Reducing Context-Dependency in Ecology: Environmental Variation Leads to Predictable Patterns of Species Associations Across Local Communities

Timothy Law

A Thesis in the Department of
Biology

Presented in Partial Fulfillment of the Requirements
For the Degree of
Master of Science (Biology)
At Concordia University
Montréal, Québec, Canada

March, 2022

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Timothy Law

Entitled: Reducing Context-Dependency in Ecology: Environmental Variation Leads to Predictable Patterns of Species Associations Across Local Communities

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

Master of Science (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. Eric Pederson

_____ Examiner
Dr. James Grant

_____ External Examiner
Dr. David Walsh

_____ Co-Supervisor
Dr. Dylan Fraser

_____ Supervisor
Dr. Pedro Peres-Neto

Approved by _____
Dr. Robert Weladji, Graduate Program Director

_____
Pascale Sicotte, Dean of Faculty

Date _____

Abstract

Reducing Context-Dependency in Ecology: Environmental Variation Leads to Predictable Patterns of Species Associations Across Local Communities

Timothy Law

Predicting how communities change in space and time requires an understanding of how mechanisms such as environmental filtering and biotic interactions shape distributions and abundances. However, ecological communities are often context dependent, so mechanisms that are important in certain environments may not be in others. To understand how context dependency affects the prediction of community structure, we asked: Can the environment predict the outcomes of community assembly? Using fish association networks estimated with Markov random fields for over 700 lakes in Ontario, Canada, we tested if species association patterns, representing potential community assembly mechanisms, varied as a function of the environment. We examined the effect of the environment at two scales: pairwise and community level, summarizing potential mechanisms between species and across whole communities. The environment was a strong predictor of community level species association patterns but not of pairwise patterns, suggesting that the cumulative outcome of mechanisms structuring communities can be explained by the environment. We then tested if community level patterns were associated with the uniqueness of a lake's species composition. We found that as species association patterns became stronger, they lead to lakes with more common species compositions. Taken together, our results show that variation in the outcome of community assembly can be explained by the environment using community level patterns, offering a way for community ecologists to study context-dependency in community structure across differing environmental gradients and species compositions.

Acknowledgements

I would like to first thank the whole lab of Community and Quantitative Ecology, both past and present for their continual support and encouragement. This thesis would not be half of what it is without all the input and engaging conversations with everyone, especially Alexandra Engler Alienor Stahl and Gabriel Khattar. Thank you for supporting me through thick and thin, even when my ramblings made you scratch your heads. For my friends, family and loved ones who are brave enough to read this, thank you for being my rock during my thesis. You have always supported me in my endeavours, and provided an environment where I could be myself. You have all helped to shape me into the person I am today. I would also like to thank, NSERC-CGS, and BIOS2 for funding my research. The connections I made at BIOS2 and opportunities I was able to take have helped me become a more well-rounded scientist.

Finally, this would none of this would have been possible without the invaluable time, commitment and support from my supervisors. Thank you to my co-supervisor Dr. Dylan Fraser, for your contributions in shaping my project, without which this would not have been possible. I leave my sincerest gratitude to my supervisor Dr. Pedro Peres-Neto for dedicating his time and for continually challenging me, and helping me think outside of the box. Your passion for science and ecology has helped me and will continue to help me grow as a scientist.

Table of Contents

List of Tables and Figures

In-text Tables and Figures

Figures in appendices

List of Symbols and Abbreviations

Symbols & abbreviations	Definition
SAP	Species association patterns
PPCA	Probabilistic principal component analysis
LASSO	Least absolute shrinkage and selection operator – variable selection and regularization
SES_{mean}	Standardized effect size of the mean species association patterns in a community
SES_{SD}	Standardized effect size of the standard deviation (SD) of species association patterns in a community
LCBD	Local Contribution to Beta Diversity
SES_{LCBD}	Standardized effect size of LCBD
Constrained	Null model framework where species were shuffled within primary watersheds
Unconstrained	Null model framework where species were shuffled across all lakes

Introduction

Unravelling the processes that assemble local ecological communities has long remained a central question in community ecology (Diamond 1975, Leibold et al. 2004). Much of the focus has centred around whether communities are assembled randomly or non-randomly (Connor and Simberloff 1979, Gotelli 2000, Chase and Myers 2011). Many of these studies are aimed at identifying the mechanisms underlying specific systems of interest (e.g., groups of communities within a landscape), such as competition, predator-prey dynamics and environmental selection (Gotelli and McCabe 2002, Sfenthourakis et al. 2006). However, the importance of a specific mechanism within the same taxa, landscape, or region varies frequently – often described as context-dependency (or contingency) where the effect of a mechanism driving species memberships in a community can change as a function of abiotic and biotic drivers (Chamberlain et al. 2014, see Leibold et al. 2021 and Catford et al. 2021 for discussions). Instead of focusing on identifying general mechanisms that structure specific communities, focusing on the categorization of the contexts that mechanisms are contingent on, can help us understand the "local" nature of community ecology (Simberloff 2004). As local communities are contingent on a combination of interacting factors that lead to a large but finite number of states, identifying and understanding different drivers, their interactions and how they modulate the strength of different mechanisms should improve our understanding of community assembly theory.

A framework that could increase our ability towards generalization is the analysis of species co-occurrence patterns, where the quantification of the levels of aggregation, segregation, or randomness among species are used extensively to investigate community assembly mechanisms (See box 1 for details). There are at least two forms of important contingencies in community ecology. One refers to the outcomes of different mechanisms underlying patterns of species associations in which similar species association patterns (positive or negative; herein denoted SAPs) can be attributed to environmental selection or competition, depending on the environment (Connor and Simberloff 1983, Peres-Neto et al. 2001, Cadotte and Tucker 2017). Although the challenges of disentangling multiple mechanisms is frequently used as a criticism against describing community structure based on patters of co-occurrence (i.e., inferring mechanisms from patterns; Freilich et al. 2018, Blanchet et al. 2020), the ability to determine how multiple factors interact to influence patterns of species associations may provide a useful generalization. It allows for these patterns to be contrasted within and across landscapes and/or taxa and can assist in uncovering general versus context-

dependent predictions and patterns. For example, the emergence of a common set of predictors across multiple landscapes or taxa could be useful for generalization. The second form of context-dependency across landscapes that we bring to light here is that the strength of species associations may vary across environments (i.e., non-stationarity in species associations across environmental gradients). Newer methods like Markov network models with covariates (e.g., environmental predictors) allow for modelling context-dependency in species associations as a function of environmental variation in space and time (Clark et al. 2018).

Here, we build a quantitative framework to determine whether and how environmental variables and landscape properties explain local pairwise SAPs and community-level SAPs. To describe community-level SAPs, we propose the use of two statistical moments, mean and variance, to aggregate and quantify patterns across all species-pairs associations within single local communities. Mean and variance can be contrasted across communities within the same region and across regions offering a way to understand how environmental and landscape features influence communities regardless of ecosystem and taxa. The ability to improve on generalization also contributes to uncovering the mechanisms underlying community assembly. For example; Do extreme environments lead to smaller variation in the strength of species associations found within local communities and do average environments lead to weaker average species associations but with large variance? That is, are strong positively and negatively associated species found within the same local community in average environments? As such, strong (negative or positive) and random SAPs can be used to quantitatively assess which models based in one landscape with a set of specific species can predict SAPs in other landscapes with similar environments but different species compositions (i.e., communities sharing no species in common can have similar SAPs). By using SAPs, we can compare communities not by their species compositions but by how they are structured. This increases our ability to generalize across communities composed of different taxa and across different environments. Here, we modelled how environmental factors could predict (a) SAPs across all pairs, and (b) the mean and variance of SAPs across individual communities. Lastly, we assessed whether uniqueness (or commonness) in species compositions could predict our metrics of community structure (mean and variance) built from SAPs reflecting different community assembly mechanisms. This is relevant because it allows us to further determine the deterministic nature of species associations. One should expect that lakes composed of weaker (more random) SAPs, should result in more random combinations of species and thus be more uncommon than lakes composed of stronger species associations.

We used a comprehensive dataset of over 700 lakes with taxonomic information of entire fish communities, spread out over a large latitudinal gradient in Ontario, Canada. The discrete nature of lakes has made them important study systems for macroecologists looking at metacommunity patterns, community structure and ecosystem functioning (Magnuson et al. 1998, Hortal et al. 2014). Well-defined dispersal barriers, discrete boundaries, and knowledge of colonization history (see Mandrak and Crossman 1992 and Mandrak 1995), help make it easier to define the context upon which community structure may depend (Olden et al. 2001, Peres-Neto 2004). A variety of factors, both abiotic (e.g., pH, temperature) and biotic (e.g., competition, predation) structure fish communities (Jackson et al. 2001, Sharma et al. 2011, Giam and Olden 2016), making these communities good study systems for evaluating how environmental factors can drive many potential mechanisms. By identifying the environments where similar combinations of SAPs can be found, we reframe the environment as a property which can be used to predict species associations across communities regardless of composition.

Methods

Study system

The fish community dataset comprising of 706 lakes, was obtained through the Ontario Broad-Scale Monitoring Program for Inland Lakes (BsM) (Sandstorm et al. 2013) conducted by the Ontario Ministry of Northern Development, Mines, Natural Resources and Forestry (OMNRF). Lakes spanning a latitudinal and longitudinal gradient from 43.06 ºN – 54.52 ºN and -95.06 ºE – -74.50 ºE were sampled once during the summer months (June to September) between 2008 and 2012, using a stratified-random, spatial sampling design (Lester et al. 2020) (Fig 1). Stratification levels were based on geographic area (Fisheries Management Zones), lake surface area and the presence of three recreationally important species: walleye (*Sander vitreus*), lake trout (*Salvelinus namaycush*) and brook trout (*Salvelinus fontinalis*) (Lester et al. 2020). Sampling was standardized following Sandstorm et al. (2013). A combination of two mesh gill net types was used to sample lakes: a small mesh net, developed in Ontario (Sandstorm et al. 2013) and a large mesh net following the standards set by the American Fisheries Society for detecting angler harvested fishes (Bonar et al. 2009) (see Sandstrom et al. 2013 for full mesh, gang, length, and height details). A spatially and depth stratified design was

used to place both net types and ensured that effort was evenly allocated across a lake in area and depth. Secchi depth, temperature and oxygen profiles of the lakes were also recorded. Water for chemical analysis was collected in the spring. Human activity on lakes was estimated using aerial counting of the number of boats (recreational vs angling), number of anglers in a boat, ice huts, and open ice fisherman. Together with environmental data from Environment Canada, a total of 89 environmental variables were recorded to describe the lakes (Table 1).

Prior to use in our proposed analytical framework, environmental data were summarized into fewer variables using a probabilistic principal component analysis (PPCA) as implemented in the R package pcaMethods (Stacklies et al. 2007). PPCA uses the same concepts as a more traditional principal component analysis, but differs in the way it combines an expectation-maximization (EM) algorithm with a probabilistic model to deal with missing data (Stacklies et al. 2007). Across the environmental dataset, 6.17% contained missing values, spread over 80 of the 89 environmental variables. Cross validation was used to determine the estimation error and optimal number of axes for missing value estimation. Estimation error for variables with missing values was calculated using the normalized root mean square of prediction (NRMSEP). Briefly, NRMSEP is the square difference between real and estimated values for a variable, normalized by within variable variance (Stacklies et al. 2007). For variables with a high NRMSEP (> 0.8), prediction error is high, so missing values were simply estimated by the variable mean value. By using the mean value in these cases, lakes with missing values likely became more similar and, as such, less uninformative, making our models less predictable and more conservative in their results. The optimal number of axes for missing value estimation was selected by minimizing the average NRMSEP across all estimated environmental variables. The final PPCA model was fit using 40 axes, with missing values for 5 variables (sulphate percentile, winter fishing hut count, winter open ice fishing count, summer shore fishing count and conservation land status) estimated using their means. With these parameters, the first 12 axes explaining 81% of the variation in the environmental dataset were retained. Prior to PPCA, variables were visually analyzed for normality. As distributions were found to be non-normal, variables were scaled and transformed systematically using the bestNormalize R package (Peterson 2018). Briefly, variables are transformed using a suite of included transformations, and the best transformation was selected based on the Pearson P statistic for Gaussianity (divided by its degrees of freedom). Variables with a value for the P statistic closer to 1 prior to transformation were not transformed (Table 1).

Estimating local species co-occurrence coefficients with Markov networks

We used Markov (binary) network models with covariates to estimate species co-occurrence coefficients for our fish communities. Markov models estimate the conditional relationships between species pairs (i.e., SAPs) while considering other covariates and represent these relationships as graphs. Here, species are represented by nodes and their associations are represented by edges. Co-occurrence strength is assigned to each edge to describe the strength and type of association (aggregated or segregated) between pairs of nodes. The absence of an edge signals an absence of a detected species pair association (i.e., species association is random - their association cannot be predicted by environmental features or by their co-distribution). We implemented Markov network models with covariates as described by Clark et al. (2018 and references within). This framework describes the increase in log-odds of observing species *j* given the presence-absence of species *k* and covariate *x,* which can be modelled as a logistic function:

$$\log\left[\frac{P(y_j = 1 \,|y_{\setminus j}, x)}{1 - P(y_j = 1 \,|y_{\setminus j}, x)}\right] = \alpha_{j0} + \beta_j x + \sum_{k:k\neq j} (\alpha_{jk} + \beta_{jk} x) y_k$$

where y_j is a vector of binary presence-absences for species *j*, and $y_{\setminus j}$ the vectors of binary presence-absences for all other species. α_{j0} is the species-level intercept and $\beta_j x$ are the coefficients estimating the effects of environmental covariates on the occurrence probability of species *j* (i.e., purely abiotic component). α_{jk} represents the regression coefficient of species *j* on the k^{th} species (i.e., the conditional relationship between species *j* and species *k*) and β_{jk} coefficients estimate the effects of the (statistical) interaction between each environmental predictor and the k^{th} species. Combined, parameters α_{jk} and $\beta_{jk} x$ describe the biotic components represented as conditional relationships between species *j* and the k^{th} species, as well as the effects of covariate *x* (environmental factors) on these relationships. If $\alpha_{jk} = 0$, the probability of occurrence of species *j* and *k* are conditionally independent after controlling for other species and environmental covariates. If $\alpha_{jk} \neq 0$ but $\beta_{jk} x = 0$, then the probability of occurrence of species *j* and the k^{th} *species* are conditionally dependent but the strength of their dependence (α_{jk}) does not vary as a function of the environmental covariate x. Assume there are 20 species and 15 environmental predictors (covariates); then, the model for a single

species would contain one intercept α_{j0}, 15 β_j coefficients (one for each environmental covariate), 19 α_{jk}coefficients (species *j* on each *k*th species) and 19 × 15 $\beta_{jk}x$ (statistical interactions between each species and each environmental covariate), adding to 320 predictors. To generate a more intuitive understanding of this Markov network model in estimating species co-occurrence coefficients, we created a few small, simulated examples (see Appendix I).

Model parameterization was estimated using linear logistic regression. Because the number of estimated coefficients grows quickly with the number of species and covariates, regularization was used to control for overfitting. Using LASSO regularization (Tibshirani 1996), model coefficients were forced to zero depending on the regularization parameter λ, adding sparsity while maintaining similar predictive power across possible combinations of predictors. 10-fold cross validation was used to identify optimal λ values. Note that conditional relationships in Markov models are symmetric (undirected); that is the conditional relationship of one species on another is the same in both directions. However, given the large number of species and/or covariates in our data, a common method for avoiding exponentially growing parameter estimation is to estimate parameters from a series of single-species regressions (as above) and combining them in a common matrix to approximate the Markov network (graph) (Cheng et al. 2014). Consequentially, higher order interactions (e.g., three-way interactions such as interactions between two or more species, and interactions between two species and environmental predictors) are not considered. Additionally, because parameters were estimated from separate regressions (i.e., one regression for each species), symmetry in model coefficients is not guaranteed between any given two species. Here, conditional relationships were made symmetric between any two given species (say *j* and *k*) by retaining coefficients with the larger absolute value between the separate logistic regressions having species *j* and *k* as responses, respectively (Meinshausen and Bühlmann 2006). Retaining the larger absolute value instead of the smallest makes estimates less conservative (less zero coefficients; Cheng et al. 2014). It is worth mentioning here some of the advantages of a Markov (binary) random field approach over the now common joint species distribution model (JSDM). First, it is technically more challenging to consider variation that is non-stationarity in space and/or time between pairs of species (i.e., $\beta_{jk}x$ coefficients). Second, because JSDMs use selected axes from latent models (e.g., ordination) of residual variation, it is not clear how comparable pairwise coefficients of species associations are (Tikhonov et al. 2017). One of the associate challenges is estimating the appropriate degrees of freedoms for penalizing parameters to make species pairs coefficients comparable among each other. Finally, in a review of different methods to

estimate species-pair coefficients by Popovic et al. (2019), Markov (binary) models was found to be very robust and overperform JSDMs.

