 2011; Ceballos *et al.* 2015). This rapid decline in biodiversity threatens ecosystem services essential for human well-being, including food security, water purification, and climate regulation (Cardinale *et al.* 2012). The clustering of extinction patterns within specific regions and taxa (Leung *et al.* 2020), highlights the need for targeted conservation efforts to protect vulnerable species and habitats. Understanding ecological relationships, such as how species interact through food webs, mutualism, competition and energy flow, is essential for mitigating biodiversity loss. By comprehending how species interact with each other and their environment, we can inform sustainable practices and preserve the health and resilience of ecosystems (Cardinale *et al.* 2012; Díaz *et al.* 2006). As global challenges intensify, ecological research remains indispensable for developing strategies that balance human development with environmental stewardship.

Consequently, identifying key variables for measurement has become crucial for accurately evaluating ecosystems and the species they support (Failing *et al.* 2007; Pereira & David Cooper 2006).

One such effort resulted in the definition of Essential Biodiversity Variables (EBVs), a set of critical measurements required to understand, monitor, and manage biodiversity changes over time (Pereira *et al.* 2013). Conceptualized by the Group on Earth Observations Biodiversity Observation Network (GEO BON), EBVs provide a standardized framework that enables consistent and comprehensive biodiversity monitoring. They are intended to bridge the gap between data collection and policy, enabling scientists, conservationists, and policymakers to effectively assess biodiversity trends, the success of conservation actions, and understand the impacts of environmental changes (Kissling *et al.* 2018; Navarro *et al.* 2017). The EBV framework covers a wide range of biodiversity dimensions, from genetic composition to ecosystem function, providing a comprehensive view of ecosystem health. Within this framework, species population EBVs hold particular importance. By tracking changes in species abundance and distribution, these variables provide key insights into how species respond to pressures such as habitat loss, climate change, and other human activities (Parmesan 2006). These variables not only reveal long-term trends in species numbers but also serve as indicators of broader ecosystem health and stability (Kissling *et al.* 2018). They can pinpoint areas where ecosystems are most threatened or where restoration efforts are succeeding (Tittensor *et al.* 2014). This makes species population EBVs indispensable for prioritizing conservation efforts, informing environmental policy, and enhancing our understanding of the broader impacts of environmental change on biodiversity.

1.2. The significance of species abundance in ecological research

Understanding species abundance (i.e., the number of individuals of a species within a given area) is a cornerstone of ecological and biological research (Buckland *et al.* 2005; McGill *et al.* 2006). Species abundance provides critical insights into population dynamics, community structure, and environmental conditions (Mace & Baillie 2007; Magurran & McGill 2011). By examining abundance patterns, researchers can infer the availability of resources, the presence and intensity of competition and predation, and the overall stability and resilience of the ecosystem (Loreau *et al.* 2002; Paine 1966). Species abundance data are vital for effective conservation and management strategies, as it helps identify species at risk of decline and the factors driving these changes (Gaston & Fuller 2008).

Moreover, studying species abundance contributes to our understanding of biodiversity patterns and ecosystem functioning (Loreau & de Mazancourt 2013; Tilman *et al.* 2014). The distribution of abundances among species within a community (often described by species abundance distributions) reflects underlying ecological processes such as niche partitioning, competitive interactions, and environmental selection (Chesson 2000; McGill *et al.* 2007). Higher species richness and evenness in abundance are generally associated with greater ecosystem stability and resilience, as diverse communities can better withstand environmental changes and recover from disturbances (Chapin *et al.* 2000). Conversely, low abundance can signal ecological imbalances resulting from anthropogenic stressors, such as habitat degradation, pollution, or overexploitation, necessitating immediate conservation actions (Pimm *et al.* 2014).

The importance of species abundance extends to understanding community interactions, as it reveals patterns of species coexistence and competition, shedding light on the underlying mechanisms that structure ecological communities (Chesson 2000; MacArthur & Levins 1967;

MacArthur 1965). For instance, higher abundances of a keystone species can promote biodiversity by maintaining the structure and integrity of the ecosystem by controlling populations of other species, thereby promoting biodiversity (Estes *et al.* 2011; Mills & Doak 1993; Paine 1969). Conversely, the overabundance of invasive species can lead to declines in native populations through competition for resources or predation (Mack *et al.* 2000).

In addition, species abundance data also aid in predicting the responses of ecological communities to environmental changes. Shifts in abundance due to climate change, habitat alteration, or other anthropogenic factors can alter community composition, competitive dynamics, and food web structures (Walther *et al.* 2002). Such shifts can ultimately influence the resilience of ecosystems to disturbances and the sustainability of natural resources (Hughes *et al.* 2003; Loreau *et al.* 2001; Tilman 1996).

While species abundance offers insights into biodiversity, it may not fully capture ecosystem productivity. For instance, two species may have similar abundances but differ greatly in size, biomass, and ecological roles (McGill *et al.* 2007; White *et al.* 2007). Therefore, incorporating additional metrics such as biomass (i.e., which accounts for organism size and energy flow) provides a more comprehensive understanding of trophic dynamics and ecosystem function that abundance data alone cannot offer.

1.3. The interplay of species abundance and biomass in ecosystem dynamics and stability

Biomass refers to the total mass of living organisms within a given area or ecosystem and is a critical metric in ecological research. Unlike species abundance, which counts the number of individuals, biomass accounts for the size and number of organisms, providing a more holistic measure of ecosystem functionality, productivity, and energy flow (Cebrián 1999; Odum 1969). Biomass reflects the amount of organic material available in each trophic level, influencing the

distribution of energy among producers, consumers, predators, and decomposers (Reiss *et al.* 2009; Trebilco *et al.* 2013). Changes in biomass often signal shifts in resource availability or environmental conditions, highlighting its ecological importance in maintaining balance and sustainability (Cebrián & Duarte 1995; Chapin *et al.* 2011).

Analyzing community biomass, or the total biomass of all species within an ecosystem, provides valuable insights into community interactions and management strategies (Hilborn & Walters 1992). Biomass dynamics reflect competition, predation, mutualism, and other interspecific interactions, influencing overall community structure and function (Loreau 2010; Tilman 2020). As such, community biomass is often used in the context of ecosystem management to maintain biodiversity and ecosystem services (Hooper *et al.* 2005). For instance, biomass-based approaches help manage fisheries, forests, and agricultural systems by focusing on the sustainability of resource extraction while ensuring long-term ecosystem productivity (Pauly *et al.* 1998). Monitoring changes in community biomass allows for early detection of overexploitation or environmental degradation, making it an essential tool for adaptive management (Noss 1990; Pimm 1984).

The relationship between species abundance and biomass is complex and varies among ecosystems. Species with high abundance may not dominate in terms of biomass, particularly if they are small-bodied species, while larger species may have significant biomass despite lower abundances (Cyr & Pace 1993). Therefore, integrating both species abundance and biomass provides a more nuanced understanding of community dynamics and ecosystem function. Together, these metrics are essential for developing effective conservation and management strategies, as changes in abundance or biomass can have cascading effects on ecosystem stability and resilience (Cardinale *et al.* 2012; Paine 1966).

By monitoring species abundance and community biomass, ecologists can develop models to predict ecological outcomes and devise strategies to mitigate negative impacts on biodiversity. These insights are crucial for managing ecosystems sustainably and preserving the ecological balance.

1.4. Challenges in sampling species abundance and the need for predictive models

Species abundance data are fundamental for constructing and validating ecological models that predict population trends and ecosystem responses to various stressors, such as climate change, habitat fragmentation, and invasive species (Brook *et al.* 2008). Accurate abundance data ensure that these models can reliably forecast future scenarios, aiding policymakers and conservationists in making informed decisions (Cardinale *et al.* 2012). However, sampling the abundance of species in ecosystems is both time-consuming and financially demanding.

Accurate assessments often require extensive fieldwork involving deploying numerous survey methods such as quadrat sampling, transect lines, and mark-recapture techniques (Sutherland 2006). These methods necessitate significant investments of time to ensure that sampling is representative and unbiased, often requiring multiple visits to different sites under varying conditions. Additionally, the costs associated with fieldwork, including travel expenses, equipment, and personnel, can be substantial (Lindenmayer & Likens 2010). Advanced technologies, such as remote sensing and genetic barcoding, though potentially more efficient, add further to the financial burden due to their high initial setup and operational costs (Dickinson *et al.* 2010). The logistical challenges of accessing remote or difficult terrain further escalate both time and budget requirements (Yoccoz *et al.* 2001).

To overcome these challenges, scientists have developed predictive models to estimate species abundance without the need for extensive fieldwork. These models aim to provide accurate

abundance estimates of species abundance by relying on a variety of ecological and environmental variables (Bradley 2016; Brosse *et al.* 1999; VanDerWal *et al.* 2009), based on the widely held but debated assumption that community assembly mechanisms operate through three selection processes: environmental conditions, biotic interactions, and dispersal (Figure 1.1, HilleRisLambers et al. 2012; Kraft et al. 2015). For instance, habitat characteristics, climate data, land use patterns, and species-specific traits such as reproductive rates and dispersal abilities are frequently incorporated into these models (Elith & Leathwick 2009; Guisan & Thuiller 2005). Advanced statistical techniques, including machine learning algorithms and Bayesian approaches, are often employed to enhance predictive accuracy (Franklin 2010).

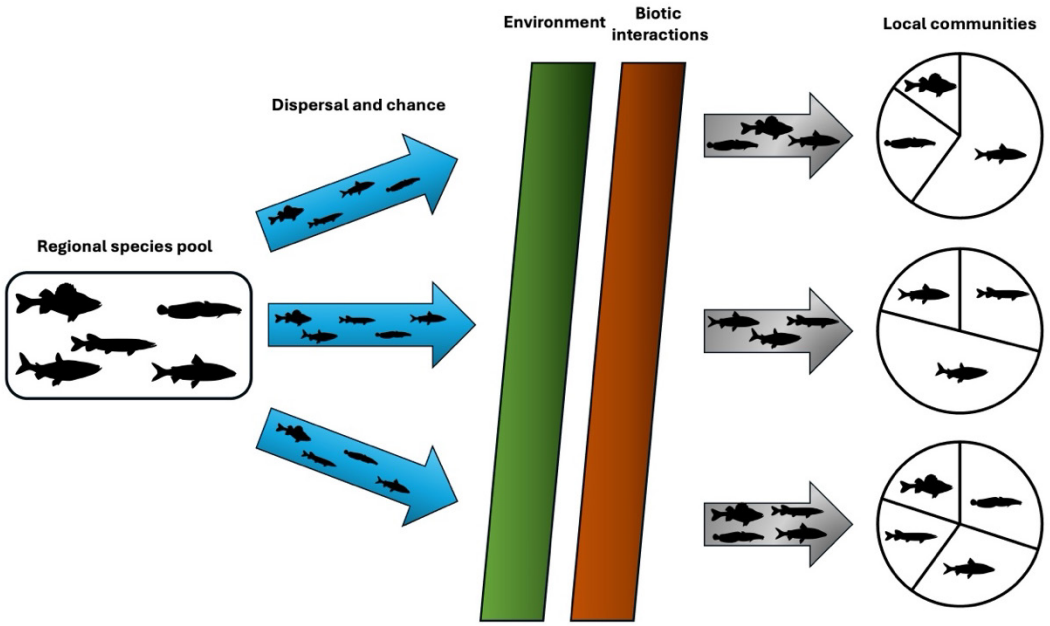

Figure 1.1: Conceptual diagram of community assembly and the mechanisms assumed to be involved. This diagram illustrates the process of community assembly, starting from the regional species pool, which consists of five species. Species disperse across the landscape through natural movement and chance (depicted by blue arrows). As they disperse, species encounter environmental selection (green) that select based on abiotic conditions, and biotic filters (brown), which involve species interactions such as competition or predation. These selection processes shape the composition of local communities, both in terms of presence-absence (represented by the grey arrows) and relative abundance (illustrated by the pie charts, with species proportions). The outcome reflects how regional dynamics and local selection combine to create unique communities at different locations. This figure has been adapted from HilleRisLambers (2012).

Despite their potential, these models face several challenges. One significant issue is the quality and resolution of input data; as inaccurate or coarse data can lead to unreliable predictions (Araújo & Guisan 2006). Models may also struggle to account for complex interactions between species and their environments, such as biotic interactions and feedback mechanisms, which can significantly influence abundance patterns (Kissling *et al.* 2012). Temporal variability also poses a challenge, as environmental conditions and species dynamics can change over time, necessitating models that can adapt to these fluctuations (Thorson *et al.* 2016). A particularly challenging aspect is the lag time between environmental changes and the corresponding response in species abundance. This delay can complicate model predictions, as immediate responses may not be evident, requiring long-term data to accurately capture these dynamics (Lindenmayer *et al.* 2010). Moreover, while models can reduce the need for direct sampling, they still require initial data for calibration and validation, which can be resource intensive. Thus, while predictive models are valuable tools for estimating species abundance, they must be carefully constructed and continuously refined to ensure their reliability and applicability across different contexts.

Previous research has demonstrated that the presence-absence data can serve as effective predictors for species abundance (Brotons *et al.* 2004; Pearson & Dawson 2003; Tyre *et al.* 2003). This approach leverages two assumptions: that community composition can significantly influence individual species' abundances (MacKenzie *et al.* 2002), and that it can serve as an indirect measure of environmental conditions. Indeed, community composition, which encompasses the variety of species within an ecosystem, provides a comprehensive snapshot of ecological interactions and environmental conditions (Holyoak & Leibold 2006). Species distributions are often influenced by specific environmental factors, such as temperature, humidity, or salinity (Guisan & Thuiller 2005). When direct measurements of these environmental variables are unavailable, presence-absence

data can act as a proxy, reflecting the tolerances and preferences of species to different environmental conditions. For instance, the presence of a species known to thrive in a particular habitat condition can indicate the prevalence of that condition in the study area, even if they are not directly measured. Conversely, the absence of a species might signal unfavorable environmental conditions or the presence of competing species.

By analyzing presence-absence data, researchers can infer both biotic interactions and abiotic environmental factors, providing a more comprehensive understanding of ecosystem dynamics (MacKenzie *et al.* 2002; Tyre *et al.* 2003). Including community composition as a predictor in abundance models enhances their accuracy by accounting for biotic interactions such as predation, competition, and mutualism, which are critical drivers of species abundance (Gaston, 2003). This approach allows for a more nuanced understanding of the factors influencing species abundance and can improve the predictive performance of ecological models.

However, there are limitations to using presence-absence data as predictors of species abundance. One primary challenge, often referred to as the Eltonian shortfall, is the lack of detailed knowledge about species interactions and ecological functions (Hortal *et al.* 2015). Focusing on a limited number of species with well-established interactions can constrain the model's scope and generalizability (Elith & Leathwick 2009). Alternatively, including a broader range of species leads to high-dimensional datasets, complicating the modelling process (Guisan & Thuiller 2005). High dimensionality not only increases computational demands but also introduces the risk of overfitting, where the model becomes tailored to the specific dataset rather than providing generalizable predictions (Franklin 2010).

To address these challenges, researchers are developing new methods to manage high-dimensional data and complex interactions without compromising model efficiency (Dormann *et al.* 2013).

Advances statistical modelling is becoming increasingly important. In particular, the incorporation of latent variables offers a promising solution by capturing the underlying structure of complex ecological datasets, thus improving model performance and predictive accuracy.

1.5. Latent variables: concept, origins, and applications in ecology

Latent variables are unobserved or hidden variables that are inferred from observed data. Originating from statistical theory, they are used to account for hidden or unmeasured factors that influence the relationships among observed variables but are not directly measured. Essentially, latent variables capture underlying structures within a dataset that impacts the observed data (Bollen 1989). They were first utilized in psychology and social sciences, particularly in Structural Equation Modelling (SEM), to model complex relationships between observed variables by introducing unobservable constructs (latent variables) that influence these relationships (Kline 2015). For instance, in psychology, latent variables represent abstract concepts such as intelligence or satisfaction, which cannot be directly measured but are inferred from responses to various tests or survey items (Bartholomew *et al.* 2011).

In ecology, the concept of latent variables has been gradually adopted to address the complexity of ecological data, where numerous environmental factors and species interactions often remain unmeasured or unknown (Chakraborty *et al.* 2011; Clark *et al.* 2017). Latent variables in ecological models serve as proxies for these unobserved factors, helping to explain variability in species distribution, abundance, and community composition that cannot be accounted for by observed variables alone (Warton *et al.* 2015a). One of the early applications of latent variables in ecology was through ordination methods like correspondence analysis, where latent variables helped to reduce the dimensionality of ecological data and to identify underlying environmental gradients influencing species distributions (Legendre & Legendre 2012; Popovic *et al.* 2022). More recently,

latent variables have been incorporated into more complex models such as joint species distribution models (JSDMs) and Gaussian copula models. In these models, latent variables help to account for species co-occurrence patterns driven by unmeasured environmental factors or species interactions (Hui *et al.* 2017; Ovaskainen *et al.* 2017). For example, Hui (2013) demonstrated that clustering species by environmental affinities (i.e., captured through latent variables), could significantly improve the predictive performance of species distribution models. The use of latent variables in these models allows ecologists to handle high-dimensional data more effectively, providing a more accurate representation of ecological processes and improving predictions of species distributions and community dynamics (Niku *et al.* 2019).

As ecological datasets continue to grow in complexity and size, latent variables offer a powerful tool to unravel the intricate relationships between species and their environments, aiding in the development of more robust ecological models. However, their application is not without challenges.

In highly dynamic systems, the predictive power of latent variables often decreases because the relationships between species, their environment, and the community structure are continuously changing (Dormann 2007b; Legendre & Legendre 2012). These systems include environments such as tropical forests with highly seasonal rainfall, intertidal zones influenced by tidal fluctuations, or temperate ecosystems experiencing significant seasonal shifts in resources like temperature and light (Clark *et al.* 2014). In these cases, latent variables, which typically capture static or semi-static ecological processes (Warton *et al.* 2012), may not effectively account for rapid temporal changes in species distributions and interactions. For example, in intertidal ecosystems, species distributions are driven by constantly changing abiotic factors like water levels and salinity, which vary on hourly or daily scales (Harley & Helmuth 2003). Latent variables capturing more

stable, underlying processes, such as long-term environmental gradients, might not predict species' responses to these rapid changes accurately (Warton *et al.* 2015a). Similarly, in ecosystems with distinct wet and dry seasons, rapid shifts in resource availability can alter species interactions and abundance patterns in ways that static latent variables may fail to capture (Ovaskainen *et al.* 2016b). This inability to model temporal variability can lead to inaccuracies, especially when dynamic fluctuations strongly influence species occurrence, abundance, or biomass.

Despite these challenges, in certain cases, it is acceptable and even advantageous to assume ecosystems to be static, particularly when the goal is to establish baseline relationships between species and their environment (Guisan & Thuiller 2005). Static models offer a way to simplify complex ecological systems, providing essential insights when developing or testing a framework. By focusing on a fixed moment in time, researchers can isolate environmental variables and species interactions without the confounding influence of temporal dynamics (Bahn & McGill 2013). This simplification is also helpful when datasets provide only snapshots, lacking the temporal resolution necessary for dynamic modelling (Guisan & Zimmermann 2000).

Static models serve as an important initial step for comparing model performance and refining predictions under idealized conditions. Assuming static conditions allow researchers to identify limitations in the model and make precise adjustments in later stages when dynamics are incorporated (Wisz *et al.* 2008). Although ecosystems are naturally dynamic (i.e., subject to seasonal changes, resource availability, anthropogenic disturbances), testing a framework in a static context is a practical and essential step before introducing complexity (Beale & Lennon 2012; Dormann *et al.* 2013).

1.6. The necessity of model testing: from simulations to empirical validation

Studying a model using simulated data is an invaluable approach for understanding its mechanics, potential failures, and inherent biases (Grimm & Railsback 2013). Simulated data provide a controlled environment where true parameters and underlying distributions are known, allowing researchers to systematically test the model's performance and validate its outputs. This controlled setting makes it easier to identify how well the model captures the relationships within the data and to pinpoint specific scenarios where the model may falter (Gelman & Hill 2006). Furthermore, working with simulated data enables the exploration of a wide range of conditions and edge cases (i.e., scenarios with extreme conditions) that might be difficult to encounter in real-world datasets (Efron & Tibshirani 1994). This helps in identifying biases, such as overfitting or underfitting, and in understanding the limitations of the model. Additionally, simulated data can be generated to include known levels of noise and variability, providing insights into the model's robustness and sensitivity to different types of data imperfections (Harrell 2015). Overall, using simulated data as a preliminary step allows for a thorough and rigorous evaluation of a model, facilitating improvements before applying it to real-world data where such control and clarity are not possible.

Once a model has been validated through simulations, the next crucial step is its application to empirical data. This process is essential for several reasons. First, while simulations offer controlled environments to test model behavior under a variety of scenarios, they often fail to capture the full complexity and variability present in real-world systems (Bansal *et al.* 2007). Second, empirical evaluation allows researchers to assess how well the model performs when confronted with the stochasticity, noise, and potential biases inherent in observational data (Dietze *et al.* 2018). Moreover, testing on empirical datasets can reveal unanticipated model limitations or dependencies that were not evident during simulation-based testing, thus offering valuable insights into model

generalizability (Grimm & Railsback 2013). Selecting relevant and appropriately challenging empirical datasets is essential to ensure that they account for the known limitations of the model, such as missing variables or assumptions about species interactions (Evans *et al.* 2013). Ultimately, this step is crucial for verifying the model's utility in practical applications and for guiding further refinement.

The overarching goal of this thesis is to improve predictive models of species abundance and biomass by incorporating community data and environmental information. The three chapters are designed to build upon each other, progressively expanding the scope and application of these models (Figure 1.2). Chapter 2 establishes the foundation by focusing on simulated data to refine single-species abundance models, specifically examining how varying levels of information (i.e., true environmental drivers versus latent variables based on species co-occurrence) affect model accuracy. This chapter lays the groundwork for understanding how the inclusion of latent variables can enhance the model's predictive capabilities. Chapter 3 extends this work by applying the developed framework to empirical data from lakes, with a focus on sport fish species. This chapter addresses whether latent variables and the inclusion of different fish assemblages improve abundance predictions and explores how specific lake characteristics influence the predictive performance of the models. By transitioning from simulations to real-world aquatic ecosystems, Chapter 3 allows for a deeper evaluation of the framework in natural conditions, particularly where dispersal is limited, and species interactions are mediated by environmental selection. Finally, Chapter 4 builds on the abundance predictions from Chapter 3 to compare stacked and community models for predicting biomass across different spatial scales and levels of species richness. This chapter examines whether a stacked model, which predicts biomass by first estimating species abundance, or a community model that predicts biomass directly from environmental data and

community composition performs better. By progressing from single-species abundance models to multi-species predictions and finally to biomass modelling, these chapters are sequentially linked, each addressing broader ecological questions while refining and testing the models at increasing levels of complexity.

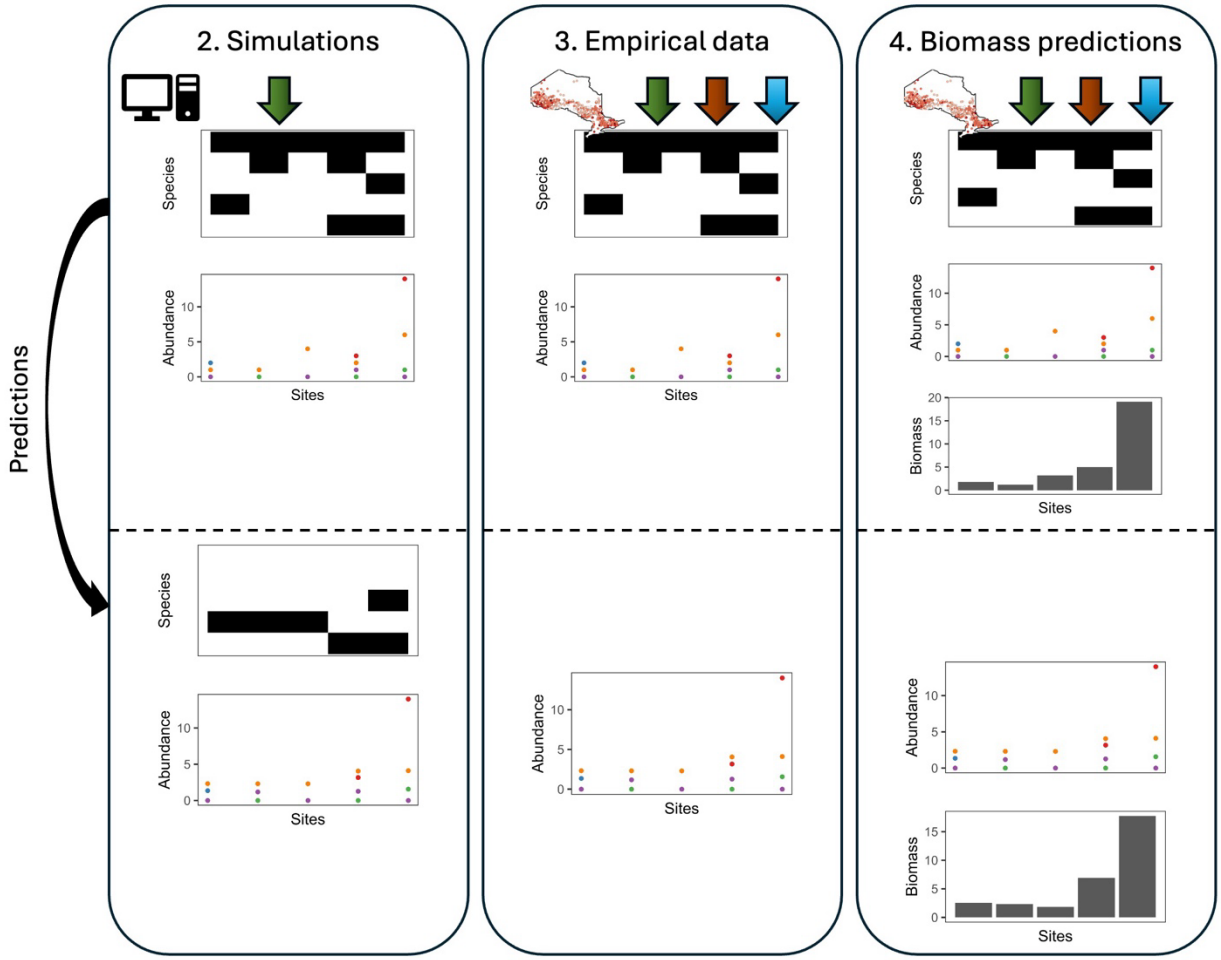

Figure 1.2: Conceptual diagram illustrating the commonalities and distinctions across chapters. Chapter 2 focuses on simulated communities where environmental selection (green arrow) is the sole assumed mechanism driving community assembly. In contrast, Chapters 3 and 4 utilize an empirical dataset that incorporates all three assumed selection processes: environmental (green arrow), biotic (brown arrow), and dispersal (blue arrow). Across all chapters, presence-absence data are used to generate latent variables which are then used as predictors. Chapter 2 predicts both presence-absence and abundance, while Chapter 3 focuses solely on predicting abundance. Finally, Chapter 4 expands on this by using abundance predictions to estimate community biomass.

1.7. Chapters overview and novelty of research

1.7.1. Chapter 2: Advancing single species abundance models: robust models for predicting abundance using co-occurrence from communities

Chapter 2 introduces a novel framework that integrates environmental variables with species co-occurrence data to predict single-species abundance distributions. Environmental variables represent ecological filters that shape species distributions, while co-occurrence data offer insights into potential species interactions or unmeasured environmental influences. By employing latent variable models, we reduce the complexity of these datasets and uncover hidden structures that would otherwise remain elusive. Building on existing approaches, this framework utilizes dimensionality reduction techniques to combine environmental and latent variables, enabling more accurate predictions of species abundance.

We assessed the model's performance through detailed simulations, evaluating its ability to account for predictive errors caused by unmeasured drivers. In particular, we aimed to: (1) derive guidelines for determining the number of latent variables used in modelling single species abundances; (2) contrast model performance containing varying levels of information on the true underlying drivers of abundance (environment) versus containing latent variables (environmental proxies based on co-occurrence patterns of species sharing variable levels of environmental affinities); and (3) assess how predictive performance varies as a function of sample size. To test these objectives, we simulated assemblages of species with varying strengths of relationships to environmental variables and without biotic interactions. We assessed the conditions that improved model predictions for a target species by using co-occurrence data from the remaining species as proxies for missing environmental predictors. This approach allowed us to evaluate how well the model

could predict the abundance of the target species by incorporating indirect information provided by the presence and absence of other species within the community.

Our results showed that incorporating presence-absence latent predictors generally improved model performance when compared to models lacking relevant environmental predictors, although there was considerable variation in performance across simulations. Notably, all models tended to have greater error rates when predicting abundant species compared to rare species.

One important aspect of the proposed framework is its versatility. It is highly flexible in terms of parameter estimation, as it can accommodate any regression style approach. This allows one to predict both presence-absence and abundance and demonstrates strong performance in predicting low-abundance species. Moreover, the framework is not limited to Gaussian copulas; other latent modelling procedures can be employed. The framework could also be used to predict biomass rather than abundance by replacing the family of the Generalised Linear Model (GLM) used, depending on the variable of highest interest for management.

1.7.1. Chapter 3: Advancing single species abundance models: leveraging multi-species data to uncover lake-specific predictive patterns and improve fisheries predictions

Chapter 3 builds directly upon the foundation laid in Chapter 2 by applying the developed framework to an empirical dataset. This transition from controlled simulations to real-world data is a critical step in validating the model's effectiveness. Species abundance is influenced by a combination of three key ecological filters: environmental selection, species interactions, and dispersal (HilleRisLambers *et al.* 2012; Vellend 2010). While the simulations in Chapter 2 were designed to isolate the effects of environmental selection (i.e., intentionally excluding species interactions and dispersal to focus on environmental drivers and the role of latent variables), the

exclusion of two fundamental ecological processes that are essential in shaping community structure in natural ecosystems introduced a significant limitation.

To address this limitation, it was imperative to select an empirical dataset for Chapter 3 that inherently minimized the influence of dispersal, thereby aligning with the assumptions and constraints of the framework established in Chapter 2 (i.e., as one cannot control for species interactions). Lakes were chosen as the study system because they are relatively closed environments with limited opportunities for species dispersal (de Bie *et al.* 2012; Shurin 2000). This characteristic makes lakes an ideal natural laboratory for testing the framework's predictions. Although it is impossible to entirely exclude species interactions in real-world datasets, selecting a system with constrained dispersal allows for a more direct assessment of environmental selection's role in shaping species abundance. The insights gained from this empirical evaluation will not only test the robustness of the framework but also provide a clearer understanding of its applicability in complex, real-world ecological scenarios.

Several adaptations were necessary to adjust to the available data. First, we selected models capable of handling non-linearity and the relative abundance of species were selected to better capture the complexities of the system. Additionally, we implemented specific adjustments to manage the wide array of environmental variables at our disposal, ensuring that the models could accurately account for these diverse influences. These modifications were critical for aligning the framework with the empirical dataset, ultimately leading to a more robust and reliable analysis. These considerations led to the formulation of our main research questions, aimed at understanding how different lake types and species interactions influence predictive ability. Specifically, we set out to investigate: (1) whether including latent variables improved predictions; (2) whether the predictions of sport fish abundances are better improved using sport fish, non-sport fish, or all fish species as predictors;

(3) which types of lakes significantly increase or decrease predictive ability and whether lakes that affect predictive ability are rare or ubiquitous in their environment and/or species composition; (4) how much species share lakes in which they increase or decrease predictive ability; and (5) whether sport fish abundances are better predicted by including all lakes or only lakes where the species is present.

We discovered that low abundance species were better predicted by models based solely on environmental variables, while high abundance species were better predicted by models incorporating latent variables derived from species co-occurrence data. We also found that the contributions of lakes to predictive models were correlated within species groups based on their occurrence levels. Species with low abundance exhibited different patterns of lake contributions compared to those with high abundance. Notably, these patterns were independent of the lakes' environmental uniqueness, species community uniqueness, or specific environmental variables. These findings highlight the complexity of predicting species abundance, aligning with Chapter 2's results, where incorporating presence-absence latent predictors improved model performance. The lack of consistent patterns across environmental factors further underscores the need for flexible models that can account for diverse influences on species distributions.

1.7.2. Chapter 4: Predicting biomass: a comparative analysis of community models and stacked abundance models across spatial scales

In Chapter 4, we shift focus from species abundance to community biomass, a critical metric for understanding ecosystem productivity and health. While the framework developed in the previous chapters demonstrated its utility in predicting species abundance using environmental data and latent variables derived from presence-absence data, biomass prediction introduces new challenges and opportunities. Building on the framework foundation laid, we introduce a stacked model that

first predicts the abundance of individual species and then, after weighting each by the average species weight, sums these predictions to estimate community biomass. We set out to investigate the following questions: (1) whether a stacked abundance model or a community model (i.e., directly predicting community biomass from environmental data and latent variables) would better predict total biomass; (2) whether diverse lakes (i.e., in terms of species richness) are better predicted than non-diverse lakes; and (3) whether regional biomass is better predicted than local biomass.

The results indicate that the stacked model tends to overestimate biomass, particularly in ecosystems where a few highly abundant species dominate. This overprediction is likely due to correlated errors in species abundance, which become amplified in the biomass estimates. In contrast, the community model demonstrated a narrower distribution of predictive errors, suggesting it may better capture community-level processes and offer more reliable biomass predictions. Furthermore, we observed that the predictive accuracy of both models varied with species richness, with more diverse and evenly distributed lakes (i.e., more equitable distribution of individuals across species) generally showing higher accuracy. These findings provide valuable insights into refining biomass prediction models, with significant implications for applications in fisheries management and biodiversity monitoring.

Chapter 2: Advancing single species abundance models: robust models for predicting abundance using co-occurrence from communities[1]

> "*The smallest, most seemingly insignificant event is part of an intricate whole and to understand why one particular mote of dust falls in one particular path, and lands in one particular location, is to understand the will of Amaat. There is no such thing as "just a coincidence"*."
>
> Ann Leckie, *Ancillary Justice*

2.1. Abstract

Accurate estimates of abundance are crucial for successful conservation and management. However, gathering abundance data is costly. Species Abundance Models (SAMs) are increasingly used to predict variation in abundance for resource management for single species, but collecting enough relevant environmental information to build effective SAMs can often be challenging. Species co-occurrence patterns may provide additional information on missing environmental predictors, and data on presence-absence species co-occurrence are typically easier to collect than abundance or detailed environmental data. However, it is still not clear when supplementing abiotic data with co-occurrence data should improve abundance predictions, as co-occurrence data itself represents a noisy indicator of the local environment. Using simulated data where we manipulated the strength of relevant environmental predictors across multiple species, we assessed the conditions that improve model predictions of a target species by using co-occurrence data on the remaining species as a proxy for missing environmental predictors. Because species often share

[1] A version of chapter 2 is currently in review at Ecosphere; it is also available as a preprint on EcoEvoRxiv. Stahl, A., Pedersen, E. J., & Peres-Neto, P. R. (2024). Advancing single species abundance models: robust models for predicting abundance using co-occurrence from communities. *EcoEvoRxiv [Preprint]*. https://doi.org/10.32942/X2S32J

environmental preferences in nature, an aspect simulated in our data, latent variables are expected to summarize important environmental gradients across co-occurring species. We employed Gaussian copulas to generate presence-absence co-occurrence-based latent variables as proxies. These latent variables, along with various combinations of environmental predictors, were subsequently used as predictors in SAMs. We evaluated the accuracy of these models in predicting the presence and abundance of target species through model validation exercises. Our results showed that incorporating presence-absence latent predictors generally improved model performance when compared to models lacking relevant environmental predictors, although there was considerable variation in performance across simulations. All models tended to have greater error rates when predicting abundant species compared to rare species. The goal of our proposed framework is to offer a novel and easy to implement method for accurately predicting abundance from both biotic and environmental information.

2.2. Introduction

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

scale management to be able to predict the local abundance of specific species based on easier-to-sample data such as the presence or absence of other species.

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

In many cases, however, the environmental data gathered and used for predicting abundance variation in space (e.g., across sites) may stand as the primary source of low predictive accuracy, rather than other additional factors. For instance, relevant environmental variables may be missing or subject to measurement errors, or there could be time lags in environmental fluctuations and related changes in abundances (Bengtsson *et al.* 1997; Dornelas *et al.* 2013; Myers 1998); and these lags may vary spatially and temporally (i.e., non-stationarity in lag-responses) even for the same species. If an unmeasured driver affects the abundance of at least two species, whether positively, negatively, or even in opposite directions between the species, one can expect that information about the distribution of one of these two species would improve the prediction of the other. This

is especially expected when the probability of a species' presence or absence is related to its abundances, and when the presence or absence of other species act as proxies for unmeasured quantitative factors (e.g., low versus high values), or qualitative factors (e.g., presence or absence of the missing factor). Indeed, several studies have shown that, for certain species, the most accurate predictor of abundance was information regarding the presences and absences of other species (González-Salazar *et al.* 2013; Lewis *et al.* 2017; Öğlü *et al.* 2020; Olkeba *et al.* 2020). While pairwise comparisons can be somewhat effective when studying single species, the interactions among multiple species can be complex and may not be adequately captured by pairwise comparisons alone.

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

A solution to incorporating complex co-occurrence data while retaining a low dimensionality is to employ latent variable models (Walker & Jackson 2011). Latent variables are unobservable variables or factors that are not directly measured but rather estimated based on the associations (covariation) among species. These latent variables aim to estimate the joint model probability distribution of species presences-absences and represent the underlying structure or patterns in the

data by specifying how data points (e.g., species composition across local communities or sites) are likely to be generated. Several methods exist to estimate latent variables from abundance or presence-absence data, including non-model-based (e.g., classic ordination methods such as principal component analysis) and model-based (e.g., mixed-model ordinations) methods (Popovic *et al.* 2019, 2022; Walker & Jackson 2011). The power of latent variable methods stems from their ability to capture hidden variation in a dataset in low dimensionality (ter Braak 1985; ter Braak & Prentice 1988). Our contribution here is to demonstrate the robustness of modelling the abundances of single target species as function of latent variables that model the co-occurrence (presence-absence patterns) of the other species. This aspect is particularly important for the management and conservation programs tailored to specific species. We introduce this general modelling framework and evaluate its ability to represent sources of predictive error caused by unmeasured drivers through detailed simulations.

The goal of this study is to assess the robustness of our proposed framework for advancing single species abundance distribution models using species co-occurrence data of other species in their communities. We used detailed simulations to contrast the performance of models containing various levels of information on the environment and community composition. Moreover, because we generate abundance distributions for all species in our simulations, we can contrast our model performance between abundance-based and species-co-occurrence based. Specifically, using comprehensive simulations, we set out to assess the performance of our proposed species-abundance framework by: (1) deriving guidelines for determining the number of latent variables used in modelling single species abundances, (2) contrasting model performance containing varying levels of information about the true underlying drivers (environment) versus latents (i.e., environmental proxies based on co-occurrence patterns of species sharing variable levels of

environmental affinities; Figure 2.1), and (3) assessing how predictive performance varies as a function of sample size (i.e., number of sites or local communities used as input into the model). In this study, we focused on scenarios in which species and their communities are influenced solely by environmental variation, without considering the impact of species interactions or dispersal, which can either enhance or diminish model performance (i.e., increase or decrease predictive accuracy, respectively).

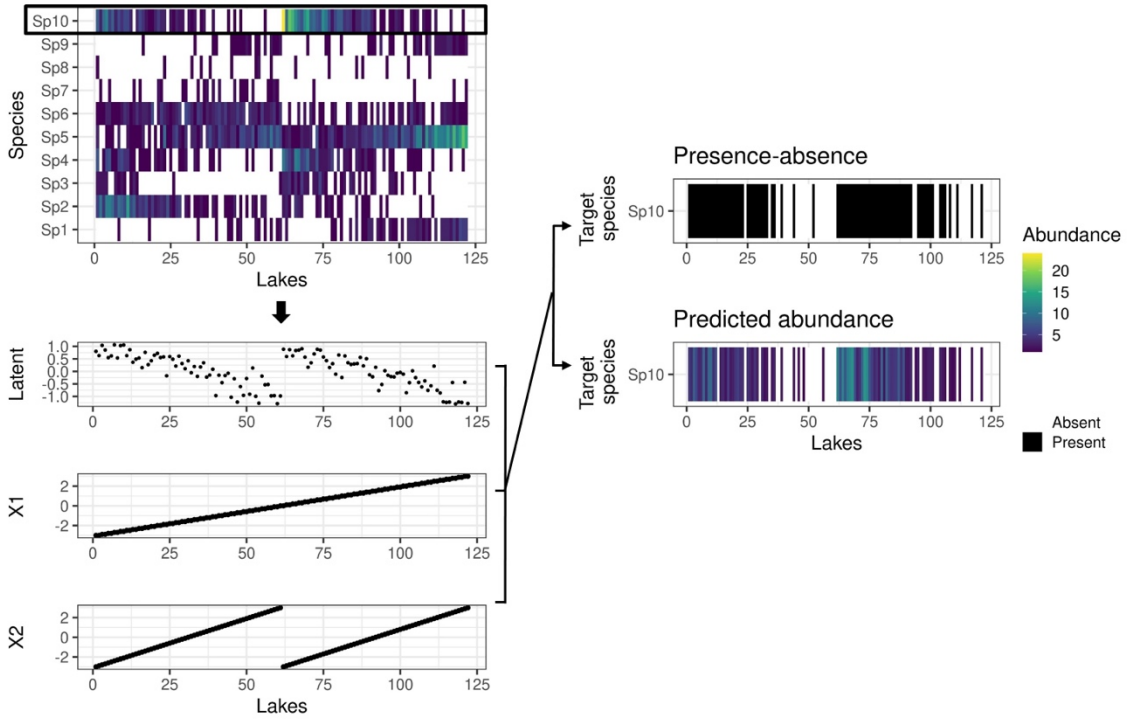

Figure 2.1: The rationale underlying our model framework and simulation workflow to assess its performance. First, species abundances were simulated for all species (top left panel) as a function of multiple environmental factors. In this example, two environmental variables were used to simulate species abundances ($\mathbf{X}_1$ and $\mathbf{X}_2$; bottom left panel). Species abundances are then transformed into presence-absence data and used to derive latent variables (bottom left panel). Here, only one latent variable is presented for simplicity, allowing one to more easily associate it with the abundances of the original simulated species. Variation in species abundances (target species) across sites is then modeled against latent and environmental variables or reduced combinations (e.g., removing an environmental variable and assess the conditions that affect latent performances), depending on specific simulation scenarios. The model can produce either abundance or presence-absence predictions for each site. The black rectangular outline highlights the target species (species 10) that the model aims at predicting.

2.3. Materials and methods

The simulations to test our framework followed the subsequent steps (see Figure 2.1 for an illustration of how this general workflow for a single simulated landscape):

1. Use stochastic simulations to generate landscape-scale environmental variation for each site in a landscape, and to generate coefficients for each species determining how average species abundance should vary as a function of environmental variables.
2. Simulate the abundance of species in each site, based on the environmental variables and coefficients generated in step 1.
3. Calculate latent variables from the presence-absence data of the previously generated abundance using Gaussian Copulas (statistical tools used to model the dependence structure between normal variables while allowing each variable to maintain its own marginal distribution, Popovic et al. 2022).
4. Using a subset of the data generated, train a set of statistical models for each species to predict local abundance. Trained models varied in the number of included environmental variables and whether the model included latent variables.
5. Use a suite of metrics to evaluate the ability of each model to predict patterns of presence-absence and abundance for the sites that were not used to estimate the models.

2.3.1. Steps 1 and 2: simulating communities

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

$$A_{s,j,u} \sim Poisson(\mu_{s,j,u})$$ Equation 2.1a

$$\mu_{s,j,u} = \exp\left(b_{0,s,u} + b_{1,s,u}X_{1,j,u} + b_{2,s,u}X_{2,j,u} + \cdots + b_{E,s,u}X_{E,j,u}\right)$$ Equation 2.1b

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

Table 2.1: Variable symbols and indexes, and their associated values and distributions used in the simulation study. Bold letters indicate that the variable is a vector or a matrix.

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

We simulated environmental covariates by drawing *J* independent, normally distributed values for each of the *E* environmental variables for each landscape (step 1). Thus, values for each covariate were statistically independent, with each environmental covariate having a mean of 0 and a

variance of 1 across sites. These environmental covariates can be interpreted as environmental gradients given that they were generated independently. The coefficients ($b_{0,s,u}, b_{1,s,u}, \ldots b_{E,s,u}$) for each species were drawn from a uniform distribution with a range of -2.4 to 1.2 for the intercept, and -0.8 to 0.8 for the slopes. The ranges for the coefficients were determined through simulation trials where we identified the minimum and maximum coefficients that allowed for all species to be present in at least 10% of sites and at most in 90% of sites. The selected parameters allowed to generate species with different levels of strength between abundance and environment variables (e.g., narrow versus broad niche breadths; step 2). Table 2.1 summarizes how each variable in Eq. 2.1 was generated. The distribution across species of spatially averaged species abundance within each landscape was approximately log-normally distributed (Figure 2.2), resembling common patterns found in natural communities.

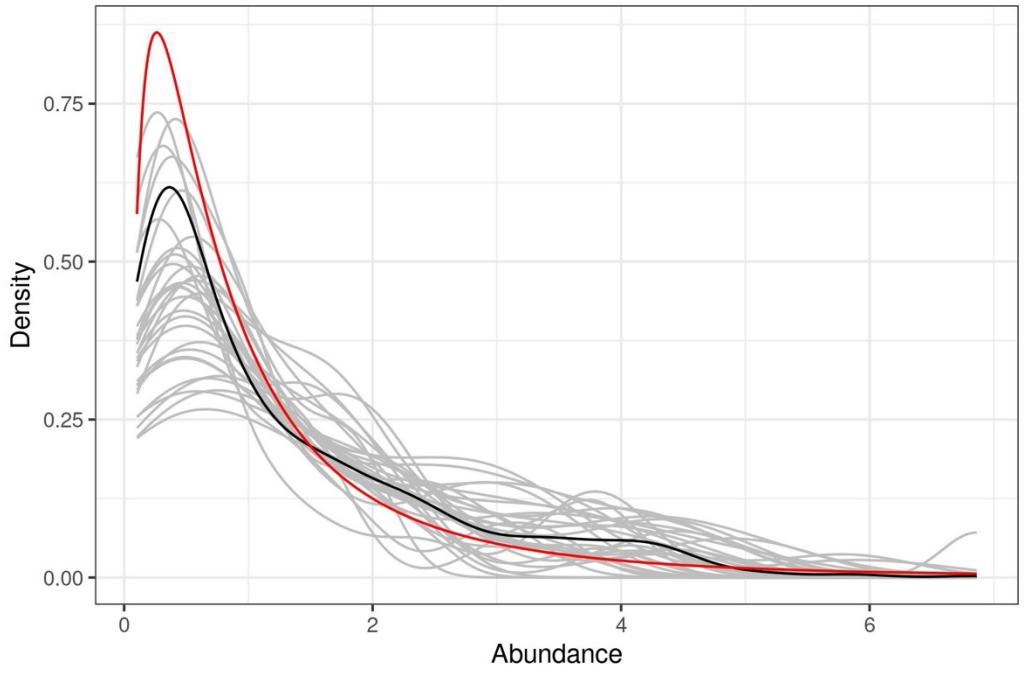

Figure 2.2: The density of average species abundance across sites within each landscape. For each landscape, we calculated the average abundance of each species and plotted the density of abundances in each of the 30 landscapes (grey lines). We also plotted the density of abundances across all landscapes to represent the average landscape (black line). The red line is a reference line indicating the probability density function of a log-normal distribution with the same log-mean and log-standard deviation of the average abundance distribution across replicates.

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

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

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

was always lowest for models with a single latent variable, regardless of the number of environmental predictors used to simulate species abundances. As such, we conducted a preliminary trial to evaluate the number of latent variables needed to best approximate the environmental gradients in our simulated landscapes.

Using Eq. 2.1, we simulated U landscapes of size J (number of sites), containing S species and a varying E number of environmental predictors ($U = 450$, $J \in \{100, 200, 300\}$, $S \in \{10, 20, 30\}$, $E \in [1,5]$; Table 2.1). To evaluate the optimal number of latent parameters (axes) needed to best approximate the environmental gradients in our simulated landscapes and compare the impact of adding or removing latent variables, we generated several numbers of latent variables for each possible combination of parameter values. Therefore, for each possible combination of parameter values, we fitted the presence-absence data into a stacked species regression model before using a model-based ordination with Gaussian copulas by using the functions *stackedsdm* and *cord* from the package *ecoCopula* (Popovic et al. 2019, version 1.0-2) with L different numbers of latent factors to model them ($L \in [1,5]$).

We extracted the BIC value of each of these models and subtracted from them the BIC of the best model from any given simulation set (i.e., lowest BIC for the species considered in the current landscape). To evaluate the effectiveness of the latent variables in representing (i.e., serve as a proxy) environmental variation, we conducted a redundancy analysis (RDA) of the original environmental variables used to simulate species abundance regressed against the extracted latents using the function *rda* from the package *vegan* (Oksanen et al. 2024, version 2.6-2). Ability of latents to represent environmental variation was measured via the RDA adjusted R^2 (Peres-Neto *et al.* 2006). We determined from this trial that, regardless of the number of sites J or species S in the simulation, BIC was always lowest with a single latent variable (Figure S2.2), but adjusted R^2 did

increase with the number of latent predictors, until the number of latents equaled E, after which the adjusted R^2 did not increase with more latent variables (Figure S2.2), so there is no reason to extract more than *E* latent variables for any given simulation.

2.3.3. Step 4: Contrasting the performance of abundance models

We compared the models containing only the environmental variables used to generate species abundances (Eq. 2.1) against the ones containing selected environmental variables and the latent variables (community composition). This allowed us to compare model performance under ideal conditions because we used the true environmental drivers used to simulate species abundances against models from which we removed various combinations of environmental variables (scenarios) and replaced them with latent variables (proxies) to represent the missing sources of variation. Note, however, that ideal conditions do not imply perfect model performance, as different species were simulated with varying degrees of strength and associated errors relative to environmental variables (e.g., narrow versus broad niche breadths).

For this contrast, we created *U* landscapes, and for each landscape *u*, we generated *K* replicates (*U* = 30, *K* = 10 replicates per landscape). For each replicate *k*, we simulated abundances for each *s* species in each site *j* using Eq. 2.1, using three environmental variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ per landscape containing multiple sites. We simulated 20 species and 1000 sites per landscape. We fixed the number of latent factors to 3 as we had three environmental variables (see RDA results in previous section). Replicates (i.e., landscapes using the same coefficients but had varying values of environmental gradients) were used to allow a reasonable estimate of the metrics used to contrast model performances.

We randomly sampled 100 sites (out of the 1000 simulated) from each landscape *u* (referred here as to the training set), and for each training set we estimated abundance models with different

combinations of environmental and latent predictors (step 4). Each model was estimated using a Generalized Linear Model (GLM), using a Poisson distribution with a log-link function (Kéry & Royle 2015; Nelder & Wedderburn 1972). We used the *manyglm* function from the R package *mvabund* (Wang et al. 2022, version 4.2-1) to fit separate models for each replicate landscape simultaneously for all species separately.

We were interested in comparing models containing different combinations of environmental variables and latent variables. The complete list of model scenarios considered is described in Table 2.2. As each species had different strengths of relationship with each environmental variable (i.e., different coefficient values in Eq. 2.1 were used to simulate each species), we ordered the models based on the absolute decreasing values of the environmental coefficients used to simulate the species' abundance. For instance, if species *A* had the values of -0.5, 0 and 0.8 as coefficients for the environmental variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$, respectively, $\boldsymbol{X}_3$ had the largest influence on driving abundance values, followed by $\boldsymbol{X}_1$ (i.e., importance is given by absolute decreasing coefficient values) and $\boldsymbol{X}_2$. But if species *B* had values of 0.7, -0.5 and 0.3 as coefficients for the environmental variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$ respectively, its abundance was mostly driven by variations of $\boldsymbol{X}_1$, then $\boldsymbol{X}_2$ and finally $\boldsymbol{X}_3$. When removing $\boldsymbol{X}_1$ from the predictors of a model, species *A* and *B* were not impacted in the same way due to the lesser influence $\boldsymbol{X}_1$ had on the abundance of species *A*. We predicted that including latent variables should increase predictive ability more when added to a model that only included environmental predictors that weakly predicted the abundance of an individual species. To test this, we compared model performance with and without latent variables for models including different combinations of strengths of environmental variables.

For models containing one environmental variable as predictor, we labeled the predictors as "high", "intermediate", and "low", corresponding to the decreasing values of coefficients of the

environmental variables. For models incorporating two environmental variables, we designated the model with the two highest coefficients as "high", the model with the highest and lowest coefficient as "intermediate", and the model with the two lowest coefficients as "low".

Table 2.2: All models considered in this study based on combinations of environmental variables and community composition (latents). The best model is expected to be the "true" model considering all three environmental variables. **A** refers to the abundance matrix, $\mathbf{X}_1$ to $\mathbf{X}_3$ to the environmental variables, and $\mathbf{Z}_1$ to $\mathbf{Z}_3$ to the community composition (latent variables).

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

2.3.4. Step 5: comparison of model performance

For each model estimated for each replicate within the same landscape, we generated predictions for species abundances at the remaining 900 sites in the landscape from which the sites were sampled from (the test set). To establish baselines for optimal model performance, we also calculated predicted abundances in the test set using the oracle model: i.e., the model employing the true coefficients used to simulate each species' abundances to predict the conditional expected abundance for each species in each site. The oracle model represents the best possible model for estimating the simulated abundances in each test set that could be derived using data from the training set. Two other models were singled out: (i) a benchmark model containing all three environmental variables, to identify in which scenarios having access to all environmental variables

(drivers of the abundance) did not suffice to properly estimate the environmental coefficients (by comparing the performance of the benchmark model to that of the oracle model), and (ii) a latent model containing only the latent variables, to study how species co-occurrence patterns performed as predictors of their own. We assessed how effectively the different models, including the oracle model, predicted the pattern of presences and absences as well as the true abundances in the test set.

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

Metrics for evaluating presence-absence predictions

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

Therefore, instead of using a fixed probability threshold to convert the probabilities into presence-absence predictions, we used a prevalence-based approach. For each species, we set a threshold

equal to the true occurrence (prevalence) rate of the species across a given landscape (Liu *et al.* 2005). We used this threshold to generate a predicted presence-absence matrix for each site and each species in each landscape for a given model. This was achieved by determining whether the expected abundance by the model for that site was greater (present) or lower (absent) than the threshold value. We then compared the performance of each model to the oracle model using a range of metrics, the equations for which are provided in Table 2.3. Using the predicted presence-absence matrices, we calculated the True Skill Statistic (TSS, Peirce 1884; Table 2.3) for each model, species and landscape replicate. The TSS, which ranges from -1 to +1, measures the difference between the sensitivity and specificity of the model. A score of +1 indicates a perfect agreement between the model's predictions and the true presence-absence, while a score of 0 or lower signifies performance no better than random (Allouche *et al.* 2006). We calculated the ratio of the TSS of the model over the TSS of the oracle and computed the mean for each model, species and landscape. Then, we grouped species into bins based on occurrence rates across different landscapes. A TSS ratio of ≥ 1 indicates that the model performed as well or better than the oracle, while a TSS ratio of ≤ 0 means that the model predicted presence as badly or worse than random chance.

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

Table 2.3: Metrics used for assessing model predictive performance based on presence-absence and abundance of target species. J represents the number of sites, $\boldsymbol{A}_s$ the true abundance of the (target) species, $\boldsymbol{P}_s$ the predicted abundance, TP the true positives, FP the false positives, TN the true negatives, and FN the false negatives. Bold letters indicate that the variable is a vector or a matrix. The True Skill Statistic (TSS), sensitivity, and specificity are calculated for all sites of the landscape. Having evaluated the presence-absence predictions of the models and to avoid artificially inflating the error rate of the abundance metrics, the Mean Absolute Percentage Error (MAPE), Root Mean Squared Percentage Error (RMSPE), Relative Mean Squared Error (RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), and Root Mean Ratio Percentage Error (RMRPE) are calculated for sites where the species is truly present (i.e., abundance of 1 or more).

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

Metrics for evaluating abundance predictions

When evaluating how each model predicted species abundance, we limited comparisons to sites where the species was present (i.e., abundance of 1 or higher) and calculated the following prediction metrics for each model, species and landscape replicate: Mean Absolute Percentage Error (MAPE), Root Mean Squared Percentage Error (RMSPE), Relative Mean Squared Error

(RMSE), Symmetric Mean Absolute Percentage Error (SMAPE), and Root Mean Ratio Percentage Error (RMRPE) (see Table 2.3 for definitions of these metrics). We calculated the ratio of each metric to the corresponding metric calculated for the oracle model (i.e., best possible scenario) and calculated the average ratio for each model, species and landscape (referred to as the ratio metric in the results). We also calculated the delta metric, defined as the metric calculated for a model containing only environmental variables minus the metric calculated for a model with the same environmental variables as well as latent variables. As above, a negative delta metric indicated that the latent model performed better than the same model lacking latent variables.

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

2.4. Results

2.4.1. Number of latent variables needed to capture environmental variation

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

The adjusted R^2 consistently increased with the number of latent dimensions until it equaled the actual number of environmental variables used to simulate the data, at which point it plateaued (Figure 2.3, Figure S2.2). This indicates that additional latent variables did not improve the model's ability to predict the environmental state of a given location. The maximum fraction of variance explained was not significantly affected by the number of true environmental variables used to generate (simulate) species abundances; capturing variation from one environmental gradient was as feasible as capturing it from three or four environmental gradients (i.e., variables). Note, again,

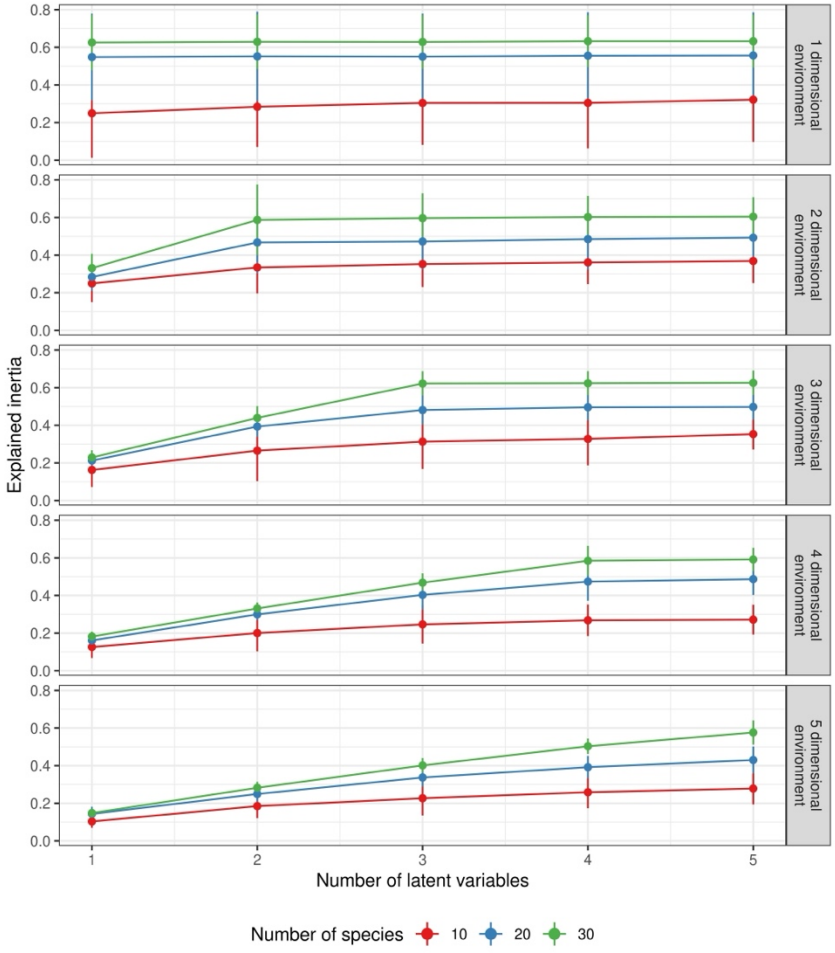

Figure 2.3: Variation in adjusted R^2 as a function of the number of latent variables used, as well as the true dimensions of the environment and the number of species in the landscape. Here we used 500 sites, and variations according to other number of sites are presented in Figure S2.2. Colors represent the varying number of species in the landscape, and each panel indicates the true dimension of the environment (i.e., number of environmental variables used to simulate the abundance of a given target species).

that the interpretation here as gradients is possible because environmental variables were generated independently. The adjusted R^2 was not sensitive to the number of sites in the landscape used to estimate the latent variables, but it was sensitive to the number of species used: models based on 10 species could only explain about 30% of the variation in environmental variables, regardless of the number of latent variables used, whereas models based on 30 species could explain ~60% of variation in the environmental matrix (Figure S2.2).

In contrast, the Bayesian Information Criterion (BIC) consistently increased with the number of latent dimensions, without showing any signs of reaching a plateau (Figure S2.1). While models with lower BIC are generally expected to have better predictive ability for unobserved data - suggesting that the best model would always retain one latent variable regardless of the environmental dimension - this expectation did not align with our observations for the adjusted R^2. This discrepancy indicates that BIC (as calculated by *ecoCopula*) is not a good metric of the predictive performance of the latent model, at least when applied to gradients driving abundances while their latents were extracted from presence-absence data. Therefore, we did not report BIC of the estimated latent models for the remainder of our simulations.

2.4.2. Models' performance

Presence-absence predictions

We now focus on the models' performance in predicting presence-absence, including the ratio TSS (representing how well each model performed compared to the oracle model) and delta TSS (represented how well models without latent variables performed relative to models including latent variables). The ratio of the TSS had a mean of 0.7 and ranged from -1.6 to 1.7 (recall that any value below 0 indicates that the model did not perform better than random, while any value above 1 represents better performance compared to the oracle). Initially examining the TSS across

species occurrence percentiles, there were no obvious patterns (Figure 2.4). In this case, the number of occurrences of a target species did not influence model's performance. When comparing models, those containing two environmental variables performed better on average than those with only one, regardless of whether latents are included or not.

When comparing models with and without latent variables, any delta TSS value above 0 indicates that the environmental model performs better, while any negative value indicates a better performance by the latent model. Models containing latent variables generally performed better on average across all (target) species, especially for those with high occurrence and in models containing only one environmental predictor (Figure 2.4). The differences are less pronounced when comparing models that contain two environmental variables (i.e., where only one environmental predictor is missing from the model). Reducing the number of sites used to fit the model did not affect the performance of the TSS, sensitivity, or specificity (Figure S2.3).

When comparing the TSS as performance of the oracle (i.e., a model using the true coefficients of the environmental variables to generate the species' conditional expectations), benchmark (i.e., a model containing all three environmental variables), and latent models (i.e., a model containing only the latent variables), we can notice that they are very correlated across species occurrence percentiles (Figure 2.5). The benchmark and oracle models have extremely similar performances. Regarding sensitivity, the benchmark and oracle models are also highly correlated, while the latent model demonstrates good correlation for species with low occurrence. For specificity, the benchmark and oracle models are correlated for high occurrence species, while the benchmark and latent models are correlated for low occurrence species.

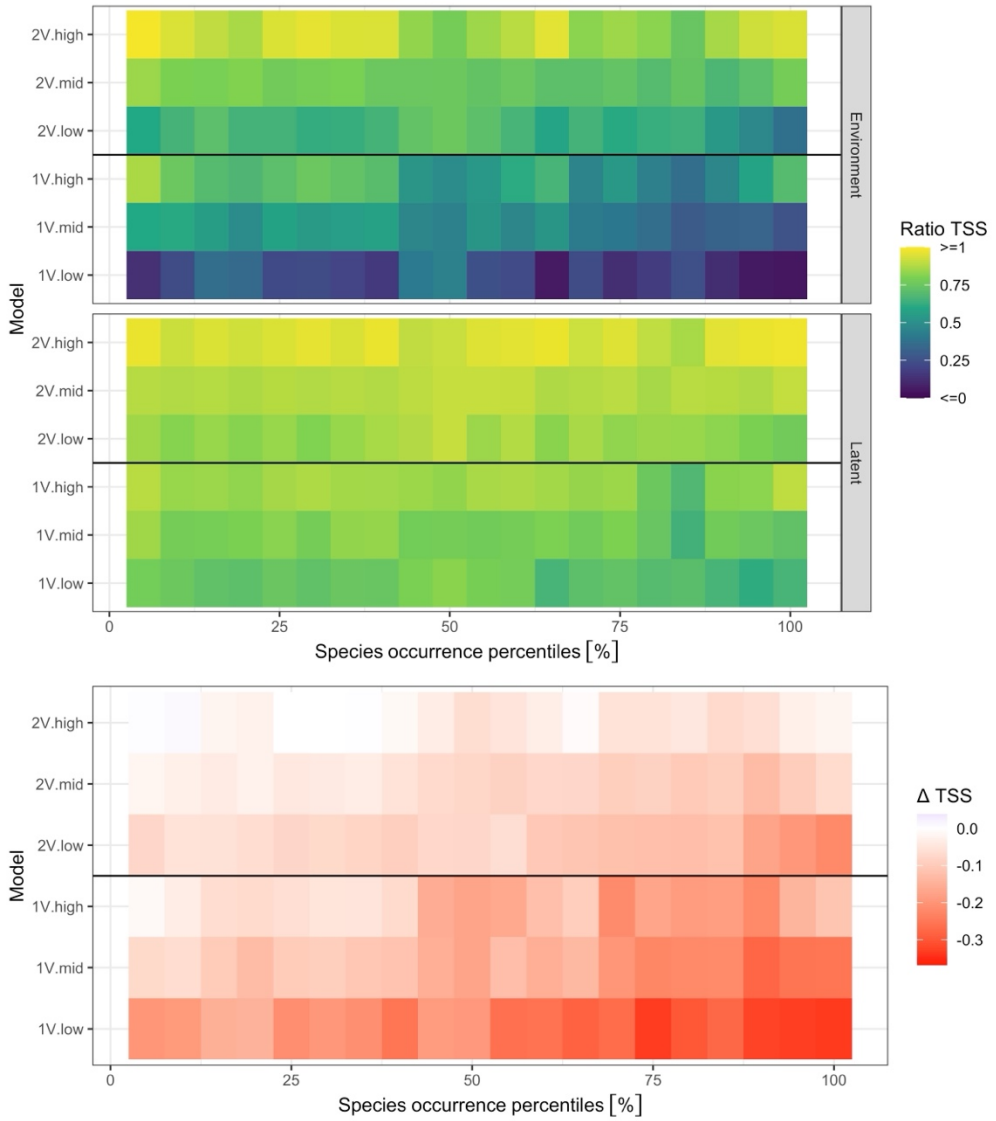

Figure 2.4: Ratio TSS and delta TSS for each model and bin of species occurrence percentiles. The ratio TSS was averaged across all landscapes and replicates per model and species, with species binned by percentile of occurrence (percentage of sites occupied) and divided by the TSS of the oracle model. A value of 1 for the ratio TSS indicates an identical performance between the model and the oracle model, while a value below 0 represents a performance similar to that of a random model. To improve contrast between colors, we confined the color scheme between 0 and 1. Any value below 0 indicates a prediction of presence-absence no better than a random model, and any value above 1 indicates a better prediction than the oracle model. The environment panel represents models containing only environmental variables, while the latent panel is for models containing latent variables (mix of latent and environmental predictors); the models were then ordered from bottom to top as fewest to the greatest number of environmental variables included and sorted by coefficients relative to each environmental variable (see Methods for more information, note that the "mid" model refers to the "intermediate" model). The delta TSS was measured as the TSS of the model with environmental variables minus the TSS of the model with the same combination of environmental variables and latent variables. A negative value indicates that the model with latent predicts the presence-absence of the species better than the model containing only environmental variables.

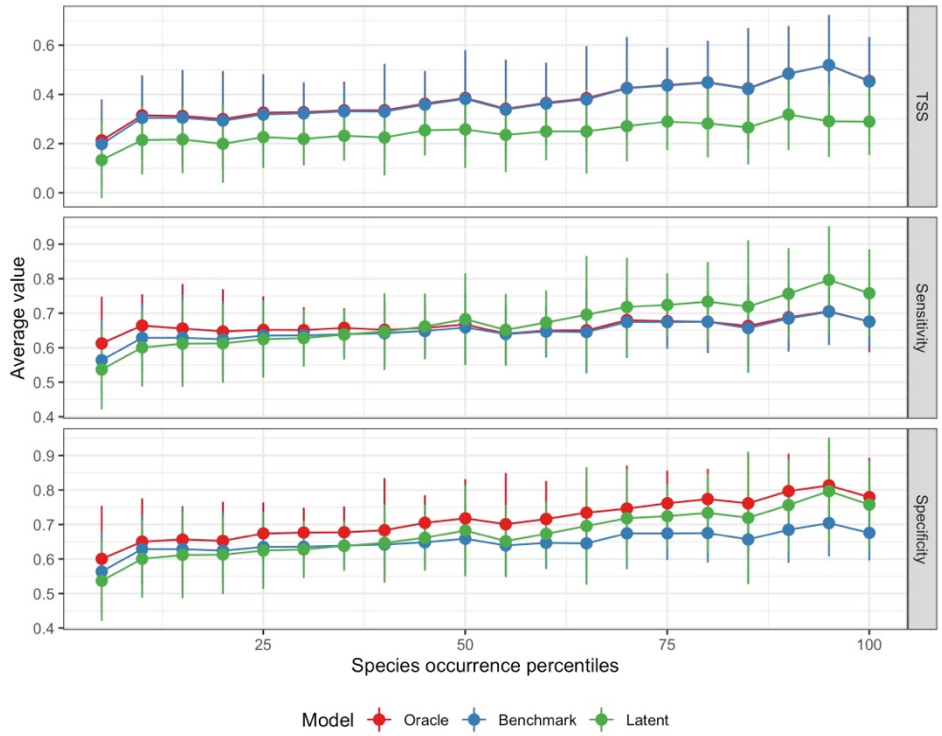

Figure 2.5: Correlation between the metrics studied (TSS, sensitivity, and specificity) depending on the model across species occurrence percentiles. The vertical panels indicate the different metrics, with models represented in different colors. The oracle model refers to the model using the true environmental coefficients, while the other models were fitted using all environmental variables (benchmark) or latent variables (latent). The True Skill Statistic (TSS) measures the difference between sensitivity and specificity of the model and ranges from -1 to +1. A score of +1 indicates a perfect agreement between the predictions of the model and the true presence-absence, while a score of 0 or less represents a performance no better than random. Sensitivity represents the ability to correctly classify a species as "present", while specificity represents the ability to correctly classify a species as "absent". Their values can be interpreted as a percentage, with values of 1 indicating perfect classification of either presence or absence, and values of 0.5 no better than random. Here we used 500 sites, and variations according to other number of sites are presented in Figure S2.3.