This modelling framework was carried out using the R package MRFCov (Clark et al. 2018). Prior to model fitting, very common species (>95% prevalence) and rare species (<2.5% prevalence) were removed from the data as cross-validation typically results in large errors for species with a small and large prevalence. In total, 43 out of 87 observed species were considered in the final analyses. PPCA axes (12 in total explaining 81% of the environmental variation) calculated from the environmental variables (see above) were used as environmental covariates. Environmental covariates were standardized (mean of zero and standard deviation of one). To account for model uncertainty in the process of fitting the Markov model (i.e., estimation of coefficients and LASSO regularization), we randomly subsampled the data (without replacement) with a proportion of 90% and refitted the model as described earlier. This was repeated 1000 times and the final estimated parameters in the model are the mean values taken from all the bootstrapped models. If a particular predictor (environmental factor or species) was not included in the selection process via LASSO for a particular subsample, then the coefficient for that predictor was set to zero for that subsample when averaging over across all subsamples. To test the final (averaged over 1000 subsamples) global model performance (i.e., all single species models), a 10-fold cross validation, repeated 500 times, of the global model with and without environmental covariates was performed. This allowed us to estimate model performance and whether environmental covariates improved the fit of the model over the distribution of all species used to estimate conditional relationship between species pairs.

An advantage of estimating species associations using Markov models with covariates, is that we can use the global model to estimate SAPs for each community (lake) separately. The estimated parameters from the global model, describes species co-occurrence coefficients for average environmental conditions, and relationships between species co-occurrence coefficients and environmental covariates (factors; Fig. 2). For a given pair of species (*j* and *k*), their association is estimated as the log-odds of observing species *j* given species *k* and the (statistical) interactions between species *k* and all environmental covariates and the intercept, after controlling for all environmental predictors, all other species not in that pair (i.e., species ≠ *j* and *k*) and their interactions of these other species with the environmental covariates; and vice-versa (i.e., a model of *j* on *k* and *k* on *j*). By simply entering the observed values of the environmental covariates and species in a particular lake, SAPs can be predicted for observed

species pairs in that lake. As such, we were able to estimate SAPs for every observed species pairs across all lakes. To better understand the way in which species associations were estimated, consider three species *j*, *k* and *i*, and environmental covariates x_1 and x_2. To estimate the association between *j* and *k*, we start by estimating these two logistic regression models:

$$\log\left[\frac{P(y_j = 1 \,|y_{\setminus j}, x)}{1 - P(y_j = 1 \,|y_{\setminus j}, x)}\right] = \alpha_{j0} + \beta_j x_1 + \beta_j x_2 + \alpha_{jk} + \alpha_{ji} + \beta_{jk} x_1 y_k + \beta_{jk} x_2 y_k + \beta_{ji} x_1 y_i + \beta_{ji} x_2 y_i$$

$$\log\left[\frac{P(y_k = 1 \,|y_{\setminus k}, x)}{1 - P(y_k = 1 \,|y_{\setminus k}, x)}\right] = \alpha_{k0} + \beta_k x_1 + \beta_k x_2 + \alpha_{kj} + \alpha_{ki} + \beta_{kj} x_1 y_j + \beta_{kj} x_2 y_j + \beta_{ki} x_1 y_i + \beta_{ki} x_2 y_i$$

The conditional random fields (CRF) predicted species associations between species *j* and *k* for a given local community *c* is then:

$$CRF_{jk} = \max\left(abs(\alpha_{jk}, \alpha_{kj})\right) \times x_{c1} + \max\left(abs(\beta_{jk} x_2 y_k, \beta_{kj} x_2 y_j)\right) \times x_{C2}$$

where x_{C1} and x_{c2} are the values for environmental predictors 1 and 2, respectively, at the local community *c*. As already discussed, to assure symmetry, we pick the maximum absolute value between coefficients while maintaining its original sign (not shown directly in the equation for the sake of brevity). Note that: a) α_{kj} (or α_{jk}) are not multiplied by local conditions as it contributes to the overall expected changes in log of odds of observing species *j* and *k* together; b) the coefficients used to estimate CRF_{jk} are conditional on all other species (except the two species of interest), environmental covariates and statistical interactions between all other species (except the two species of interest) and environmental covariates. CRF_{jk} can be then interpreted as an estimate of the (additional) contribution to the expected changes in log of odds of observing species *j* and *k* together given their overall spatial association across all communities (α_{kj} or α_{jk}) plus the association given the local environmental conditions. As such, CRF_{jk} is composed of a global species associations component (i.e., α coefficients) and a local component (β coefficients). If the contribution of the local components is high over the global, this suggests that the association between the two species in question (*j* and *k*) changes as a function of the environment. Although we did not explore this venue for the sake of brevity, non-stationarity in species associations across environmental gradients is a form of context-dependence. Thus, dividing the local over the global contribution could serve as a potential

metric of context-dependency and an indicator of how “generalizable” a model is to other systems.

Linking species level (pairwise) patterns of species associations to environmental and historical factors

We used gradient boosted regression trees (see Hastie et al. 2009 for an introduction) to estimate the relationship between the environment and SAPs across all lakes. Gradient boosted regression trees were chosen over more traditional linear models for their relaxation of assumptions and other benefits – they can handle different types of predictor variables, require no prior data transformation or removal of outliers, and can fit complex non-linear relationships and automatically consider complex and multiple interaction effects (Elith et al. 2008). Gradient boosted regression tree models improve on the poor predictive performance of individual regression trees by fitting multiple regression trees sequentially, minimizing error using a loss function (Elith et al. 2008). Here, we used gradient boosted regression trees as implemented in the R package GBM (Greenwell et al. 2019) and used the R package caret (Kuhn 2008) to optimize regression tree parameters. We varied the number of trees to fit, shrinkage (i.e., learning rate), interaction depth and the minimum number of observations in each terminal node. Because our goal was to assess whether there was a relationship between environment factors and all conditional relationships, we optimized models for the highest model R^2 value. All environmental variables were used, and missing values were replaced with the mean of the respective variable. Unlike our Markov models that used PPCA to represent environmental variation and reduce model parameters, with the tree regressions we aimed to more directly interpret the environmental predictors and used all the environmental variables instead of PPCA axes.

Linking community-level patterns of species associations to environmental and historical factors

To investigate whether and how the measured environmental factors predict community-level patterns of SAPs, we first calculated the mean and spread (standard deviation SD) of the conditional relationships for observed species of each lake estimated using the Markov network models (i.e., the mean or SD of estimated SAPs for each lake) (Fig. 2). Communities with positive or negative means are comprised, respectively, mostly of species pairs that aggregate or segregate in a specific environment. Communities with means close to zero are comprised

mostly of either species pairs that have no clear association patterns in a specific environment, or by a mixture of positive and negative species pairs. To differentiate communities wherein associations are not observed from communities with a mix of positive and negative associations, we relied on the SD of association patterns around the mean. A community with a mean close to zero and a small SD is comprised of pairs with no association patterns. Conversely, a large SD would suggest that a wide range of positive and negative species associations exists, and that the community was not comprised mostly of species pairs with no association patterns.

Because the mean and standard deviation (SD) can be biased by the number of species in a lake (e.g., lakes with large number of species will tend to converge to similar mean values and low standard deviations), a null model was used to standardize means and SDs so that they could be contrasted across lakes (Ulrich et al. 2017, 2018) (Fig. 2). To generate an expected outcome based on species richness, the species matrix was randomized, keeping species and lake totals fixed using the curveball algorithm (Strona et al. 2014). Keeping species and lake totals fixed during randomization has been demonstrated to have appropriate Type I error rates, and is a preferred algorithm for data that are island-like (ex. lakes), where species-area effects are strong (Gotelli 2000). After each randomization, the global species co-occurrence model with environmental covariates was re-estimated. We then used the estimated parameters from the global model to predict conditional relationships for species pairs per lake, allowing us to then calculate an adjusted (by number of species) expected mean and SD of conditional relationships for each lake. Repeating this procedure 999 times, we generated a null distribution of means and SDs for each lake. As in traditional permutation procedures, the observed value was made part of the null distribution to estimate null means and SDs. For lake *j*, the observed mean and SD was scaled by estimating a standardized effect size using the null distribution of the mean and SD generated for lake *j*:

$$SES_{j.mean} = \frac{\overline{mean}_{obs} - \overline{mean}_{null}}{SD_{mean.null}} \qquad SES_{j.SD} = \frac{\overline{SD}_{obs} - \overline{SD}_{null}}{SD_{SD.null}}$$

where mean and SD are the metrics of interest, and $SD_{mean.null}$ and $SD_{SD.null}$ are standard deviations of each metrics' null distribution for lake *j*, respectively. By comparing the distribution of observed values against the distribution of expected values under the null model, we evaluated whether observed values were similar to those generated by mechanisms structuring random communities in respect to the chosen permutation algorithm. With the fixed-fixed

algorithm, communities are the result of the random colonization of sites with respect to species. Thus, standardized effect sizes that were close to 0 represented observed communities that have community structure expected by communities that were randomly colonized by species. Because the size of species pools can influence the ability to separate the effect of evolutionary and historic processes, analysing standardized effects instead of observed ones should increase our ability to describe more contemporary (ecological) processes on SAPs among species (Carstensen et al. 2013). In our study system, historical processes have shaped the distribution of fishes on the landscape, likely influencing the formation of species pools and subsequently local communities (Mandrak and Crossman 1992, Mandrak 1995, Henriques-Silva et al. 2013). To determine whether these historical factors may have influenced SAPs, we repeated the null model analysis framework, constraining the shuffling processes so that row and column totals within distinct biogeographic areas remained (again) fixed. In our study, we used primary watersheds as biogeographic areas, as they are representative of past dispersal and historical processes (Mandrak 1995, Olden et al. 2001).

To estimate the relationship between the community summaries (i.e., standardized community means and SDs) and the environment, a linear multiple regression model was used. All environmental variables were used, and missing values were replaced with the mean of the respective variable. Environmental variables were all scaled (mean of 0 and SD of 1) prior to analysis. Because of the large number of variables, LASSO regularization was used and a 5-fold cross validation, repeated 100 times, was used to identify optimal λ values for determining the shrinkage penalty.

Using community-level patterns of species associations to predict rare and common species compositions

Lastly, we assessed whether uniqueness (or commonness) in species compositions could predict mean and variance of species associations. This is relevant because it allows us further determining the deterministic nature of species associations. We estimated how unique or common the composition of a particular lake found at one lake was compared to all others by its Local Contribution to Beta Diversity (LCBD; Legendre and De Cáceres 2013). LCBD can also be calculated (interpreted) as the mean distance between a community and all other communities. The LCBD for lake *l*, given all lakes *s* and species *p*, can be estimated as:

$$LCBD_l = \frac{1}{s}\sum_{k=1}^{s} D_k \qquad D_k = \sqrt{\sum_{j=1}^{p}(x_{lj} - x_{kj})^2}$$

where D_k is the Euclidean distance between lake *l* and lake *k*. LCBD for lake *l* is then, the mean Euclidean distance between itself and all other communities, where communities with identical compositions have a distance equal to zero. Because the Euclidean distance is not an appropriate dissimilarity metric for comparing differences in species compositions among communities (Legendre and Legendre 1998, Legendre and Gallagher 2001), we transformed the species matrix using a Hellinger transformation prior to calculating LCBD. Lakes with high values of LCBD have more unique combinations of species than lakes with lower values. In the unrealistic case where all communities have identical compositions, LCBD would be 0 for all communities. Similarly, to the mean and SD of the conditional relationships of each lake, LCBD values can covary with species richness (Ulrich et al. 2017, 2018). As such, a null model was used to standardize LCBD values. We followed the same procedure as described for standardizing the mean and SD of conditional relationships for lakes, except with LCBD as the metric of interest. As before, the species matrix was randomized using the curveball algorithm (Strona et al. 2014) where after each randomization, LCBD values for each lake were estimated. The randomization process was repeated 999 times to generate a null distribution of LCBD values for each lake. Again, the observed value was made part of the null distribution to estimate the null mean and SD LCBD values. For lake *j*, the observed LCBD was scaled by calculating the standardized effect size using the null distribution of LCBD value lake *j*;

$$SES_{j.LCBD} = \frac{\overline{LCBD}_{obs.j} - \overline{LCBD}_{null.j}}{SD_{LCBD.null.j}}$$

where LCBD is the metric of interest and SD_{LCBD} is the standard deviation of the null distribution for lake *j*. LCBD values generated with the null model represent communities where lakes are randomly colonized by species. If a lake has a standardized effect size of 0, the lake has beta diversity patterns similar to a lake that was colonized randomly. Non-zero standardized effect sizes represent lakes that have beta diversity patterns that are not produced through random colonization, with positive and negative effect sizes representing lakes that are more or less unique than expected under our null expectations, respectively. Like for the standardized effect size of the community mean and SD, we repeated the null model analysis framework, constraining the shuffling processes within primary watersheds to determine the influences of historical factors on the distribution of fishes across the landscape. To determine whether our metrics of community structure (the standardized community means and SDs) predicts spatial patterns of beta diversity, we estimated the relationship between the standardized LCBD, and

the standardized community means and SDs with a linear model. All variables were standardized prior to analysis.

Results

Global Markov network for predicting species associations

Model performance of the estimated network model in predicting species associations evaluated using cross-validation resulted in a mean total prediction accuracy of 89.5% across all species pairs across all lakes. Comparing the mean total prediction accuracy of the models with and without environmental covariates, we found little evidence to support differences in prediction accuracy between the two models (Appendix II). This indicates that species have strong patterns of association; the presence or absence of a species is just as good a predictor as environmental features. Note, though, that the model with environmental covariates had lower sensitivity (i.e., poorer at predicting presences) but higher specificity (i.e., better at predicting absences) than the model without covariates. In other words, the model with covariates did not improve total prediction accuracy but was more precise at predicting presences. The global Markov network with environmental covariates was then used to estimate local species associations (as described in section *Linking species level (pairwise) patterns of species associations to environmental and historical factors*) for each of the 706 lakes. Because rare and very common species were removed, nine lakes had no species pairs or just one and were removed, resulting in a total of 697 lakes with predicted SAPs.

Predicting variation in species-pairs associations across lakes

Overall, the boosted gradient regression trees indicated that the environment was a poor predictor of variation in species-pairs associations across lakes. Across all combinations of parameters, R^2 was low (Appendix III). The final selected model had an R^2 of 0.0395. Out of the 89 environmental predictors, 77 had a non-zero influence, with the relative importance of environmental predictors quickly decreasing after the first three variables (mean lake depth, dissolved organic carbon and maximum lake depth) (Appendix III).

Predicting variation in community-level patterns of species associations

While predicting variation in species-pairs associations across lakes was challenging (see above), predicting community summaries based on species associations generated very strong predictive models. By showing that lakes with similar environments select for similar SAPs within communities, our results support the idea that SAPs are context dependent. There was a strong relationship between the environment and SES_{mean} and SES_{SD} (Table 2). The strength of the relationship between SES_{mean} and the environment were similar regardless of the null model used to estimate SES_{mean}. However, a stronger relationship was detected when SES_{SD} values were estimated using the constrained null model (species shuffling within primary watersheds) than with the unconstrained null model (species shuffling across all lakes). Thus, accounting for historical processes through the null model analyses increased the ability to detect relationships between the environment and the variance of SAPs across communities. As such, we only report here the results for values estimated using the constrained null model (see Appendix IV for results using the unconstrained null model). The strongest predictors of SES_{mean} were climate variables (Fig. 3). SES_{mean} increased with the number of ice-free days and decreased with maximum surface temperature and the proportion of days that were cold during ice free days. Stronger species associations on average were more likely to be found in lakes with longer and warmer ice-free periods but with cooler maximum temperatures. Climate variables were also strong predictors of SES_{SD} (Fig. 4). SES_{SD} decreased with the date of spring and increased with the amount of solar radiation received during cold dates, and the number of ice-free days. Lakes with more variation in species associations were more likely to be those where spring arrives earlier and the ice-free period is longer, and when more radiation is received on cold days.