Abundance predictions

To assess the goodness of fit for abundance-based models (i.e., target species include abundance information while latents are based on presence-absence of the other species), we calculated six metrics to assess the extent to which the models mispredict species abundances. Again, we used the ratio of each metric over the same metric calculated for the oracle model (i.e., representing the

best possible predictive scenario), along with the delta metric to compare models that differ in composition due to the inclusion or exclusion of latent variables.

To assess across all species the impact on model performance of removing any given environmental predictor, we had to consider the varying strengths in the relationship between each species abundance and each environmental variable to compare the predictive ability of latents. As a reminder, in models containing one environmental variable as predictor, we labeled the predictors as "high", "intermediate", and "low", corresponding to the decreasing coefficients of the environmental variables. For models incorporating two environmental variables, we designated the model with the two highest coefficients as "high", the model with the highest and lowest coefficient as "intermediate", and the model with the two lowest coefficients as "low". Regardless of the metric considered, we observe the following patterns: prediction error increases as species abundance increases, and models containing two environmental variables outperform models containing only one environmental variable (Figure 2.6, Figure S2.4). When comparing models with or without latent variables, highly abundant species were best predicted by models containing latent variables (Figure 2.6, Figure S2.4). For species with low and medium abundances, the inclusion or exclusion of latent did not impact the performance of the models; they exhibited very similar values of error.

When comparing the metrics in relation to the performance of the oracle (i.e., a model using the true coefficients of the environmental variables to generate the species' conditional expectations), benchmark (i.e., a model containing all three environmental variables) and latent models (i.e., a model containing only the latent variables), we observe identical trends across all metrics. The performance of the three models was very similar for low abundance species; however, the latent

model diverged when the abundance percentile was higher than 70%, with an increase in predictive error (Figure S2.5). The metrics were not sensitive to the number of sites in the landscape used to fit the models (Figure S2.6).

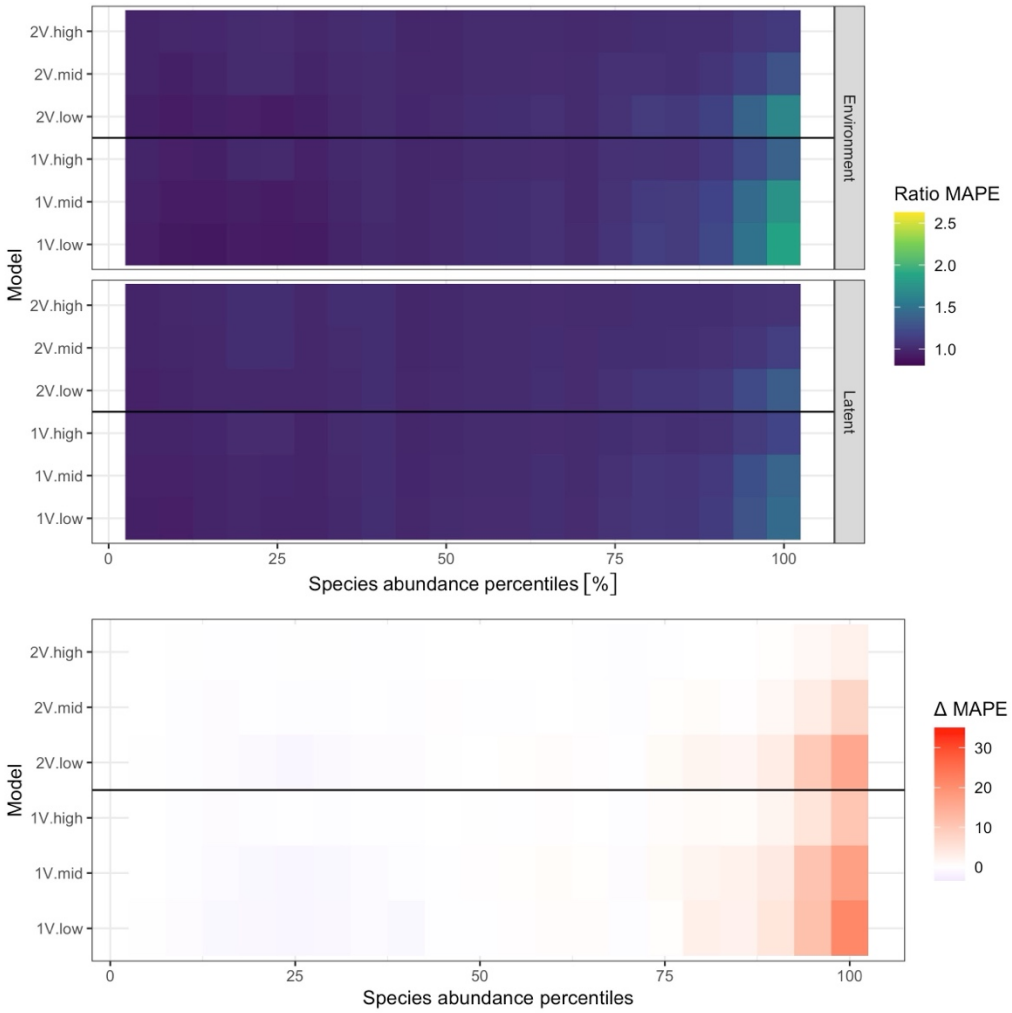

Figure 2.6: Ratio Mean Absolute Percentage (MAPE) and delta MAPE are presented for each model and bins of species abundance percentiles. The MAPE is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance and divided by the MAPE of the oracle model to derive the ratio MAPE. The environment panel represents models containing only environmental variables, while the latent panel depicts models containing latent predictors. The models are then ordered from bottom to top, from the fewest to the greatest number of environmental variables included and sorted by coefficients relative to each environmental variable. See Methods for more information, note that the "mid" model refers to the "intermediate" model. Delta MAPE was measured as the MAPE of the model with environmental variables only minus the MAPE of the model with the same combination of environmental and latent predictors. A positive value indicates that the model with latent predicts the abundance of the species better than the model containing only environmental variables.

2.5. Discussion

2.5.1. Number of latent variables needed to capture environmental variation

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

strategy when using latent models based on co-occurrence data to predict abundance of single (target) species.

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

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

2.5.2. Model performance

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

Presence-absence predictions

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

a sensitivity analysis after including measurement errors either in the way environment (e.g., spatial variation within sites, temporal lags in species responses to environments) or abundance (e.g., estimates based on mark-recapture) are measured.

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

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

Abundance predictions

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

As expected, adding environmental variables improved the abundance predictions. Since no measurement error was included in the simulations for either environmental variables or species abundances, the inclusion of any environmental variable is likely to improve predictive accuracy. However, it is interesting to note that adding community composition only improved predictions for the high abundance species. One possible explanation for this is that the way we generated species abundances resulted in low-abundance species also being only weakly predictable from environmental variation (and thus only weakly predictable from community composition). In our simulations, a species would have low average abundance if it either had a small intercept (b_0) and values of the environmental slopes (b_1 to b_E values) close to zero (so it would be roughly equally distributed across the landscape), or if it had a very small intercept value (b_0) and one large environmental slope value, so it was well-predicted by a single environmental variable. As such, the low predictive power of latent variables for rare species observed in our results may not generalize to species in natural settings. In fact, one might expect that species with intermediate abundances are likely to be best predicted due to the positive relationship typically observed between occupancy (number of sites occupied) and abundance (Gaston 1996; Wright 1991).

Species with low abundances may not occupy all suitable habitats, while those with high abundances could be generalists, occupying an excess of environments. Additionally, many other non-environmental factors (e.g., biogeography, dispersal limitation, species interactions, species introductions) may plays an important role in shaping patterns of species distributions and biodiversity in local communities and regionally (Boulangeat *et al.* 2012; Guisan & Thuiller 2005; Lewis *et al.* 2017). We suggest that future research could extend these simulations to incorporate nonlinear and non-stationary environmental gradients, given the growing interest in how such dynamics influence species abundance (Doser *et al.* 2024).

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

modelling procedure, whether it is based on maximum likelihood, Bayesian or machine learning models.

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

Additionally, increasing the number of sites sampled did not influence predictive performance, a result we anticipated since we sampled uniformly across the landscapes and captured the entire range of variation when fitting the model. However, such uniform sampling across landscapes is unlikely to be realistic when using empirical data, particularly in complex and patchy landscapes in which environmental features are clumped and spatially autocorrelated. This issue extends beyond our study. Various approaches have been proposed to mitigate the impact of complex landscapes on the predictive performance of species distribution models based on environmental features. Different sampling methods (Christianson & Kaufman 2016; Fortin *et al.* 1989), model validation techniques (Wenger & Olden 2012), and modelling frameworks (e.g., Dormann 2007a

for a review, Guélat & Kéry 2018) are among these proposed solutions and could, in principle, be incorporated into our modelling framework given its flexibility.

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

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

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

2.6. Supplementary Information

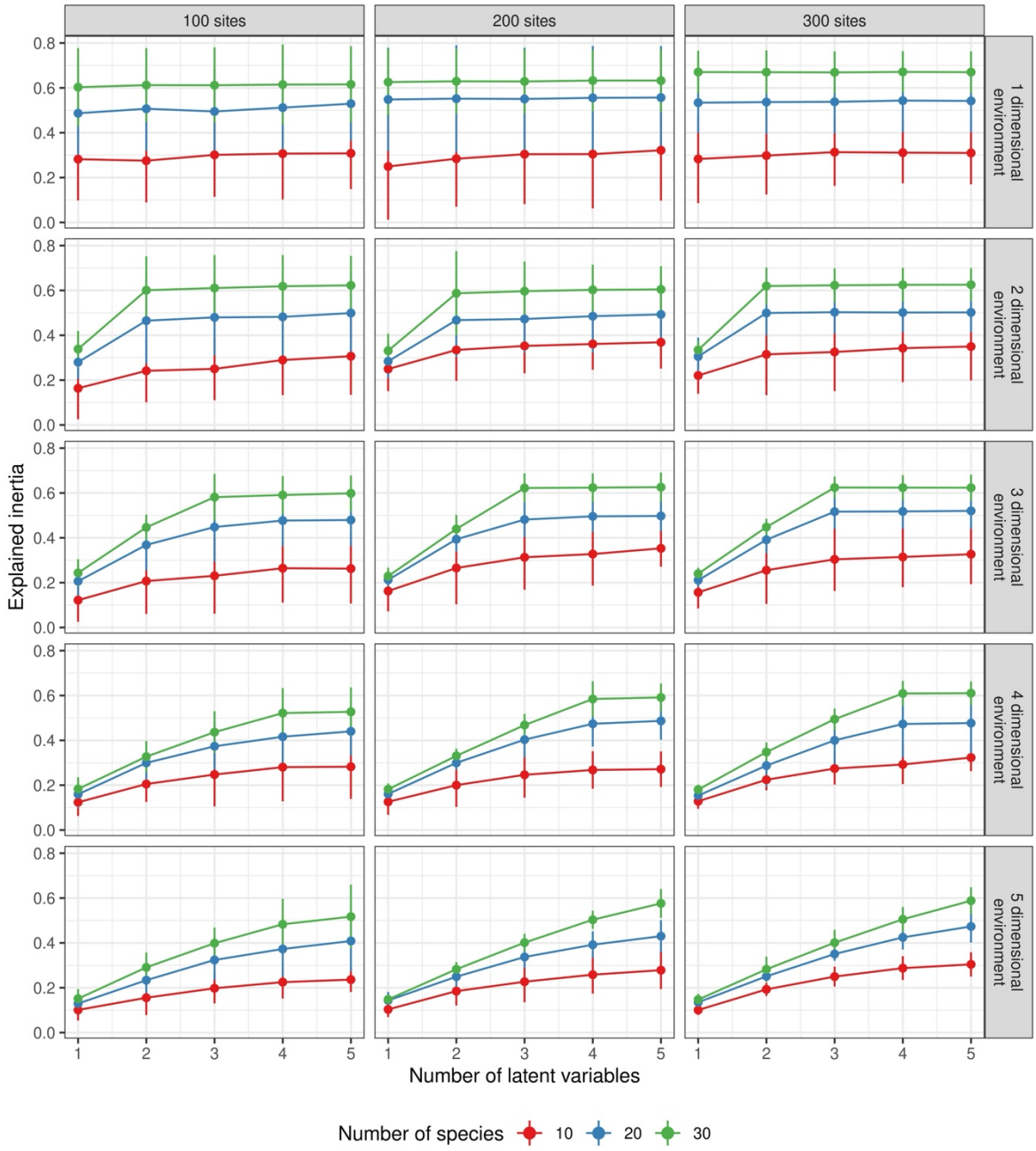

Figure SI 2.1: Variation in delta BIC as a function of the number of latent variables used, as well as the true dimensions of the environment, the number of species in the landscape and the number of sites. Horizontal panels represent the number of sites, and each vertical panel indicates the true dimension of the environment (i.e., number of environmental variables used to simulate the abundance of a given target species). Colors represent the varying number of species in the landscape. The delta BIC is calculated as the BIC of the model minus the BIC of the best model for the ongoing simulation.

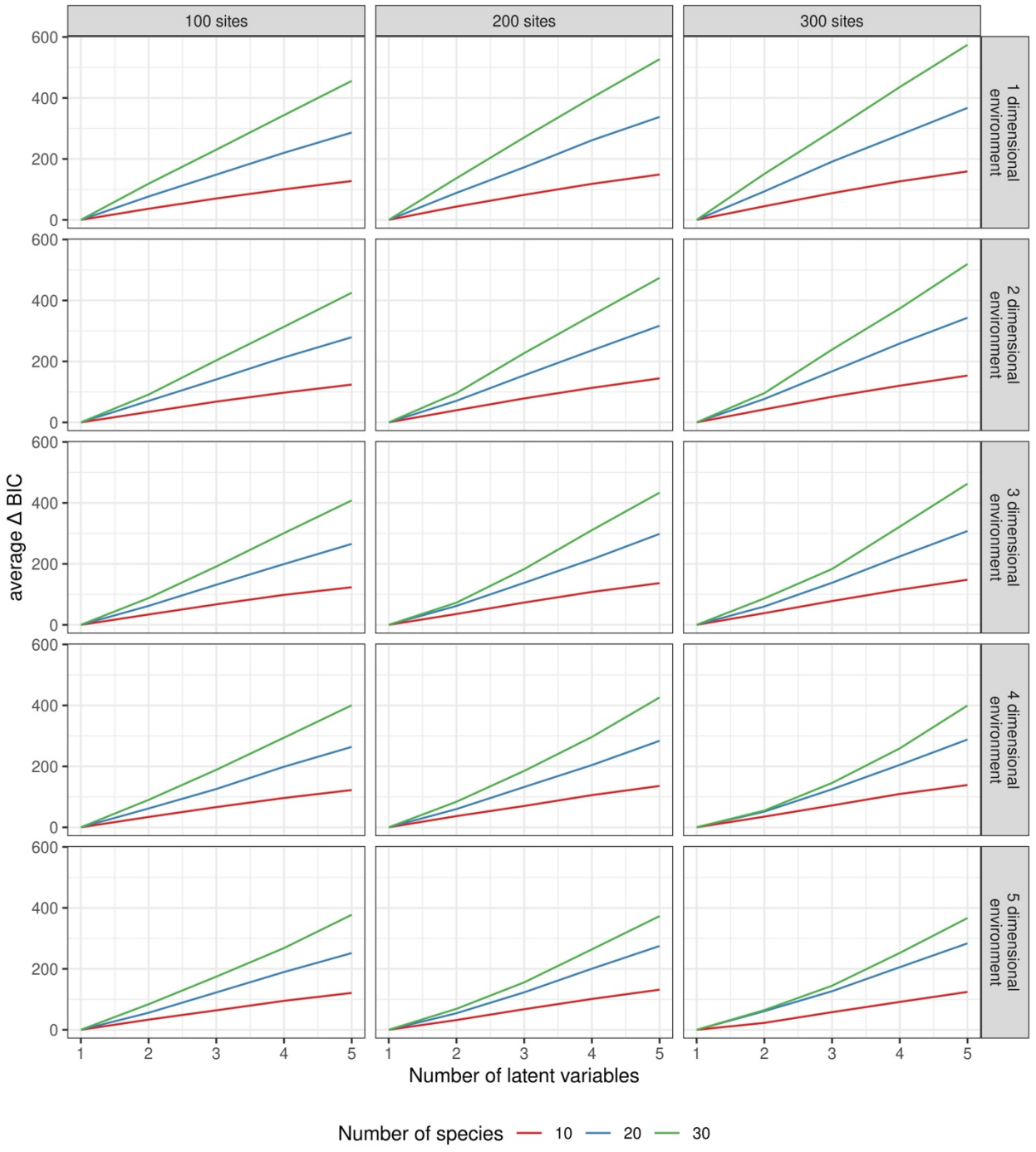

Figure SI 2.2: Variation in adjusted R^2 as a function of the number of latent variables used, as well as the true dimensions of the environment, the number of species in the landscape and the number of sites. Horizontal panels represent the varying number of sites, and each vertical panel indicates the true dimension of the environment (i.e., number of environmental variables used to simulate the abundance of a given target species). Colors represent the varying number of species in the landscape.

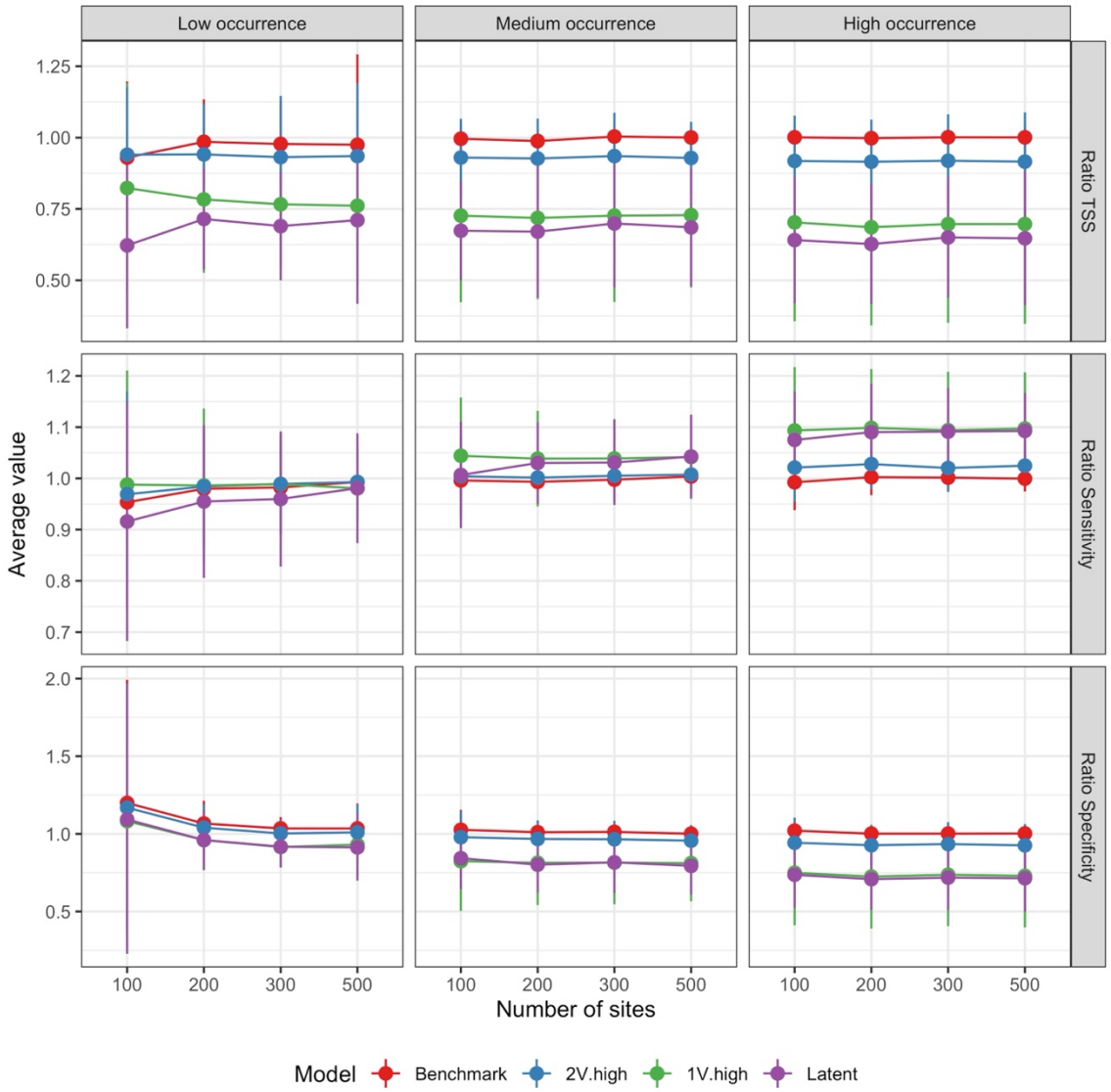

Figure SI 2.3: Average value of the studied metrics (Ratio TSS, ratio sensitivity, and ratio specificity) depending on the number of sites used to fit the models, the model used, and the occurance of species. Horizontal panels represent the different occurrence: species with low, medium and high occurrence corresponding respectively to bins of 15, 50, and 80 percentiles of occurrence. Vertical panels indicate the metrics considered, with the models represented in different colors. The ratio metric is calculated as the metric for the predictions of a model for a species of the landscape divided by the same metric calculated for the oracle model. For the ratio TSS, a score of 1 indicates a perfect agreement between the predictions of the considered model and the oracle model, while a score of 0 or less represents a performance no better than random. For the ratio sensitivity, it represents the ability to correctly classify a species as "present", while the ratio specificity represents the ability to correctly classify a species as "absent". For both metrics, values above 1 indicate a better performance than the oracle model and values below 1 indicate a lesser performance. The benchmark model refers to the model containing all environmental variables, 2V.high the model with the two environmental variables with the highest coefficients, 1V.high the model with the environmental variable with the highest coefficient, and Latent the model containing the latent variables.

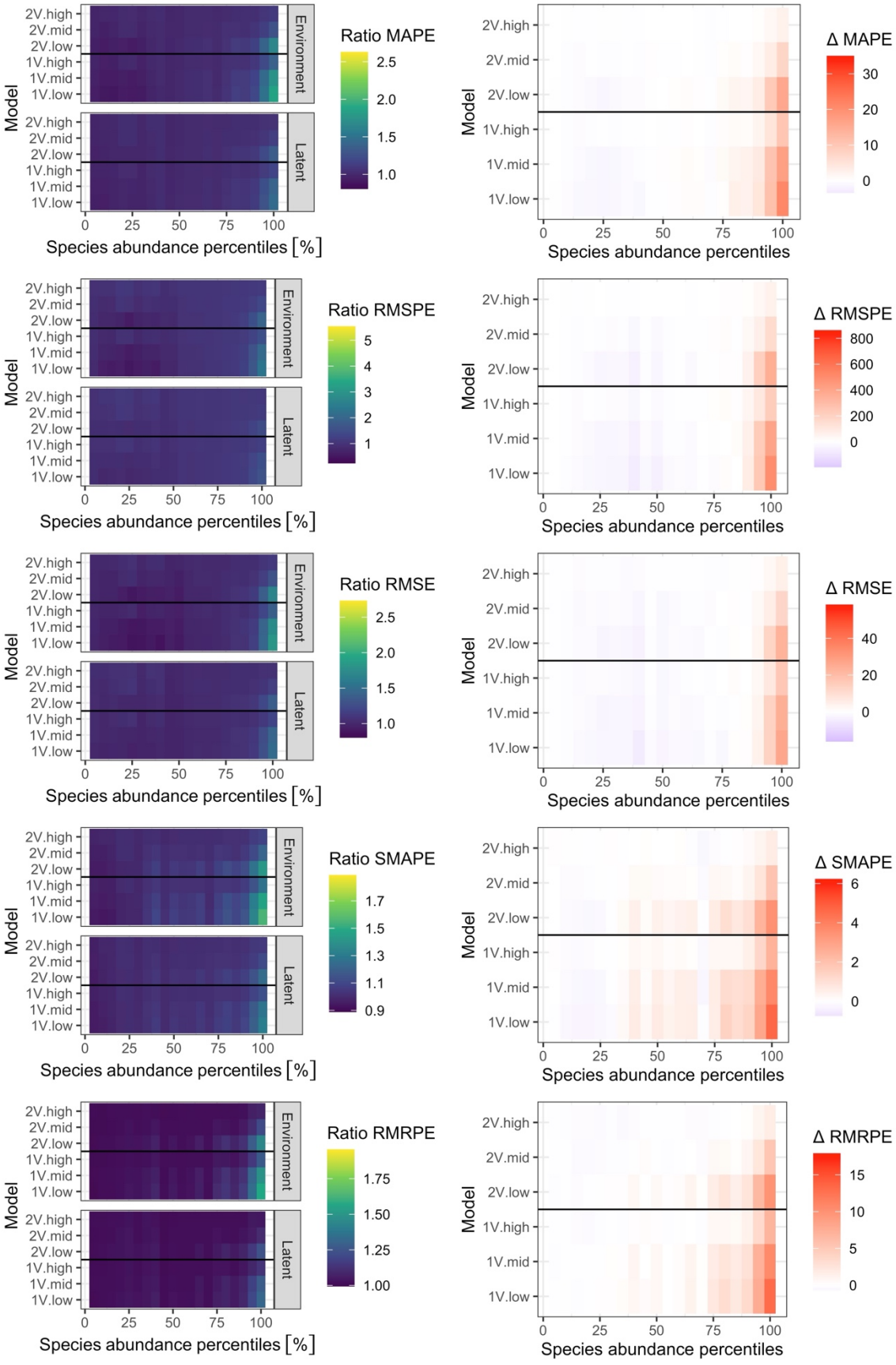
Model
2V.high
2V.mid
2V.low
1V.high
1V.mid
1V.low
Environment
Latent
0
25
50
75
100
Species abundance percentiles [%]
Ratio MAPE
2.5
2.0
1.5
1.0
Δ MAPE
30
20
10
0
Ratio RMSPE
5
4
3
2
1
Δ RMSPE
800
600
400
200
0
Ratio RMSE
Δ RMSE
40
20
0
Ratio SMAPE
1.7
1.5
1.3
1.1
0.9
Δ SMAPE
6
4
2
0
Ratio RMRPE
1.75
1.50
1.25
1.00
Δ RMRPE
15
10
5
0

Figure SI 2.4: Abundance metrics and the comparison of performance between environmental models and latent models measured as delta metrics. Each metric is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance, and divided by the metric of the oracle model to give the ratio metric. The environment panel represents models containing only environmental variables, while the latent panel depicts models containing latent predictors. The models are then ordered from bottom to top, from the fewest to the greatest number of environmental variables included and sorted by coefficients relative to each environmental variable. See Methods for more information, note that the "mid" model refers to the "intermediate" model. The delta metric was measured as the metric of the model with environmental variables only minus the metric of the model with the same combination of environmental and latent predictors. A positive value indicates that the model with latent predicts the abundance of the species better than the model containing only environmental variables.

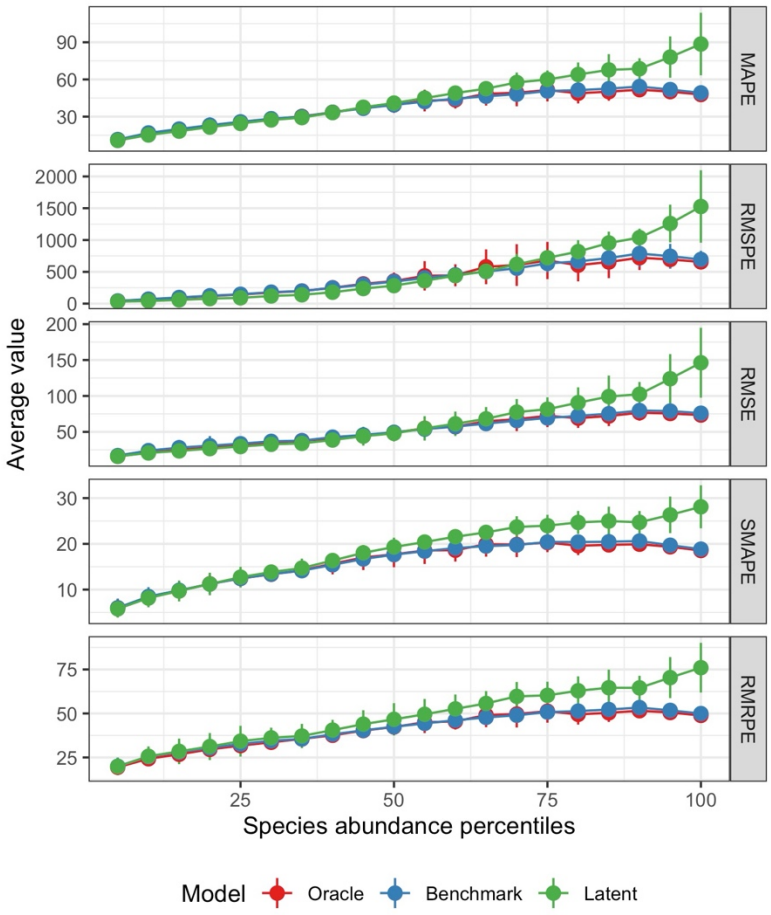

Figure SI 2.5: Correlation between the metrics studied (MAPE, RMSPE, RMSE, SMAPE, and RMRPE) depending on the model across species abundance percentiles. The vertical panels indicate the different metrics, with models represented in different colors. Each metric is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance. The oracle model refers to the model using the true environmental coefficients while the other models were fitted using all environmental variables (benchmark) or latent variables (latent).

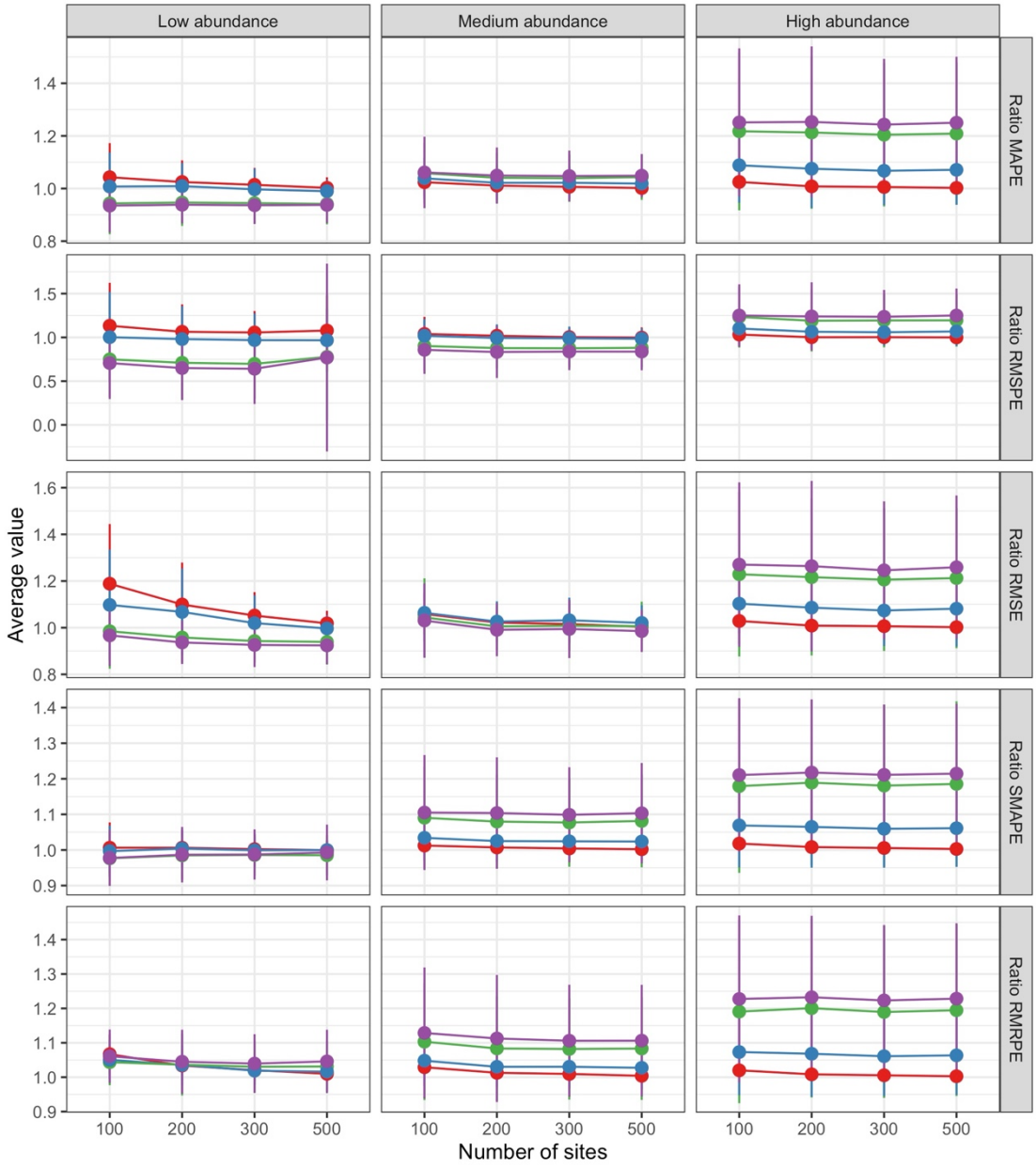

Figure SI 2.6: Average value of the studied metrics (MAPE, RMSPE, RMSE, SMAPE, and RMRPE) depending on the number of sites used to fit the models, the model used, and the abundance of species. Horizontal panels represent the different abundances: species with low, medium and high occurrence corresponding respectively to bins of 15, 50, and 80 percentiles of occurrence. Vertical panels indicate the metrics considered, with the models represented in different colors. Each metric is averaged across all landscapes and replicates per model and species, with the species binned by percentile of abundance and divided by the metric of the oracle model to give the ratio metric. The benchmark model refers to the model containing all environmental variables, 2V.high the model with the two environmental variables with the highest coefficients, 1V.high the model with the environmental variable with the highest coefficient, and Latent the model containing the latent variables.

Chapter 3: Advancing single species abundance models: leveraging multi-species data to uncover lake-specific patterns for improved fisheries predictions[2]

> "*Deep in the human unconscious is a pervasive need for a logical universe that makes sense. But the real universe is always one step beyond logic.*"
>
> Frank Herbert, *Dune*

3.1. Abstract

Predicting species abundance is critical for understanding ecological dynamics and informing conservation and management strategies. Traditional species abundance models (SAMs) often rely on environmental variables and the presence or absence of key species to predict abundance. However, these models may overlook the broader community context and cannot account for unmeasured environmental variation. Community composition at a location can serve as a proxy for both the effects of unobserved environmental variables and biotic interactions on the abundance of a focal species. In this study, we tested whether incorporating community composition information improved the ability of SAMs to predict the observed abundance of sport fish in a landscape-scale lake dataset. We used a recently developed modelling framework that uses latent variables derived from community composition as proxies for unmeasured environmental factors. We assessed the impact of varying the number of latent variables and the subset of the community used for constructing latents on the prediction accuracy of the models. We also examined whether lakes contributed similarly across species, attempted to identify specific types of lakes that significantly influence predictive ability, and evaluated whether including lakes where the species

[2] We plan to submit this chapter to the Canadian Journal of Fisheries and Aquatic Sciences (CJFAS) for publication.

was absent increased predictive error. We found that low abundance species were better predicted by models based solely on environmental variables, while high abundance species were better predicted by models incorporating latent composition variables. Additionally, we observed that lake contributions to predictive models were correlated within species groups based on their occurrence levels, with low abundance species showing a different pattern of lake contributions compared to high abundance species. Importantly, these patterns were not related to the lakes' environmental or ecological distinctiveness or any specific environmental variables. Finally, we identified that the best latent model for predicting sport fish abundance varied by species, with no clear pattern correlating with trophic level, occurrence, or abundance. These findings emphasize the importance of tailoring predictive models to specific species and contexts, recognizing the complex interplay between species abundance, environmental variables, and community composition.

3.2. Introduction

Species abundance serves as a key indicator of population health and viability within ecosystems. It offers critical insights into a species' vulnerability to local extinction, detectability, and potential impact on their local communities, thereby guiding conservation strategies and promoting sustainable management practices. Understanding distributional patterns of species abundance is essential for assessing whether local populations or a species within a large region are at risk and require conservation efforts, or if they can be sustainably harvested without compromising their long-term viability (Degnbol & Jarre 2004). Such information is invaluable for policymakers, conservationists, and resource managers responsible for balancing ecological integrity with societal demands. However, accurately estimating species abundance presents considerable challenges. Data collection is often resource-intensive, requiring extensive fieldwork, all of which can be costly

and time-consuming (Dickinson *et al.* 2010; Lindenmayer & Likens 2010; Yoccoz *et al.* 2001). Ethical concerns also arise, as many methods for estimating fish abundance involve some degree of fish handling.

Sampling constraints limit the frequency and spatial coverage of abundance assessments (e.g., across multiple lakes, streams or watersheds), making it challenging to generate comprehensive data across large geographic areas, over extended time periods (Jackson & Harvey 1997), and for multiple species. These constraints are especially challenging when rapid conservation or management actions are required at the species or lake level. To mitigate these limitations, fisheries researchers often reduce sampling efforts (e.g., number of waterbodies) and develop predictive models to estimate abundance across broader regions, notably using Species Abundance Models (SAMs, Waldock et al. 2022). Many conventional SAMs incorporate local and regional environmental variables as proxies to estimate abundance (Boyce *et al.* 2016; Brosse *et al.* 1999; Lek *et al.* 1996; Sobrino *et al.* 2020; VanDerWal *et al.* 2009). Variables such as temperature, habitat quality, and substrate, among many others, are typically relatively easy to measure and can serve as predictors for species abundances in space and time. While these models can offer useful abundance estimates, they often lack the precision and accuracy needed for fine-scale management decisions and may not fully account for complex biotic interactions, such as competition and predation, that also influence abundance distributions (Gaston 2003; Mack *et al.* 2000; MacKenzie *et al.* 2002). Consequently, there is a continuing need to refine predictive models by integrating new data sources and quantitative frameworks to better capture the multifaceted drivers of species abundances.

Stahl et al. (2024) proposed a framework that advances species abundance prediction by integrating both environmental variables and co-occurrence data. While prior SAMs have incorporated

presence-absence data as predictors, these models typically only included the presence of 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). Stahl et al.'s (2024) framework addresses this limitation by incorporating presence-absence data for the entire local community as predictors of local abundance of a target species, thus offering a more comprehensive perspective. This modelling framework offers at least two advantages over traditional approaches. First, by incorporating patterns of co-occurrence across multiple species, it can use these patterns as proxies for unmeasured environmental predictors. Second, it allows integrating species interaction networks at both local and regional scales, serving as predictors of abundance variation for a given target species. The framework employs Gaussian copulas to generate latent variables from species covariation, enabling the identification and characterization of more complex covariation patterns within multispecies data (Popovic *et al.* 2018). Latent variables were initially introduced to address the challenge of describing high dimensional species patterns and (e.g., indirect gradient analysis) and, more recently, to represent unobservable factors inferred from species covariation in ecological models (e.g., species interactions, missing environmental predictors; see Walker & Jackson 2011). Latent factors are a small set of variables that have been estimated from co-occurrence data to capture as much variation in community composition as possible; as such, if local community composition is primarily shaped by species responses to large-scale environmental gradients and local inter-specific interactions, then these latent variables can serve as proxies for these missing factors when included in abundance models. Stahl et al. (2024) demonstrated that copula-based latent variables serve as robust proxies for unmeasured environmental variables, improving abundance predictions in simulations where local species abundances were modelled as a linear function (Poisson regression) of environmental conditions and location-specific process error (i.e., without simulating species interactions and non-linear

responses to environmental gradients were not simulated). In this context, the latent abundance-predictive approach effectively identified and represented the simulated missing environmental gradients that underlined the target species' abundance distribution. This improvement in performance was consistent across various scenarios, highlighting the framework's robustness across diverse ecological contexts.

However, in real-world ecosystems, species interactions, such as competition, predation, and mutualism, often play a critical role in shaping community structure and species abundance (Chase & Leibold 2003; Tylianakis *et al.* 2008). These interactions introduce additional complexities that latent variables can uncover. By capturing both environmental influences and species interactions, latent variables offer a more comprehensive representation of the factors driving species abundances. This dual capacity to reflect environmental conditions and species interactions highlights the potential of latent variables to improve the accuracy and robustness of ecological models when applied to empirical data. Here, we apply a latent abundance-predictive approach to an empirical dataset of lake fishes, providing a valuable opportunity to evaluate the model's accuracy and assess the ability of latent variables without the potential confounding effects of both species' interactions and dispersal. Lake fish communities, being more isolated systems compared to riverine and terrestrial systems, should be more strongly influenced by environmental factors 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.

This study aims to expand the framework introduced by Stahl et al. (2024) by testing its performance on a large empirical dataset of multiple species across almost 600 lakes and across very diverse environmental conditions. Specifically, we seek to determine whether incorporating

latent variables improves predictive accuracy of species abundance in real-world ecosystems, where interactions and habitat specificity are key factors. We focused on predicting sport fish abundances because of their important role in ecological systems (e.g., large biomass), their cultural and economic significance, and their increased vulnerability to fishing pressure. Sport fishes are key targets of resource management strategies, and applying our models in this context allows us to demonstrate the practical utility of our framework in real-world management scenarios. Three modelling scenarios were considered: one where each sport fish species is modelled using latent variables derived solely from other sport fishes, a second where each sport fish species is modelled using latent variables derived solely from non-sport fishes, and another where latent variables are based on the entire species community, including both sport and non-sport fishes.

To evaluate model performance, we developed a set of assessment tools that analyze both individual species predictions and community-level patterns. These tools will also be valuable to future users of our species abundance modelling framework. We began by evaluating which lake types significantly affect predictive performance and whether these lakes exhibit rare or common environmental conditions and/or species compositions. This approach should provide valuable insights into the generalizability of our models across diverse multispecies ecological contexts. Additionally, we analysed shared patterns in species' predictive errors, as correlated errors may indicate these species are shaped by similar interactions and habitat use – key considerations for developing conservation strategies that incorporate community dynamics. Finally, we evaluated whether sport fish abundances are better predicted by models that include all lakes or only those where the species is present, addressing the trade-off between model generality and specificity. This comprehensive assessment approach allows us to evaluate the framework's robustness in capturing the complexities of lake ecosystems, identify areas for further refinement in predictive

modelling, and provide a roadmap for future applications by modellers and fisheries managers. Notably, our proposed modelling and assessment frameworks are flexible and can be readily adapted to various modelling approaches.

3.3. Materials and methods

3.3.1. Dataset

Fish abundance was collected in 707 lakes by the Ontario Broadscale Monitoring Program (Lester *et al.* 2021; Sandstrom *et al.* 2011) 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

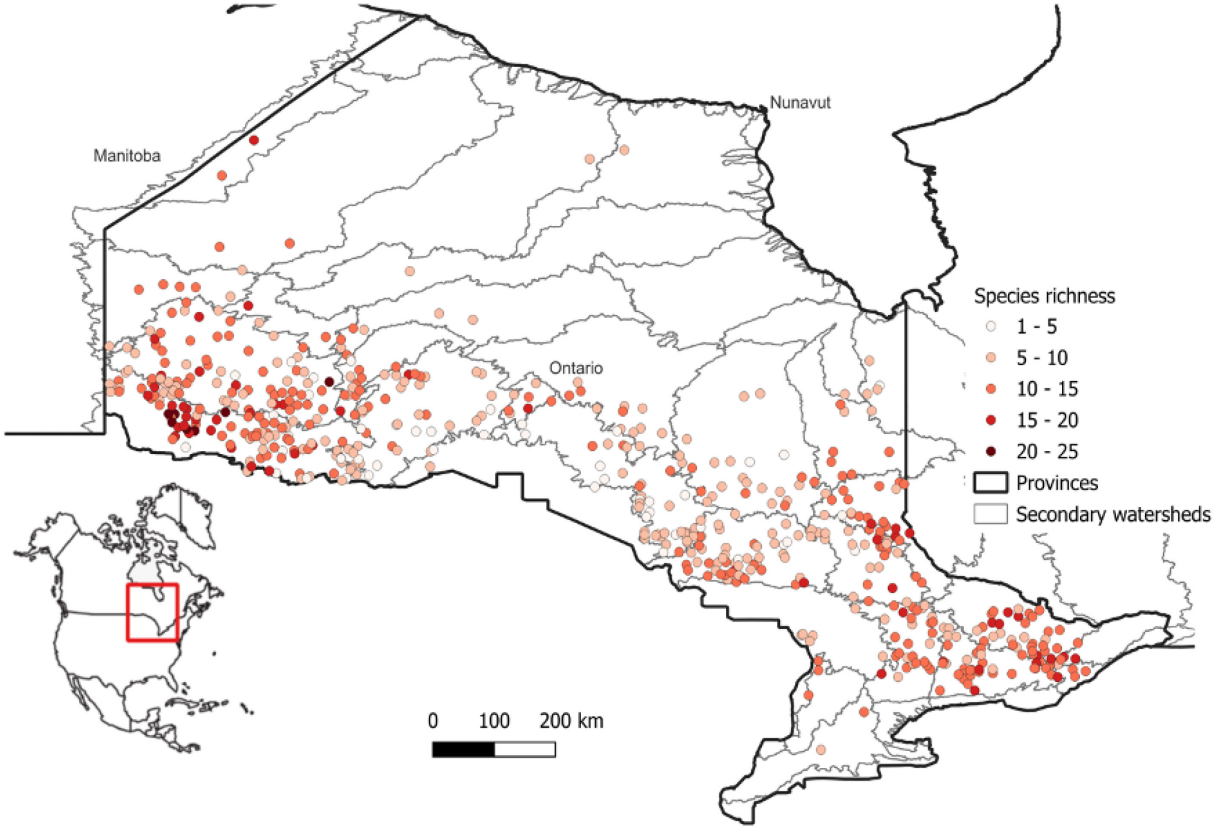

Figure 3.1: Map of the 594 lakes in Ontario, Canada, included in our models. Each point is color-coded to represent the number of species present in the lake (i.e., species richness). Black lines delineate the provincial political boundaries, while grey lines delineate the secondary watersheds (Ontario Ministry of Natural Resources and Forestry - Provincial Mapping Unit 2024).

during the summers (June to September) from 2008 to 2012. The 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 3.1, 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, a number of small mesh gillnets (stretch mesh size between 13 and 38 mm) and a number of large mesh gillnets (stretch mesh size 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 18-hour period, with the number of species per lake ranging 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. To streamline the analysis and reduce computational time, 39 species that occurred in fewer than 10 lakes were excluded, as they had minimal impact on the abundance of the remaining 48 species (i.e., present in less than 2% of our dataset, McGarigal et al. 2000). For abundance predictions, we focused on 14 sport fish species, which are critical for management purposes (see Table 3.1 and Figure S3.2; selection of sport fish species was made following personal correspondence with Dr. Dylan Fraser, Concordia University,

Montreal, Canada). After applying these filters, we retained 34 non-sport fish species, resulting in a final dataset of 48 species across 594 lakes (Figure 3.1).

Table 3.1: List of species included in the dataset, with both common and scientific 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

Category	Common name	Scientific name	Incidence (%)
	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

3.3.2. 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 S3.2). 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 varimax 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 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 a calibration and validation sets, 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. To identify the optimal number of PC 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 III 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 S3.3 and Figure S3.3).

3.3.3. 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 by first producing a stacked species regression model with a binomial family, followed by a model-based ordination with Gaussian copulas using the functions *stackedsdm* and *cord* from the R package *ecoCopula* (Popovic et al., 2019, version 1.0-2). This method was selected due to its robustness for binomial data and computational speed (Popovic *et al.* 2022). The stacked species regression model is fitted as a null model specifically to generate Dunn-Smyth residuals (Dunn & Smyth 1996). These residuals, which approximate standard normal residuals, are particularly useful for models with non-normal data, such as Generalized Linear Models (GLMs). They are well-suited for non-Gaussian responses, including binary, count, and Poisson-distributed data. The Gaussian copula model is then fitted on these residuals. To account for bias due to lake size, we included the log-transformed area of each lake as a predictor 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 sport fishes. 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 interest in these respective species. 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. 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 III 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 S3.4).

3.3.4. 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.

3.3.5. 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 using Tweedie distribution (Tweedie 1984) with a log-link function within a Generalized Additive Model (GAM) framework, using the functions *tw* and *gam* from the R package *mgcv* (Wood 2004, 2017, version 1.9-1). Each predictive variable was modelled with a 2^{nd} 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. Our focus was on predicting the abundance of 14 economically important species, commonly referred as sport fish (Table 3.1).

3.3.6. 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 of predicted abundance as a measure of the bias of model prediction (Eq. 3.1).

$$LE_{s,m,l} = log_{10}\left(\frac{\hat{Y}_{s,m,l}}{Y_{s,m,l}}\right) \quad \text{Equation 3.1}$$

where s, m, l are indexes for individual species, model, and individual lakes. Y refers to the observed abundance and $\hat{Y}$ to the predicted abundance.

This metric assesses whether the model overestimated or underestimated the species' abundance in that lake. A positive log error quantitatively indicates the model overpredicting abundance, whereas a negative log error reflects an underprediction. 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 for a given observation (species in a lake) is defined as the $\log_{10}$ of the ratio of predicted abundance to observed abundance.

The log error measures the relative magnitude of the difference between predictions and observations, rather than the absolute difference between the two. Again, the log error 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.

3.3.7. 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. 3.1) of the model that incorporated latent variables (Eq. 3.2).

$$\Delta LE_{s,m} = Q2\left(\left|LE_{s,l,m_0}\right|\right) - Q2\left(\left|LE_{s,l,m}\right|\right) \quad \text{Equation 3.2}$$

where *s*, *m*, *l* are indexes for individual species, models, and individual lakes. *Q2* 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 with 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. 3.1) 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 lake as its Local Contribution to Beta Diversity (LCBD, Whittaker 1960). 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. 3.3). 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.

$$Contribution_{l,s} = \mathrm{Q2}_{l \in C_j}\left(\left|LE_{s,j}\right|\right) - \mathrm{Q2}_{l \in V_j}\left(\left|LE_{s,j}\right|\right) \qquad \text{Equation 3.3}$$

where *l, s*, *j* are indexes for individual lakes, individual species, and replicates. The median (referred to as Q2 in Eq 3.3) $LE_{j,s}$ was calculated for the lakes in the validation set for species *s* in replicate *j*. V_j in Eq. 3.3 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 (Q2) log error closest to zero.

Unlike the Euclidean distance, the Mahalanobis distance accounts for correlations between variables (De Maesschalck *et al.* 2000; Mahalanobis 1936). The pairwise Mahalanobis distance between lakes was calculated over the first 62 axes of a PCA based on the 64 environmental variables. We applied Principal Component Analysis (PCA) instead of using the original variables because their correlation structure exhibited rank deficiency: 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 & 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. 3.3). 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, highlighting the importance of accounting for species-specific ecological requirements and interactions when modelling species abundance across landscapes. This method is essential for determining whether a lake's contribution to model performance for one species can be generalized to other species which is key for developing robust and transferable ecological models. 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 (Legendre 1993; Legendre & Fortin 1989).

(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. 3.1) across replicates, and then 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. 3.4).

$$\Delta SLE_{s,m} = \left| \frac{1}{M} \sum_{l \in M} LE_{l,s} \right| - \left| \frac{1}{L_s} \sum_{l \in L_s} LE_{l,s} \right| \quad \text{Equation 3.4}$$

where *s*, *m*, *l, M, L_s,* are indexes for individual species, models, individual lakes, all lakes of the dataset, and lakes where species *s* is present.

A positive ΔSLE indicates that the model using only lakes where the species is present perform better, while a negative value suggests that the model fitted with all lakes performs better.

3.4. Results

Our first goal was to identify whether including latent variables improved our predictions by evaluating the ΔLE between the environmental model and the latent models. Not all species benefitted from the inclusion of latent variables (Figure 3.2). Importantly, the method used to generate these latent variables did not affect the direction of the ΔLE values; this consistency indicates that including latent variables, regardless of the method employed, produced the same overall effect on predictive ability, whether that be an improvement or decline compared to the environmental model. A distinct trend emerged: species with low occurrences were predicted more accurately by the environmental model, whereas those with higher occurrences were better predicted by models incorporating latent variables. We then evaluated the impact of different species groups on predictive performance by comparing the effectiveness of latent models, contrasting those based on sport fish species, non-sport fish species, or all fish species combined. Our analysis revealed that the best-performing model varied by species, although the differences in LE densities across models were relatively modest, indicating that the variation in predictive accuracy between models was not necessarily substantial (Figure 3.3, Table S3.4). Cisco, lake whitefish, largemouth bass, northern pike, and smallmouth bass were best predicted by the model

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.

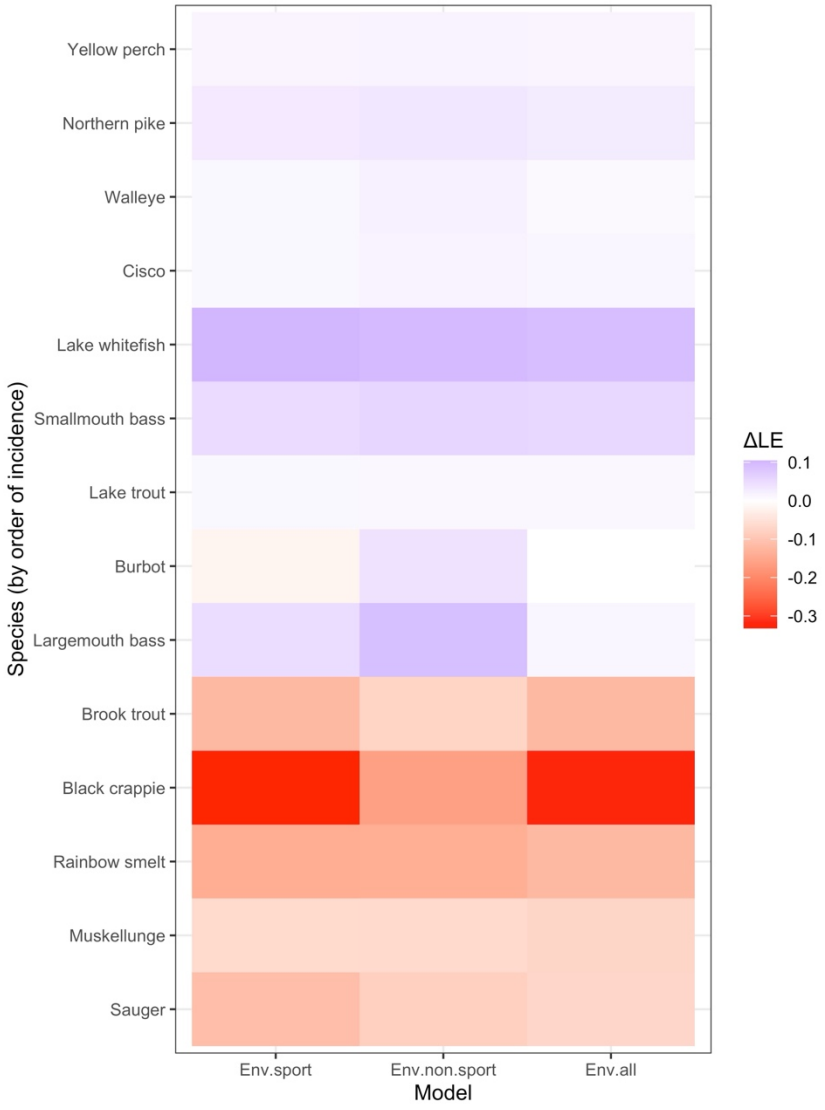

Figure 3.2: ΔLE as a function of model and species. The ΔLE was calculated as the median absolute log error of the model with only environmental variables, minus the median absolute log error of the model incorporating latent predictors (Eq. 3.2). Positive values (in blue) indicate that the model with 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.

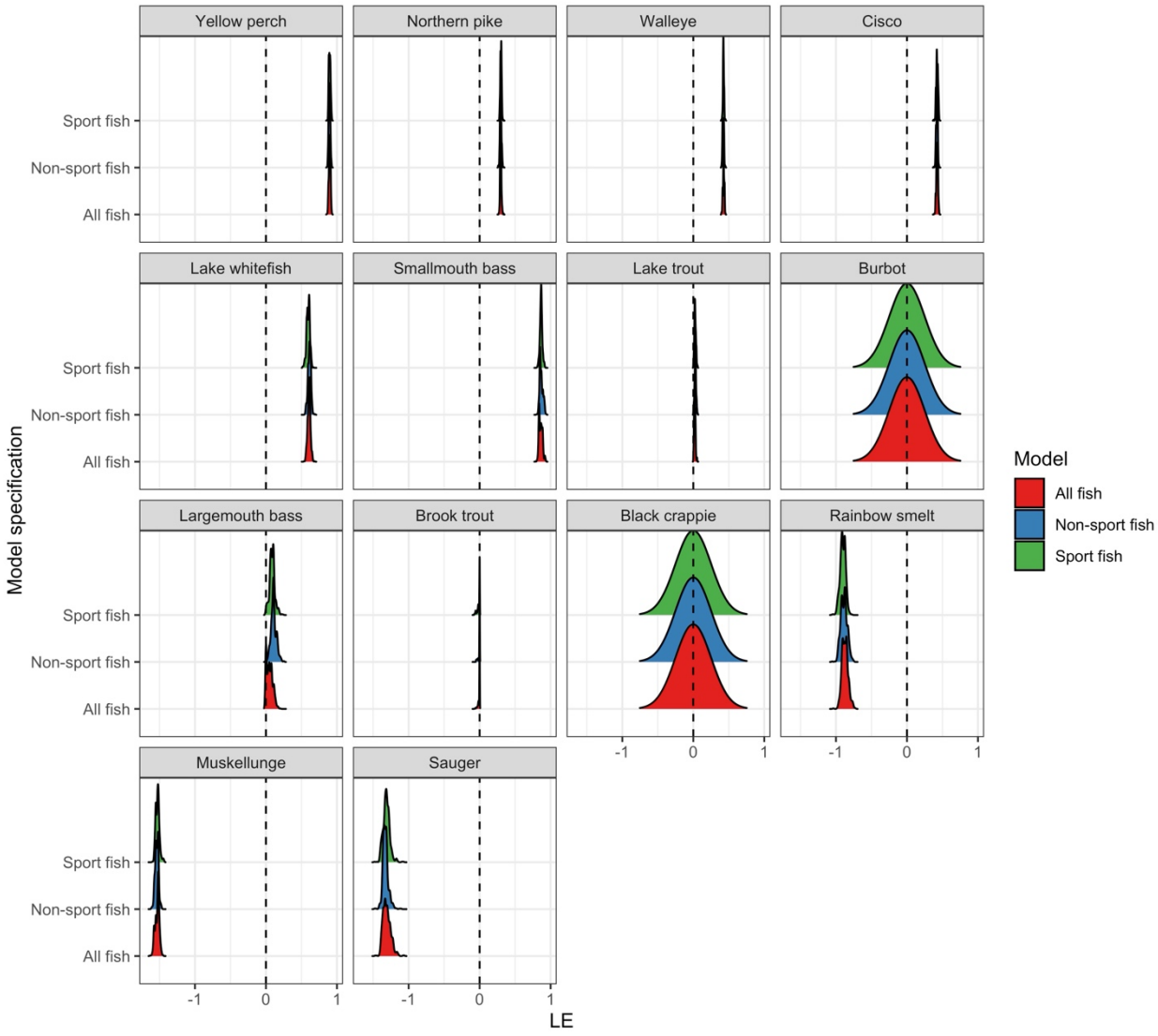

Figure 3.3: Density plot of the log error as a function of species and model. The log error was calculated following Eq. 3.1, and for each lake, the median log error was taken across replicates for each species and model. Latent variables were generated using three groups: (1) sport fish species (green), (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.

Next, we focused on identifying which type of lakes influenced predictive ability by analysing their contribution to LE and determining whether lakes that affected predictive ability were rare or common in terms of their environmental characteristics and/or community composition (Figure

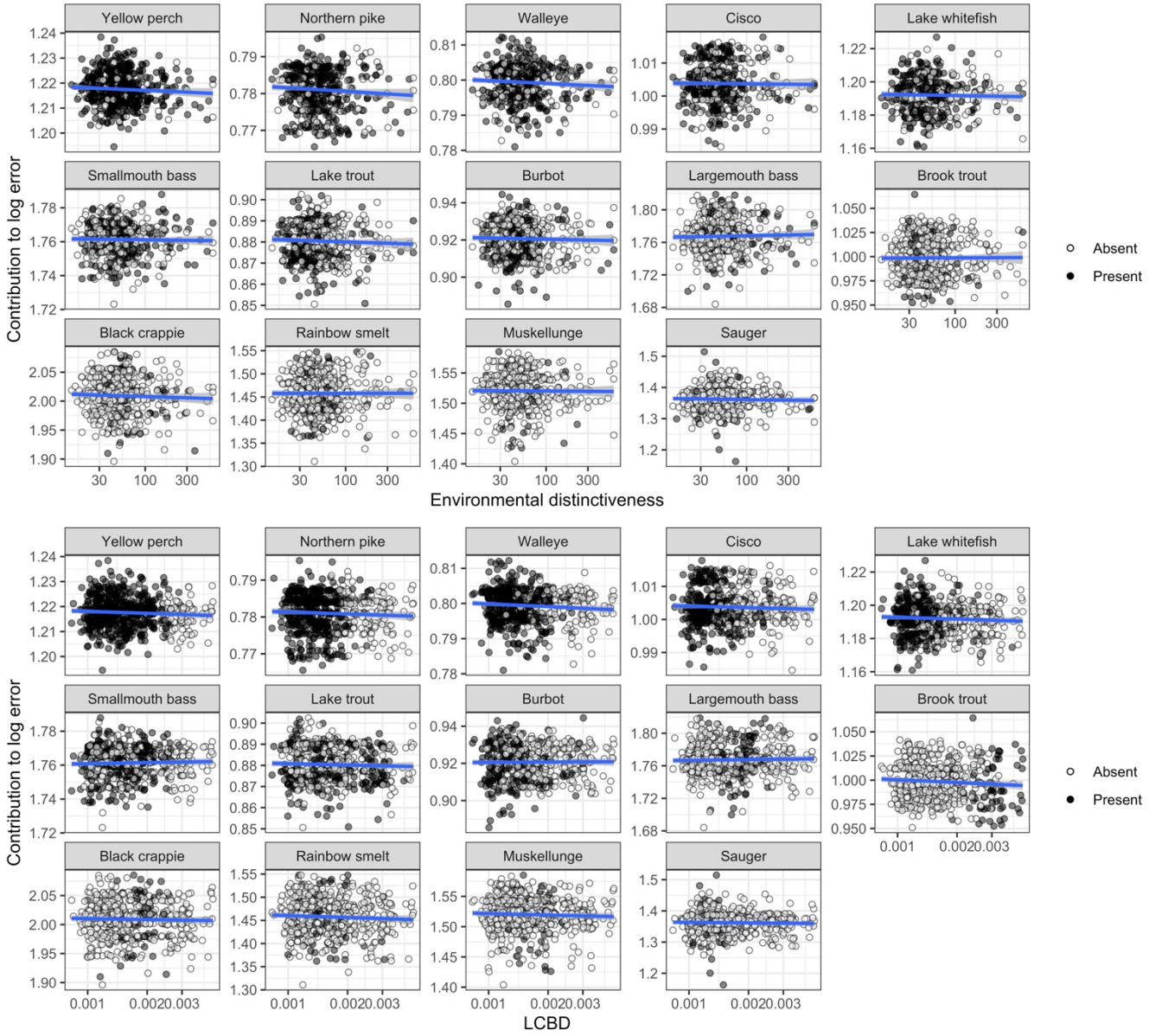

Figure 3.4: Contribution of each lake to the log error as a function of environmental distinctiveness and Local Contribution to Beta Diversity (LCBD) per species (see methods how these values were calculated). The lake's contribution was measured as the median across replicates of the difference between 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. 3.3). 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.

3.4). The LE metric showed no correlation with how rare or common a lake was in terms of its environmental characteristics (Mahalanobis distance) or its species composition (LCBD). To further identify types of lakes that 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 S3.5). 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 S3.6). 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 3.2). In contrast, species in the second group exhibited idiosyncratic responses, with no significant correlations either within or between groups. Additionally, the species groups 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.

Table 3.2: 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 S3.6). 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

Finally, we examined whether sport fish abundances 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 3.5). 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.

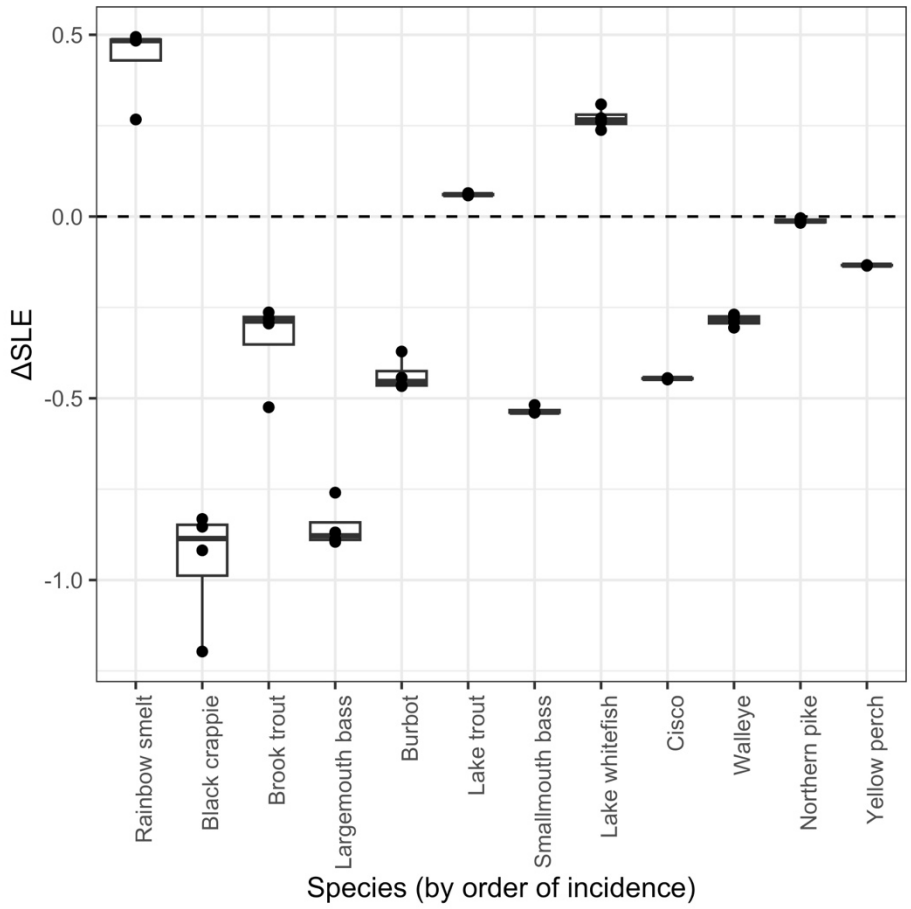

Figure 3.5: Boxplot of the ΔSLE per species. The ΔSLE is calculated as the absolute mean log error fitted using all lakes minus the absolute mean log error of the model fitted using only where the species is present (Eq. 3.4). A positive ΔSLE indicates better performance when using the reduced lake pool, 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.

3.5. Discussion

Our first goal was to assess whether a latent-abundance model, 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 show that while latent variables improved predictions for high-occurrence species, they did not consistently improve predictive ability across all species, particularly for low occurrence species. 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.

These above findings are consistent with the original framework assessment on simulated data, which also found better prediction for higher occurrence species when latent variables were included. They also align with the broader literature, which suggest that low-occurrence species are generally more vulnerable to stochastic environmental fluctuations and demographic instability (Brown *et al.* 1995; Gaston 1994), while high-occurrence species tend to engage in more complex biotic interactions (Araújo & Luoto 2007; Mouquet *et al.* 2003). 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 particular 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 significantly affect 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 power is not influenced on the specific environmental attributes or species compositions of the lakes. 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 could be seen as less predictable due to the greater likelihood of them containing more microhabitats and local environmental variation (Strayer & Findlay 2010). 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 correlation 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 & Guisan 2006; Legendre & Legendre 2012; Ovaskainen *et al.* 2010). These results underscore the complexity of ecosystem dynamics and the need for sophisticated models that account for diverse 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 - emphasizes the need for nuanced approaches that consider the unique ecological contexts

of each species. This complexity presents a challenge when developing broad conservation and management strategies for multiple species, where balancing species-specific needs with general models for multiple species may be essential.

Our study reveals a series of interconnected patterns across 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 (Brown *et al.* 1995; Gaston 1994), high abundance species are likely responding to more complex, community-level interactions that are better captured by latent variables (Araújo & Luoto 2007; Mouquet *et al.* 2003). 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 its reliance on a dataset from lake ecosystems, where dispersal is relatively constrained. 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). 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 latent variables and co-occurrence data into predictive models for species abundance, particularly in lake ecosystems. We found that low abundance species were better predicted by environmental models, while high abundance species benefited more from models incorporating latent variables. Additionally, lake contributions to predictive accuracy were correlated with species occurrence patterns, suggesting distinct ecological processes at play for low- and high-abundance species. However, these contributions were not linked to environmental or community distinctiveness, suggesting that other, yet unidentified factors may be influencing variation among lakes in predictive performance. Alternatively, this could indicate that our modelling framework is robust to variation among lakes. 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. Future research should

aim to identify the specific factors driving the observed patterns of lake contributions and further exploring the role of latent variables in capturing species interactions. Additionally, grouping species based on other ecological characteristics, such as trophic levels or life history traits, could offer deeper insights into the underlying mechanisms governing species distributions and abundance in aquatic ecosystems.

3.6. Supplementary Information

3.6.1. 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 2^{nd} 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 & Wood 2011). This procedure was repeated 100 times and for six species with different occurrence rates representative of the whole dataset (Table S3.1). The out-of-sample average prediction was calculated across replicates, and the median across species of the Mean Squared Error (MSE) was derived.

Table SI 3.1: 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 S3.1). 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 & 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.