We found a strong positive linear relationship between the SES_{SD} and SES_{mean} (R^2= 0.443; Fig. 5). Lakes with stronger species associations on average, have more variation in their SAPs. The residuals from this model can also be useful for exploring further patterns of context dependency. Residuals represent lakes that have more (positive residuals) or less variance (negative residuals) in species associations than one would expect given the average strength of species associations in a lake. Using a linear model with LASSO, we found a modest relationship between the residual variation and the environment (R^2= 0.338). Although this could have been done in a single model with SES_{SD} as a predictor of SES_{mean}, and vice-versa, we decided to do it in this way to improve on the narrative. Given that we have a high number of degrees of freedom, not penalizing by SES_{SD} on SES_{mean} (or vice versa) does not affect

estimates. Like with SES_{mean} and SES_{SD}, climate variables were important, with a negative relationship between the residuals and maximum monthly radiation and a positive relationship with minimum monthly radiation (Fig. 6). However, mean lake depth was also a strong predictor and was negatively correlated with the residuals. Taken together, an increase in minimum monthly radiation and mean lake depth, but a decrease in maximum monthly radiation results in communities with more variation in SAPs than one would expect given the average strength of SAPs in a community.

Given how important climate was for predicting SES_{SD}, SES_{mean} and their mutual residuals, we tested if the predictive performance of these models could be equally or better achieved by using a simpler predictor. For example, latitude may serve as a proxy of environmental variation in a system where latitudinal variation in climate is very strong. We fitted the same models again, with the addition of latitude and longitude as predictors and used variation partitioning to assess the proportion of variation explained by just geographic location and what proportion is shared by geographic location and the environment. Overall, geographic location did not improve the model as it explained the same variation that was captured by the environmental variables (Appendix V). This suggests that we did not miss any potentially relevant environmental predictor that varies across this broad environmental gradient (at least latitudinally).

Predicting rare versus common species compositions with community-level patterns of species associations

We found a decrease in SES_{LCBD} with increasing SES_{mean} and SES_{SD}, however the relationship was stronger with SES_{mean} than with SES_{SD} (Table 3). Communities with more varied and stronger SAPs were less likely to be more unique in their combination of species. Contrasting the linear models built using values from the unconstrained and constrained null models, our results show some evidence of how considering historical effects may influence the predictability of beta-diversity patterns. Models performed worse when using SES_{LCBD}, SES_{mean} and SES_{SD} values from the constrained null model (Table 3). The difference was the greatest when using SES_{SD} values as a predictor, where SES_{SD} values calculated from the constrained null model explained little of the variation in beta-diversity. By constraining species-shuffling within watersheds, we controlled for the effect of species that were unable to disperse into the

watershed. That is, we evaluated the influence of using species pools as shaped by historical dispersal patterns on predicting the relationship between community-level SAPs and LCBD.

Discussion

Our main goals were to determine if the environment could predict SAPs over a large environmental gradient; and outline a quantitative framework to do so. Our study system contains a wide combination of species association patterns and environments and our results show that communities in similar environments have similar SAPs. Our results contribute to growing evidence that the outcome of different processes (i.e., species interactions, environmental selection) is context dependent (Chamberlain et al. 2014, Bar-Massada and Belmaker 2017, MacDougall et al. 2018) by demonstrating that the environment can determine the expected range of results of community assembly, across a range of species compositions. We show that the ability to detect context-dependency is dependent on the scale of observation, with a stronger relationship with the environment detected at the community-scale than at the pairwise-scale. Climate variables were the strongest predictors, with lakes having colder winters exhibiting patterns more similar to those expected if communities were assembled randomly. Community-level species patterns can change across environmental gradients because of (1) the environment selecting for species pairs with different competitive abilities and/or (2) an interaction between the environment and the relative importance or outcome of deterministic and stochastic mechanisms (see below).

Environmental conditions can select for species that can colonize a site, altering the types of potential interactions (i.e. competition) (Kelt et al. 1995, Belyea and Lancaster 1999). While environmental selection is generally considered a deterministic process, random community-level SAPs within a given environment can arise if the environment selects for species with low competitive ability, increasing the importance of other processes such as dispersal (Chase 2007, Lepori and Malmqvist 2009). For example, harsh winter conditions can select for communities comprised of species with wider environmental tolerances (Magnuson et al. 1998). According to the competition-colonization hypothesis, species that can colonize a wider variety of sites are predicted to have less competitive ability (Skellam 1951, Tilman 1994). As such, it is possible that in these more extreme lakes, communities are comprised of species that are less likely to compete resulting in more random communities, assuming the absence of other deterministic mechanisms. This is exemplified with the better performing constrained null

model, where species pools are defined for each watershed so that species in southern watersheds cannot appear in the north where the environment is generally more extreme.

Alternatively, instead of the environment selecting for species with strong or weak competitive abilities, the environment can alter the outcome or importance of mechanisms like competition and predation, leading to different patterns in different contexts. Biotic interactions can be context-dependent, so that the outcome of them can change depending on the environment (Chamberlain et al. 2014). For example, variables like water temperature can alter life history traits such as growth and prey capture rate, determining when species may be excluded due to predation (Hein et al. 2013). Moreover, other variables like lake size can determine the amount of refuge from predation, increasing or decreasing the impact of mechanisms like predation on structuring a lake (Post et al. 2000, Jackson et al. 2001, MacDougall et al. 2018). Our results are consistent with these trends where smaller lakes are more structured in the variation of SAPs than for larger lakes.

However, it can also be argued that historical dispersal patterns from freshwater refugia produce predictable SAPs. Species pools in this area are shaped by environmental tolerances and dispersal capacity of species (Mandrak and Crossman 1992, Mandrak 1995). Predictable patterns can be driven by interactions between species pools and community assembly history (Fukami 2004). Indeed, our models generally performed better when watersheds were considered (via null model constrains; see also Peres-Neto et al. 2001), reflecting the influence of species pools in generating predictable SAPs. However, SAPs were still predictable even when watersheds were not considered. This is likely due to the cumulative effect of both historical and contemporary processes on structuring fishes in this region and the random stratified design of the dataset. While dispersal post-deglaciation may have also selected for species following the competition-colonization hypothesis, human mediated dispersal in our study area is strong, introducing novel fish to watersheds and homogenizing species pools (Cazelles et al. 2019). Additionally, sampling of lakes in the dataset are stratified using fish management zones, lake size and recreationally important fishes (Walleye, Lake Trout and Brook Trout) helping to control for effects between watersheds (Lester et al. 2020).

Summarizing community-level SAPs using the mean and standard deviation was not just powerful for detecting context dependency but also for understanding the deterministic nature of these communities. For more unique communities, SAPs on average were more like those

observed in randomly assembled communities. A likely explanation is that stochasticity in community assembly can lead to more unique communities. Multiple stable states can arise in randomly assembled communities through the priority effect if environments are colonized differentially (Belyea and Lancaster 1999, Chase 2003, Fukami 2015). Common communities had SAPs on average more different than those observed in randomly assembled communities, suggesting that more deterministic processes may be responsible for more common communities (see also Arranz et al. 2022). However, these communities also had more variation in SAPs. This suggests that these communities are not only driven more deterministic but may also have a greater variety of potential processes underlying them. By considering not only just the mean of a community which can show the relative importance of stochastic and deterministic processes in structuring a community, consideration of the variance can also reveal potential variation in processes structuring a community. A step forward would be to determine whether the sources of greater variance in SAPs are due to a greater contribution of the non-stationarity component in species associations across environmental gradients (i.e., species associations that change as a function of the environment; measure by the influence of the statistical interaction between species and their environments).

Our results show that relationships with the environment are also dependent on scale. When using just pairwise SAPs, as opposed to community-level SAPs, we were only able to find a weak relationship with the environment. The weaker relationship does not necessarily mean that pairwise patterns lack context dependency. Instead, the mixing of opposite pairwise patterns within the same lake could blur the signal of the environment on SAPs. For systems like fish communities where a variety of mechanisms structure communities, it is not uncommon to find both aggregated and segregated patterns in the same environment (Giam and Olden 2016, McGarvey and Veech 2018, Cordero and Jackson 2019). Furthermore, difficulty in finding a common relationship between pairwise SAPs and the environment can be further exacerbated by the differential sensitivity of different types of biotic interactions and the environment (Chamberlain et al. 2014).

Instead of taking a species approach to pairwise patterns, analyzing the pairwise association patterns of traits may better reveal context-dependency in association patterns. Traits which can be a proxy for different interactions can be used to disentangle opposite patterns (Peres-Neto 2004). Alternatively, the SAPs themselves can be used to reveal traits about fishes. Grouping pairs that maximize the relationship with the environment can be used to

explore groups of species with common and context-dependent relationships with the environment (see Borthagaray et al. 2014). While communities may be structured by combinations of processes that have outcomes which change predictability with the environment, clearly the analysis of context-dependency using pairwise patterns requires more consideration, necessitating the consideration of possible mechanisms and interactions.

Our study offers a different perspective for addressing context dependency in community ecology. Isolating the effects of the environment on specific processes can lead to misleading predictions if interacting processes lead to non-linear outcomes, resulting in unexpected conclusions (Kolasa et al. 2021). This extrapolation problem results in the frequent use of context-dependency as an explanation for differences between study systems (Catford et al. 2022). By focusing on the observed patterns instead of the mechanisms, we were able to describe the variation in the outcomes of community assembly across a large environmental gradient. Borrowing from evolutionary ecology, we can describe this as the "reaction norm", where a range of phenotypic changes (or in our case, SAPs) can be observed for a specific environmental variable (Sultan and Stearns 2005). This type of thinking within community ecology is not new. For example, the environment has been used to categorize lake communities into turbid and clear water states, with an expected range of communities for each site (Carpenter 2003). Our results suggest the same can be done with community-assembly. While our methods were unable to determine the exact importance of each community assembly mechanisms for creating each observed pattern, we were still able to describe a common relationship with the environment and the resulting SAPs across all communities.

Conclusion

We show that context-dependency can be explained statistically by studying how SAPs change across large environmental gradients. By summarizing SAPs at the community level, we described a framework where communities – regardless of species compositions – can be compared across environments. While our framework does not necessarily provide a mechanistic understanding of context-dependency, we showed that our metric can predict patterns in diversity across the landscape. We argue that variation in SAPs across environments can be better understood by using the reaction norm as a framework for reconciling differences in outcomes across study systems. By framing SAPs as a trait, a range of expected outcomes can be described for different environments.

For study systems such as lake fish communities where the role of top-down and bottom-up processes can be dependent on the environment (MacDougall et al. 2018) and are facing an increasing number of environmental stressors (Comte et al. 2013, Hansen et al. 2017, Cazelles et al. 2019), our approach can be important for studying how communities respond to changing conditions. Our consideration of the environment and the relationship with community assembly processes offers another perspective on how to reconcile the problem of context-dependency, using existing tools in community ecology like SAPs.

Works Cited

Arranz, I., B. Fournier, N. P. Lester, B. J. Shuter, and P. R. Peres-Neto. 2022. Species compositions mediate biomass conservation: The case of lake fish communities. Ecology 103:e3608.

Bar-Massada, A., and J. Belmaker. 2017. Non-stationarity in the co-occurrence patterns of species across environmental gradients. Journal of Ecology 105:391–399.

Belyea, L. R., and J. Lancaster. 1999. Assembly Rules within a Contingent Ecology. Oikos 86:402.

Blanchet, F. G., K. Cazelles, and D. Gravel. 2020. Co-occurrence is not evidence of ecological interactions. Ecology Letters 23:1050–1063.

Borthagaray, A. I., M. Arim, and P. A. Marquet. 2014. Inferring species roles in metacommunity structure from species co-occurrence networks. Proceedings of the Royal Society B: Biological Sciences 281:1–7.

Cadotte, M. W., and C. M. Tucker. 2017. Should Environmental Filtering be Abandoned? Trends in Ecology and Evolution 32:429–437.

Carpenter, S. R. 2003. Regime Shifts in Lake Ecosystems: Pattern and Variation. Excellence in Ecology. Ecology Institute.

Carstensen, D. W., J. P. Lessard, B. G. Holt, M. Krabbe Borregaard, and C. Rahbek. 2013. Introducing the biogeographic species pool. Ecography 36:1310–1318.

Catford, J. A., J. R. U. Wilson, P. Pyšek, P. E. Hulme, and R. P. Duncan. 2022. Addressing context dependence in ecology. Trends in Ecology and Evolution 37:158–170.

Cazelles, K., T. Bartley, M. M. Guzzo, M.-H. Brice, A. S. MacDougall, J. R. Bennett, E. H. Esch, T. Kadoya, J. Kelly, S. Matsuzaki, K. A. Nilsson, and K. S. McCann. 2019. Homogenization of freshwater lakes: recent compositional shifts in fish communities are explained by gamefish movement and not climate change. Global Change Biology 25:4222–4233.

Chamberlain, S. A., J. L. Bronstein, and J. A. Rudgers. 2014. How context dependent are species interactions? Ecology Letters 17:881–890.

Chase, J. M. 2003. Community assembly: When should history matter? Oceologia 136:489-498.

Chase, J. M. 2007. Drought mediates the importance of stochastic community assembly. Proceedings of the National Academy of Sciences of the United States of America 104:17430–17434.

Chase, J. M., and J. A. Myers. 2011. Disentangling the importance of ecological niches from stochastic processes across scales. Philosophical Transactions of the Royal Society B: Biological Sciences 366:2351–2363.

Cheng, J., E. Levina, P. Wang, and J. Zhu. 2014. A sparse ising model with covariates. Biometrics 70:943–953.

Clark, N. J., K. Wells, and O. Lindberg. 2018. Unravelling changing interspecific interactions across environmental gradients using Markov random fields. Ecology 99:1277–1283.

Comte, L., L. Buisson, M. Daufresne, and G. Grenouillet. 2013. Climate-induced changes in the distribution of freshwater fish: Observed and predicted trends. Freshwater Biology 58:625-639.

Connor, E. F., and D. Simberloff. 1979. The Assembly of Species Communities: Chance or

Competition? Ecology 60:1132-1140.

Connor, E. F., and D. Simberloff. 1983. Interspecific Competition and Species Co-Occurrence Patterns on Islands: Null Models and the Evaluation of Evidence. Oikos 41:455-465.

Cordero, R. D., and D. A. Jackson. 2019. Species-pair associations, null models, and tests of mechanisms structuring ecological communities. Ecosphere 10:e02797.

Diamond, J. M. 1975. Assembly of species communties. Pages 342–444 *in* C. L. Martin and J. M. Diamond, editors. Ecology and evolution of communities. Harvard University Press, Cambridge, Massachusettes, USA.

Diniz-Filho, J. A. F., L. M. Bini, and B. A. Hawkins. 2003. Spatial autocorrelation and red herrings in geographical ecology. Global Ecology and Biogeography 12:53–64.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77:802–813.

Freilich, M. A., E. Wieters, B. R. Broitman, P. A. Marquet, and S. A. Navarrete. 2018. Species co-occurrence networks: Can they reveal trophic and non-trophic interactions in ecological communities? Ecology 99:690–699.

Fukami, T. 2004. Community assembly along a species pool gradient: Implications for multiple-scale patterns of species diversity. Population Ecology 46:137–147.

Fukami, T. 2015. Historical Contingency in Community Assembly: Integrating Niches, Species Pools, and Priority Effects. Annual Review of Ecology, Evolution, and Systematics 46:1–23.

Giam, X., and J. D. Olden. 2016. Environment and predation govern fish community assembly in temperate streams. Global Ecology and Biogeography 25:1194–1205.

Giraudoux, P. 2021. pgirmess: Spatial Analysis and Data Mining for Field Ecologists.

Gotelli, N. J. 2000. Null Model Analysis of Species Co-Occurrence Patterns. Ecology 81:2606.