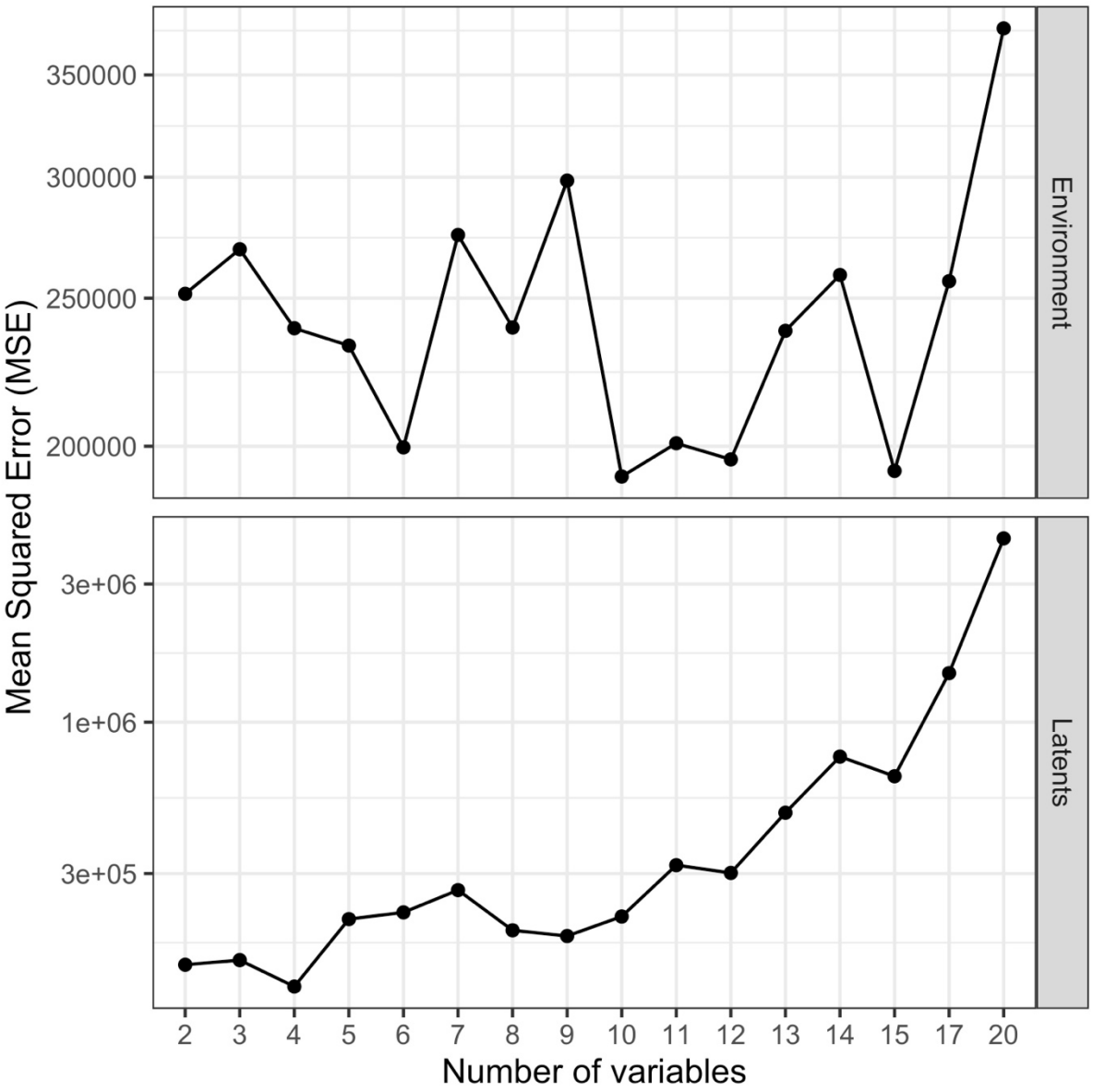

Figure SI 3.1: Median Mean Squared Error (MSE) as a function of number of composite environmental and latent variables. The figure shows the median Mean Squared Error (MSE); with the MSE calculated for out-of-sample abundance predictions across replicates and the median calculated across species. The 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.

3.6.2. Extended results

Table SI 3.2: 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)
	Maximum water temperature (°C)

Category	Environmental variable
	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 SI 3.3: 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 SI 3.4: 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

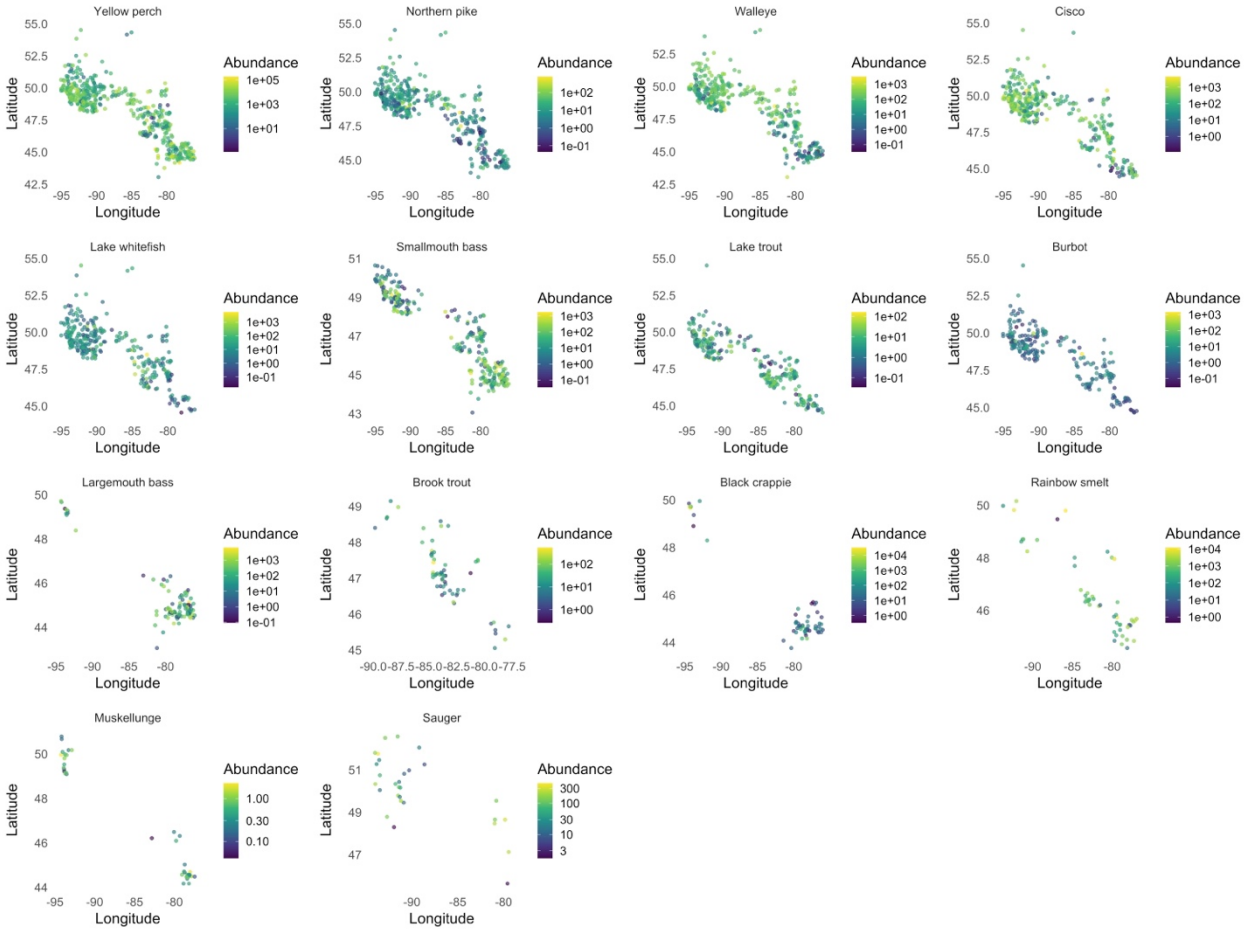

Figure SI 3.2: Maps showing the abundance distribution of each sport fish species. Species are organized by incidence within the dataset, with the most common species at the top and the least common at the bottom. Each point represents a lake where the species was observed. Abundance values are represented on a $\log_{10}$ scale, providing a clearer depiction of the wide range of abundance levels across the lakes.

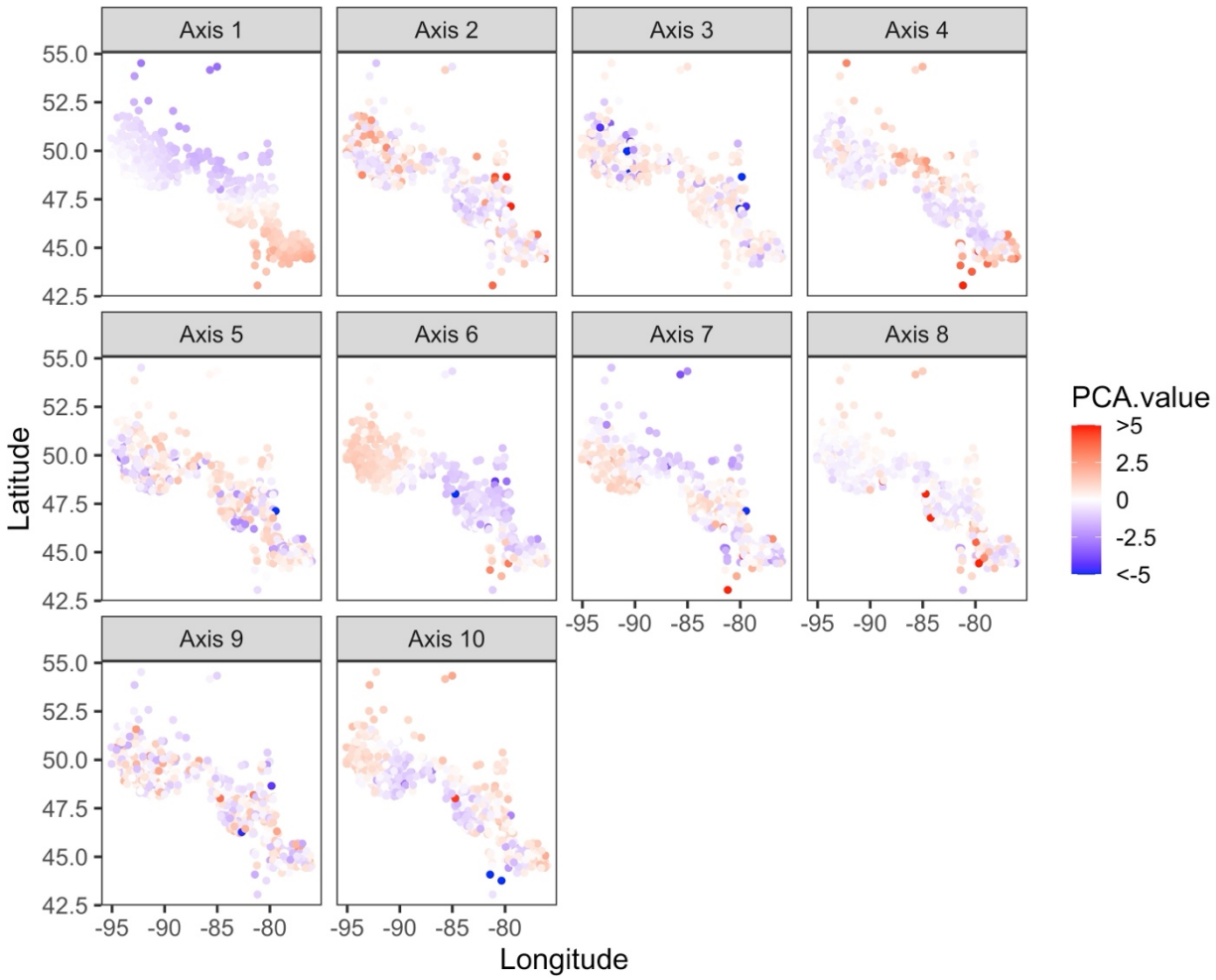

Figure SI 3.3: Maps illustrating the spatial patterns for the first 10 axes of the Principal Component Analysis (PCA) conducted on 64 environmental variables. These axes capture the major gradients in environmental variation across the study area, with each map representing one of the top 10 PCA axes.

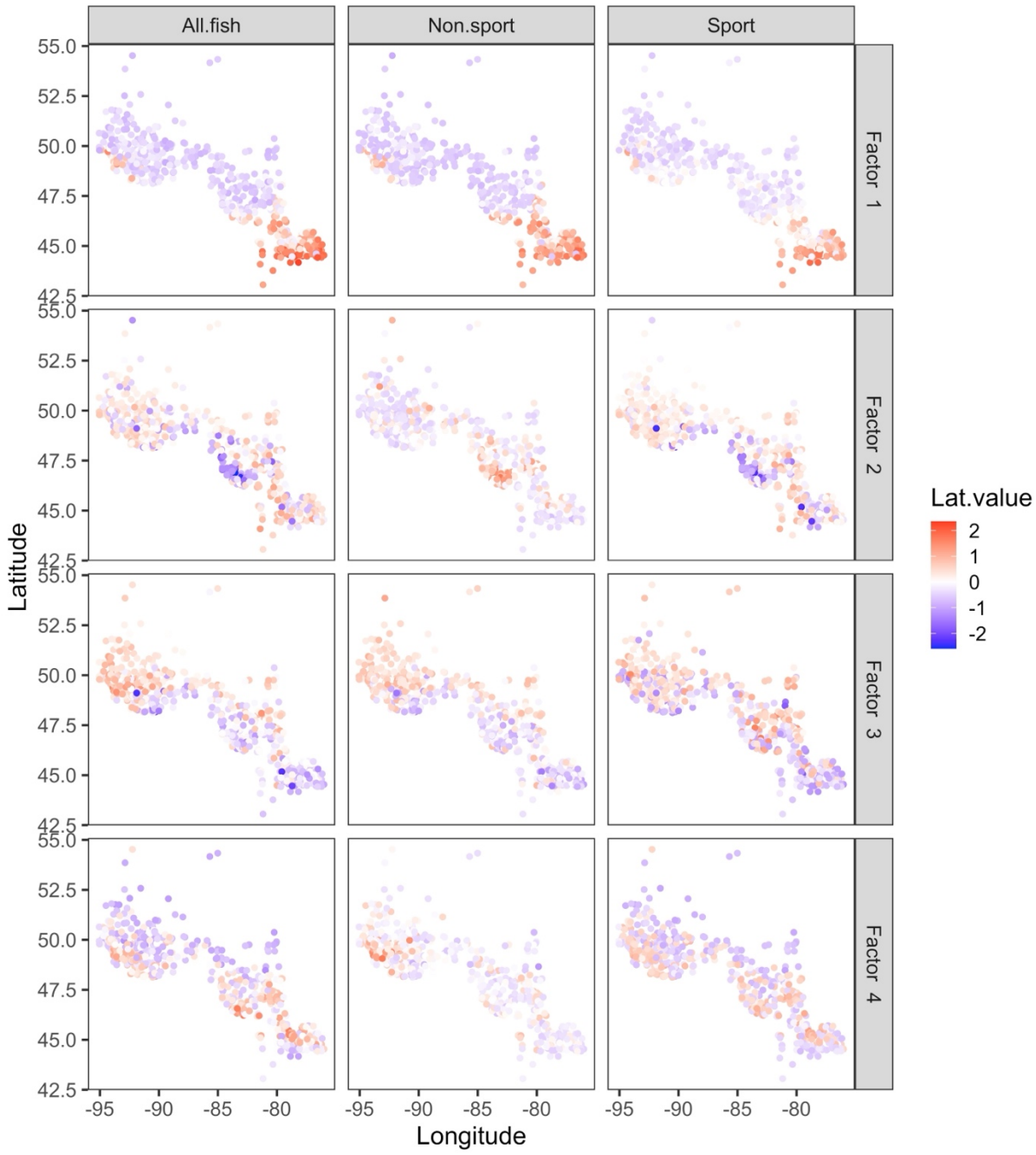

Figure SI 3.4: Maps showing the spatial distribution of latent variables derived from three different fish assemblages. We generated the latent variables using (1) sport fish species, labeled as 'Sport,' (2) non-sport fish species, labeled as 'Non.sport,' and (3) all fish species, labeled as 'All.fish.' 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.

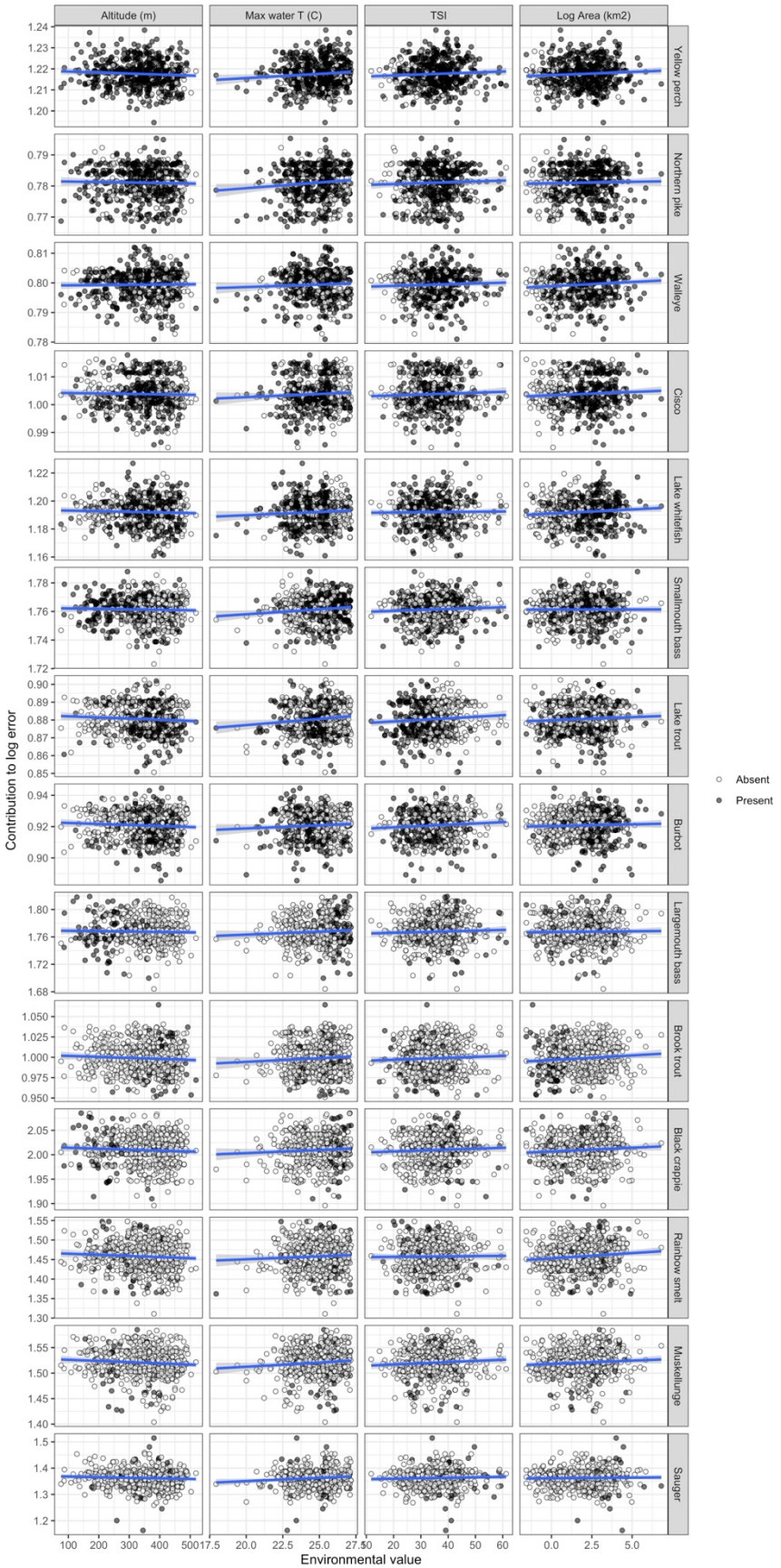
Altitude (m)
Max water T (C)
TSI
Log Area (km2)
Yellow perch
Northern pike
Walleye
Cisco
Lake whitefish
Smallmouth bass
Lake trout
Burbot
Largemouth bass
Brook trout
Black crappie
Rainbow smelt
Muskellunge
Sauger
Contribution to log error
Environmental value
Absent
Present

Figure SI 3.5: 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

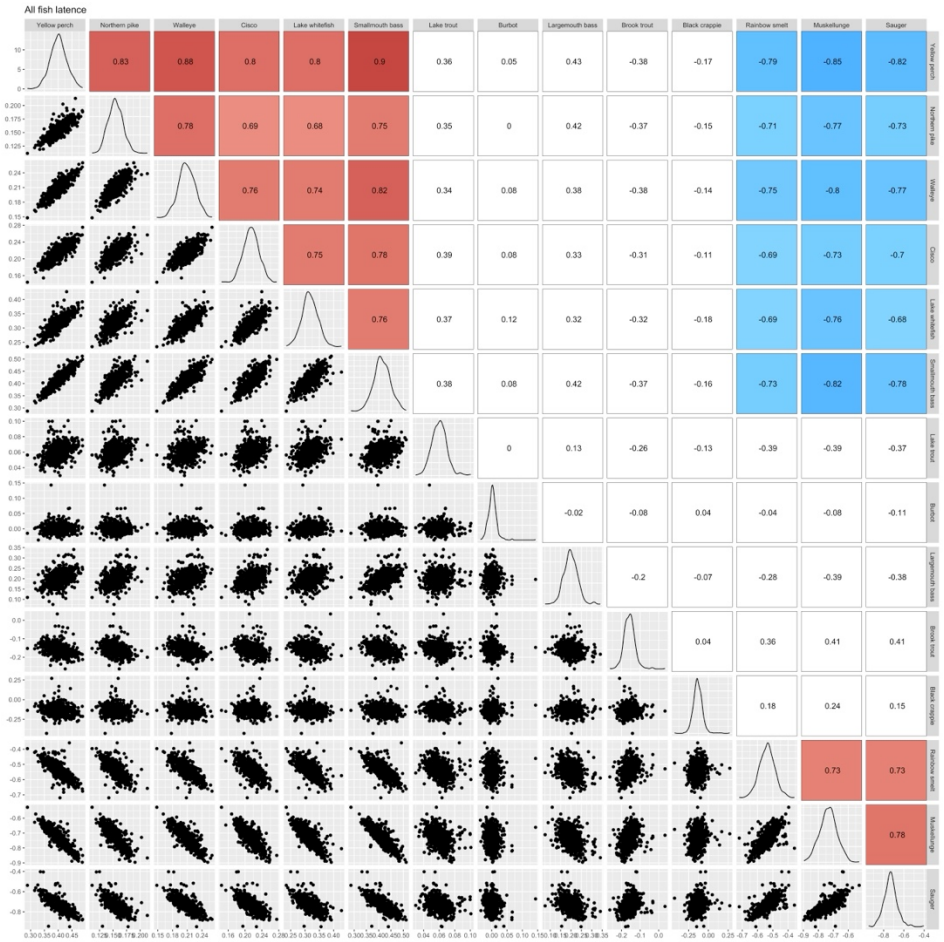

Figure SI 3.6: Correlation of lake contributions between species for model containing latent variables generated from all fish species. The patterns observed allowed us to group species in the following manner: (Group 1) rainbow smelt, muskellunge, and sauger; (Group 2) burbot, lake trout, black crappie, 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.

Chapter 4: Comparative assessment of community and stacked abundance models for predicting biomass across spatial scales[3]

"*Even worthless things can become valuable once they become rare. This is the grand lesson of my life.*"

Micaiah Johnson, *The Space Between Worlds*

4.1. Abstract

Accurately predicting biomass in aquatic ecosystems is essential for advancing fisheries management and biodiversity conservation. Biomass models provide critical insights into ecosystem productivity, trophic dynamics, and energy flow – key components for sustaining ecosystem resilience under environmental changes. This study evaluates two common modelling approaches - stacked abundance models and community models - for predicting fish community biomass in lake ecosystems. The stacked model estimates biomass by aggregating species-specific abundance predictions, while the community model directly predicts biomass at the community level, incorporating environmental variables and latent variables derived from species co-occurrence patterns. We assess the predictive performance of both approaches across gradients of species richness, biomass, and latitude. Our results show that the stacked model tends to overestimate biomass, particularly in lakes dominated by a few highly abundant species. In contrast, the community model yields narrower error distributions, indicating improved predictive accuracy by capturing community-level processes. Additionally, predictive accuracy for both models varies with species richness, with more diverse and evenly distributed communities achieving better predictions. These findings highlight the strengths and limitations of each

[3] We plan to submit this chapter to Oikos for publication.

modelling approach, offering valuable insights for refining biomass predictions to support fisheries management and biodiversity monitoring efforts.

4.2. Introduction

Biomass, a fundamental metric of community health, is tightly coupled within ecosystem functioning, providing essential insights into energy flow and productivity within and among ecosystems (White *et al.* 2007). As a result, there is an increasing demand for predictive models capable of estimating biomass in response to environmental variation across ecosystems and changes within ecosystems. The demand for such models stems from their potential to guide conservation efforts, optimize resource management, and forecast the impacts of human-driven activities on biodiversity and ecosystem services (Pecl *et al.* 2017). These models enable decision-makers to implement proactive strategies for preserving biodiversity, rather than responding to population declines or species extinctions after they occur (Leung *et al.* 2020).

Despite its importance, predicting biomass at various spatial scales remains a significant challenge in ecology. Traditional methods of estimating biomass have primarily relied on environmental variables such as temperature, nutrient availability, and habitat structure to explain ecosystem productivity and health (Smith 1998). These variables serve as proxies for energy input, growth rates, and carrying capacity, playing a crucial role in biomass predictions, particularly in aquatic environments (Pauly & Christensen 1995; Ware & Thomson 2005). However, these models often face limitations when dealing with high-dimensional data and complex ecological interactions, leading to potential predictive inaccuracies (Leung *et al.* 2020). Additionally, ecosystem heterogeneity, temporal variability, and the dynamic nature of species interactions introduce complexities that are difficult to capture using environmental variables alone (Guisan & Thuiller 2005; McGill *et al.* 2006). A further challenge arises from the limitations of presence-absence data,

which, while valuable for predicting species distribution and abundance, may be insufficient for accurately estimating biomass, particularly in diverse ecosystems with intricate interspecies relationships (Hébert & Gravel 2023). These challenges underscore the need for innovative approaches that integrate multiple sources of ecological data and account for the uncertainty inherent in ecological predictions (Ferrier & Guisan 2006).

A promising approach to improving biomass prediction is the application of stacked abundance models. These models first estimate the abundance of individual species based on environmental variables and presence-absence data, and subsequently aggregate these predictions to derive community biomass (Ovaskainen *et al.* 2017). This framework offers advantages over traditional community-level models by accounting for species-specific responses to environmental factors and propagating uncertainty throughout the modelling process (Hébert & Gravel 2023; Leung *et al.* 2020). However, the effectiveness of stacked abundance models is not guaranteed; inaccuracies in abundance predictions can accumulate, potentially increasing variance in biomass estimates, particularly for rare or hard-to-predict species (Warton *et al.* 2015a). Therefore, evaluating the performance of stacked abundance models relative to direct community biomass models is essential for determining which approach provides more accurate and reliable predictions (Warton *et al.* 2015a). Furthermore, given the role of species compositions in influencing biomass (Arranz *et al.* 2022), it is essential to investigate how these compositions affect the outcomes of both modelling strategies.

Beyond selecting between stacked and community models, a key consideration is whether predictive accuracy is influenced by species richness and community composition. Species diversity plays a well-established role in shaping ecosystem stability and functioning, and it may also affect the accuracy of biomass predictions (Arranz *et al.* 2022). Species-rich lakes,

characterized by more complex food webs and interactions, present greater modelling challenges compared to less diverse systems (Tunney *et al.* 2017). Accurately predicting biomass in these diverse communities may require models that incorporate species interactions and community structure, adding further complexity. Determining whether diversity enhances or hinders predictive accuracy is essential for improving model performance and informing conservation strategies in complex ecosystems.

Lastly, predictive models must demonstrate predictive accuracy across different spatial scales, ranging from local to regional levels. Ecological processes often operate at different scales, making it essential to assess whether models that perform well at the lake level can be reliably extended to broader scales, such as watersheds or regions (Cumming *et al.* 2017). Accurate biomass predictions at regional scales are particularly valuable for large-scale management and policymaking, as they can inform decisions on fisheries management, pollution control, and biodiversity conservation, among other concerns. In this study, we address three key research questions: (1) whether a stacked abundance model or a community model (i.e., one that directly predicts community biomass from environmental data and latent variables) generates better biomass predictions, (2) whether lakes with higher species richness are predicted more accurately than less diverse lakes, and (3) whether regional biomass predictions outperform (i.e., greater accuracy) local biomass predictions. Through these investigations, we aim to generate insights into the effectiveness of predictive models for supporting conservation and management efforts across spatial scales.

4.3. Materials and methods

4.3.1. Dataset

We used fish abundance and biomass data from 707 lakes, generated by the Ontario Broad-scale Monitoring Program (Lester *et al.* 2021; Sandstrom *et al.* 2011) and conducted by the Ontario

Ministry of Natural Resources and Forestry (OMNRF, 2012) in Canada (Figure 4.1 and S4.1). The sampled lakes ranged from 43° to 54° latitude and -95° to -76° longitude, with surface areas between 0.21 to 905 km^2 and maximum depths from 1.2 to 213 m. Sampling occurred during the summer months (June to September) from 2008 to 2012. Lakes were selected using a stratified random sampling design, with strata based on geographic zones and lake surface area. A depth-stratified sampling approach was implemented to ensure accurate estimates of both fish abundance and biomass (see Lester et al. 2021 and Sandstrom et al. 2011 for more details on methods). The dataset covers 25 secondary watersheds and 82 tertiary watersheds.

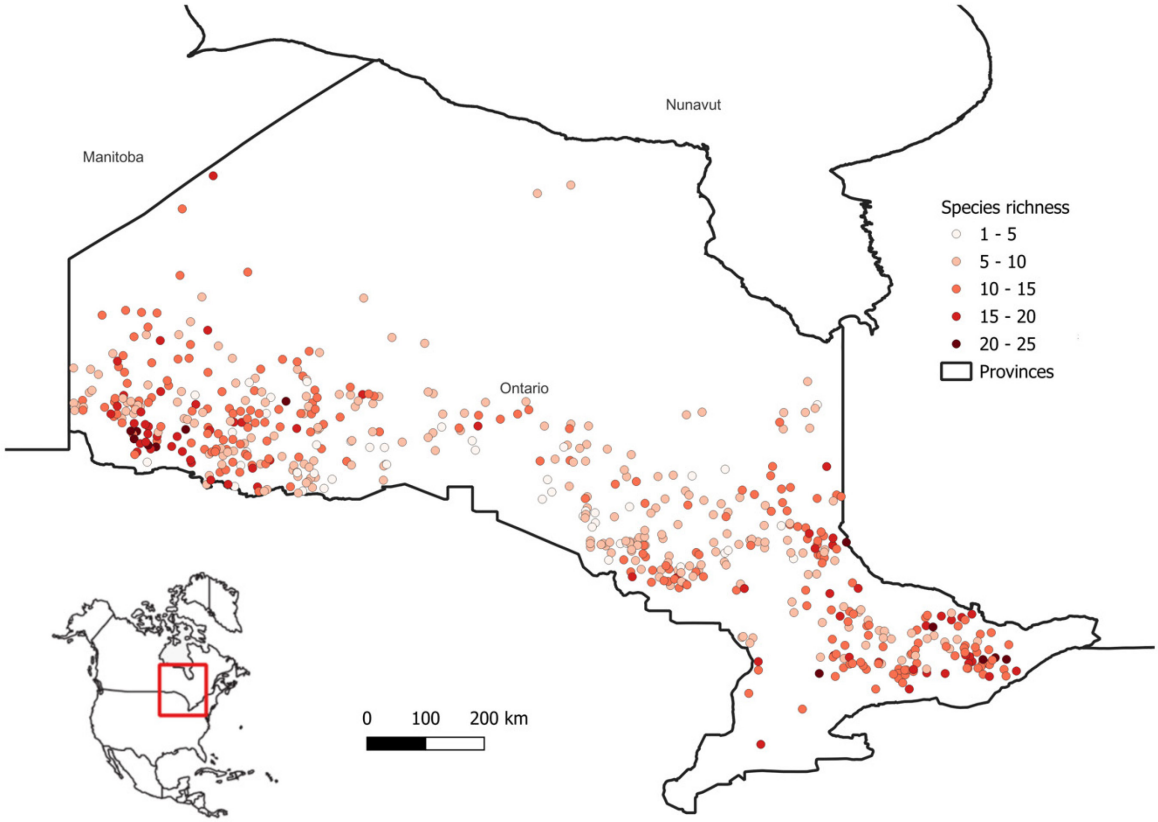

Figure 4.1: Map of the 583 lakes surveyed in Ontario, Canada. Each point is color-coded to indicate species richness (i.e., the number of species present in the lake). Black lines denote provincial boundaries within Canada.

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, two types of gillnets were deployed overnight for 18 hours: a number of small mesh gillnets with stretch mesh size ranging between 13 and 38 mm, and a number of large mesh gillnets of stretch mesh sizes ranging between 38 and 127 mm (Appelberg 2000; Arranz *et al.* 2022). The catch from these nets was used to calculate the index of fish abundance density (number of fish caught per unit effort, CPUE) and the index of fish biomass density (weight of fish caught per unit effort, BPUE) for each species in each lake. These indices reflect the expected catch per 100 meters of each net type over an 18-hour period. We assumed that CPUE was an accurate proxy for local density of each species in each lake (Olin *et al.* 2009).

The dataset originally included 87 species, with species richness per lake ranging from 2 to 25. To streamline the analysis and reduce computational demands, species occurring in fewer than six lakes (i.e., less than 2% of the dataset; McGarigal et al. 2000) were excluded. After applying this threshold, a total of 54 species across 583 lakes were retained for further analysis (Table 4.1, Figure 4.1). All fish were identified at the species level.

Table 4.1: List of species included in the dataset, with common and scientific names, along with their average weight (in kg). The average adult weight was calculated as the mean of the minimum and maximum weights reported in the Ontario Freshwater Fishes Life History Database by Eakins (version 5.31, 2024).

Common name	Scientific name	Average weight (kg)
Black crappie	*Pomoxis nigromaculatus*	0.295
Blackchin shiner	*Miniellus heterodon*	0.002
Blacknose dace	*Rhinichthys atratulus*	0.003
Blacknose shiner	*Notropis heterolepis*	0.002
Bluegill	*Lepomis macrochirus*	0.195
Bluntnose minnow	*Pimephales notatus*	0.003
Bowfin	*Amia ocellicauda*	1.950
Brook stickleback	*Culaea inconstans*	0.001
Brook trout	*Salvelinus fontinalis*	0.810
Brown bullhead	*Ameiurus nebulosus*	0.350

Common name	Scientific name	Average weight (kg)
Burbot	*Lota lota*	2.350
Central mudminnow	*Umbra limi*	0.006
Cisco	*Coregonus artedi*	0.260
Common carp	*Cyprinus carpio*	3.300
Common shiner	*Luxilus cornutus*	0.022
Creek chub	*Semotilus atromaculatus*	0.051
Deepwater sculpin	*Myoxocephalus thompsonii*	0.015
Emerald shiner	*Notropis atherinoides*	0.005
Fallfish	*Semotilus corporalis*	0.360
Fathead minnow	*Pimephales promelas*	0.003
Golden shiner	*Notemigonus crysoleucas*	0.029
Greater redhorse	*Moxostoma valenciennesi*	1.350
Iowa darter	*Etheostoma exile*	0.001
Johnny darter	*Etheostoma nigrum*	0.002
Lake chub	*Couesius plumbeus*	0.038
Lake trout	*Salvelinus namaycush*	3.650
Lake whitefish	*Coregonus clupeaformis*	2.450
Largemouth bass	*Micropterus nigricans*	1.250
Logperch	*Percina caprodes*	0.015
Longnose dace	*Rhinichthys cataractae*	0.007
Longnose sucker	*Catostomus catostomus*	0.650
Mimic shiner	*Paranotropis volucellus*	0.002
Mottled sculpin	*Cottus bairdii*	0.008
Muskellunge	*Esox masquinongy*	8.950
Ninespine stickleback	*Pungitius pungitius*	0.001
Northern pike	*Esox lucius*	3.200
Northern redbelly dace	*Chrosomus eos*	0.002
Pearl dace	*Margariscus nachtriebi*	0.009
Pumpkinseed	*Lepomis gibbosus*	0.145
Rainbow smelt	*Osmerus mordax*	0.050
Rock bass	*Ambloplites rupestris*	0.205
Round whitefish	*Prosopium cylindraceum*	0.280
Sauger	*Sander canadensis*	0.825
Shorthead redhorse	*Moxostoma macrolepidotum*	1.100
Silver redhorse	*Moxostoma anisurum*	1.350
Slimy sculpin	*Cottus cognatus*	0.005
Smallmouth bass	*Micropterus dolomieu*	0.750
Spoonhead sculpin	*Cottus ricei*	0.003
Spottail shiner	*Hudsonius hudsonius*	0.009
Trout perch	*Percopsis omiscomaycus*	0.008
Walleye	*Sander vitreus*	2.200
White sucker	*Catostomus commersonii*	0.800
Yellow bullhead	*Ameiurus natalis*	0.320
Yellow perch	*Perca flavescens*	0.160

4.3.2. Variables and transformation

At each lake, a comprehensive set of environmental variables was recorded alongside fish abundance and biomass sampling. A total of 65 environmental variables were recorded for each lake, categorized into hydro-morphology, fishing activities, productivity, climate, watershed, and water chemistry (Table S4.1, see Sandstrom et al. 2011 for details on sampling methods). To reduce redundancy, all variables were standardized to a common scale (mean = 0 and variance = 1) and processed through Principal Component Analysis (PCA) with varimax rotation to generate a smaller set of composite environmental variables (Zou *et al.* 2006). This approach streamlined the dataset while preserving essential variability. The analyses were conducted using the *prcomp* and *varimax* functions from the R package *stats* (R Core Team 2017). Since the dataset was divided into calibration and validation sets (see the *Model fitting* section for details), the PCA dimension-reduction step was first performed on the calibration set. The resulting environmental axes were then applied to project the validation set, ensuring the orthogonality of the axes was preserved as new lakes (validation) were incorporated. This method preserved the data structure while reducing dimensionality and was repeated for each validation replicate. Ultimately, 10 principal components were retained as aggregate variables for subsequent analysis (see section 3.6.1 for details); they accounted for about 18% of the total environmental variation (i.e., depending on the replicate).

We converted species presence-absence data into latent variables using a stacked species regression model, followed by a model-based ordination with Gaussian copulas (Stahl *et al.* 2024). This analysis was performed using the *stackedsdm* and *cord* functions from the R package *ecoCopula* (Popovic et al. 2019, version 1.0-2). This method was selected due to its robustness for binomial data and computational speed (Popovic *et al.* 2022). The stacked species regression model is fitted as a null model specifically to generate Dunn-Smyth residuals (Dunn & Smyth 1996). These

residuals, which approximate standard normal residuals, are particularly useful for models with non-normal data, such as Generalized Linear Models (GLMs). They are well-suited for non-Gaussian responses, including binary, count, and Poisson-distributed data. The Gaussian copula model is then fitted on these residuals. To account for bias due to lake size, we included the $\log_{10}$-transformed area of each lake as a predictor in the stacked species regression model. Latent variables were generated using all fish species included in the analysis, with the number of latent variables set to 4. To determine the optimal number of composite environmental variables (PCA axes) and latent variables, we performed an analysis in which one was held constant while the other was varied (see section 3.6.1 for details). The combination yielding the lowest out-of-sample error was selected, leading to the use of 10 composite environmental variables and four latent variables.

4.3.3. Model fitting

To address the research questions outlined in the Introduction, we followed the framework detailed below:

Latent variable generation: latent variables were derived using presence-absence data from all fish species (i.e., the ones that passed the threshold; presence in at least six lakes) across the entire dataset.

Dataset partitioning: the dataset was randomly divided into a calibration set (70%) and a validation set (30%).

Sparse PCA: sparce PCA was applied to the environmental variables in the calibration set, with the resulting axes used to project the environmental variables of the validation set. This ensured that the PCA axes remained orthogonal between the two sets.

Species abundance prediction: The calibration set was used to train models predicting the local abundance of each fish species, and the validation set was used to predict out-of-sample abundance. This process was repeated 1,000 times, and the median predicted abundance of each species and lake was calculated.

Biomass prediction: species abundances per lake were multiplied by species' average weight and summed to estimate total biomass for each lake. Data on species' weight ranges were obtained from the Ontario Freshwater Fishes Life History Database (Eakins 2024), and the average of the provided weight range was used to represent each species. Similar to the previous step, the abundance data were randomly split into calibration and a validation sets to predict out-of-sample biomass. This process was repeated 1,000 times and the median predicted biomass for each lake was calculated.

Direct community biomass model: a direct community biomass model was fitted to the calibration set using both composite environmental variables and latent variables. The model was used to predict out-of-sample biomass, and this process was repeated 1,000 times to calculate the median predicted biomass for each lake.

We compared two strategies to predict biomass: (1) a direct model that predicts community biomass using composite environmental variables and latent variables derived from presence-absence, and (2) a stacked model that first predict species abundance using composite environmental variables and latent variables generated from presence-absence. In the stacked approach, predicted species abundances for each lake were multiplied by average weight of each species, and the resulting values were summed to estimate community biomass. We modelled variation in local abundance and biomass using a Tweedie distribution (Tweedie 1984) with a log-link function within a Generalized Additive Model (GAM), using the *tw* and *gam* functions from the R package *mgcv*

(Wood 2004; Wood et al. 2016, version 1.9-1). Each explanatory variable was fitted with a second-order thin-plate regression spline smoother (Wood 2003) with three basis functions ($k = 3$) and additional penalty on the null space for each smoother, allowing them to shrink to zero if necessary (Marra & Wood 2011). All smoothing parameters were estimated using Restricted Maximum Likelihood (REML, Wood 2011) based on data from the calibration set. The Tweedie distribution was selected for its flexibility in modelling various mean-variance relationships, making it well-suited for the abundance and biomass data, which were represented as densities (i.e., number of fish caught per unit effort, CPUE; weight of fish caught per unit effort, BPUE). This approach was more appropriate than count-based distributions, such as Poisson or negative binomial, given the continuous nature of our data. We used three basis functions to balance flexibility and parsimony, allowing the GAM to capture the main trends without overfitting, as our goal was to identify large-scale patterns while preserving generalizability. Separate models were fitted for each species to predict abundance, and this process was repeated across multiple folds of calibration and validation to ensure robust model performance across different data subsets.

Species abundances were adjusted by multiplying each species' predicted abundance by its average weight to account for the disproportionate contributions of larger species to total biomass, ensuring that biomass predictions reflected their greater contribution to biomass. Predicted abundance was only considered for lakes where the species was known to occur, as presence-absence data served as model input. This approach leverages occurrence information to minimize potential error, prevent predictions where species are absent, and reduce model bias, thereby enhancing predictive accuracy. Species weight data were obtained from the Ontario Freshwater Fishes Life History Database (Eakins 2024), with the average of the reported weight range used to represent each species.

4.3.4. Analysis

As a reminder, we set out to answer three questions: (1) whether a stacked abundance model or a community model (i.e., directly predicting community biomass from environmental data and latent variables) would better predict total biomass, (2) whether diverse lakes (i.e., in terms of species richness) are better predicted than non-diverse lakes, and (3) whether regional biomass is better predicted than local biomass.

To evaluate lake diversity, we calculated both species richness per lake (i.e., the number of species present per lake) and Shannon's diversity index (Shannon 1948). Shannon's Index (also known as the Shannon-Wiener Index) quantifies community diversity incorporating for both species richness (i.e., the number of species) and species evenness (i.e., the relative biomass of each species). Higher values indicate greater diversity, reflecting both a higher number of species and a more equitable distribution of individuals among them. Shannon's diversity index was calculated using the *diversity* function from the R package *vegan* (Oksanen et al. 2024, version 2.6-6.1). Here we used biomass to calculate Shannon's Index as it better represented the energy flow and functional roles of species, linking diversity with ecosystem productivity and resilience (Fung *et al.* 2013; White *et al.* 2007).

We hypothesized that species richness would correlate with latitudinal gradients, as biodiversity typically increases toward lower latitudes (Hillebrand 2004; Willig *et al.* 2003). To further explore the relationship between predictive error and biodiversity metrics, including species richness and Shannon's Diversity Index, we analyzed spatial patterns in predictive error for both models. For these analyses, we fitted a Generalized Additive Model (GAM) using a Gamma distribution with a log-link function. A tensor product smoother was applied on latitude and longitude, using five basis functions to capture potential nonlinear spatial patterns (Wood 2003). The latitude and longitude

of each lake were recorded at its geographic center. The tensor product smoother was chosen for its ability to apply separate smoothing penalties to latitude and longitude, ensuring that their distinct effects on the response variable were accurately represented. Unlike thin-plate regression splines, which assume equal effects of latitude and longitude, tensor product smoothers treat each dimension independently, accommodating differences in units or scales and providing greater flexibility and accuracy in modelling spatial variation (Wood 2006, 2017). Smoothing parameters were estimated using restricted maximum likelihood (REML), known for its robustness in smoothing estimation (Wood 2011). We used the *te* and *gam* functions from the R package *mgcv* to fit the models (Wood 2004; Wood et al. 2016, version 1.9-1) and the *draw* function from the R package *gratia* to visualize the partial plots of latitude and longitude (Simpson 2024, version 0.9-2). The *te* function captures flexible relationships between predictors without assuming explicit interactions between them. Partial plots show the relationship between each predictor and the response variable while holding other predictors constant, allowing us to isolate and understand each predictor's specific influence on prediction error in our models (Wood 2017).

To evaluate prediction accuracy across spatial scales, lakes were grouped lakes by secondary and tertiary watersheds. Predicted and observed biomass were aggregated at three levels: (1) individual lakes, (2) tertiary watersheds, and (3) secondary watersheds. Specifically, the biomass of lakes within the same watersheds was summed, resulting in a single predicted and observed value for each watershed. The predictive performance of both models was assessed by calculating the ratio error (Eq. 4.1) at each spatial scale, allowing a comparison of the models' effectiveness at local and regional levels. The ratio error measures the relative magnitude of the discrepancy between predictions and observations, rather than the absolute difference.

$$Error_{m,l} = \frac{\hat{Y}_{m,l}}{Y_{m,l}}$$ Equation 4.1

where m represents the model and l the lake or aggregation of lakes. Y refers to the observed biomass and $\hat{Y}$ to the predicted biomass.

We also evaluated two other aggregation methods: grouping by the nearest neighboring lakes and by lakes within a fixed distance. For the nearest neighboring method, pairwise distances were calculated to identify the 10 and 50 nearest neighboring lakes. We aggregated the predictions from each model, along with the observed biomass, for these groups (Eq. 4.2) and calculated the predictive error (Eq. 4.1) at three spatial levels: (1) individual lakes, (2) groups of the 10-nearest-lakes level, and (3) groups of the 50-nearest-lakes. This approach yielded a predictive error for each lake. The group sizes of 10 and 50 lakes were arbitrarily chosen to represent approximately 2% and 9% of the dataset, respectively.

$$Value_{m,l,a,g} = \sum_{k} \hat{Y}_{m,k}$$ Equation 4.2

where m denotes the model, l the focal lake, a the aggregation method (e.g., nearest neighboring lakes), g the group (e.g., the 10 nearest neighboring lakes), and k an index representing the lakes within the selected group, including a the focal lake (i.e., for the 10 nearest neighboring lakes, k refers to each of these 10 lakes as well as the focal lake). $\hat{Y}$ represents the predicted biomass. The same equation was applied to calculate the observed biomass, by substituting $\hat{Y}$ with Y.

For the fixed distance method, we identified all lakes within a 50 km and 100 km radius of each focal lake by calculating pairwise distances. Predictions from each model, along with the observed biomass, were then aggregated for the lakes within each radius to assess prediction accuracy across spatial scales (Eq. 4.2). Predictive error was calculated at three levels: (1) individual lake level, (2)

within a 50 km radius around the focal lake, and (3) within a 100 km radius around the focal lake. This approach evaluates the models' performance as predictions are scaled up from local to broader spatial levels and yielded a predictive error for each lake. The 50 km and 100 km distances correspond to approximately the 2.5th and 7.5th quantiles of the pairwise geographics distances between lakes, respectively.

4.4. Results

Our primary objective was to assess whether the stacked abundance model or the community model provided better predictions of total biomass in a large lake-fish ecosystem. To achieve this, we compared the predicted biomass from both models with the observed biomass (Figure 4.2). These plots confirm that the stacked model tends to predict biomass with less variability, with predictions clustering closer to the mean values, indicating constrained estimates. In contrast, the community model exhibits greater flexibility, providing a wider range of predictions that align more closely

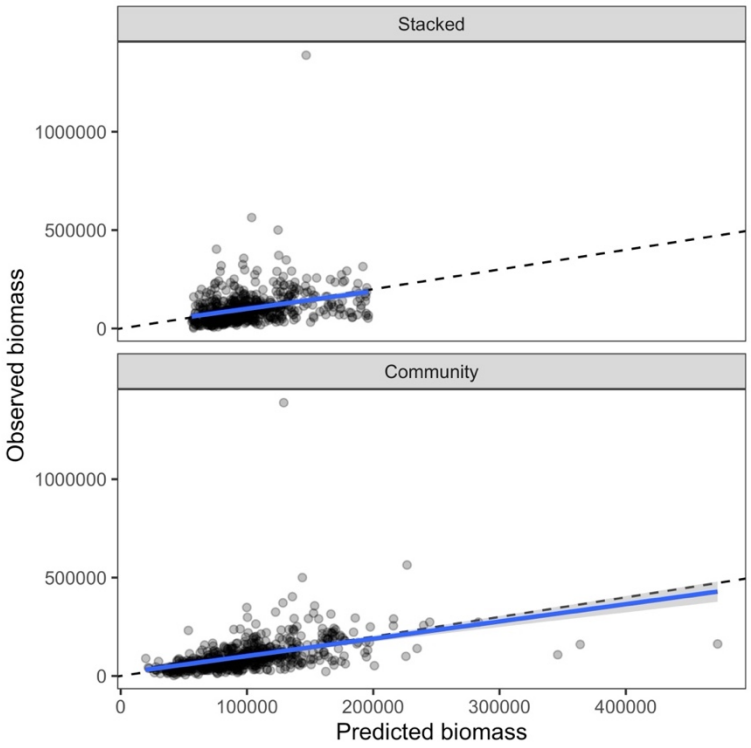

Figure 4.2: Comparison of predicted versus observed biomass for the two models. The *stacked model* (top panel) predicts species abundance using composite environmental variables and community composition, multiplies the predicted abundance by the average species weight, and fits a second model to estimate community biomass by summing these values. The *community model* (bottom panel) directly predicts community biomass using composite environmental variables and species composition. The dashed line represents the 1:1 line, indicating perfect agreement between predicted and observed biomass. The blue line represents the trend across all lakes (i.e., linear regression between predictive and observed biomass). A version of this figure using $\log_{10}$ scale is available in Figure S4.2.

with the observed biomass. This suggests that the community model may perform better at capturing variation, particularly for higher biomass values, while the stacked model appears more conservative in its predictions. This is evident in the stacked model's limited ability to predict biomass values outside a specific range, resulting in a narrower range of predictions overall (Figure S4.2). In contrast, the community model captured a wider spectrum of biomass values, reflecting greater variability across lakes (Figures 4.2 and S4.2). Comparing the tendencies of the two models, we found that the stacked model frequently overpredicted biomass, leading to a pronounced skew toward overprediction (Figure 4.3). While the community model also showed a tendency to overpredict, its distribution was more balanced, with a longer tail reflecting a greater mix of both under- and overpredictions.

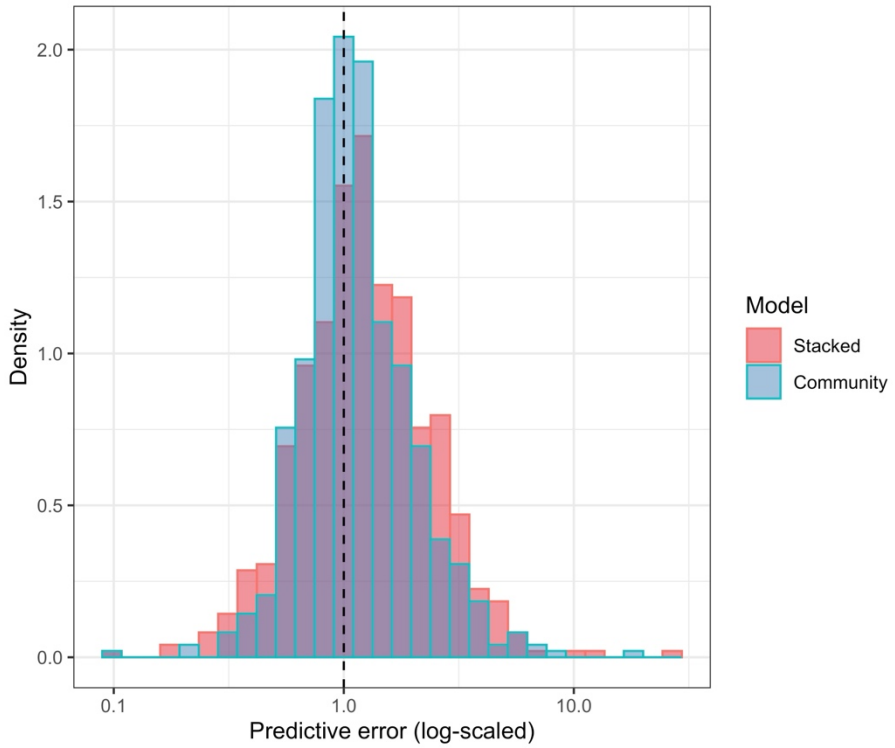

Figure 4.3: Histogram of predictive error by model type. Predictive error is calculated as the ratio of predicted biomass to observed biomass and displayed on a $\log_{10}$ scale. The stacked model (red) predicts species abundance and then estimates biomass, while the community model (blue) directly predicts biomass using composite environmental variables and community composition (see methods for more details). The dashed line indicates perfect prediction, where predicted biomass matches observed biomass.

For the second objective, we evaluated how lake diversity, measured by species richness and Shannon's index, affected prediction accuracy. In the stacked model, no clear relationship emerged between prediction accuracy and species richness (Figure 4.4). However, predictive error (log_{10}-scale) showed a positive trend with and Shannon's index, indicating that lakes with lower richness were generally underpredicted, while those with higher richness tended to be overpredicted. The community model followed a similar trend with Shannon's index, though the trend was less pronounced. Interestingly, for species richness, the community model exhibited an inverse relationship in which lakes with lower richness were more often overpredicted, whereas those with higher richness were typically underpredicted. An analysis of the spatial smooths of predictive errors reveals a clear latitudinal trend in both the stacked and community models. Lakes in the southern regions consistently showed a positive partial effect, indicating overpredictions of biomass, while northern lakes exhibited a negative partial effect, reflecting underpredictions (Figure S4.3). Notably, the stacked model displayed a slightly different pattern, with localized hotspots of overprediction in southern parts of Eastern and Western Ontario. Despite these spatial trends, the explained deviance of the models remained relatively low, at 9.7% for the stacked model and 4.1% for the community model.

Our final objective was to determine whether regional biomass predictions provided greater accuracy than local biomass predictions. To do so, again, we evaluated three aggregation methods: grouping by watershed levels (Figure 4.5), using a fixed number of nearest neighbors (Figure S4.4), and grouping neighbors within a fixed distance (Figure S4.5). Across all methods and models, the results consistently indicated that increasing the level of aggregation level reduced prediction error. This trend is reflected in the density distributions where higher aggregation levels produced

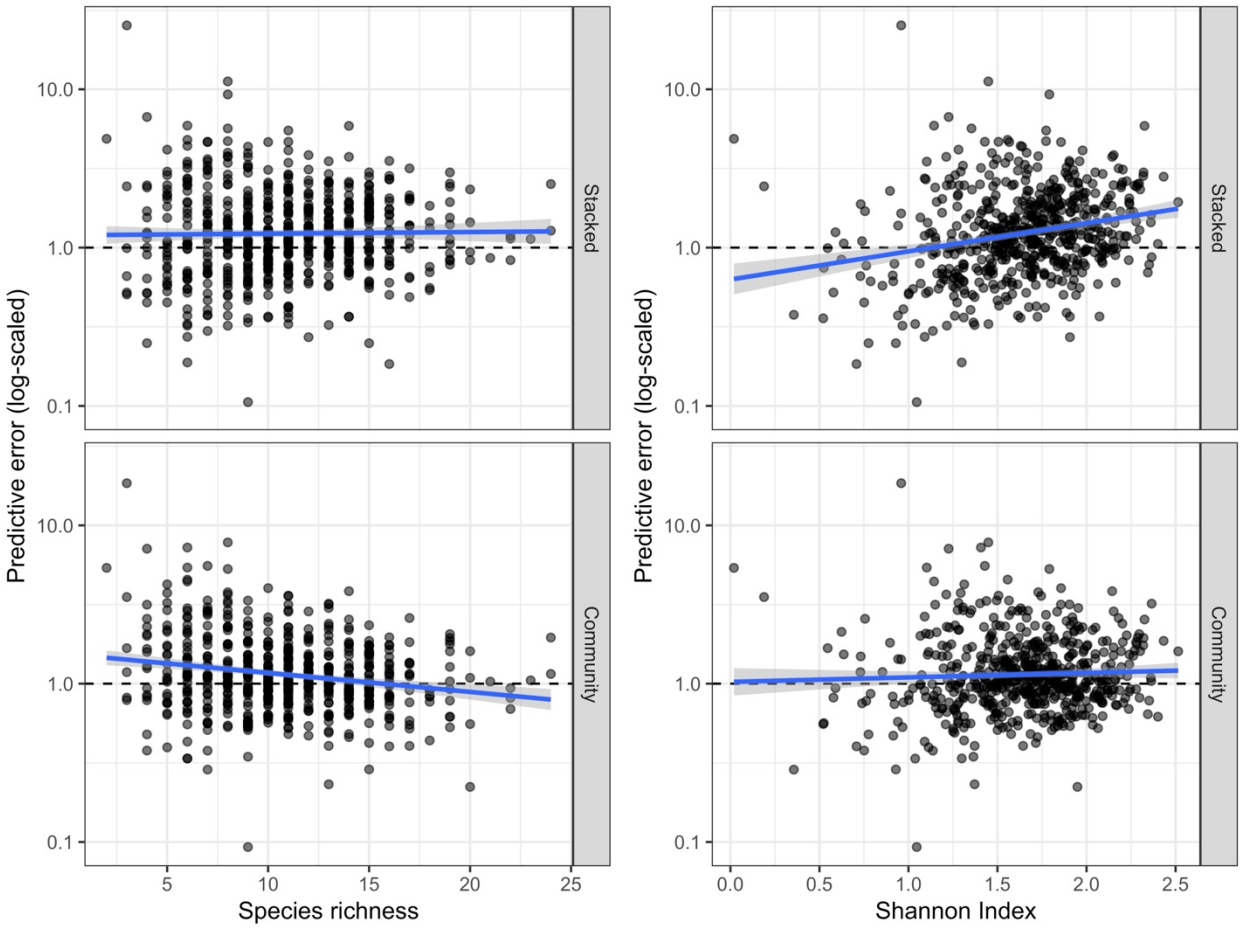

Figure 4.4: Predictive error of biomass per lake plotted against community diversity. Predictive error is calculated as the ratio of predicted biomass to observed biomass and presented on a $\log_{10}$ scale. Community diversity is measured using both species richness (i.e., number of species per lake; left panels) and Shannon's index (i.e., species diversity weighted by biomass; right panels). The stacked model results are shown in the top panels, while the community model results are in the bottom panels. The dashed line indicates perfect prediction, where predicted biomass matches observed biomass. The blue line represents the trend across all lakes, obtained from linear regression between predictive error and community diversity for each model and diversity metric.

sharper, narrower peaks centered around a predictive value of 1 on the $\log_{10}$ scale (Figures 4.5, S4.4 and S4.5).

Comparing the two models across aggregation methods, the community model exhibited lower variability in predictive errors than the stacked model (Figure 4.5). In contrast, the stacked model showed a more dispersed distribution with a slight bias toward overprediction. This pattern is further highlighted in the empirical cumulative distribution function (ECDF), where the stacked

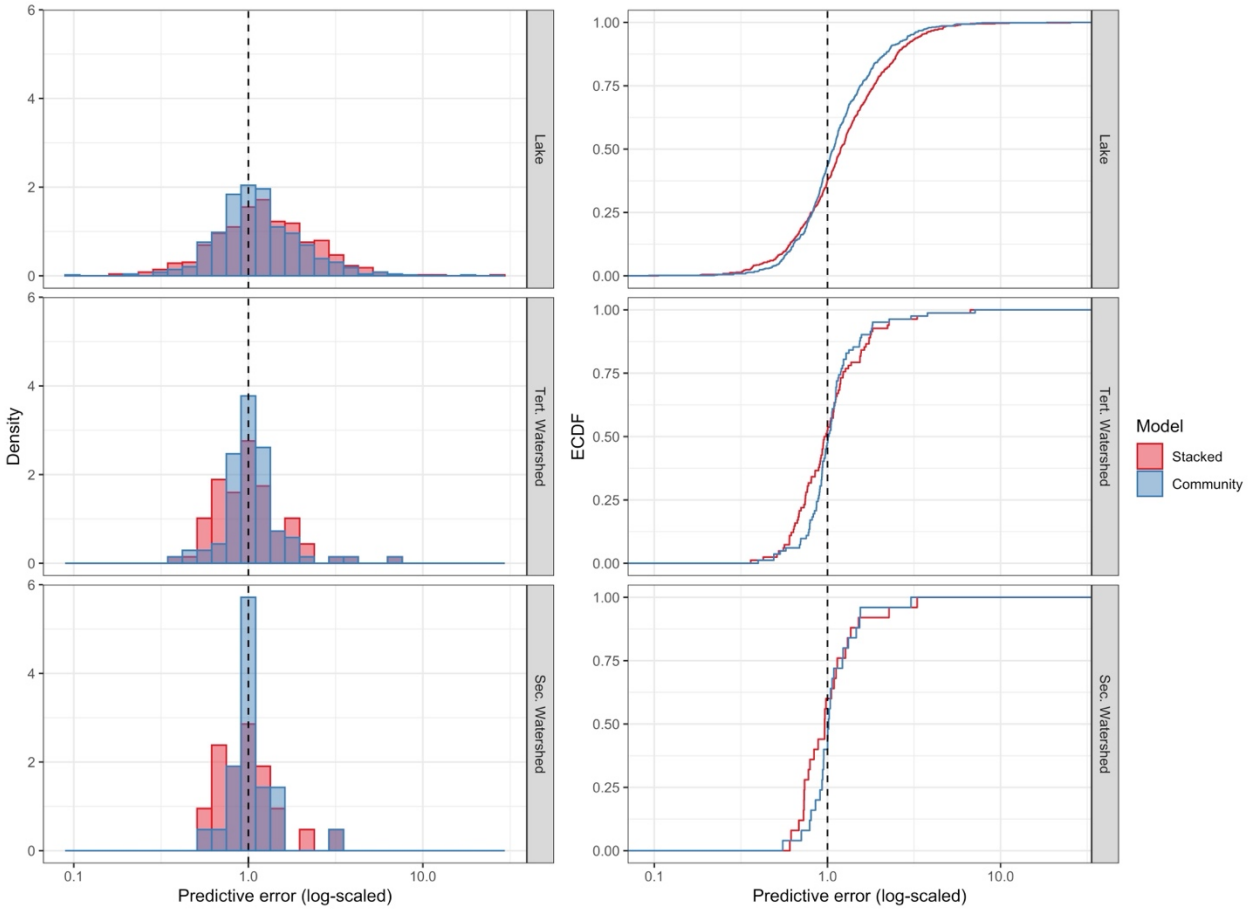

Figure 4.5: Histogram and Empirical Cumulative Distribution Function (ECDF) of predictive error across different aggregation levels and models. Predictive error is calculated as the ratio of predicted biomass to observed biomass and presented on a $\log_{10}$ scale. Aggregation levels are shown on the vertical panel, with the lowest aggregated level at the top (lake level) and the most aggregated level (secondary watershed) at the bottom. The stacked model (red) predicts species abundance and then estimates biomass, while the community model (blue) directly predicts biomass using composite environmental variables and species composition. The left panels show the density of predictive error, while the right panels display the ECDF of predictive error. The dashed line indicates a perfect prediction, where predicted biomass matches observed biomass. Alternative aggregation methods are shown in Figure S4.4 and S4.5.

model reached a predictive error of 1 on the $\log_{10}$ scale more rapidly than the community model, reflecting greater variability and reduced precision in its predictions.

4.5. Discussion

The comparison between the stacked and community models reveals key differences in their predictive tendencies. Notably, the stacked model's predictions exhibit shrinkage toward the mean,

resulting in more conservative predictions for extreme biomass values. This phenomenon, known as “shrinkage,” is a characteristic of hierarchical or multi-level models, where partial pooling of information across groups helps to regularize predictions and reduce overfitting (Gelman *et al.* 2013). This results in more stable but potentially biased predictions, particularly when capturing extreme values is crucial. In this study, the stacked model struggled to predict biomasses below or above certain thresholds, indicating that it may smooth over important ecological variability that influences biomass dynamics. This shrinkage arises from the model’s structure, where species-specific abundance predictions are generated independently and subsequently aggregated to estimate total biomass (Clark *et al.* 2014).

Since the stacked model relies on individual species predictions, errors in these predictions can propagate, contributing to the overall tendency to compress extreme values. While shrinkage can be beneficial in preventing overfitting, it may pose challenges in ecological contexts where rare or highly abundant species play a disproportionate role in ecosystem functioning. In such cases, this limitation could obscure important ecological patterns (Royle & Dorazio 2008). In contrast, the community model produced a broader range of predictions, capturing greater variability in biomass across lakes. By treating the entire community as a single unit of analysis, community models better capture complex inter-species dynamics – such as competition, facilitation, and shared resource use - by directly incorporating these interactions within the model (Warton *et al.* 2015b). This approach offers greater flexibility in representing ecological heterogeneity, which likely explains the model’s wider range of predictions in this study. However, it also increases the risk of overfitting, particularly when species interactions are complex or not well understood or captured by the model (Ovaskainen *et al.* 2017).

The skewed tendency of the stacked model to overpredict biomass more often than it underpredicts suggests a potential bias inherent to the stacking process. When species abundances are summed, correlated errors in individual predictions can inflate overall biomass estimates (Latimer *et al.* 2006). As demonstrated in Chapter 3 (Figure 3.3), frequently occurring species were generally overpredicted, while rare species were often underpredicted, indicating that dominant species may be distorting the final biomass predictions. In contrast, the community model exhibited a longer tail toward overprediction, but its distribution was more balanced. This balance likely reflects the ability of community models to account for both environmental and biotic factors simultaneously, integrating information across species and environmental conditions simultaneously to minimize systematic bias (Ovaskainen *et al.* 2017; Warton *et al.* 2015a).

To assess whether lake diversity, measured by species richness and Shannon's index, influences biomass prediction, we compared the predictive errors across models. In the stacked model, the absence of a clear trend with species richness, coupled with a significant trend with Shannon's Index, suggests that community evenness, rather than richness alone, affects predictive accuracy. Sites with lower richness but more uneven species distributions were often underpredicted, likely because to the dominance of a few species distorted the model's ability to generalize effectively (Hillebrand *et al.* 2008).

In the community model, the weaker trend with Shannon's Index and the reverse trend with species richness - where low-richness lakes were often overpredicted and high-richness lakes underpredicted – suggest that the model may perform better in more homogeneous communities with fewer or simpler species interactions (Ovaskainen *et al.* 2017). This reversal of trends with richness implies that the community model captures broader ecological dynamics, including interspecies interactions, which may be less prominent in lower-diversity ecosystems dominated

by only a few species (Warton *et al.* 2015a). Similar patterns have been observed in other studies comparing stacked species distribution models with joint models, where community-level approaches often excel at capturing complex ecological interactions (Ovaskainen & Abrego 2020).

Both the stacked and community models revealed a clear latitudinal trend in the spatial smooths of predictive errors. These spatial patterns likely reflect the large latitudinal scale of the dataset, which introduces substantial variability in environmental conditions and species compositions across regions. It is important to note that the latent and environmental predictors included in the models did have clear spatial trends (see Chapter 3, Figures S3.3 and S3.4), meaning that the remaining spatial trends in error occur even after including highly spatially structured predictors. Northern lakes, typically characterized by lower species richness and biomass, are inherently more challenging to predict accurately, contributing to discrepancies between observed and predicted values (Villéger *et al.* 2017). Notably, the stacked model displayed regional spots of overprediction in both Eastern and Western Ontario. This may be due to the stacked model's sensitivity to species-level abundances, which can amplify errors when dominant species are mispredicted (Dormann *et al.* 2018). However, this sensitivity could also enable the model to capture variations in species interactions across the latitudinal gradient. Indeed, interactions such as competition and predation often shift along latitudinal gradients in response to changing environmental conditions and species richness (Schemske *et al.* 2009). This variation can significantly influence ecosystem dynamics, particularly in fish communities, where predation pressure and interspecific competition vary between northern and southern lakes (Garvey *et al.* 2003; Law 2022; Roesti *et al.* 2020). Such biotic interactions may introduce complexities that the community model does not fully capture. The low explained deviance of the spatial smoothers for predictive error in both models suggest that spatial gradients are not the primary drivers of predictive error. It is possible that unmeasured

key environmental drivers are influencing model performance (Ovaskainen *et al.* 2017). Although including latitude and longitude (or smooth terms) could help account for spatial autocorrelation – serving as proxy for unmeasured factors - the low explained deviance suggests that this adjustment would likely yield only marginal improvements in biomass predictions (Guisan *et al.* 2013). Finally, our investigation of the impact of aggregating predictions at different scales revealed that increasing the aggregation level (i.e., whether by watershed levels, nearest neighbors, or distance) consistently reduced prediction error for both the stacked and community models. This finding aligns with previous research, which shows that spatial aggregation improves prediction accuracy by smoothing out local variations and accounting for larger-scale ecological processes (Ay *et al.* 2017; Chardon *et al.* 2016). As aggregation increases, local noise and fine-scale variability diminish, enabling the model to capture broader patterns that drive biomass predictions (Jackson & Fahrig 2015; de Knegt *et al.* 2010). Note however, that we averaged predictions across individual lakes and did not fit the models at the different aggregated levels, unlike the literature. When comparing the two models, the stacked model exhibited slightly more dispersed prediction errors and a tendency toward overprediction, consistent with research indicating that stacking species-level predictions can introduce biases (Ovaskainen *et al.* 2010). In contrast, community models, which inherently accounts for species co-occurrence and shared environmental conditions, demonstrated more accurate predictions, as reflected in its narrower predictive error distribution (Clark *et al.* 2014; Pollock *et al.* 2014). The cumulative distribution function (ECDF) further supports this, showing that the stacked model reached a predictive error of 1 more rapidly, highlighting its bias toward overprediction. This bias likely arises from treating species independently, which can overlook the complex interactions among species within communities. To improve model performance, several studies suggest that aggregation approaches, such as watershed- or nearest-neighbor-based methods, are effective tools for ecological predictions by

accounting for spatial autocorrelation and expanding the spatial scope of predictions (Dormann *et al.* 2007; Ferrier *et al.* 2002; Jackson & Fahrig 2015; Wagner & Fortin 2005). These approaches help reduce local error by considering spatial relationships. However, further refinements - such as incorporating latent spatial predictors or using latitude and longitude as covariates - could enhance predictive accuracy (Ay *et al.* 2017). Additionally, by accounting for spatial autocorrelation, we could mitigate the observed biases across regions, particularly in regions where the models tend to over- or underpredict.