Gotelli, N. J., and D. J. McCabe. 2002. Species co-occurrence: A meta-analysis of J. M. Diamond's assembly rules model. Ecology 83:2091–2096.

Greenwell, B., B. Boehmke, J. Cunningham, and G. Developers. 2019. gbm: Generalized boosted regression models. R package version 2.

Hansen, G. J. A., J. S. Read, J. F. Hansen, and L. A. Winslow. 2017. Projected shifts in fish species dominance in Wisconsin lakes under climate change. Global Change Biology 23:1463–1476.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. 10. Boosting and Additive Trees. Pages 337–384 The Elements of Statistical Learning. Springer, New York.

Hein, C. L., G. Öhlund, and G. Englund. 2013. Fish introductions reveal the temperature dependence of species interactions. Proceedings of the Royal Society B: Biological Sciences 281:20132641.

Henriques-Silva, R., Z. Lindo, and P. R. Peres-Neto. 2013. A community of metacommunities: Exploring patterns in species distributions across large geographical areas. Ecology 94:627–639.

Hortal, J., J. C. Nabout, J. Calatayud, F. M. Carneiro, A. Padial, A. M. C. Santos, T. Siqueira, F. Bokma, L. M. Bini, and M. Ventura. 2014. Perspectives on the use of lakes and ponds as model systems for macroecological research. Journal of Limnology 73:46–60.

Jackson, D. A., P. R. Peres-Neto, and J. D. Olden. 2001. What controls who is where in

freshwater fish communities - The roles of biotic, abiotic, and spatial factors. Canadian Journal of Fisheries and Aquatic Sciences 58:157–170.

Kelt, D. A., M. L. Taper, and P. L. Meserve. 1995. Assessing the impact of competition on community assembly: A case study using small mammals. Ecology 76:1283–1296.

Kolasa, J., M. P. Hammond, and J. Yan. 2021. Metacommunity research can benefit from including context-dependency. bioRxiv:2021.09.26.461405.

Kuhn, M. 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28.

Legendre, P., and M. De Cáceres. 2013. Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. Ecology Letters 16:951–963.

Legendre, P., and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129:271–280.

Legendre, P., and L. Legendre. 1998. Numerical Ecology. 2nd Englis. Elsevier, Amsterdam.

Leibold, M. A., E. P. Economo, and P. Peres-Neto. 2010. Metacommunity phylogenetics: Separating the roles of environmental filters and historical biogeography. Ecology Letters 13:1290–1299.

Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and A. Gonzalez. 2004. The metacommunity concept: A framework for multi-scale community ecology. Ecology Letters 7:601–613.

Lepori, F., and B. Malmqvist. 2009. Deterministic control on community assembly peaks at intermediate levels of disturbance. Oikos 118:471–479.

Lester, N. P., S. Sandstrom, D. T. de Kerckhove, K. Armstrong, H. Ball, J. Amos, T. Dunkley, M. Rawson, P. Addison, A. Dextrase, D. Taillon, B. Wasylenko, P. Lennox, H. C. Giacomini, and C. Chu. 2020. Standardized Broad-Scale Management and Monitoring of Inland Lake Recreational Fisheries: An Overview of the Ontario Experience. Fisheries 46:107–118.

Lyons, S. K., K. L. Amatangelo, A. K. Behrensmeyer, A. Bercovici, J. L. Blois, M. Davis, W. A. Dimichele, A. Du, J. T. Eronen, J. Tyler Faith, G. R. Graves, N. Jud, C. Labandeira, C. V Looy, B. McGill, J. H. Miller, D. Patterson, S. Pineda-Munoz, R. Potts, B. Riddle, R. Terry, A. Tóth, W. Ulrich, A. Villaseñor, S. Wing, H. Anderson, J. Anderson, D. Waller, and N. J. Gotelli. 2016. Holocene shifts in the assembly of plant and animal communities implicate human impacts. Nature 529:80–83.

MacDougall, A. S., E. Harvey, J. L. McCune, K. A. Nilsson, J. Bennett, J. Firn, T. Bartley, J. B. Grace, J. Kelly, T. D. Tunney, B. McMeans, S. I. S. Matsuzaki, T. Kadoya, E. Esch, K. Cazelles, N. Lester, and K. S. McCann. 2018. Context-dependent interactions and the regulation of species richness in freshwater fish. Nature Communications 9:973.

Magnuson, J. J., W. M. Tonn, A. Banerjee, J. Toivonen, O. Sanchez, and M. Rask. 1998. Isolation vs. extinction in the assembly of fishes in small northern lakes. Ecology 79:2941–2956.

Mandrak, N. E. 1995. Biogeographic patterns of fish species richness in Ontario lakes in relation to historical and environmental factors. Canadian Journal of Fisheries and Aquatic Sciences 52:1462–1474.

Mandrak, N. E., and E. J. Crossman. 1992. Postglacial dispersal of freshwater fishes into Ontario. Canadian Journal of Zoology 70:2247–2259.

McGarvey, D. J., and J. A. Veech. 2018. Modular structure in fish co-occurrence networks: A comparison across spatial scales and grouping methodologies. PLoS ONE 13:1–20.

Meinshausen, N., and P. Bühlmann. 2006. High-dimensional graphs and variable selection with the Lasso. Annals of Statistics 34:1436–1462.

Morales-Castilla, I., M. G. Matias, D. Gravel, and M. B. Araújo. 2015. Inferring biotic interactions from proxies. Trends in Ecology and Evolution 30:347–356.

Myers, B. J. E., A. J. Lynch, D. B. Bunnell, C. Chu, J. A. Falke, R. P. Kovach, T. J. Krabbenhoft, T. J. Kwak, and C. P. Paukert. 2017. Global synthesis of the documented and projected effects of climate change on inland fishes. Reviews in Fish Biology and Fisheries 27:339–361.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner. 2020. vegan: Community Ecology Package.

Olden, J. D., D. A. Jackson, and P. R. Peres-Neto. 2001. Spatial isolation and fish communities in drainage lakes. Oecologia 127:572–585.

Peres-Neto, P. R. 2004. Patterns in the co-occurrence of fish species in streams: The role of site suitability, morphology and phylogeny versus species interactions. Oecologia 140:352–360.

Peres-Neto, P. R., J. D. Olden, and D. A. Jackson. 2001. Environmentally constrained null models: Site suitability as occupancy criterion. Oikos 93:110–120.

Peterson, R. A. 2018. bestNormalize: normalizing transformation functions. R package version 1:573.

Popovic, G. C., D. I. Warton, F. J. Thomson, F. K. C. Hui, and A. T. Moles. 2019. Untangling direct species associations from indirect mediator species effects with graphical models. Methods in Ecology and Evolution 10:1571–1583.

Post, D. M., M. L. Pace, and N. G. Halrston. 2000. Ecosystem size determines food-chain length in lakes. Nature 2000 405:6790 405:1047–1049.

Sandstorm, S., M. Rowson, and N. Lester. 2013. Manual of instructions for broad-scale fish community monitoring using North American (NA1) and Ontario small mesh (ON2) gillnets. Peterborough, Ontario.

Sfenthourakis, S., E. Tzanatos, and S. Giokas. 2006. Species co-occurrence: The case of congeneric species and a causal approach to patterns of species association. Global Ecology and Biogeography 15:39–49.

Sharma, S., P. Legendre, M. de Cáceres, and D. Boisclair. 2011. The role of environmental and spatial processes in structuring native and non-native fish communities across thousands of lakes. Ecography 34:762–771.

Simberloff, D. 2004. Community Ecology: Is it time to move on? The American Naturalist 163:787–799.

Skellam, J. G. 1951. Random Dispersal in Theoretical Populations. Biometrika 38:196.

Stacklies, W., H. Redestig, M. Scholz, D. Walther, and J. Selbig. 2007. pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics 23:1164–1167.

Strona, G., D. Nappo, F. Boccacci, S. Fattorini, and J. San-Miguel-Ayanz. 2014. A fast and

unbiased procedure to randomize ecological binary matrices with fixed row and column totals. Nature Communications 5:4114.

Sultan, S. E., and S. C. Stearns. 2005. Environmentally Contingent Variation: Phenotypic Plasticity and Norms of Reaction. Pages 303–332 Variation. Academic Press.

Tibshirani, R. 1996. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58:267–288.

Tikhonov, G., N. Abrego, D. Dunson, and O. Ovaskainen. 2017. Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. Methods in Ecology and Evolution 8:443–452.

Tilman, D. 1994. Competition and biodiversity in spatially structured habitats. Ecology 75:2–16.

Ulrich, W., A. Baselga, B. Kusumoto, T. Shiono, H. Tuomisto, and Y. Kubota. 2017. The tangled link between β- and γ-diversity: a Narcissus effect weakens statistical inferences in null model analyses of diversity patterns. Global Ecology and Biogeography 26:1–5.

Ulrich, W., Y. Kubota, B. Kusumoto, A. Baselga, H. Tuomisto, and N. J. Gotelli. 2018. Species richness correlates of raw and standardized co-occurrence metrics. Global Ecology and Biogeography 27:395–399.

Tables and Figures

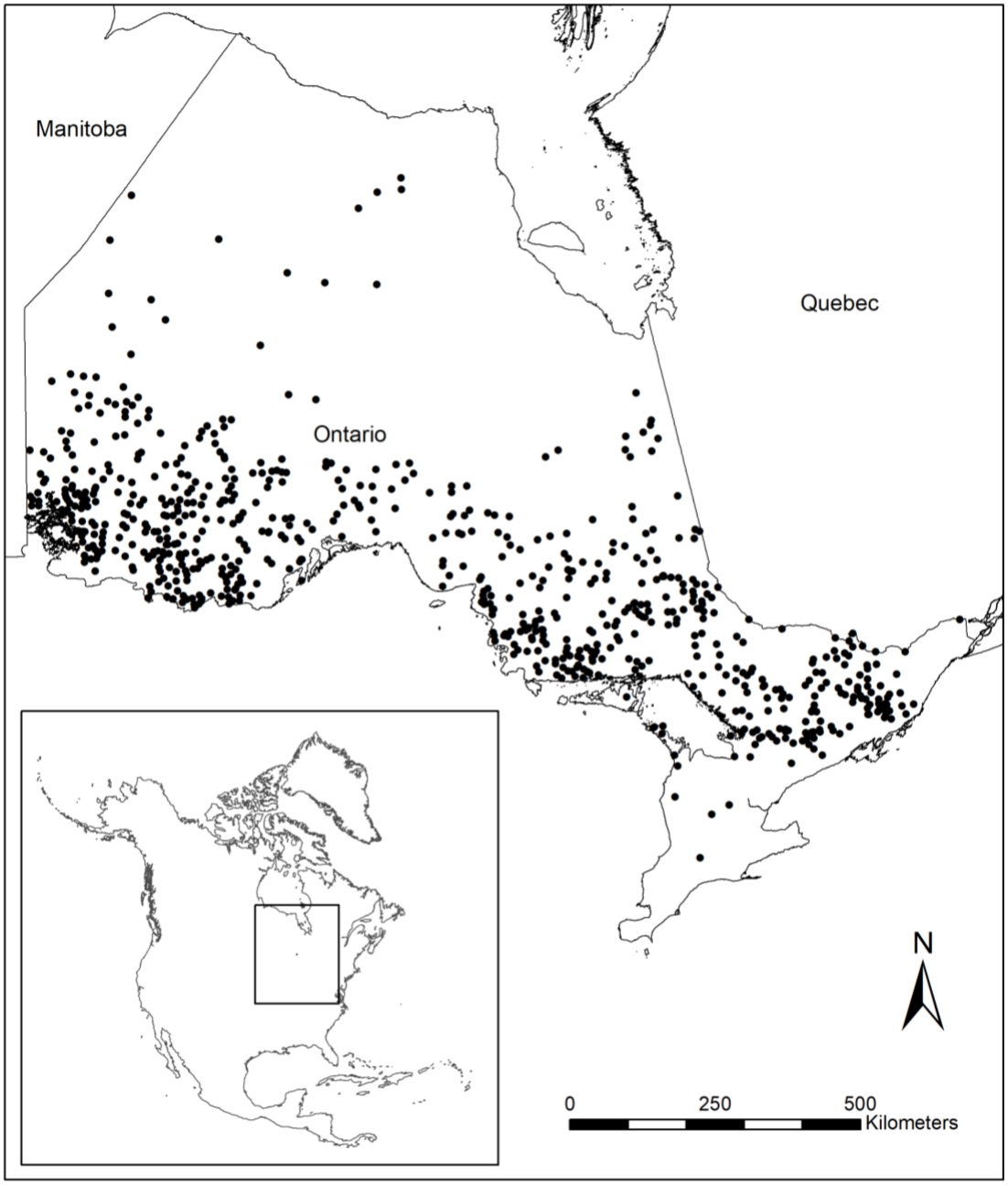

Figure 1. Locations of the 706 in-land lakes surveyed by the Ontario Ministry of Natural Resources and forestry that consisted of our study system. Surveyed lakes are all located in Ontario, Canada.

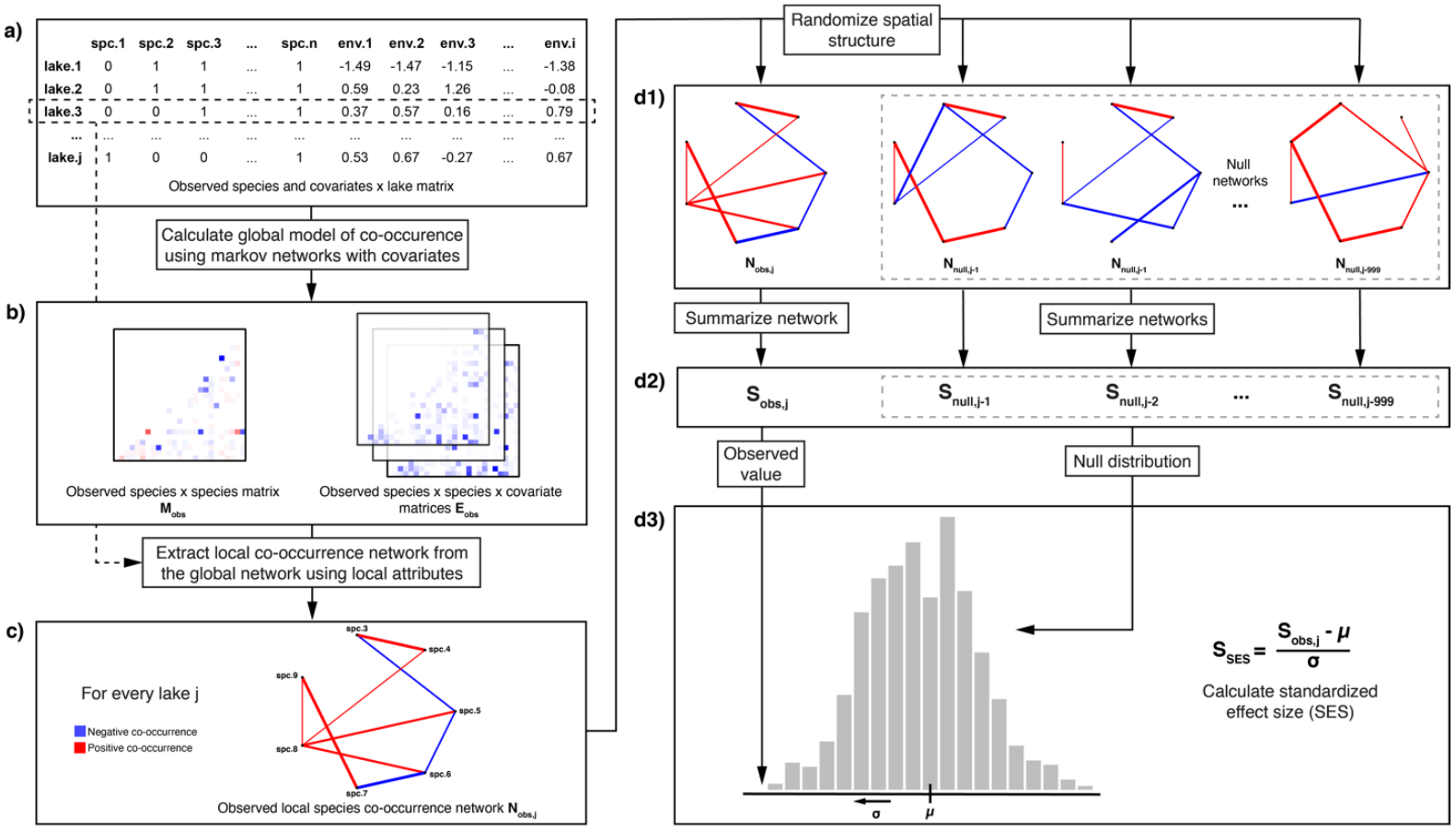

Figure 2. Procedure used to calculate species association networks for each lake and community-level metrics of community structure. A matrix (a) containing the presence-absence of species and associated environmental variables for each lake was used to calculate a global model of SAPs (b) using Markov networks with covariates, summarizing SAPs between species across all lakes and how they change across different environments. Local co-occurrence networks for each lake were then extracted from the global model by inputting local species presence-absence and environmental data (c). The process was then repeated with the presence-absence of species shuffled, resulting in one observed association network and 999 “null” networks for each lake (d1). Observed and

null networks were then summarized by calculating the mean and standard deviation of observed SAPs (d2). The difference between the observed metric and the null expectation was then calculated using a standardized effect size (d3).