In this study, we evaluated the performance of stacked and community models in predicting biomass across a range of lake ecosystems. Using composite environmental variables and presence-absence data, we demonstrated that the community model consistently outperformed the stacked model, producing a narrower, and less skewed distribution of predictive errors. However, the community model exhibited sensitivity to species richness, tending to overpredict biomass in low-richness sites and underpredict in high-richness ones. In contrast, the stacked model introduced biases into biomass predictions, likely driven by errors in predicting the abundances of dominant species, which disproportionately affect total biomass. For practitioners aiming to predict biomass, our results suggest that directly modelling biomass through community models could provide more robust predictions, especially in systems with balanced species compositions. However, due to potential dataset-specific dynamics, we recommend further validation across diverse ecosystems.

Future research should also explore more refined modelling approaches that better account for the influence of dominant species in biomass predictions. Additionally, testing models across various ecosystems and incorporating dispersal dynamics could further improve these models. Evaluating the potential of hybrid approaches that combine elements of stacked and community models would also be valuable, as such methods could capture both species-specific contributions and broader

ecosystem dynamics, improving the robustness and applicability of biomass predictions in diverse ecological contexts.

4.6. Supplementary Information

Table SI 4.1: 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 phosphorous (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

Category	Environmental variable
	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)
	Maximum water temperature (°C)
	Mean annual air temperature for 1981 and 2010 (°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 16 and 20°C) during ice free period
	Proportion of cold days (between 22 and 26°C) during ice free period
	Proportion of cold days (between 8 and 12°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)

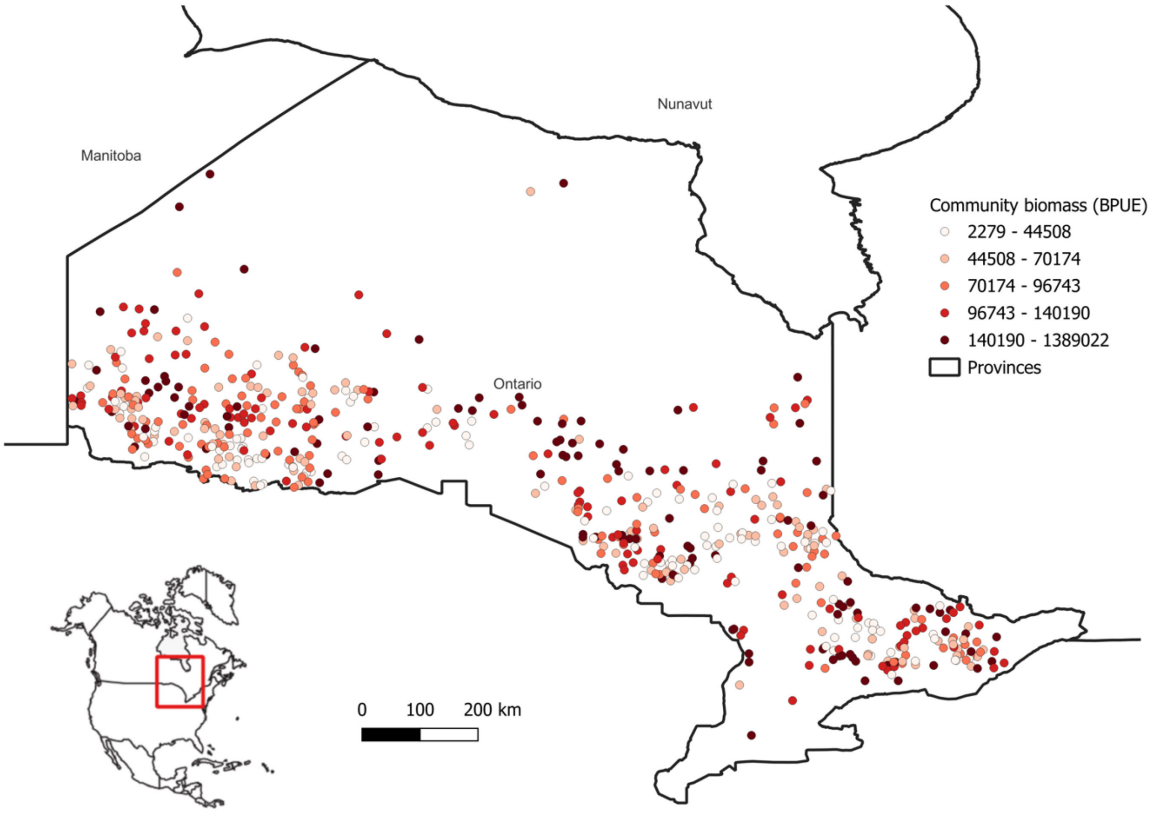

Figure SI 4.1: Map of the 583 lakes surveyed in Ontario, Canada. Each point is color-coded to indicate community biomass (i.e., total weight of fish caught per unit effort (BPUE) for each lake). Black lines denote provincial boundaries within Canada.

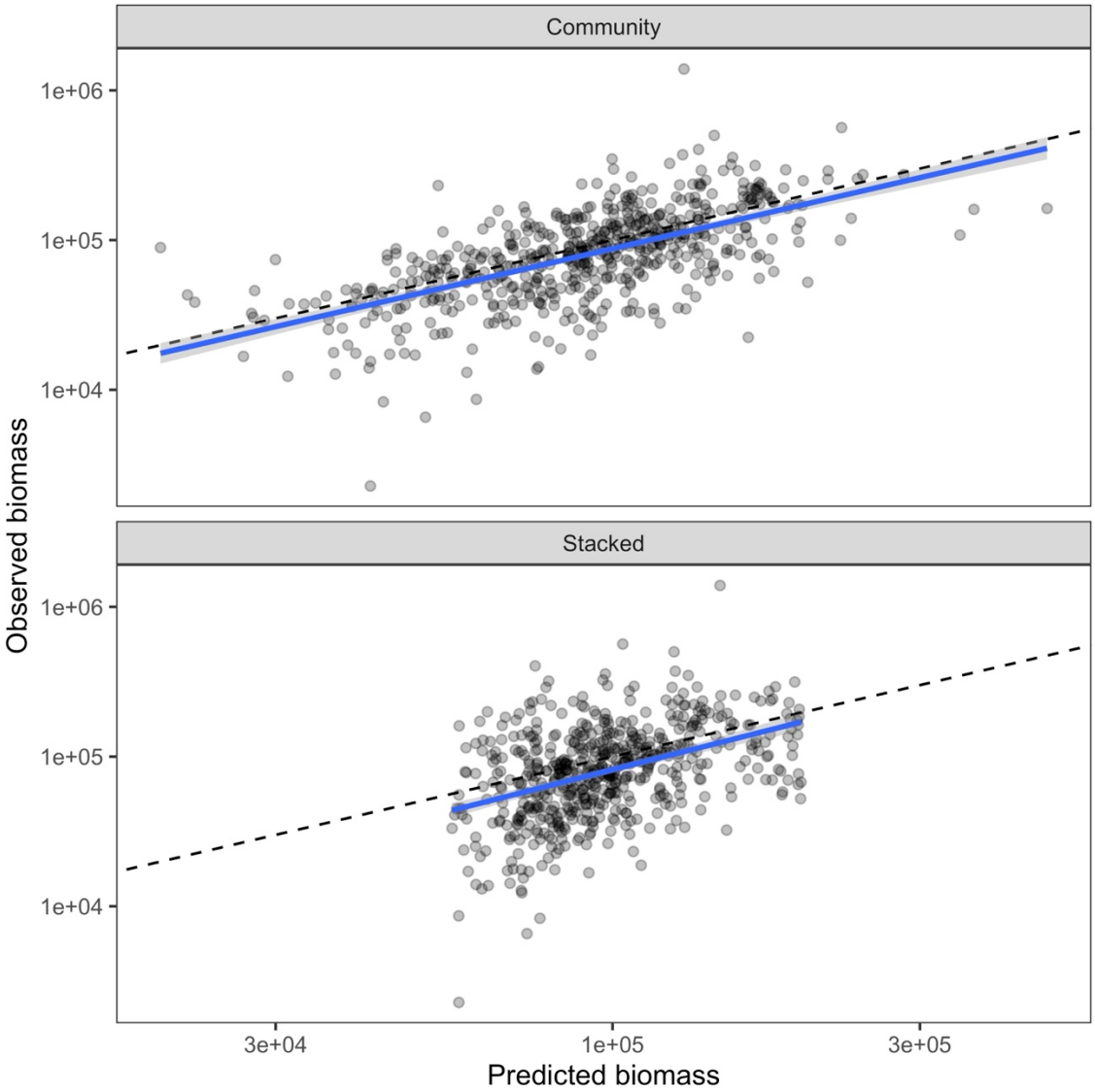

Figure SI 4.2: Comparison of predicted versus observed biomass for the two models on the $\log_{10}$ scale. The *stacked model* (top panel) predicts species abundance using composite environmental variables and community composition, multiplies the predicted abundance by the average species weight, and fits a second model to estimate community biomass by summing these values. The *community model* (bottom panel) directly predicts community biomass using composite environmental variables and species composition. The dashed line represents the 1:1 line, indicating perfect agreement between predicted and observed biomass. The blue line represents the trend across all lakes (i.e., linear regression between predictive and observed biomass).

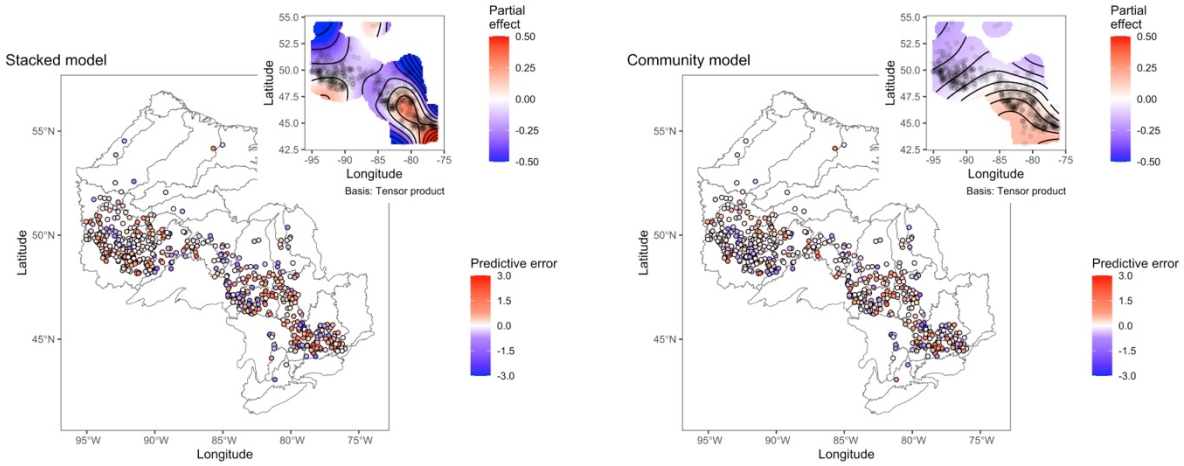

Figure SI 4.3: Map and estimated spatial smooths of the prediction errors for the two models. The stacked model (left panel) predicts abundance and then estimates biomass, while the community model (right panel) directly predicts biomass from composite environmental variables and community composition. The predictive error is calculated as the $\log_{10}$ of predicted biomass over observed biomass. The maps show underprediction in blue and overprediction in red. The black lines represent the delimitations of the secondary watersheds. The spatial smooths show in blue, areas where lake biomass tend to be more underestimated and in red, areas where lake biomass tends to be more overestimated.

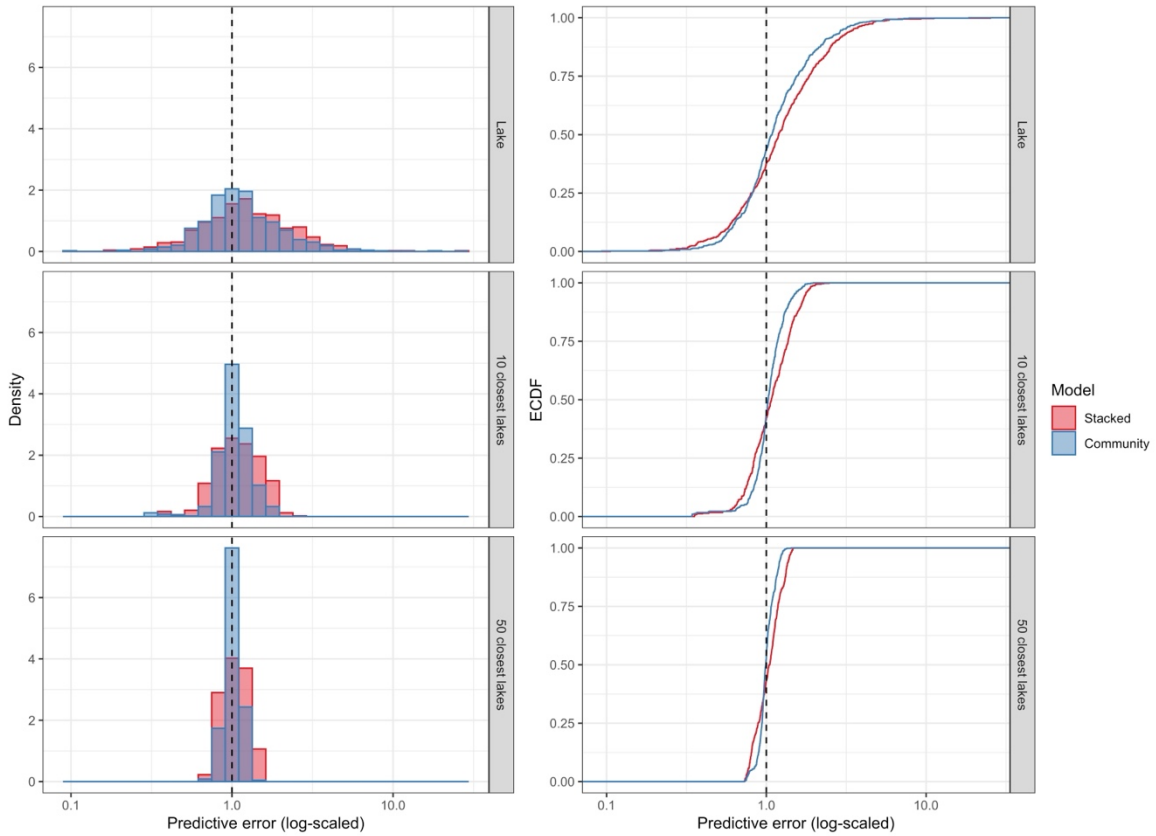

Figure SI 4.4: Histogram and Empirical Cumulative Distribution Function (ECDF) of predictive error across different aggregation levels and models using the nearest neighbor method. Predictive error is calculated as the ratio of predicted biomass to observed biomass and presented on a $\log_{10}$ scale. Aggregation levels are shown on the vertical panel, with the lowest aggregated level at the top (lake level) and the most aggregated level (50 closest lakes to the focal lake) at the bottom. The stacked model (red) predicts species abundance and then estimates biomass, while the community model (blue) directly predicts biomass using composite environmental variables and species composition. The left panels show the density of predictive error, while the right panels display the ECDF of predictive error. The dashed line indicates a perfect prediction, where predicted biomass matches observed biomass.

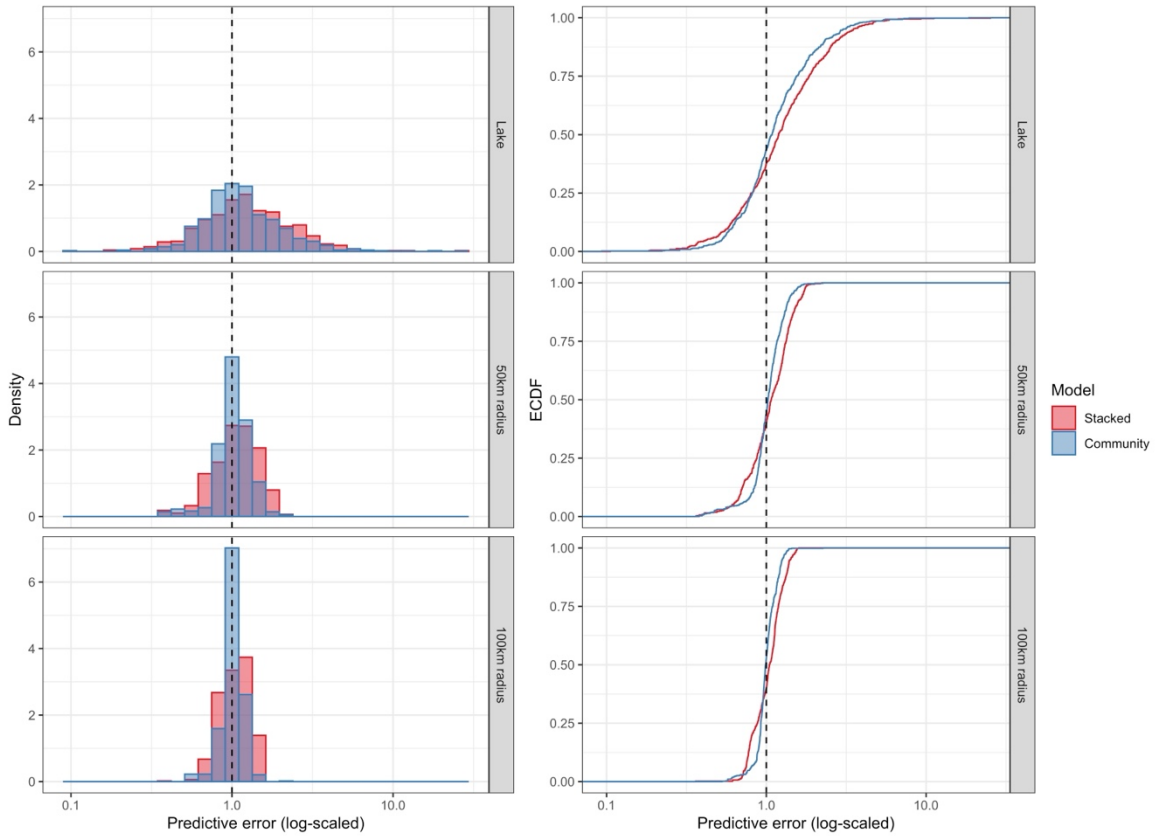

Figure SI 4.5: Histogram and Empirical Cumulative Distribution Function (ECDF) of predictive error across different aggregation levels and models using the distance-based method. Predictive error is calculated as the ratio of predicted biomass to observed biomass and presented on a $\log_{10}$ scale. Aggregation levels are shown on the vertical panel, with the lowest aggregated level at the top (lake level) and the most aggregated level (all lakes within a 100 km radius around the focal lake) at the bottom. The stacked model (red) predicts species abundance and then estimates biomass, while the community model (blue) directly predicts biomass using composite environmental variables and species composition. The left panels show the density of predictive error, while the right panels display the ECDF of predictive error. The dashed line indicates a perfect prediction, where predicted biomass matches observed biomass.

Chapter 5: Concluding remarks, assumptions, and future directions

> "*I don't pretend to see the path, but I know it's there all the same. One day, we'll look back and wonder how we ever missed it*."
>
> Peter V. Brett, *The Warded Man*

The overarching goal of this thesis was to improve predictive models of species abundance and biomass using community data and environmental proxies to contribute to more informed ecosystem management and conservation practices. The three chapters build upon each other, progressively expanding the scope of the models and their applications. Chapter 2 establishes the foundation by focusing on simulated data to refine single-species abundance models, exploring how different levels of information (true environmental drivers versus latent variables based on species co-occurrence) affect model accuracy. This simulation framework is essential for understanding the robustness of the models before their application to real-world data. Chapter 3 extends this work by applying the developed framework to empirical data from lakes, focusing on sport fish species. It investigates the role of latent variables and different fish assemblages in improving abundance predictions and explores how lake-specific characteristics influence model performance. This real-world application allows for a deeper understanding of how the framework operates under natural conditions, particularly in aquatic ecosystems. Finally, Chapter 4 takes the abundance predictions from Chapter 3 and uses them to build a stacked model for predicting community biomass. It compares the effectiveness of the stacked model versus a community model in predicting biomass across varying spatial scales and species richness levels. By progressing from single-species models to multi-species and finally to biomass prediction, the chapters are sequentially linked, each addressing a broader ecological question while refining and testing the models at different levels of complexity. This cohesive approach enhances both the predictive

accuracy of species abundance and the practical applications of these models for ecosystem management.

Throughout this thesis, we made the simplifying assumption that the ecosystem under study was static, i.e., that the environmental conditions, species interactions, and community structure remained constant over time. While this assumption facilitated model development and allowed us to focus on key relationships, it is well understood that ecosystems are inherently dynamic. Natural fluctuations (e.g., seasonal changes, variations in resource availability), anthropogenic influences (e.g., pollution, habitat degradation, climate change), and species interactions continuously reshape ecosystems (Levin 1998; Parmesan & Yohe 2003). Ecological dynamics influence species distribution, abundance, and community structure over time, which can affect predictive accuracy (Dormann *et al.* 2013; Grimm & Railsback 2013; Tilman 1994). While testing a framework on a static snapshot is crucial for identifying the model's strengths and limitations in a controlled environment, it is equally important to incorporate temporal dynamics moving forward (Hastings 2004). This consideration is particularly relevant in the context of climate change and anthropogenic impacts, where accounting for these dynamics will enhance the model's applicability to real-world scenarios. Future research should explore the integration of temporal data that captures ecosystem changes over time.

When simulating communities in Chapter 2, we had to make crucial decisions about which mechanisms driving species abundance to simulate. Given the computational complexities and the challenges involved in accurately simulating species interactions, we chose to focus solely on environmental selection. This approach allowed us to model species distributions based on environmental variables, but we acknowledge that this choice limits the applicability of our results to real-world scenarios. Environmental selection represents only one mechanism in the complex

processes governing species abundance, and as shown in Chapter 3, the results from these simulations might differ from empirical data, where species interactions play a significant role (Ovaskainen *et al.* 2010). This, however, does not mean that the framework is not useful; rather, it is that the conclusions of Chapter 2 had to be tested on an empirical dataset to understand the variations generated by other mechanisms driving species abundance.

Another limitation in Chapter 2 is that the relationship between environmental variables and species abundance was modeled as linear for simplicity of interpretation. However, ecological systems often exhibit more complex, non-linear relationships. For example, quadratic or other curvilinear relationships between environmental gradients and species abundance are likely closer to reality, as species tend to respond to optimal conditions within a range, rather than continuously increasing or decreasing with changes in the environment (Dormann *et al.* 2013). Though we began exploring these non-linear relationships, these investigations were not integrated into the thesis, leaving an important venue for future research. Furthermore, the results in Chapter 3 revealed that while some divergence existed between the predicted patterns in Chapter 2 and empirical observations, the primary trends remained consistent, aligning with expectations based on theoretical models and prior research.

In Chapter 3, we were able to show that latent variables did not always improve predictions, highlighting the complexity of modelling species abundance. This finding aligns with existing literature, where latent variables (often used to capture unmeasured environmental or biotic factors) have shown varying success in improving predictive accuracy depending on the context (Warton *et al.* 2015b). In our study, the results suggested that the influence of these latent variables is more pronounced in widespread species, while for rare or low-occurring species, environmental selection played a more critical role. This dichotomy between the predictors of species abundance based on

occurrence is supported by several ecological studies (Khattar *et al.* 2021; Magurran & Henderson 2003). Rare species, often more specialized and sensitive to environmental changes, are primarily influenced by specific habitat features and environmental variables (Gaston 1994). Studies have shown that environmental heterogeneity can also drive the distribution of rare species, making them more vulnerable to environmental shifts (Chesson 2000; Tilman 1994). Conversely, abundant or widely distributed species are often more influenced by biotic interactions, such as competition and facilitation, which play a critical role in shaping community structure (Götzenberger *et al.* 2012). Our results, which demonstrated that high-occurring species were more sensitive to species interactions (i.e., through being better predicted when including latent variables), are consistent with studies showing that the abundance of dominant species is often modulated by interspecific interactions, such as predation or competition, particularly in ecosystems with complex community dynamics (Kraft *et al.* 2015).

While Chapter 3 primarily focused on patterns across lakes and our analyses considered species based on their occurrence and abundance, we did not investigate how species' traits (e.g., ecological role, behavior, or life history) might affect community structure and species interactions. Looking through a functional lens, rather than a strictly taxonomic one, could reveal how trait redundancies or complementarity shape ecosystem dynamics and affect model predictions (Cadotte *et al.* 2011). For instance, species that share similar traits or ecological niches may introduce redundancies, minimizing their individual importance in structuring communities. Conversely, rare or functionally unique species could play disproportionately large roles, driving key ecosystem processes despite their low abundance (Mouillot *et al.* 2013). By not considering these functional traits in our current models, we may overlook key interpretations that could provide more nuanced insights into the patterns observed.

In Chapters 3 and 4, we chose to exclude species that were present in fewer than 10 and six sites, respectively, to facilitate faster and more reliable convergence of our models, particularly those incorporating latent variables. This approach is supported by research indicating that species with low occurrences can contribute disproportionate noise to models while providing minimal additional information beyond that more common species offer (Gauch 1982; McCune & Grace 2002). However, the exclusion of rare species is not without controversy. Several studies suggest that rare species may serve as sensitive indicators of ecosystem stress or habitat degradation (Cao *et al.* 2001; Faith & Norris 1989), implying their potential value in ecological models. As such, the removal of these species remains a topic of debate in ecological modelling (Poos & Jackson 2012). While thresholds, like excluding species occurring in fewer than 5-10% of sites are commonly recommended (Marchant 1990; McCune & Grace 2002; McGarigal *et al.* 2000), our approach was more conservative, as we excluded species found in less than 1-2% of sites. This decision aligns with some recommendations, though there is evidence suggesting that even applying these thresholds can affect model outcomes, especially when it comes to rare species' contributions (Poos & Jackson 2012). Future research could explore alternative approaches, such as weighting rare species differently rather than excluding them altogether.

In Chapter 4, we adopted a "*predict first, assemble later*" strategy for aggregating species predictions (Ferrier *et al.* 2002). This approach, while common in predictive modelling, is just one of several possible aggregation strategies (e.g., joint modelling or hierarchical approaches, Ferrier & Guisan 2006; Overton et al. 2002). Given the structure of our dataset, particularly the fact that environmental variables were measured at individual lake sites, it was logistically challenging, if not impossible, to apply an alternative aggregation method that combined variables across multiple lakes. As a result, we did not explore or compare different strategies in this study. Nevertheless, the

choice of aggregation method could influence predictive accuracy, potentially affecting error rates (Ovaskainen *et al.* 2017; Royle & Dorazio 2008). Future studies could test alternative modelling strategies on datasets that allow for aggregation across larger spatial or temporal scales, which may provide further insight into error reduction (Ferrier & Guisan 2006).

There are numerous metrics available to measure predictive error in ecological models (Fielding & Bell 1997; Piñeiro *et al.* 2008), each with strengths and limitations depending on the context and objective of the analysis (e.g., measuring out of sample error). Throughout the different chapters of this thesis, we adopted various metrics to better capture the patterns observed, shifting from True Skill Statistic (TSS) and Mean Absolute Percentage Error (MAPE) to log error. This transition was driven by the need for more nuanced approaches that could account for specific aspects of the predictions, such as separating over- and underprediction. For example, log error provided insights into relative error, allowing us to highlight the magnitude of prediction discrepancies, particularly for skewed distributions. The right choice of metric depends on several factors: the distribution of the data, the nature of the response variable, and whether the focus is on absolute or relative prediction accuracy (Willmott & Matsuura 2005). Each time, the selection of the metric was carefully considered, not only in relation to the model's objectives and the nature of the data, but also with attention to the known biases and limitations inherent to each metric (Botchkarev 2019; Fielding & Bell 1997; Lobo *et al.* 2008). Choosing the appropriate error metric is crucial because it influences the interpretation of the results (Piñeiro *et al.* 2008). For instance, signed metrics can provide information about whether models systematically over- or underpredict, while unsigned metrics are often used to assess general accuracy without distinguishing directionality. Therefore, researchers should consider not only the mathematical properties of a metric but also how it aligns with their specific research questions and the ecological patterns they aim to capture.

In conclusion, this thesis advances a novel framework for predicting species abundance, evolving from single-species models based on simulations to multi-species models and community biomass prediction using empirical data. The flexibility of the framework was demonstrated, showing that it can be tailored to specific datasets and ecological contexts. Throughout this work, we identified several strengths, including the capacity of the community model to capture complex ecosystem processes, as well as some limitations, such as biases in species-rich and species-poor environments. These limitations, however, present opportunities for refinement, particularly through the inclusion of species traits, functional roles, and improved handling of species interactions. Given the framework's adaptability, future research could expand on its potential for more robust applications in ecosystem management, offering critical insights into biodiversity monitoring and ecosystem productivity. The foundation laid here opens the door to further innovations in ecological modelling.

Bibliography

Allouche, O., Tsoar, A. & Kadmon, R. (2006). Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS). *Journal of Applied Ecology*, 43, 1223–1232.

Anderson, M.J., de Valpine, P., Punnett, A. & Miller, A.E. (2019). A Pathway for Multivariate Analysis of Ecological Communities Using Copulas. *Ecol Evol*, 9, 3276–3294.

Appelberg, M. (2000). Swedish Standard Methods for Sampling Freshwater Fish with Multi-mesh Gillnets. *Fiskeriverket Information 2000*, 1, 1–32.

Araújo, M.B. & Guisan, A. (2006). Five (or so) Challenges for Species Distribution Modelling. *J Biogeogr*, 33, 1677–1688.

Araújo, M.B. & Luoto, M. (2007). The Importance of Biotic Interactions for Modelling Species Distributions under Climate Change. *Global Ecology and Biogeography*, 16, 743–753.

Arranz, I., Fournier, B., Lester, N.P., Shuter, B.J. & Peres-Neto, P.R. (2022). Species Compositions Mediate Biomass Conservation: The Case of Lake Fish Communities. *Ecology*, 103, e3608.

Ay, J.-S., Chakir, R. & Gallo, J. Le. (2017). Aggregated Versus Individual Land-Use Models: Modeling Spatial Autocorrelation to Increase Predictive Accuracy. *Environmental Modeling & Assessment*, 22, 129–145.

Bahn, V. & McGill, B.J. (2013). Testing the Predictive Performance of Distribution Models. *Oikos*, 122, 321–331.

Bansal, S., Grenfell, B.T. & Meyers, L.A. (2007). When Individual Behaviour Matters: Homogeneous and Network Models in Epidemiology. *J R Soc Interface*, 4, 879–891.

Barnosky, A.D., Matzke, N., Tomiya, S., Wogan, G.O.U., Swartz, B., Quental, T.B., *et al.* (2011). Has the Earth's Sixth Mass Extinction Already Arrived? *Nature*, 471, 51–57.

Bartholomew, D.J., Knott, M. & Moustaki, Irini. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley.

Beale, C.M. & Lennon, J.J. (2012). Incorporating Uncertainty in Predictive Species Distribution Modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 247–258.

Bengtsson, J., Baillie, S.R. & Lawton, J. (1997). Community Variability Increases with Time. *Oikos*, 78, 249–256.

de Bie, T., de Meester, L., Brendonck, L., Martens, K., Goddeeris, B., Ercken, D., *et al.* (2012). Body Size and Dispersal Mode as Key Traits Determining Metacommunity Structure of Aquatic Organisms. *Ecol Lett*, 15, 740–747.

Blanchet, F.G., Cazelles, K. & Gravel, D. (2020). Co-occurrence is not Evidence of Ecological Interactions. *Ecol Lett*, 23, 1050–1063.

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. Wiley.

Botchkarev, A. (2019). A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, 14, 45–76.

Boulangeat, I., Gravel, D. & Thuiller, W. (2012). Accounting for Dispersal and Biotic Interactions to Disentangle the Drivers of Species Distributions and their Abundances. *Ecol Lett*, 15, 584–593.

Boyce, M.S., Johnson, C.J., Merrill, E.H., Nielsen, S.E., Solberg, E.J. & van Moorter, B. (2016). Can Habitat Selection Predict Abundance? *Journal of Animal Ecology*, 85, 11–20.

ter Braak, C.J.F. (1985). Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model. *Biometrics*, 41, 859–873.

ter Braak, C.J.F. & Prentice, I.C. (1988). A Theory of Gradient Analysis. In: *Advances in Ecological Research*. pp. 271–317.

Bradley, B.A. (2016). Predicting Abundance with Presence-only Models. *Landsc Ecol*, 31, 19–30.

Brook, B.W., Sodhi, N.S. & Bradshaw, C.J.A. (2008). Synergies Among Extinction Drivers Under Global Change. *Trends Ecol Evol*, 23, 453–460.

Brosse, S., Guegan, J.-F., Tourenq, J.-N. & 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, 299–311.

Brotons, L., Thuiller, W., Araújo, M.B. & Hirzel, A.H. (2004). Presence-Absence Versus Presence-Only Modelling Methods for Predicting Bird Habitat Suitability. *Ecography*, 27, 437–448.

Brown, J.H., Mehlman, D.W. & Stevens, G.C. (1995). *Spatial Variation in Abundance*. *Source: Ecology*.

Buckland, S.T., Magurran, A.E., Green, R.E. & Fewster, R.M. (2005). Monitoring Change in Biodiversity Through Composite Indices. In: *Philosophical Transactions of the Royal Society B: Biological Sciences*. Royal Society, pp. 243–254.

Burnham, K.P. & Anderson, D.R. (2004). *Model Selection and Multimodel Inference*. Springer New York, New York, NY.

Cadotte, M.W., Carscadden, K. & Mirotchnick, N. (2011). Beyond Species: Functional Diversity and the Maintenance of Ecological Processes and Services. *Journal of Applied Ecology*, 48, 1079–1087.

Cao, Y., Larsen, D.P. & Thorne, R.S.-J. (2001). Rare Species in Multivariate Analysis for Bioassessment: Some Considerations. *J North Am Benthol Soc*, 20, 144–153.

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., *et al.* (2012). Biodiversity Loss and its Impact on Humanity. *Nature*, 486, 59–67.

Carreira-Perpiñán, M.A. (1997). *A Review of Dimension Reduction Techniques*. *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*. Sheffield.

Ceballos, G., Ehrlich, P.R., Barnosky, A.D., García, A., Pringle, R.M. & Palmer, T.M. (2015). Accelerated Modern Human–Induced Species Losses: Entering the Sixth Mass Extinction. *Sci Adv*, 1, e1400253.

Cebrián, J. (1999). Patterns in the Fate of Production in Plant Communities. *American Naturalist*, 154, 449–468.

Cebrián, J. & Duarte, C.M. (1995). Plant Growth-Rate Dependence of Detrital Carbon Storage in Ecosystems. *Science (1979)*, 268, 1606–1608.

Chakraborty, A., Gelfand, A.E., Wilson, A.M., Latimer, A.M. & Silander, J.A. (2011). Point Pattern Modelling for Degraded Presence-Only Data Over Large Regions. *J R Stat Soc Ser C Appl Stat*, 60, 757–776.

Chapin, F.S.I., Matson, P.A. & Vitousek, P.M. (2011). *Principles of Terrestrial Ecosystem Ecology*. Springer New York, New York, NY.

Chapin, F.S.I., Zavaleta, E.S., Eviner, V.T., Naylor, R.L., Vitousek, P.M., Reynolds, H.L., *et al.* (2000). Consequences of Changing Biodiversity. *Nature*, 405, 234–242.