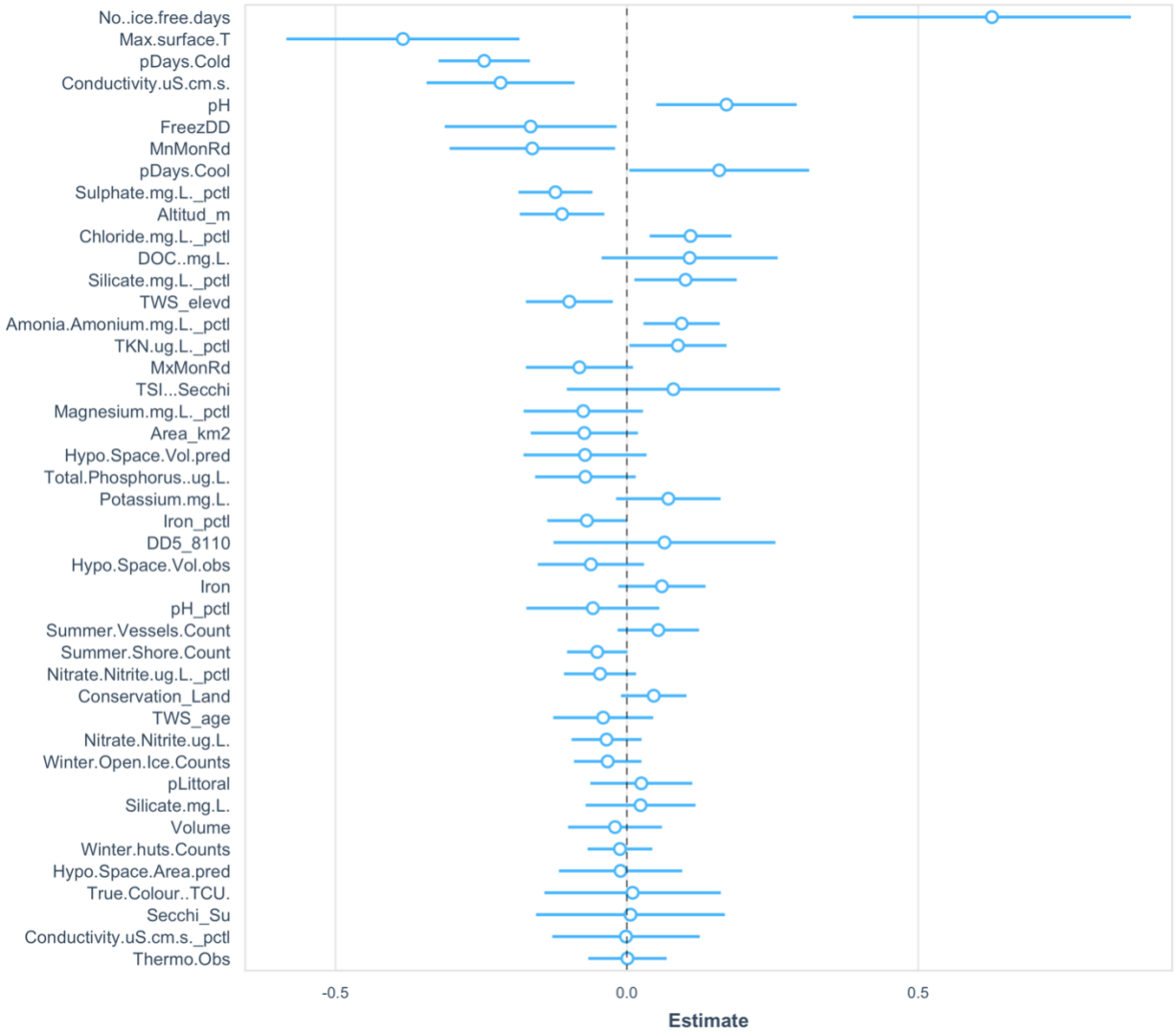

Figure 3. Estimated strength of environmental predictors of mean lake species association patterns (SAPs) in freshwater fishes from 706 freshwater fish communities in Ontario, Canada. Mean lake species association patterns were estimated using a standardized effect size by comparing mean SAPs in each observed lakes to the respective mean SAPs of each lake estimated by shuffling species within primary watersheds. Regression coefficients were then estimated using a linear model, with all environmental variables used as a predictor. Because of the large number of environmental variables used, LASSO regularization was used to add sparsity, forcing some coefficients to zero while maintaining predictive power and reducing multicollinearity.

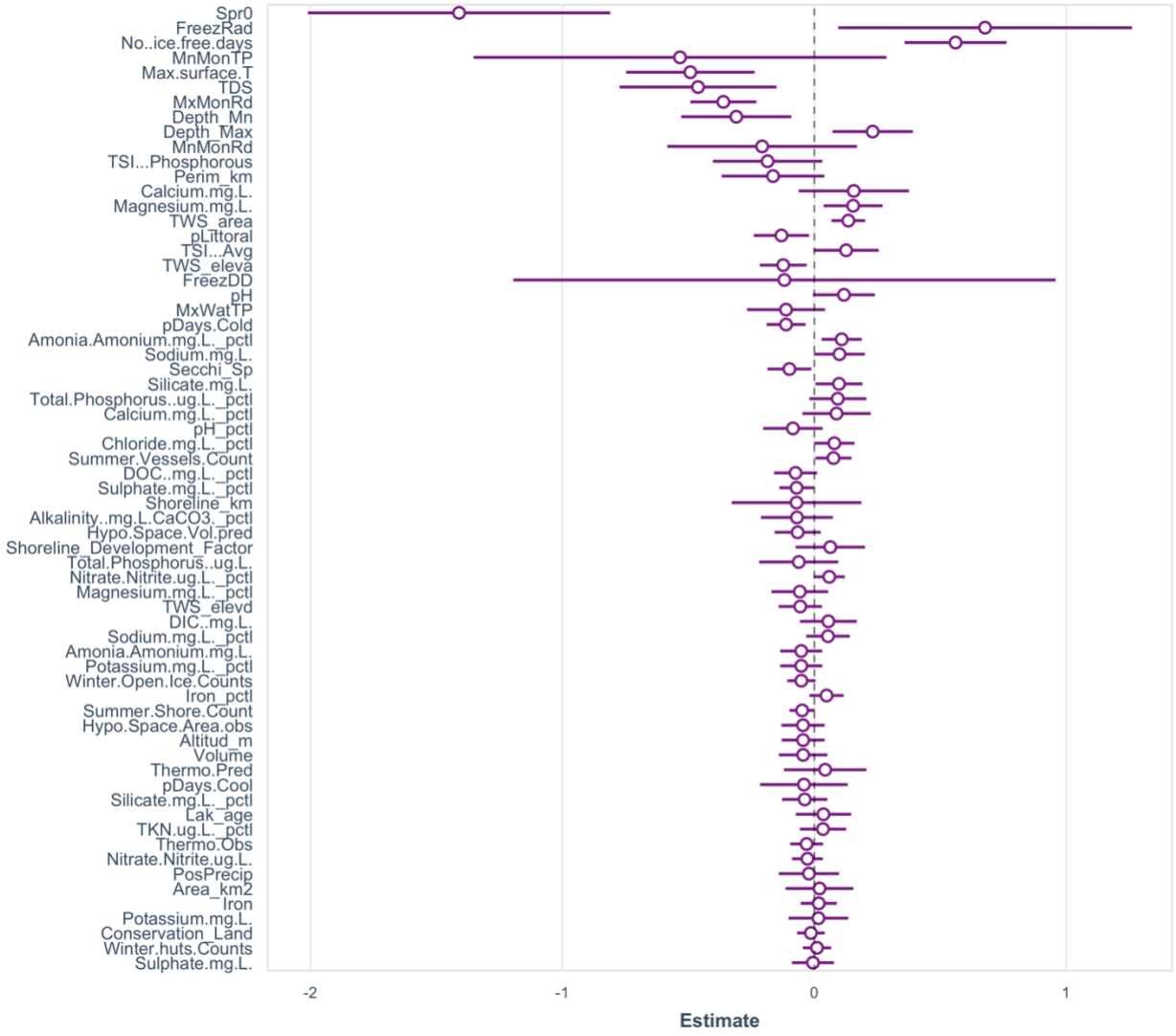

Figure 4. Estimated strength of environmental predictors of variance in lake species association patterns (SAPs) in freshwater fishes from 706 freshwater fish communities in Ontario, Canada. Variance in lake species association patterns was estimated using a standardized effect size by comparing the standard deviation (SD) of SAPs in each observed lakes to the respective SD of SAPs of each lake estimated by shuffling species within primary watersheds. Regression coefficients were then estimated using a linear model, with all environmental variables used as a predictor. Because of the large number of environmental variables used, LASSO regularization was used to add sparsity, forcing some coefficients to zero while maintaining predictive power and reducing multicollinearity.

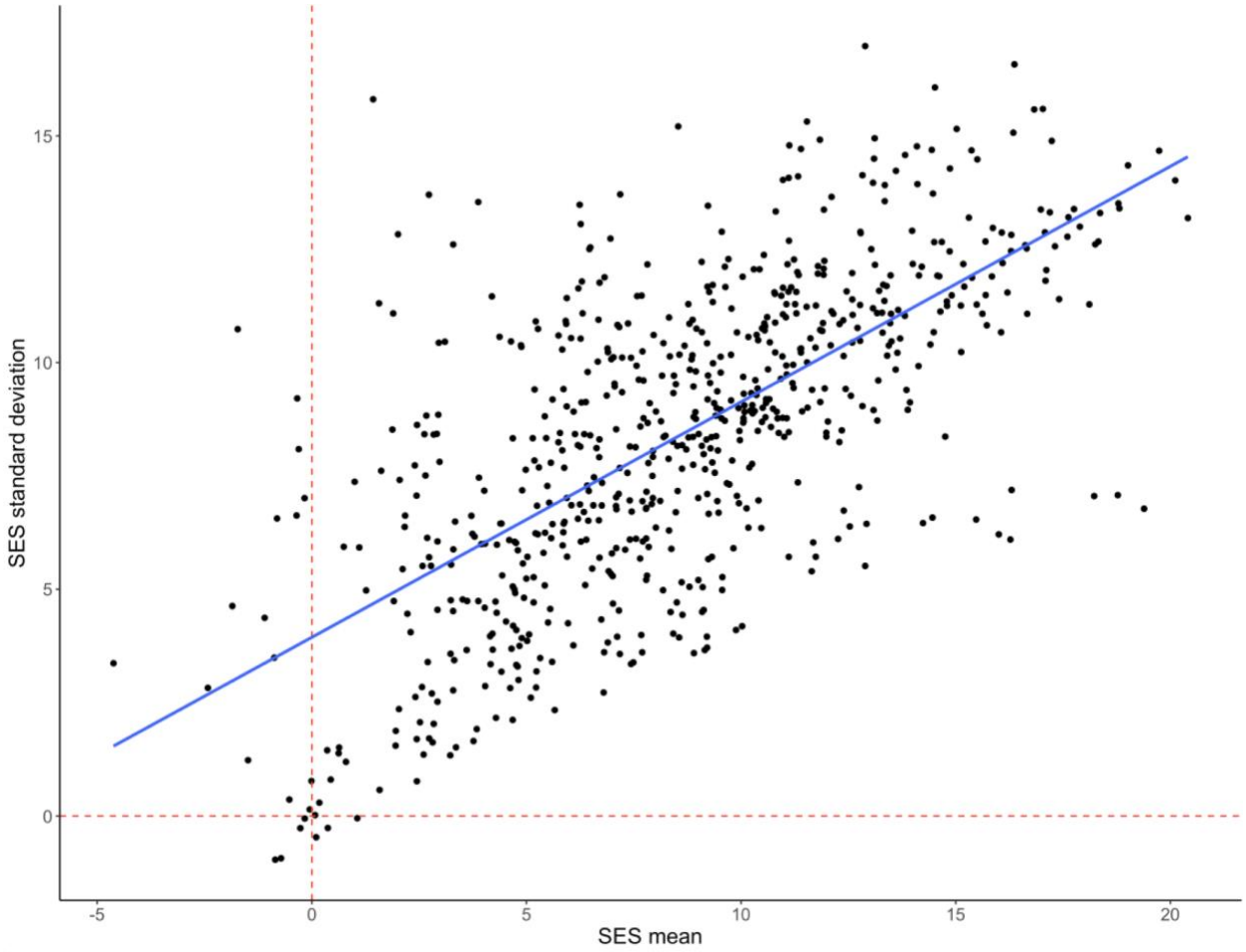

Figure 5. Relationship between the variance in lake species association patterns (SAPs) and mean lake species association patterns from 706 freshwater fish communities in Ontario, Canada. The variance and mean of lake species association patterns were estimated using a standardized effect size by comparing the standard deviation (SD) and mean SAPs in each observed lake to the respective mean and SD of SAPs of each lake estimated by shuffling species within primary watersheds.

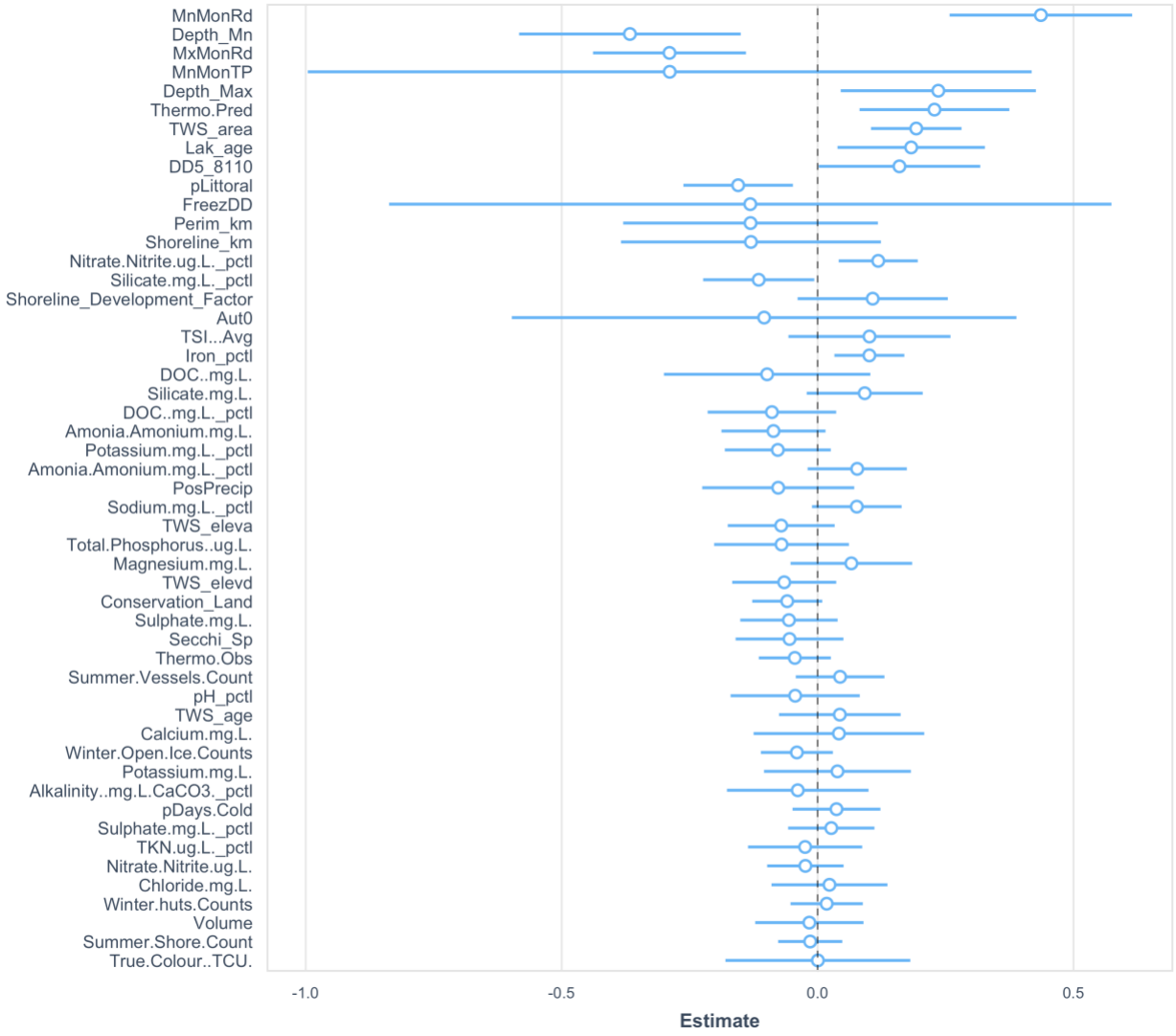

Figure 6. Estimated strength of environmental predictors for the residuals of the relationship between the variance and mean in lake species association patterns (SAPs) in freshwater fishes from 706 freshwater fish communities in Ontario, Canada. The variance and mean of lake species association patterns were estimated using a standardized effect size by comparing the standard deviation (SD) and mean SAPs in each observed lake to the respective mean and SD of SAPs of each lake estimated by shuffling species within primary watersheds. Regression coefficients were then estimated using a linear model, with all environmental variables used as a predictor. Because of the large number of environmental variables used, LASSO regularization was used to add sparsity, forcing some coefficients to zero while maintaining predictive power and reducing multicollinearity.

Table 1. Environmental variables used to fit models for estimating and predicting species association patterns (SAPs) in freshwater fishes from 706 freshwater fish communities in Ontario, Canada. Environmental variables (n = 89) were recorded and derived by the Ontario Ministry of Natural Resources and Forestry as part of the Broad Scale Management program. Variables were transformed before use in PPCA but not for other models.

Variables	Descriptor	Transformation
Waterbody_LID	Lake identifier	NA
Area_km2	Surface area of the lake (km2)	orderNorm
Lak_age	Time since glaciation of lake (Kyr from present)	No transformation
Depth_Max	Maximum lake depth (m)	yeojohnson
Depth_Mn	Mean lake depth (m)	arcsinh_x
Amonia.Amonium.mg.L.	Ammonia/Ammonium (mg/L)	orderNorm
Amonia.Amonium.mg.L._pctl	Percentile ammonia/ammonium	orderNorm
Volume	Lake volume (area*max depth)	log_x
Altitud_m	Lake altitude (m.a.s.l.)	orderNorm
Perim_km	Lake perimeter, not including islands (km)	orderNorm
TDS	Total dissolved solids (mg/L)	orderNorm
DIC..mg.L.	Dissolved inorganic carbon (mg/L)	boxcox
PosRad	Solar radiation >0C from 1981-2010	orderNorm
FreezRad	Solar radiation <0C from 1981-2010	orderNorm
Summer.Vessels.Count	Mean number of summer fishing vessels	orderNorm
Summer.Shore.Count	Mean number of summer shore anglers	No transformation
Winter.huts.Counts	Mean number of winter fishing huts	sqrt_x
Winter.Open.Ice.Counts	Mean number of winter open-ice angers	sqrt_x
DIC..mg.L._pctl	Percentile dissolved inorganic carbon	orderNorm
DOC..mg.L.	Dissolved organic carbon (mg/L)	sqrt_x
DOC..mg.L._pctl	Percentile dissolved organic carbon	orderNorm
PosDays	Average number of days >0C from 1981-2010 (days)	orderNorm
Aut0	Average date of first 0C autumn day from 1981-2010 (days)	orderNorm
No..ice.free.days	Estimated number of ice free (days)	orderNorm
MxWatTP	Estimated maximum water temperature (deg C)	orderNorm
FreezDD	Cumilative degree days <0C from 1981-2010 (days)	orderNorm
PosPrecip	Average rainfall from 1981-2010	orderNorm

PodDD	Cumilative degree days >0C from 1981-2010	orderNorm
DD5_8110	Growing degree days >5C for 1981-2010	orderNorm
MxMonTP	Maximum monthly air tempature from 1981-2010 (deg C)	orderNorm
MxMonRd	Maximum monthly radiation from 1981-2010	orderNorm
Max.surface.T	Estimated maximum surface temperature (deg C)	log_x
Airtemp_8110	Average annual temperature for 1981-2010 (deg C)	orderNorm
MnMonTP	Minimum monthly air tempature from 1981-2010 (deg C)	orderNorm
MnMonRd	Minimum monthly radiation from 1981-2010	orderNorm
Thermo.Obs	Observed thermocline depth (m)	sqrt_x
Thermo.Pred	Predicted thermocline depth (m)	arcsinh_x
pDays.Cold	Proportion of days with maximum surface temperatures between 8-12 deg C during ice free period	orderNorm
pDays.Cool	Proportion of days with maximum surface temperatures between 16-20 deg C during ice free period	orderNorm
pDays.Warm	Proportion of days with maximum surface temperatures between 22-26 deg C during ice free period	orderNorm
Hypo.Space.Area.obs	Observed hypolimnetic area (prop)	exp_x
Hypo.Space.Area.pred	Predicted hypolimnetic area (prop)	orderNorm
Hypo.Space.Vol.obs	Observed hypolimnetic volume (prop)	orderNorm
Hypo.Space.Vol.pred	Predicted hypolimnetic volume (prop)	orderNorm
pLittoral	Littoral area (prop <4.6m)	orderNorm
TSI...Avg	Average trophic state index derived from phosphorous and Secchi depth	sqrt_x
pH	pH	log_x
Conductivity.uS.cm.s.	Conductivity (uS/cm/s)	orderNorm
Alkalinity..mg.L.CaCO3.	Alkalinity (mg/L CaCO3)	arcsinh_x
Calcium.mg.L.	Calcium (mg/L)	orderNorm
Magnesium.mg.L.	Magnesium (mg/L)	orderNorm
Sodium.mg.L.	Sodium (mg/L)	orderNorm
Potassium.mg.L.	Potassium (mg/L)	orderNorm
Chloride.mg.L.	Chloride (mg/L)	orderNorm
Sulphate.mg.L.	Sulphate (mg/L)	orderNorm
Nitrate.Nitrite.ug.L.	Nitrate/Nitrite (ug/L)	orderNorm
Iron	Iron	No transformation
Nitrate.Nitrite.ug.L._pctl	Percentile nitrate/nitrite	orderNorm

TKN.ug.L.	Total Kjeldahl Nitrogen (ug/L)	sqrt_x
TKN.ug.L._pctl	Percentile total Kjeldahl nitrogen	orderNorm
Silicate.mg.L.	Silicate (mg/L)	boxcox
Silicate.mg.L._pctl	Percentile silicate	orderNorm
Secchi_Sp	Spring Secci depth (m)	orderNorm
Secchi_Su	Summer Secci depth (m)	boxcox
pH_pctl	Percentile pH	orderNorm
Conductivity.uS.cm.s._pctl	Percentile conductivity	orderNorm
Alkalinity..mg.L.CaCO3._pctl	Percentile alkalinity	orderNorm
Calcium.mg.L._pctl	Percentile calcium	orderNorm
Magnesium.mg.L._pctl	Percentile magnesium	orderNorm
Sodium.mg.L._pctl	Percentile sodium	orderNorm
Potassium.mg.L._pctl	Percentile potassium	orderNorm
Chloride.mg.L._pctl	Percentile chloride	orderNorm
Sulphate.mg.L._pctl	Percentile sulphate	orderNorm
Iron_pctl	Percentile iron	orderNorm
Total.Phosphorus..ug.L.	Total phosphorus (ug/L)	log_x
Total.Phosphorus..ug.L._pctl	Percentile total phosphorus	orderNorm
TSI...Phosphorous	Trophic state index based on phosphorus	No transformation
TSI...Secchi	Trophic state index based on Secci	arcsinh_x
True.Colour..TCU.	True colour (TCU)	boxcox
True.Colour..TCU._pctl	Percentile true colour	orderNorm
TWS_area	Tertiary watershed area (km2)	orderNorm
TWS_age	Time since glaciation of the tertiary watershed (Kyr from present)	No transformation
TWS_eleva	Tertiary watershed altitude (m.a.s.l.)	orderNorm
TWS_elevd	Difference between maximum and minimum tertiary watershed altitude (m.a.s.l.)	orderNorm
pArea_LE20	Proportion of lake area <20m in depth	orderNorm
Shoreline_km	Total lake shoreline including islands (km)	boxcox
Shoreline_Development_Factor	Shoreline development factor (Shoreline_km/(2*sqrt(pi *Area_ha/100)))	boxcox
Spr0	Average date of first >0C spring day from 1981-2010 (days)	orderNorm
Angling_Pressure	Angling pressure	orderNorm
Conservation_Land	Conservation status (1 implies a form of conservation status)	No transformation

Table 2. Estimated strength of the relationship between the environment and the variance and mean of lake species association patterns (SAPs) in freshwater fish communities in Ontario, Canada. The variance (SES_{SD}) and mean (SES_{mean}) of lake species association patterns were estimated using a standardized effect size by comparing the standard deviation (SD) and mean SAPs in each observed lake to the respective mean and SD of SAPs of each lake estimated by shuffling species across all lakes (unconstrained) or within primary watersheds (constrained). Relationships were estimated using a linear model.

SES_{mean}

Model	*R-squared*	*Adjusted R-squared*	*No. observations*
Unconstrained*	0.576	0.553	697
Constrained*	0.579	0.551	697

SES_{SD}

Model	*R-squared*	*Adjusted R-squared*	*No. observations*
Unconstrained*	0.381	0.341	697
Constrained*	0.629	0.591	697

*, indicates significance at the >99% level

Table 3. Estimated strength of the relationship between the uniqueness in species composition of a lake and the variance and mean of lake species association patterns (SAPs) in freshwater fish communities in Ontario, Canada. The variance (SES_{SD}) and mean (SES_{mean}) of lake species association patterns were estimated using a standardized effect size by comparing the standard deviation (SD) and mean SAPs in each observed lake to the respective mean and SD of SAPs of each lake estimated by shuffling species across all lakes (unconstrained) or within primary watersheds (constrained). The Local Contribution to Beta Diversity of each lake was used to estimate the uniqueness (beta) of each lake. Relationships were estimated using a linear model.

SES_{mean}

	Unconstrained			Constrained		
Predictors	*Est.*	*t*	*p*	*Est.*	*t*	*p*
(Intercept)	-0.00	-0.00	1.00	0.00	0.00	1.00
beta	-0.49	-14.95	<0.001	-0.43	-12.44	<0.001
Observations	697			697		
R^2/R^2 adjusted	0.243/0.242			0.182/0.181		

SES_{SD}

	Unconstrained			Constrained		
Predictors	*Est.*	*t*	*p*	*Est.*	*t*	*p*
(Intercept)	0.00	0.00	1.00	0.00	0.00	1.00
beta	-0.42	-12.13	<0.001	-0.30	-8.19	<0.001
Observations	697			697		
R^2/R^2 adjusted	0.175/0.174			0.088/0.087		

Box 1. Co-occurrence analyses in community ecology

Co-occurrence analyses aim to detect association patterns between species pairs that are either aggregated, segregated, or random. Aggregated species occur together more often than expected by chance alone. This pattern suggests assembly mechanisms that lead species to share the same communities such as common habitat preferences, parallel dispersal patterns and, for some taxa, mutualism and predator-prey tracking (Morales-Castilla et al. 2015). Segregated species do not occur across the same sites together and may be a result of negative assembly mechanisms like competition and predation (when preys avoid their predators in space and time), differences in habitat requirements, disparate dispersal abilities or patterns, and past biogeographic histories (Diamond 1975, Leibold et al. 2010, Morales-Castilla et al. 2015, Lyons et al. 2016). Random or null SAPs involve species where there are no distinct patterns and may be a result of equivalency in habitat and variation in dispersal abilities with no species interactions. SAPs can also vary across the landscape, with species having aggregated patterns in certain environments and segregated patterns in others (Bar-Massada and Belmaker 2017, Clark et al. 2018). For species like fishes for which several abiotic variables (e.g., pH, water temperature) have a direct effect on physiology and growth, abiotic variables can have a direct influence on the trophic interactions that can occur and the resulting observed patterns (Mandrak 1995, Hein et al. 2013, Myers et al. 2017).

Appendix I – Simulated Markov networks with covariates

One aspect of Markov network models with covariates is the ability to control for the effect of disparate or similar environmental responses on species co-occurrence patterns. For example, species with opposite responses to the same environmental gradient can appear to have a negative co-occurrence pattern. To demonstrate in a more intuitive way to the difference between a model which includes environmental covariates and one that does not, we simulated three scenarios using 1000 communities with three species, along an environmental gradient, and compared estimated species co-occurrence coefficients between a model which considered environmental covariates and one that did not. With these simulations, we hope to illustrate how the incorporation of environmental covariates reduces the strength of conditional relationships between species with shared or distinct environmental preferences and show how conditional relationships between species can change across environments.

In scenario one, we simulated the presence-absence of two species (spc1 and spc2) with opposite environmental responses, and a third random species (spc3) (Fig. S1). Estimated species pair coefficients between spc1 and spc2, were on average closer to zero when the environment was included in the model, compared to the model which did not include the environment. Without environmental covariates, a negative co-occurrence pattern was detected.

Adding covariates also allows for co-occurrence patterns to change across environmental gradients. For scenario two, we randomly simulated the presence-absence of two species (spc1 and spc3) and had the occurrence of the third species (spc2) dependent on the presence-absence of spc1. We varied this dependency along a continuous environmental gradient so that in some environments, spc1 and spc2 were positively associated with one another, and in other environments they were negatively associated with each other (Fig. S2). In the predicted local networks, the network model which includes the environment, the co-occurrence pattern between spc1 and spc2 switches from positive to negative along the gradient (Fig. S3). Contrarily, the model without the environment does not predict any changes in the co-occurrence pattern between spc1 and spc2.

Lastly, for scenario 3, we combined both scenario 1 and scenario 2, so that the presence-absence of spc1 was dependent on the env, and the presence-absence of spc2 was dependent on spc1 and the dependency changed with the environment (Fig. S4). In the model with the environment, associations between spc1 and spc2 were detected, with the pattern switching from positive to negative along the gradient (Fig S5). Without including the environment, the model did not detect any species association patterns between spc1 and spc2.