Chardon, J., Favre, A.-C. & Hingray, B. (2016). Effects of Spatial Aggregation on the Accuracy of Statistically Downscaled Precipitation Predictions. *J Hydrometeorol*, 17, 1561–1578.

Chase, J.M. & Leibold, M.A. (2003). *Ecological Niches*. University of Chicago Press.

Chesson, P. (2000). Mechanisms of Maintenance of Species Diversity. *Annu Rev Ecol Syst*, 31, 343–366.

Christianson, D.S. & Kaufman, C.G. (2016). Effects of Sample Design and Landscape Features on a Measure of Environmental Heterogeneity. *Methods Ecol Evol*, 7, 770–782.

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

Clark, J.S., Gelfand, A.E., Woodall, C.W. & Zhu, K. (2014). More Than the Sum of the Parts: Forest Climate Response from Joint Species Distribution Models. *Ecological Applications*, 24, 990–999.

Clark, J.S., Nemergut, D., Seyednasrollah, B., Turner, P.J. & Zhang, S. (2017). Generalized Joint Attribute Modeling for Biodiversity Analysis: Median-Zero, Multivariate, Multifarious Data. *Ecol Monogr*, 87, 34–56.

Cumming, G.S., Morrison, T.H. & Hughes, T.P. (2017). New Directions for Understanding the Spatial Resilience of Social–Ecological Systems. *Ecosystems*, 20, 649–664.

Cunningham, P. (2008). Dimension Reduction. In: *Machine Learning Techniques for Multimedia*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 91–112.

Cyr, H. & Pace, M.L. (1993). Magnitude and Patterns of Herbivory in Aquatic and Terrestrial Ecosystems. *Nature*, 361, 148–150.

Degnbol, P. & 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.

Díaz, S., Fargione, J., Chapin, F.S. & Tilman, D. (2006). Biodiversity Loss Threatens Human Well-Being. *PLoS Biol*, 4, e277.

Dickinson, J.L., Zuckerberg, B. & Bonter, D.N. (2010). Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annu Rev Ecol Evol Syst*, 41, 149–172.

Dietze, M.C., Fox, A., Beck-Johnson, L.M., Betancourt, J.L., Hooten, M.B., Jarnevich, C.S., *et al.* (2018). Iterative Near-Term Ecological Forecasting: Needs, Opportunities, and Challenges. *Proc Natl Acad Sci U S A*, 115, 1424–1432.

DiRenzo, G. V., Hanks, E. & Miller, D.A.W. (2023). A Practical Guide to Understanding and Validating Complex Models Using Data Simulations. *Methods Ecol Evol*, 14, 203–217.

Dormann, C.F. (2007a). Effects of Incorporating Spatial Autocorrelation into the Analysis of Species Distribution Data. *Global Ecology and Biogeography*.

Dormann, C.F. (2007b). Promising the Future? Global Change Projections of Species Distributions. *Basic Appl Ecol*, 8, 387–397.

Dormann, C.F., Calabrese, J.M., Guillera-arroita, G., Matechou, E., Bahn, V., Barto, K.N., *et al.* (2018). Model Averaging in Ecology: A Review of Bayesian, Information-Theoretic, and Tactical Approaches for Predictive Inference. *Ecol Monogr*, 4, 485–504.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., *et al.* (2013). Collinearity: A Review of Methods to Deal with it and a Simulation Study Evaluating their Performance. *Ecography*, 36, 27–46.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., *et al.* (2007). Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data: A Review. *Ecography*, 30, 609–628.

Dornelas, M., Magurran, A.E., Buckland, S.T., Chao, A., Chazdon, R.L., Colwell, R.K., *et al.* (2013). Quantifying Temporal Change in Biodiversity: Challenges and Opportunities. *Proceedings of the Royal Society B: Biological Sciences*.

Doser, J.W., Kéry, M., Saunders, S.P., Finley, A.O., Bateman, B.L., Grand, J., *et al.* (2024). Guidelines for the Use of Spatially Varying Coefficients in Species Distribution Models. *Global Ecology and Biogeography*, 33.

Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guénard, G., *et al.* (2023). adespatial: Multivariate Multiscale Spatial Analysis. *Ecol Monogr*.

Dunn, P.K. & Smyth, G.K. (1996). *Randomized Quantile Residuals*. *Source: Journal of Computational and Graphical Statistics*.

Eakins, R.J. (2024). *Ontario Freshwater Fishes Life History Database*. *Version 5.31. Online database.* Available at: https://www.ontariofishes.ca. Last accessed 5 September 2024.

Efron, B. & Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman & Hall.

Elith, J., Kearney, M. & Phillips, S. (2010). The Art of Modelling Range-Shifting Species. *Methods Ecol Evol*, 1, 330–342.

Elith, J. & Leathwick, J.R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annu Rev Ecol Evol Syst*, 40, 677–697.

Estes, J.A., Terborgh, J., Brashares, J.S., Power, M.E., Berger, J., Bond, W.J., *et al.* (2011). Trophic Downgrading of Planet Earth. *Science (1979)*, 333, 301–306.

Evans, M.R., Grimm, V., Johst, K., Knuuttila, T., de Langhe, R., Lessells, C.M., *et al.* (2013). Do Simple Models Lead to Generality in Ecology? *Trends Ecol Evol*, 28, 578–583.

Failing, L., Gregory, R. & Harstone, M. (2007). Integrating Science and Local Knowledge in Environmental Risk Management: A Decision-Focused Approach. *Ecological Economics*, 64, 47–60.

Faith, D.P. & Norris, R.H. (1989). Correlation of Environmental Variables with Patterns of Distribution and Abundance of Common and Rare Freshwater Macroinvertebrates. *Biol Conserv*, 50, 77–98.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002). Extended Statistical Approaches to Modelling Spatial Pattern in Biodiversity in Northeast New South Wales. II. Community-level Modelling. *Biodivers Conserv*, 11, 2309–2338.

Ferrier, S. & Guisan, A. (2006). Spatial Modelling of Biodiversity at the Community Level. *Journal of Applied Ecology*, 43, 393–404.

Fielding, A.H. & Bell, J.F. (1997). A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models. *Environ Conserv*, 24, 38–49.

Flecker, A.S. & Matthews, W.J. (1999). Patterns in Freshwater Fish Ecology. *Copeia*, 1999, 229–230.

Fortin, M., Drapeau, P. & Legendre, P. (1989). Spatial Autocorrelation and Sampling Design in Plant Ecology. *Vegetatio*, 83, 209–222.

Franklin, J. (2010). *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press.

Fung, T., Farnsworth, K.D., Shephard, S., Reid, D.G. & Rossberg, A.G. (2013). Why the Size Structure of Marine Communities Can Require Decades to Recover from Fishing. *Mar Ecol Prog Ser*, 484, 155–171.

Garvey, J.E., Devries, D.R., Wright, R.A. & Miner, J.G. (2003). Energetic Adaptations along a Broad Latitudinal Gradient: Implications for Widely Distributed Assemblages. *Bioscience*, 53, 141–150.

Gaston, K.J. (1994). *Rarity*. Springer Netherlands, Dordrecht.

Gaston, K.J. (1996). The Multiple Forms of the Interspecific Abundance-Distribution Relationship. *Oikos*, 76, 211–220.

Gaston, K.J. (2003). *The Structure and Dynamics of Geographic Ranges*. Oxfort University Press.

Gaston, K.J. & Fuller, R.A. (2008). Commonness, Population Depletion and Conservation Biology. *Trends Ecol Evol*, 23, 14–19.

Gauch, H.G. (1982). *Multivariate Analysis in Community Ecology*. Cambridge University Press.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., *et al.* (2013). *Bayesian Data Analysis*. 3rd edn. CRC Press.

Gelman, A. & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

González-Salazar, C., Stephens, C.R. & Marquet, P.A. (2013). Comparing the Relative Contributions of Biotic and Abiotic Factors as Mediators of Species' Distributions. *Ecol Modell*, 248, 57–70.

Götzenberger, L., de Bello, F., Bråthen, K.A., Davison, J., Dubuis, A., Guisan, A., *et al.* (2012). Ecological Assembly Rules in Plant Communities—Approaches, Patterns and Prospects. *Biological Reviews*, 87, 111–127.

Grimm, V. & Railsback, S. (2013). *Individual-based Modeling and Ecology*. Princeton University Press.

Guélat, J. & Kéry, M. (2018). Effects of spatial autocorrelation and imperfect detection on species distribution models. *Methods Ecol Evol*, 9, 1614–1625.

Guisan, A. & Thuiller, W. (2005). Predicting Species Distribution: Offering More Than Simple Habitat Models. *Ecol Lett*, 8, 993–1009.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., *et al.* (2013). Predicting Species Distributions for Conservation Decisions. *Ecol Lett*, 16, 1424–1435.

Guisan, A. & Zimmermann, N.E. (2000). Predictive Habitat Distribution Models in Ecology. *Ecol Modell*, 135, 147–186.

Harley, C.D.G. & Helmuth, B.S.T. (2003). Local- and Regional-Scale Effects of Wave Exposure, Thermal Stress, and Absolute Versus Effective Shore Level on Patterns of Intertidal Zonation. *Limnol Oceanogr*, 48, 1498–1508.

Harrell, F.E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer.

Harrison, X.A. (2014). Using Observation-level Random Effects to Model Overdispersion in Count Data in Ecology and Evolution. *PeerJ*, 2014.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY.

Hastings, A. (2004). Transients: The Key to Long-term Ecological Understanding? *Trends Ecol Evol*, 19, 39–45.

He, F. & Gaston, K.J. (2003). Occupancy, Spatial Variance, and the Abundance of Species. *Am Nat*, 162, 366–375.

Hébert, K. & Gravel, D. (2023). The Living Planet Index's Ability to Capture Biodiversity Change from Uncertain Data. *Ecology*, 104, e4044.

Hilborn, R. & Walters, C.J. (1992). *Quantitative Fisheries Stock Assessment*. Springer US, Boston, MA.

Hillebrand, H. (2004). On the Generality of the Latitudinal Diversity Gradient. *Am Nat*, 163, 192–211.

Hillebrand, H., Bennett, D.M. & Cadotte, M.W. (2008). Consequences of Dominance: A Review of Evenness Effects on Local and Regional Ecosystem Processes. *Ecology*, 89, 1510–1520.

HilleRisLambers, J., Adler, P.B., Harpole, W.S., Levine, J.M. & Mayfield, M.M. (2012). Rethinking Community Assembly Through the Lens of Coexistence Theory. *Annu Rev Ecol Evol Syst*, 43, 227–248.

Holyoak, M. & Leibold, M. (2006). *Metacommunities: Spatial Dynamics and Ecological Communities*. University of Chicago Press.

Hooper, D.U., Chapin, F.S., Ewel, J.J., Hector, A., Inchausti, P., Lavorel, S., *et al.* (2005). Effects of Biodiversity on Ecosystem Functioning: A Consensus of Current Knowledge. *Ecol Monogr*, 75, 3–35.

Hortal, J., De Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annu Rev Ecol Evol Syst*, 46, 523–549.

Hughes, T.P., Baird, A.H., Bellwood, D.R., Card, M., Connolly, S.R., Folke, C., *et al.* (2003). Climate Change, Human Impacts, and the Resilience of Coral Reefs. *Science (1979)*, 301, 929–933.

Hui, F.K.C., Warton, D.I., Foster, S.D. & Dunstan, P.K. (2013). To Mix or not to Mix: Comparing the Predictive Performance of Mixture Models vs. Separate Species Distribution Models. *Ecology*, 94, 1913–1919.

Hui, F.K.C., Warton, D.I., Ormerod, J.T., Haapaniemi, V. & Taskinen, S. (2017). Variational Approximations for Generalized Linear Latent Variable Models. *Journal of Computational and Graphical Statistics*, 26, 35–43.

Jackson, D.A. & Harvey, H.H. (1997). Qualitative and Quantitative Sampling of Lake Fish Communities. *Canadian Journal of Fisheries and Aquatic Sciences*, 54, 2807–2813.

Jackson, H.B. & Fahrig, L. (2015). Are Ecologists Conducting Research at the Optimal Scale? *Global Ecology and Biogeography*, 24, 52–63.

Kéry, M. & Royle, J.A. (2015). *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS*. Elsevier.

Khattar, G., Macedo, M., Monteiro, R. & Peres-Neto, P. (2021). Determinism and Stochasticity in the Spatial–temporal Continuum of Ecological Communities: The Case of Tropical Mountains. *Ecography*, 44, 1391–1402.

Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., García, E.A., *et al.* (2018). Building Essential Biodiversity Variables (EBVs) of Species Distribution and Abundance at a Global Scale. *Biological Reviews*, 93, 600–625.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., Mcinerny, G.J., *et al.* (2012). Towards Novel Approaches to Modelling Biotic Interactions in Multispecies Assemblages at Large Spatial Extents. *J Biogeogr*, 39, 2163–2178.

Kline, R.B. (2015). *Principles and Practice of Structural Equation Modeling*. 4th edn. Guilford Press.

de Knegt, H.J., van Langevelde, F., Coughenour, M.B., Skidmore, A.K., de Boer, W.F., Heitkönig, I.M.A., *et al.* (2010). Spatial Autocorrelation and the Scaling of Species–Environment Relationships. *Ecology*, 91, 2455–2465.

Kraft, N.J.B., Adler, P.B., Godoy, O., James, E.C., Fuller, S. & Levine, J.M. (2015). Community Assembly, Coexistence and the Environmental Filtering Metaphor. *Funct Ecol*, 29, 592–599.

Latimer, A.M., Wu, S., Gelfand, A.E. & Silander Jr., J.A. (2006). Building Statistical Models to Analyze Species Distributions. *Ecological Applications*, 16, 33–50.

Law, T. (2022). *Reducing Context-Dependency in Ecology: Environmental Variation Leads to Predictable Patterns of Species Associations Across Local Communities*.

Lawson, C.R., Hodgson, J.A., Wilson, R.J. & Richards, S.A. (2014). Prevalence, Thresholds and the Performance of Presence-absence Models. *Methods Ecol Evol*, 5, 54–64.

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

Legendre, P. & De Cáceres, M. (2013). Beta Diversity as the Variance of Community Data: Dissimilarity Coefficients and Partitioning. *Ecol Lett*, 16, 951–963.

Legendre, P. & Fortin, M.-J. (1989). Spatial Pattern and Ecological Analysis. *Vegetatio*, 80, 107–138.

Legendre, P. & Legendre, L. (2012). *Numerical Ecology*. 3rd edn. Elsevier.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., *et al.* (2004). The Metacommunity Concept: A Framework for Multi-scale Community Ecology. *Ecol Lett*, 7, 601–613.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I. & Delacoste, M. (1996). Role of Some Environmental Variables in Trout Abundance Models Using Neural Networks. *Aquat Living Resour*, 9, 23–29.

Lester, N.P., Sandstrom, S., de Kerckhove, D.T., Armstrong, K., Ball, H., Amos, J., *et al.* (2021). Standardized Broad-Scale Management and Monitoring of Inland Lake Recreational Fisheries: An Overview of the Ontario Experience. *Fisheries (Bethesda)*, 46, 107–118.

Leung, B., Hargreaves, A.L., Greenberg, D.A., McGill, B., Dornelas, M. & Freeman, R. (2020). Clustered Versus Catastrophic Global Vertebrate Declines. *Nature*, 588, 267–271.

Levin, S.A. (1998). Ecosystems and the Biosphere as Complex Adaptive Systems. *Ecosystems*, 5, 431–436.

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

Lindenmayer, D.B. & Likens, G.E. (2010). *Effective Ecological Monitoring*. CSIRO Publishing.

Lindenmayer, D.B., Likens, G.E., Krebs, C.J. & Hobbs, R.J. (2010). Improved Probability of Detection of Ecological "Surprises." *Proc Natl Acad Sci U S A*, 107, 21957–21962.

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005). Selecting Thresholds of Occurrence in the Prediction of Species Distributions, 28, 385–393.

Lobo, J.M., Jiménez-valverde, A. & Real, R. (2008). AUC: A Misleading Measure of the Performance of Predictive Distribution Models. *Global Ecology and Biogeography*.

Loreau, M. (2010). Linking Biodiversity and Ecosystems: Towards a Unifying Ecological Theory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365, 49–60.

Loreau, M. & de Mazancourt, C. (2013). Biodiversity and Ecosystem Stability: A Synthesis of Underlying Mechanisms. *Ecol Lett*, 16, 106–115.

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J.P., Hector, A., *et al.* (2001). Biodiversity and Ecosystem Functioning: Current Knowledge and Future Challenges. *Science (1979)*, 294, 804–808.

Loreau, Michel., Naeem, Shahid. & Inchausti, Pablo. (2002). *Biodiversity and Ecosystem Functioning: Synthesis and Perspectives*. Oxford University Press.

MacArthur, R. & Levins, R. (1967). The Limiting Similarity, Convergence, and Divergence of Coexisting Species. *Am Nat*, 101, 377–385.

MacArthur, R.H. (1965). Patterns of Species Diversity. *Biological Reviews*, 40, 510–533.

Mace, G.M. & Baillie, J.E.M. (2007). The 2010 Biodiversity Indicators: Challenges for Science and Policy. *Conservation Biology*, 21, 1406–1413.

Mack, R.N., Simberloff, D., Lonsdale, W.M., Evans, H., Clout, M. & Bazzaz, F.A. (2000). Biotic Invasions: Causes, Epidemiology, Global Consequences, and Control. In: *Ecological Applications*. pp. 689–710.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, A.A. & Langtimm, C.A. (2002). Estimating Site Occupancy Rates when Detection Probabilities are Less Than One. *Ecology*, 83, 2248–2255.

De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D.L. (2000). The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems*, 50, 1–18.

Magnuson, J.J., Kratz, T.K. & Benson, B.J. (2005). *Long-Term Dynamics of Lakes in the Landscape: Long-Term Ecological Research on North Temperate Lakes*. Oxford University Press.

Magurran, A.E. & Henderson, P.A. (2003). Explaining the Excess of Rare Species in Natural Species Abundance Distributions. *Nature*, 422, 714–716.

Magurran, A.E. & McGill, B.J. (2011). *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press.

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

Marchant, R. (1990). Robustness of Classification and Ordination Techniques Applied to Macroinvertebrate Communities from the La Trobe River, Victoria. *Mar Freshw Res*, 41, 493–504.

Marra, G. & Wood, S.N. (2011). Practical Variable Selection for Generalized Additive Models. *Comput Stat Data Anal*, 55, 2372–2387.

May, R.M. (1972). Will a Large Complex System be Stable? *Nature*, 238, 413–414.

McCune, B. & Grace, J. (2002). *Analysis of Ecological Communities*. MJM Software, Gleneden Beach, Oregon.

McGarigal, K., Stafford, S. & Cushman, S. (2000). *Multivariate Statistics for Wildlife and Ecology Research*. Springer New York, New York, NY.

McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. (2006). Rebuilding Community Ecology from Functional Traits. *Trends Ecol Evol*, 21, 178–185.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., *et al.* (2007). Species Abundance Distributions: Moving Beyond Single Prediction Theories to Integration Within an Ecological Framework. *Ecol Lett*, 10, 995–1015.

McInerny, G.J. & Purves, D.W. (2011). Fine-scale Environmental Variation in Species Distribution Modelling: Regression Dilution, Latent Variables and Neighbourly Advice. *Methods Ecol Evol*, 2, 248–257.

Mills, L.S. & Doak, D.F. (1993). The Keystone-Species Concept in Ecology and Conservation. *Bioscience*, 43, 219–224.

Mindrila, D. (2023). Bayesian Latent Class Analysis: Sample Size, Model Size, and Classification Precision. *Mathematics*, 11, 1–18.

Mouillot, D., Bellwood, D.R., Baraloto, C., Chave, J., Galzin, R., Harmelin-Vivien, M., *et al.* (2013). Rare Species Support Vulnerable Functions in High-Diversity Ecosystems. *PLoS Biol*, 11, e1001569.

Mouquet, N., Munguia, P., Kneitel, J.M. & Miller, T.E. (2003). Community Assembly Time and the Relationship Between Local and Regional Species Richness. *Oikos*, 103, 618–626.

Myers, R.A. (1998). When Do Environment-recruitment Correlations Work? *Rev Fish Biol Fish*, 8, 285–305.

Navarro, L.M., Fernández, N., Guerra, C., Guralnick, R., Kissling, W.D., Londoño, M.C., *et al.* (2017). Monitoring Biodiversity Change Through Effective Global Coordination. *Curr Opin Environ Sustain*, 29, 158–169.

Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized Linear Models. *J R Stat Soc Ser A*, 135, 370.

Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2019). Efficient Estimation of Generalized Linear Latent Variable Models. *PLoS One*, 14, e0216129.

Noss, R.F. (1990). Indicators for Monitoring Biodiversity: A Hierarchical Approach. *Conservation Biology*, 4, 355–364.

Nylund, K.L., Asparouhov, T. & Muthén, B.O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: a Monte Carlo Simulation Study. *Struct Equ Modeling*, 14, 535–569.

Odum, E.P. (1969). The Strategy of Ecosystem Development. *Science (1979)*, 164, 262–270.

Öğlü, B., Bhele, U., Järvalt, A., Tuvikene, L., Timm, H., Seller, S., *et al.* (2020). Is Fish Biomass Controlled by Abiotic or Biotic Factors? Results of Long-term Monitoring in a Large Eutrophic Lake. *J Great Lakes Res*, 46, 881–890.

Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., *et al.* (2024). vegan: Community Ecology Package.

Olin, M., Malinen, T. & 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, 88–94.

Olkeba, B.K., Boets, P., Mereta, S.T., Yeshigeta, M., Akessa, G.M., Ambelu, A., *et al.* (2020). Environmental and Biotic Factors Affecting Freshwater Snail Intermediate Hosts in the Ethiopian Rift Valley Region. *Parasit Vectors*, 13.

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

Ontario Ministry of Natural Resources and Forestry (OMNRF). (2012). Dataset.

Ovaskainen, O. & Abrego, N. (2020). *Joint Species Distribution Modelling*. Cambridge University Press.

Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. (2016a). Using Latent Variable Models to Identify Large Networks of Species-to-species Associations at Different Spatial Scales. *Methods Ecol Evol*, 7, 549–555.

Ovaskainen, O., Hottola, J. & Siitonen, J. (2010). Modeling Species Co-occurrence by Multivariate Logistic Regression Generates New Hypotheses on Fungal Interactions. *Ecology*, 91, 2514–2521.

Ovaskainen, O., Roy, D.B., Fox, R. & Anderson, B.J. (2016b). Uncovering Hidden Spatial Structure in Species Communities with Spatially Explicit Joint Species Distribution Models. *Methods Ecol Evol*, 7, 428–436.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., *et al.* (2017). How to Make More Out of Community Data? A Conceptual Framework and its Implementation as Models and Software. *Ecol Lett*, 20, 561–576.

Overton, J.M., Stephens, R.T.T., Leathwick, J.R. & Lehmann, A. (2002). Information Pyramids for Informed Biodiversity Conservation. *Biodivers Conserv*, 11, 2093–2116.

Paine, R.T. (1966). Food Web Complexity and Species Diversity. *Am Nat*, 100, 65–75.

Paine, R.T. (1969). A Note on Trophic Complexity and Community Stability. *Am Nat*, 103, 91–93.

Parmesan, C. (2006). Ecological and Evolutionary Responses to Recent Climate Change. *Annu Rev Ecol Evol Syst*, 37, 637–669.

Parmesan, C. & Yohe, G. (2003). A Globally Coherent Fingerprint of Climate Change Impacts Across Natural Systems. *Nature*, 421, 37–42.

Pauly, D. & Christensen, V. (1995). Primary Production Required to Sustain Global Fisheries. *Nature*, 374, 255–257.

Pauly, D., Christensen, V., Dalsgaard, J., Froese, R. & Torres, F. (1998). Fishing Down Marine Food Webs. *Science (1979)*, 279, 860–863.

Pearson, R.G. & Dawson, T.P. (2003). Predicting the Impacts of Climate Change on the Distribution of Species: Are Bioclimate Envelope Models Useful? *Global Ecology and Biogeography*, 12, 361–371.

Pecl, G.T., Araújo, M.B., Bell, J.D., Blanchard, J., Bonebrake, T.C., Chen, I.-C., *et al.* (2017). Biodiversity Redistribution under Climate Change: Impacts on Ecosystems and Human Well-being. *Science (1979)*, 355, eaai9214.

Peirce, C.S. (1884). The Numerical Measure of the Success of Predictions. *Science*, 4, 453–454.

Pereira, H.M. & David Cooper, H. (2006). Towards the Global Monitoring of Biodiversity Change. *Trends Ecol Evol*, 21, 123–129.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., *et al.* (2013). Essential Biodiversity Variables. *Science (1979)*, 339, 277–278.

Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006). Variation Partitioning of Species Data Matrices: Estimation and Comparison of Fractions. *Ecology*, 87, 2614–2625.

Peres-Neto, P.R., Leibold, M.A. & Dray, S. (2012). Assessing the Effects of Spatial Contingency and Environmental Filtering on Metacommunity Phylogenetics. *Ecology*, 93, S14–S30.

Persson, L. (2008). Community Ecology of Freshwater Fishes. In: *Handbook of Fish Biology and Fisheries, Volume 1*. Blackwell Publishing Ltd, Oxford, UK, pp. 321–340.

Phillips, S.J. & Elith, J. (2013). On Estimating Probability of Presence from Use–Availability or Presence–Background Data. *Ecology*, 94, 1409–1419.

Pimm, S.L. (1984). The Complexity and Stability of Ecosystems. *Nature*, 307, 321–326.

Pimm, S.L., Jenkins, C.N., Abell, R., Brooks, T.M., Gittleman, J.L., Joppa, L.N., *et al.* (2014). The Biodiversity of Species and their Rates of Extinction, Distribution, and Protection. *Science (1979)*, 344, 1246752.

Piñeiro, G., Perelman, S., Guerschman, J.P. & Paruelo, J.M. (2008). How to Evaluate Models: Observed vs. Predicted or Predicted vs. Observed? *Ecol Modell*, 216, 316–322.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., *et al.* (2014). Understanding Co-occurrence by Modelling Species Simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol Evol*, 5, 397–406.

Poos, M.S. & Jackson, D.A. (2012). Addressing the Removal of Rare Species in Multivariate Bioassessments: The Impact of Methodological Choices. *Ecol Indic*, 18, 82–90.

Popovic, G.C., Hui, F.K.C. & Warton, D.I. (2018). A General Algorithm for Covariance Modeling of Discrete Data. *J Multivar Anal*, 165, 86–100.

Popovic, G.C., Hui, F.K.C. & Warton, D.I. (2022). Fast Model-based Ordination with Copulas. *Methods Ecol Evol*, 13, 194–202.

Popovic, G.C., Warton, D.I., Thomson, F.J., Hui, F.K.C. & Moles, A.T. (2019). Untangling Direct Species Associations from Indirect Mediator Species Effects with Graphical Models. *Methods Ecol Evol*, 10, 1571–1583.

R Core Team. (2017). R: A Language and Environment for Statistical Computing.

Rees, H.C., Maddison, B.C., Middleditch, D.J., Patmore, J.R.M. & Gough, K.C. (2014). The Detection of Aquatic Animal Species Using Environmental DNA - a Review of eDNA as a Survey Tool in Ecology. *Journal of Applied Ecology*, 51, 1450–1459.

Reiss, J., Bridle, J.R., Montoya, J.M. & Woodward, G. (2009). Emerging Horizons in Biodiversity and Ecosystem Functioning Research. *Trends Ecol Evol*, 24, 505–514.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., *et al.* (2017). Cross-validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography*.

Roesti, M., Anstett, D.N., Freeman, B.G., Lee-Yaw, J.A., Schluter, D., Chavarie, L., *et al.* (2020). Pelagic Fish Predation is Stronger at Temperate Latitudes than near the Equator. *Nat Commun*, 11, 1527.

Royle, J.A. & Dorazio, R.M. (2008). *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. *Hierarchical Modeling and Inference in Ecology*. Academic Press.

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

Schemske, D.W., Mittelbach, G.G., Cornell, H. V., Sobel, J.M. & Roy, K. (2009). Is There a Latitudinal Gradient in the Importance of Biotic Interactions? *Annu Rev Ecol Evol Syst*, 40, 245–269.

Shannon, C.E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27, 623–656.

Shurin, J.B. (2000). Dispersal Limitation, Invasion Resistance, and the Structure of Pond Zooplankton Communities. *Ecology*, 81, 3074.

Simpson, G. (2024). gratia: Graceful ggplot-Based Graphics and Other Functions for GAMs Fitted Using mgcv.

Smith, V.H. (1998). Cultural Eutrophication of Inland, Estuarine, and Coastal Waters. In: *Successes, Limitations, and Frontiers in Ecosystem Science*. Springer New York, New York, NY, pp. 7–49.

Sobrino, I., Rueda, L., Tugores, M.P., Burgos, C., Cojan, M. & 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.

Stahl, A., Pedersen, E.J. & Peres-Neto, P.R. (2024). Advancing Single Species Abundance Models: Robust Models for Predicting Abundance Using Co-occurrence from Communities. *EcoEvoRxiv [preprint]*.

Strayer, D.L. & Findlay, S.E.G. (2010). Ecology of Freshwater Shore Zones. *Aquat Sci*, 72, 127–163.

Sutherland, W.J. (2006). *Ecological Census Techniques: A Handbook*. Cambridge University Press.

Thompson, P.L. & Gonzalez, A. (2017). Dispersal Governs the Reorganization of Ecological Networks under Environmental Change. *Nat Ecol Evol*, 1.

Thorson, J.T., Pinsky, M.L. & Ward, E.J. (2016). Model-Based Inference for Estimating Shifts in Species Distribution, Area Occupied and Centre of Gravity. *Methods Ecol Evol*, 7, 990–1002.

Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T. & Prentice, I.C. (2005). Climate Change Threats to Plant Diversity in Europe. *Proceedings of the National Academy of Sciences*, 102, 8245–8250.

Tilman, D. (1994). Competition and Biodiversity in Spatially Structured Habitats. *Ecology*, 75, 2–16.

Tilman, D. (1996). Biodiversity: Population Versus Ecosystem Stability. *Ecology*, 77, 350–363.

Tilman, D. (2020). *Resource Competition and Community Structure*. Princeton University Press.

Tilman, D., Isbell, F. & Cowles, J.M. (2014). Biodiversity and Ecosystem Functioning. *Annu Rev Ecol Evol Syst*, 45, 471–493.

Tittensor, D.P., Walpole, M., Hill, S.L.L., Boyce, D.G., Britten, G.L., Burgess, N.D., *et al.* (2014). A Mid-Term Analysis of Progress Toward International Biodiversity Targets. *Science (1979)*, 346, 241–244.

Tofallis, C. (2015). A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation. *Journal of the Operational Research Society*, 66, 1352–1362.

Trebilco, R., Baum, J.K., Salomon, A.K. & Dulvy, N.K. (2013). Ecosystem Ecology: Size-Based Constraints on the Pyramids of Life. *Trends Ecol Evol*, 28, 423–431.

Tunney, T.D., Carpenter, S.R. & Vander Zanden, M.J. (2017). The Consistency of a Species' Response to Press Perturbations with High Food Web Uncertainty. *Ecology*, 98, 1859–1868.

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. & Wardle, D.A. (2008). Global Change and Species Interactions in Terrestrial Ecosystems. *Ecol Lett*.

Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003). Improving Precision and Reducing Bias in Biological Surveys: Estimating False-Negative Error Rates. *Ecological Applications*, 13, 1790–1801.

Urban, M.C., De Meester, L., Vellend, M., Stoks, R. & Vanoverbeke, J. (2012). A Crucial Step Toward Realism: Responses to Climate Change from an Evolving Metacommunity Perspective. *Evol Appl*, 5, 154–167.

VanDerWal, J., Shoo, L.P., Johnson, C.N. & 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, 282–291.

Vellend, M. (2010). Conceptual Synthesis in Community Ecology. *Quarterly Review of Biology*, 85, 183–206.

Villéger, S., Brosse, S., Mouchet, M., Mouillot, D. & Vanni, M.J. (2017). Functional Ecology of Fish: Current Approaches and Future Challenges. *Aquat Sci*, 79, 783–801.

Wagner, H.H. & Fortin, M.-J. (2005). Spatial Analysis of Landscapes: Concepts and Statistics. *Ecology*, 86, 1975–1987.

Waldock, C., Stuart-Smith, R.D., Albouy, C., Cheung, W.W.L., Edgar, G.J., Mouillot, D., *et al.* (2022). A Quantitative Review of Abundance-Based Species Distribution Models. *Ecography*, 2022.

Walker, S.C. & Jackson, D.A. (2011). Random-effects Ordination: Describing and Predicting Multivariate Correlations and Co-occurrences. *Ecol Monogr*, 81, 635–663.

Walther, G.-R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T.J.C., *et al.* (2002). Ecological Responses to Recent Climate Change. *Nature*, 416, 389–395.

Wang, Y., Naumann, U., Eddelbuettel, D., Wilshire, J. & Warton, D. (2022). mvabund: Statistical Methods for Analysing Multivariate Abundance Data.

Ware, D.M. & Thomson, R.E. (2005). Bottom-Up Ecosystem Trophic Dynamics Determine Fish Production in the Northeast Pacific. *Science (1979)*, 308, 1280–1284.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., *et al.* (2015a). So Many Variables: Joint Modeling in Community Ecology. *Trends Ecol Evol*, 30, 766–779.

Warton, D.I., Foster, S.D., De'ath, G., Stoklosa, J. & Dunstan, P.K. (2015b). Model-Based Thinking for Community Ecology. *Plant Ecol*, 216, 669–682.

Warton, D.I., Wright, S.T. & Wang, Y. (2012). Distance-Based Multivariate Analyses Confound Location and Dispersion Effects. *Methods Ecol Evol*, 3, 89–101.

Weller, B.E., Bowen, N.K. & Faubert, S.J. (2020). Latent Class Analysis: A Guide to Best Practice. *Journal of Black Psychology*, 46, 287–311.

Wenger, S.J. & Olden, J.D. (2012). Assessing Transferability of Ecological Models: An Underappreciated Aspect of Statistical Validation. *Methods Ecol Evol*, 3, 260–267.

White, E.P., Ernest, S.K.M., Kerkhoff, A.J. & Enquist, B.J. (2007). Relationships Between Body Size and Abundance in Ecology. *Trends Ecol Evol*.

Whittaker, R.H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol Monogr*, 30, 279–338.

Willig, M.R., Kaufman, D.M. & Stevens, R.D. (2003). Latitudinal Gradients of Biodiversity: Pattern, Process, Scale, and Synthesis. *Annu Rev Ecol Evol Syst*, 34, 273–309.

Willmott, C.J. & Matsuura, K. (2005). Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Research*, 30, 79–82.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., *et al.* (2008). Effects of Sample Size on the Performance of Species Distribution Models. *Divers Distrib*, 14, 763–773.

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

Wood, S.N. (2004). Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *J Am Stat Assoc*, 99, 673–686.

Wood, S.N. (2006). Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. *Biometrics*, 62, 1025–1036.

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, 3–36.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*. 2nd edn. Chapman and Hall/CRC.

Wood, S.N., Pya, N. & Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *J Am Stat Assoc*, 111, 1548–1563.

Wright, D.H. (1991). Correlations Between Incidence and Abundance are Expected by Chance. *J Biogeogr*, 18, 463–466.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001). Monitoring of Biological Diversity in Space and Time. *Trends Ecol Evol*, 16, 446–453.

Zou, H., Hastie, T. & Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15, 265–286.

Zou, H. & Zhang, H.H. (2009). On the Adaptive Elastic-net with a Diverging Number of Parameters. *Ann Stat*, 37, 1733–1751.