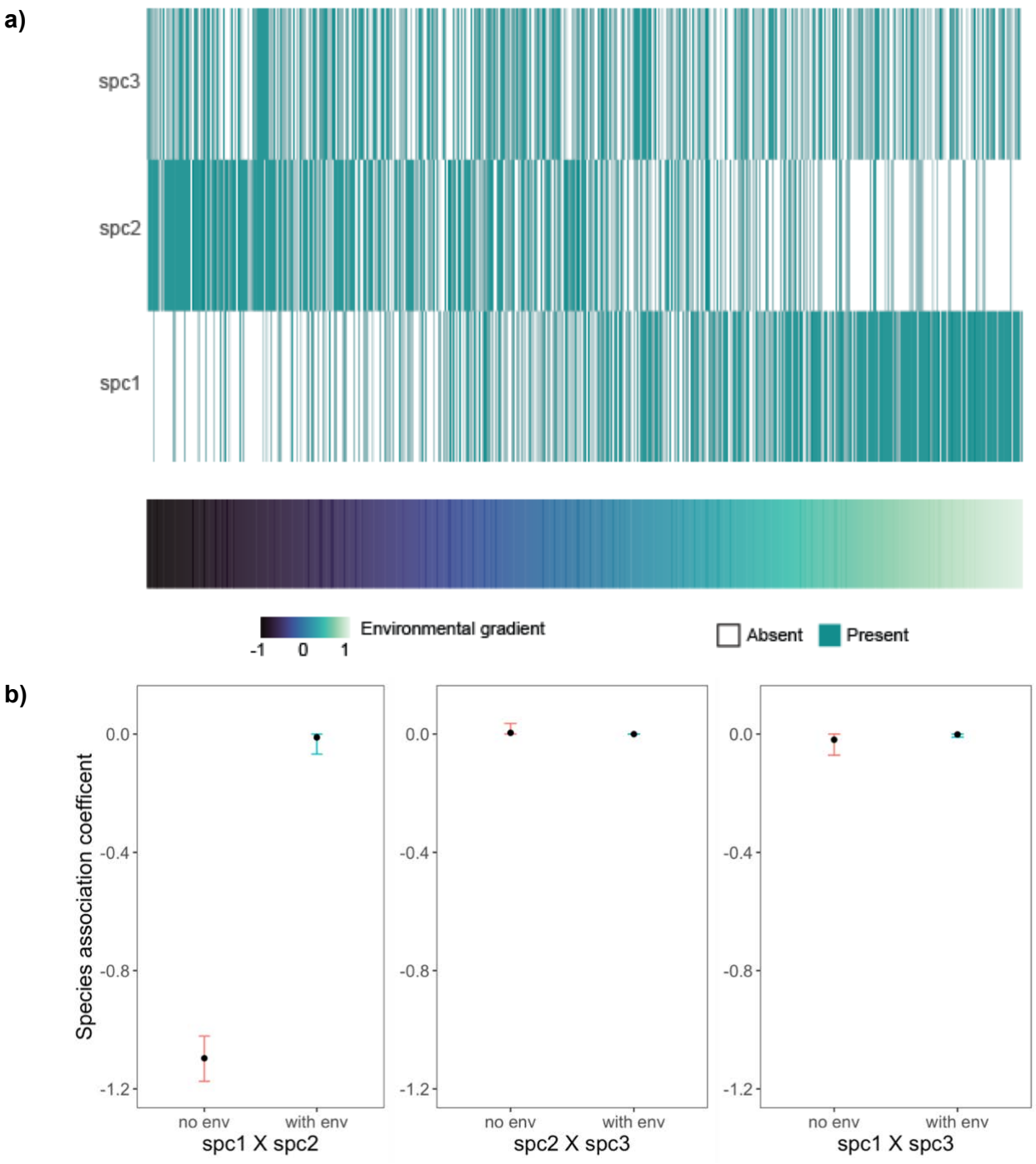

Figure S1. Scenario 1: Markov model simulation with covariates. The presence absence (a) of 2 species (spc1 and spc2) with opposite environmental responses, and a third random species (spc3) were simulated for 1000 sites along an environmental gradient. Predicted species association coefficients for species pairs (b) were predicted using a Markov model with and without the environmental gradient.

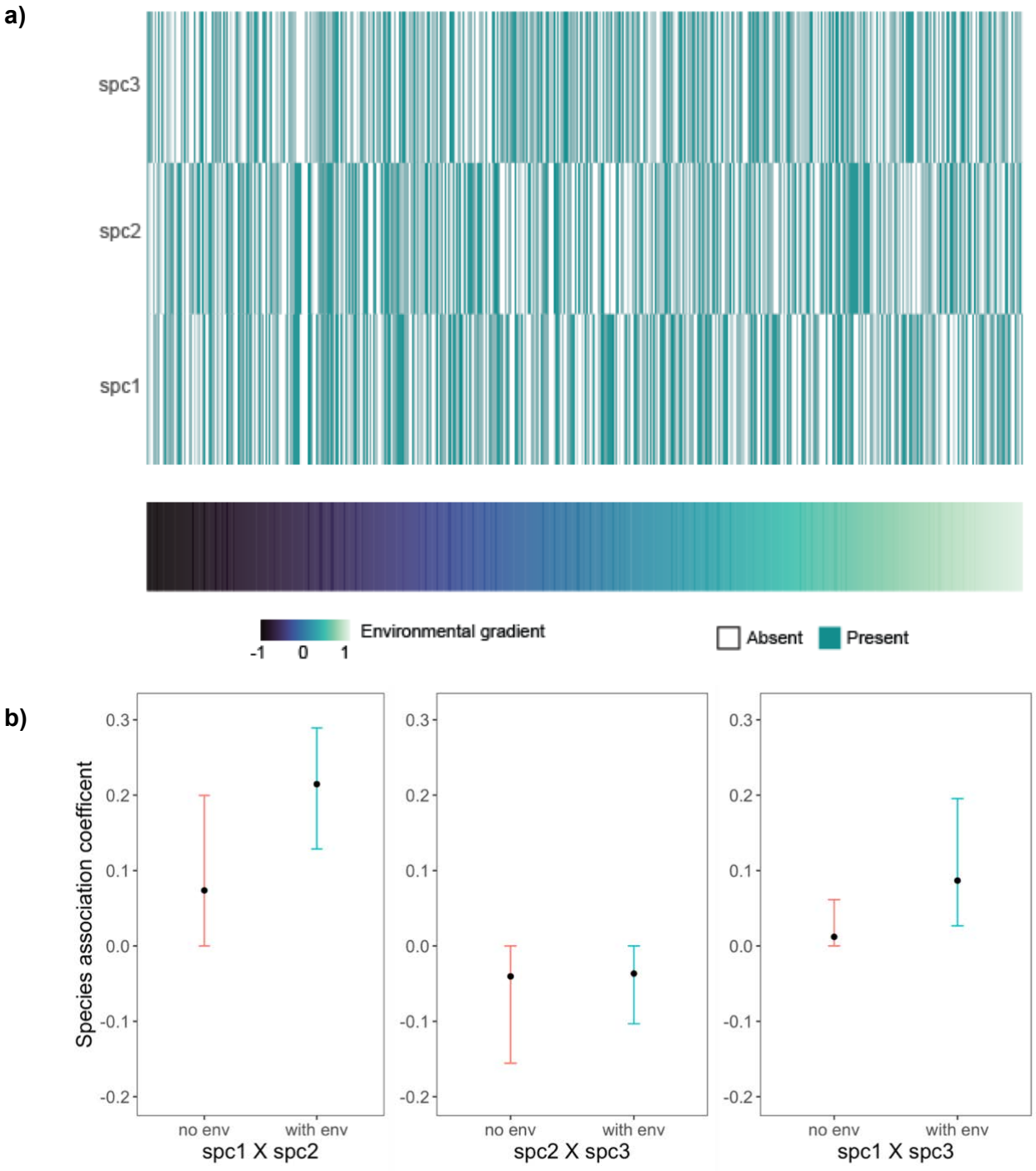

Figure S2. Scenario 2: Markov model simulation with covariates. Presence-absence (a) of spc1 and spc3 were randomly simulated for 1000 sites. Presence-absence of spc2 was dependent on spc1 and varied with the environment so that in more negative environments they were positively associated but negative in positive environments. Predicted species association coefficients for species pairs (b) were predicted using a Markov model with and without the environmental gradient.

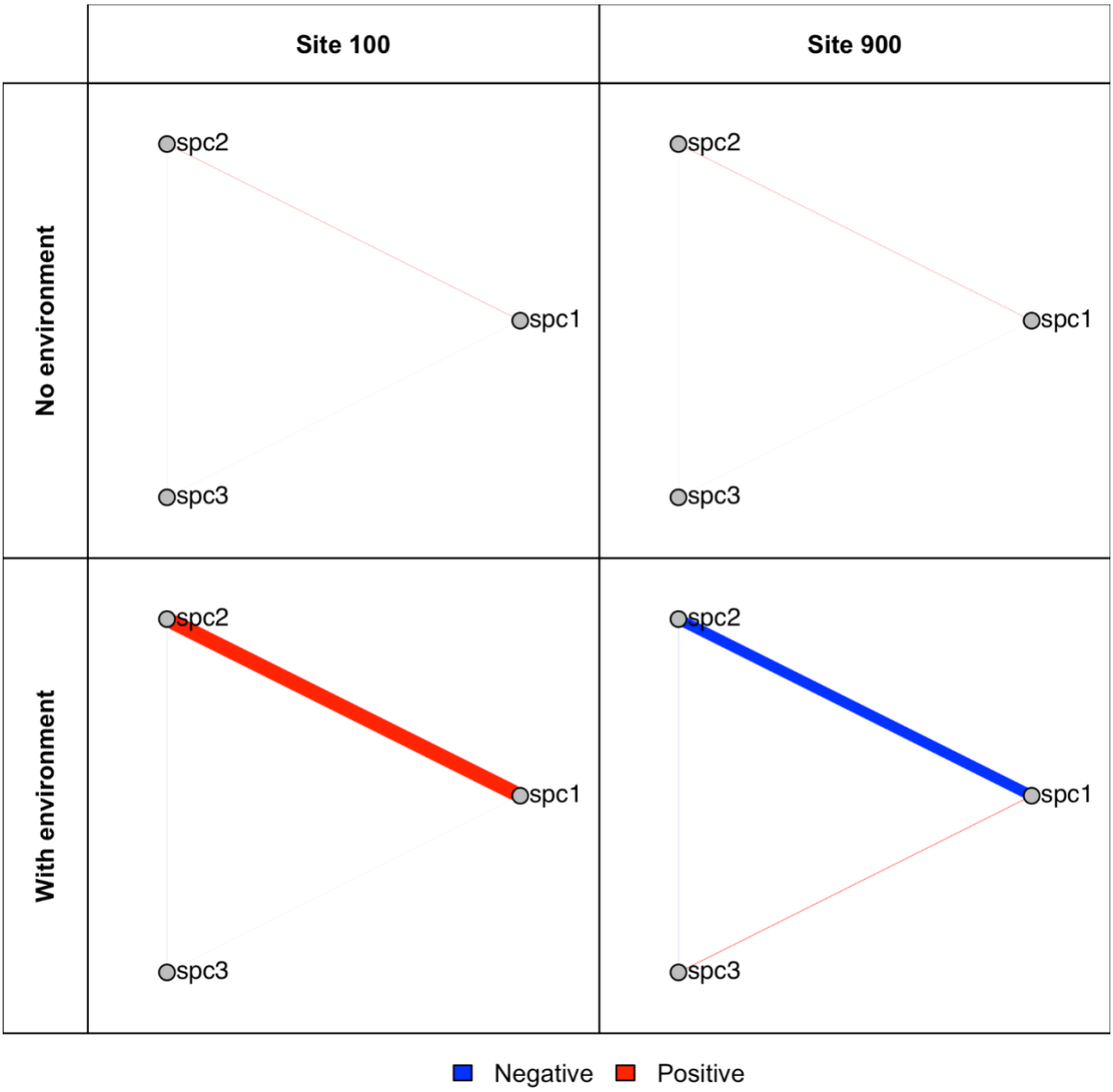

Figure S3. Scenario 2: Predicted species co-occurrence networks. Local networks were estimated for site 100 and 900, comparing estimated species associations for local sites modeled with and without the environment.

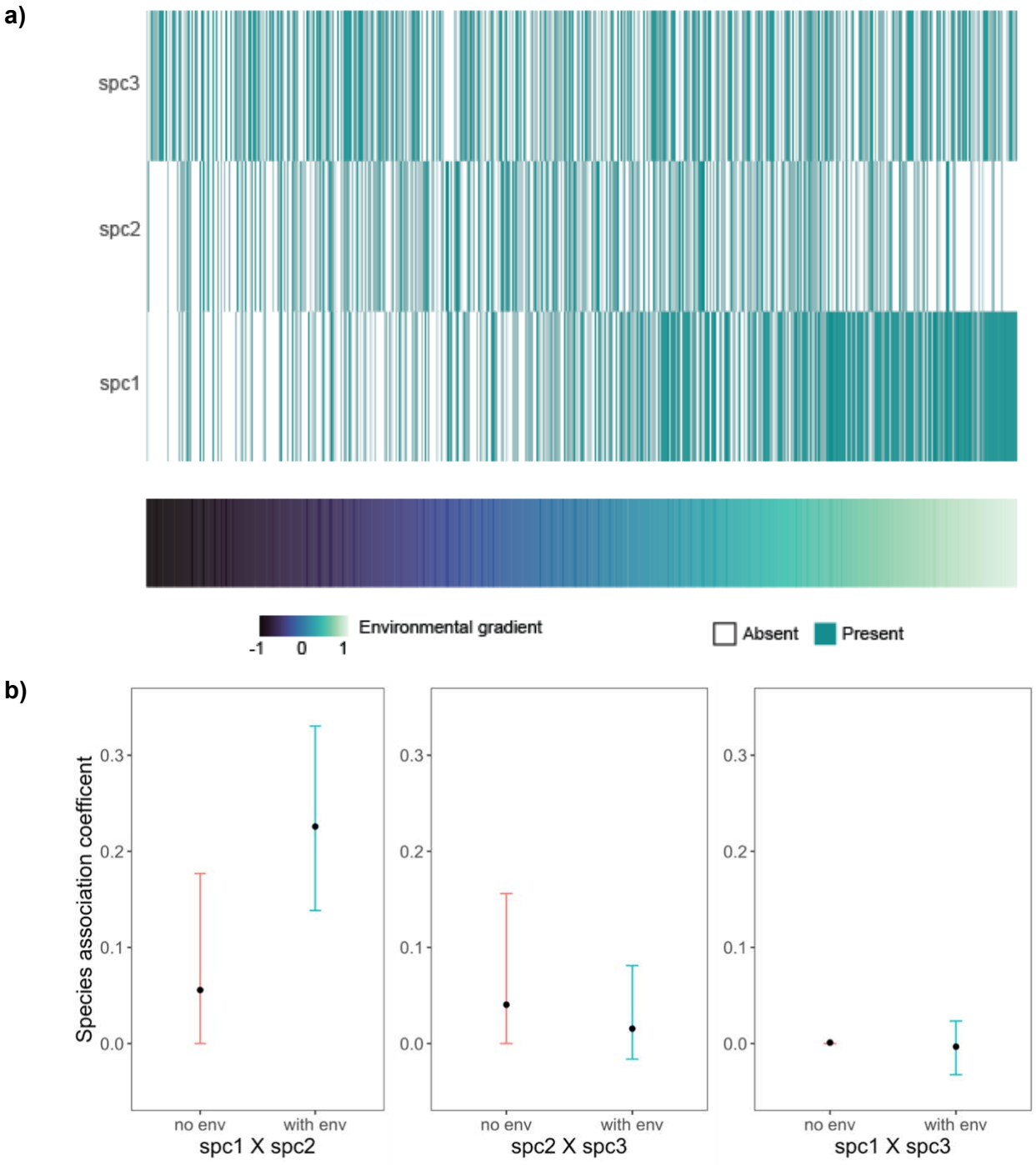

Figure S4. Scenario 3: Markov model simulation with covariates. Presence-absence (a) of spc1 and spc3 were randomly simulated over 1000 sites. Presence-absence of spc2 was dependent on spc1 and the environment (more positive in negative environments). The strength of the dependency on the environment was also dependent on the environment. Predicted species association coefficients for species pairs (b) were predicted using a Markov model with and without the environmental gradient.

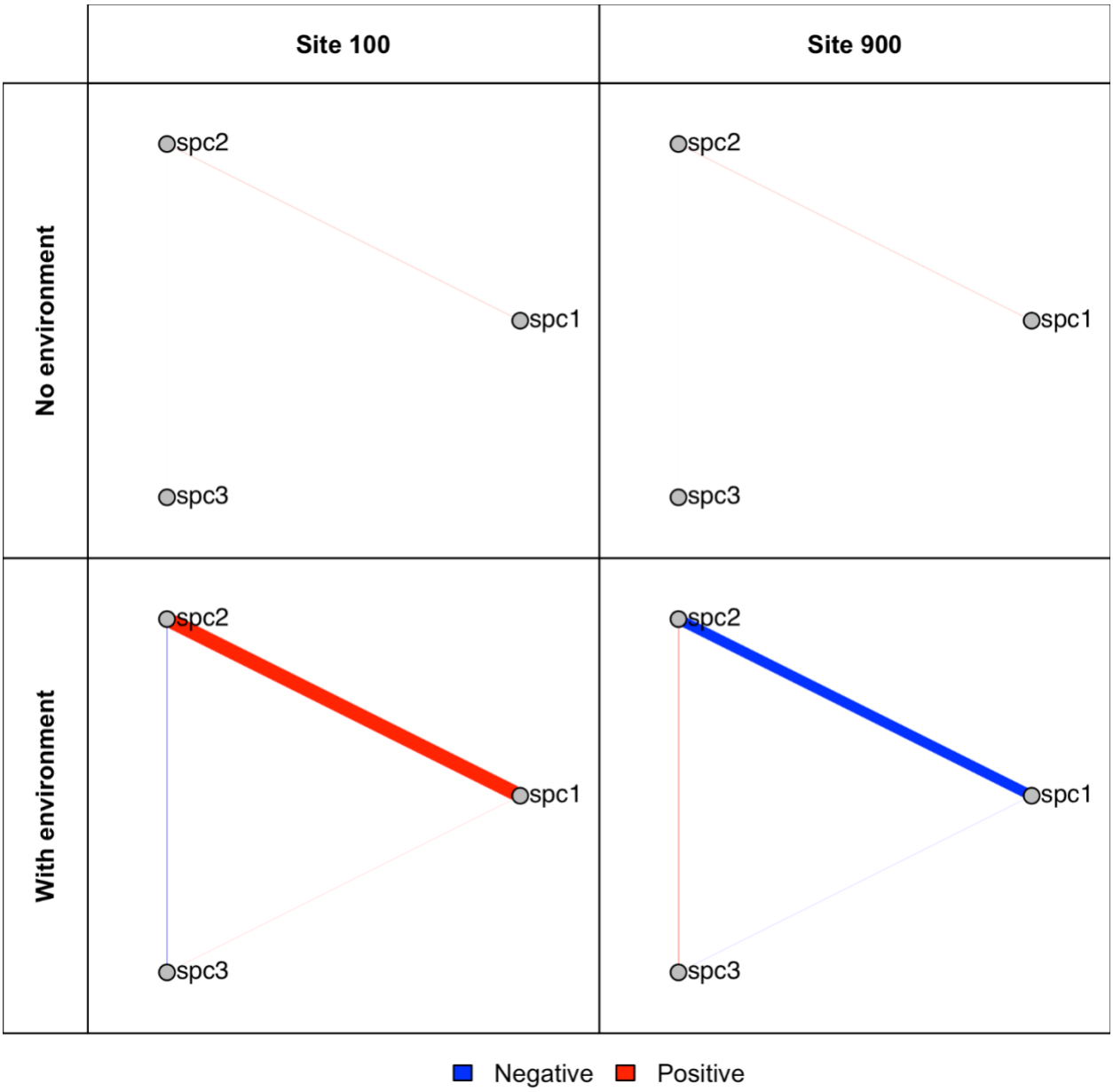

Figure S5. Scenario 3: Predicted species co-occurrence networks. Local networks were estimated for site 100 and 900, comparing estimated species associations for local sites modeled with and without the environment.

Appendix II – Cross-validation results of the final global Markov network model used to estimate species association patterns

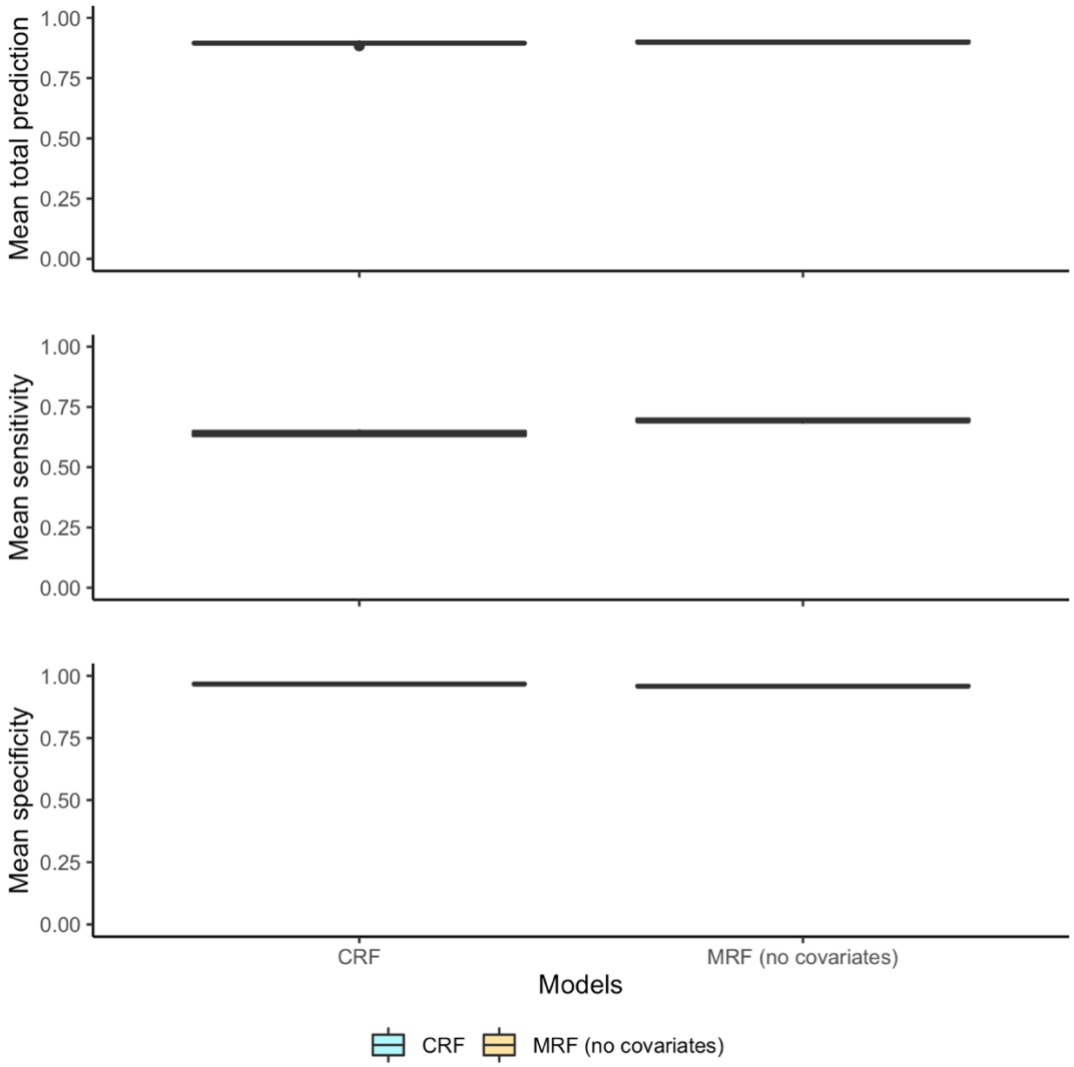

Figure S6. Mean total prediction, sensitivity and specificity of the global Markov model estimated with environmental covariates (CRF) and without (MRF). Total prediction describes the proportion of data correctly predicted. Sensitivity refers to the true positive rate and specificity refers to the true negative rate. Performance of the final global Markov network model was assessed and compared to a model without environmental covariates using a 10-fold cross validation, repeated 500 times. For each repetition, we calculated three metrics; total prediction (proportion of the data correctly predicted), sensitivity (true positive rate) and specificity (true negatives). On average, the two models did not differ greatly in total prediction, however, for the model with environmental covariates, sensitivity was lower but with higher specificity.

Appendix III – Results using boosted regression trees to predict pairwise species association patterns

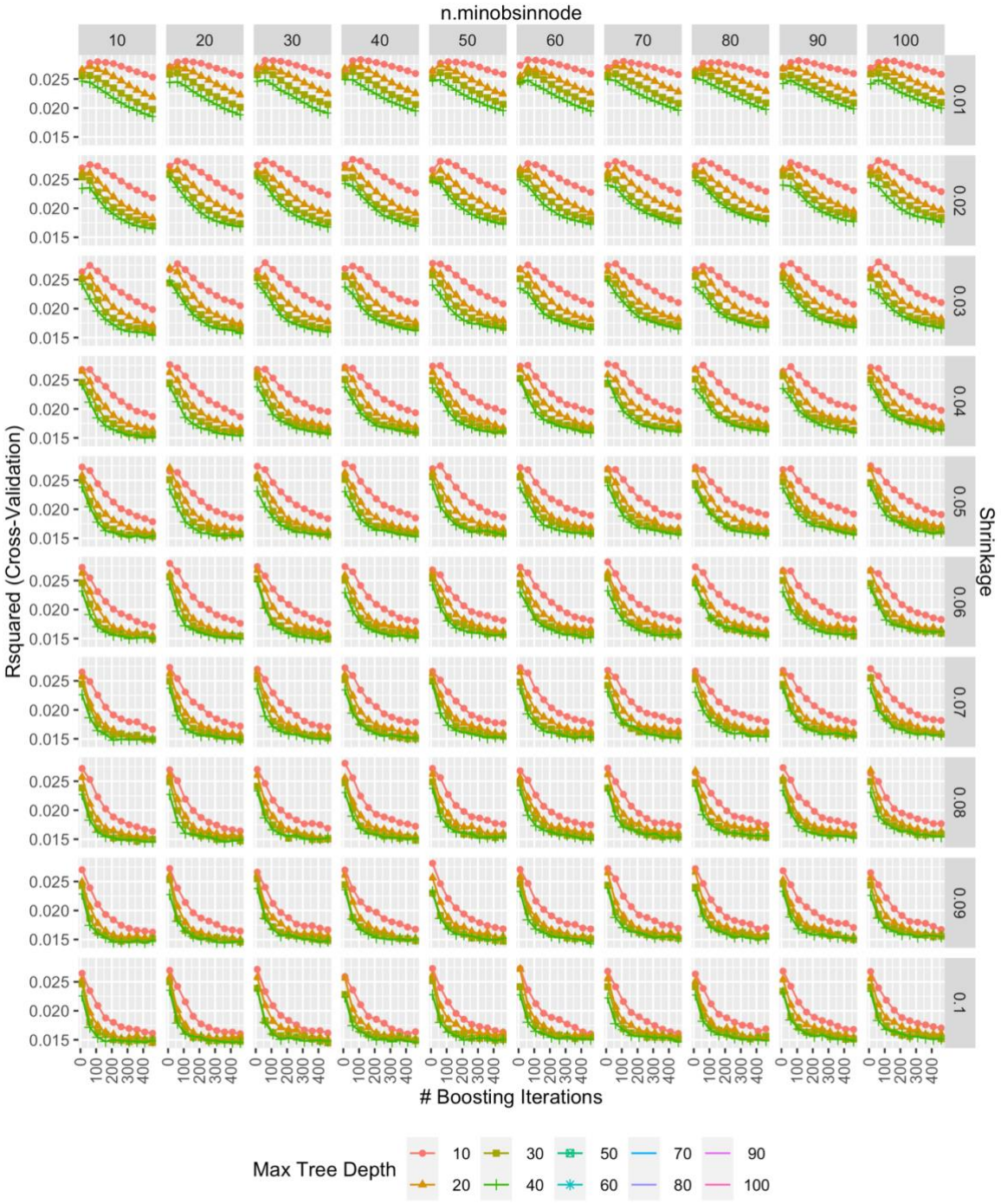

Figure S7. Parametrization of the boosted regression tree predicting pairwise species association patterns across all pairs using all environmental variables. The final model had an R^2 of 0.0395 and was fit using a shrinkage parameter of 0.02, with 60 trees with a maximum tree depth of 10, and a minimum of 40 observations per node.

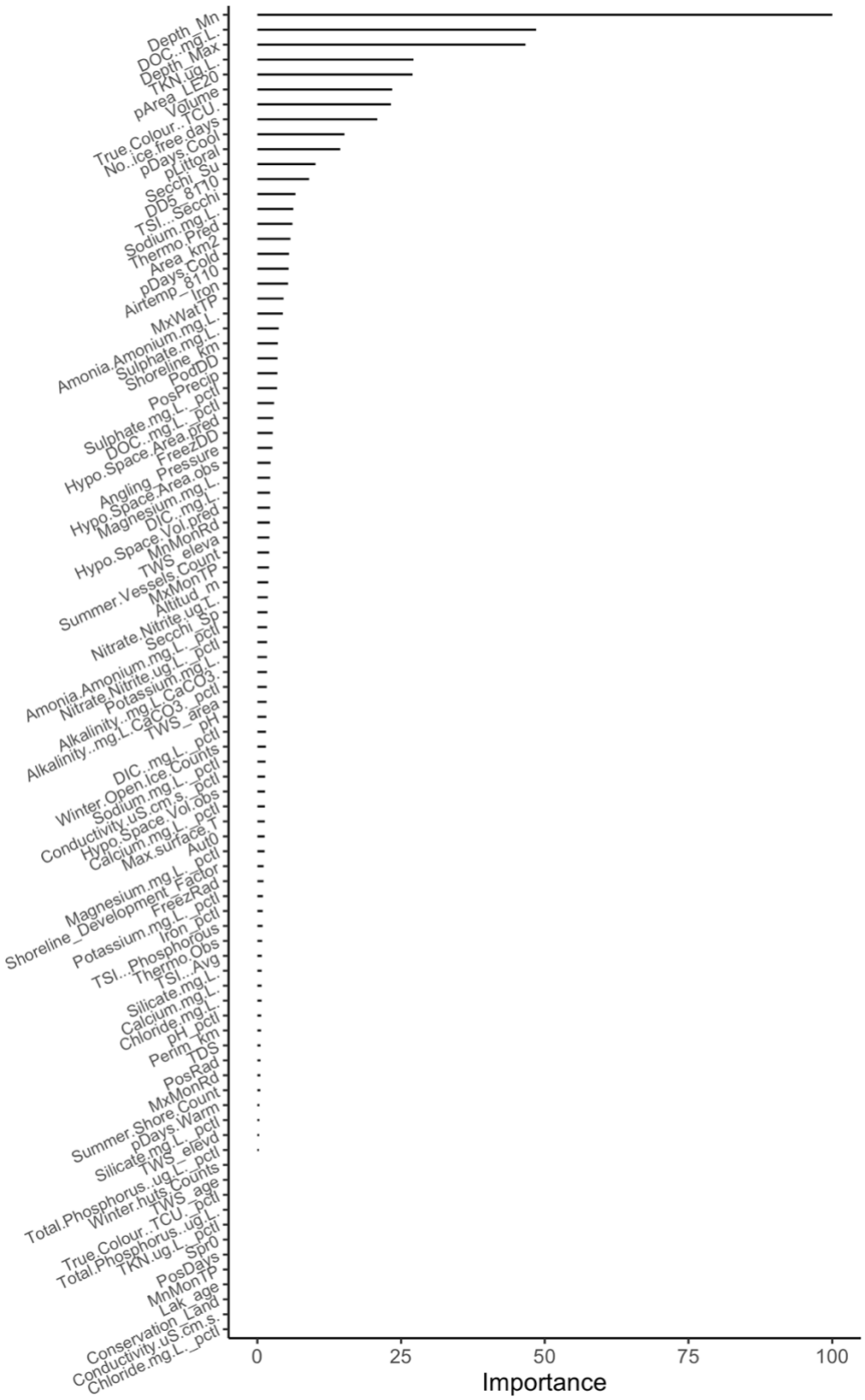

Figure S8. Environmental variables ordered by importance by the boosted regression tree for predicting pairwise species association patterns across all pairs. Of the 89 variables, 77 had a non-zero influence on the chosen boosted regression tree.

Appendix IV – Results using an unconstrained null model

Relationships between the environment and SES mean and SD values estimated using the unconstrained null model resulted in different environmental variables being selected. Like with the constrained null model, climate variables were the strongest predictors of SES mean (Fig. S9). SES mean was negatively correlated with maximum surface temperature and proportion of days that were cold during ice free days, and positively correlated with the average number of degree days above 5°C. While the other two variables were also strong predictors in the constrained null model (Fig. 3), degree days above 5°C was a weaker predictor. Results differed more with SES SD, where the strongest predictors, degree days above 5°C, predicted thermocline depth, and shoreline length (Fig. S10), were not strong predictors in the constrained model (Fig. 4). This suggests that the effect of historical processes, as controlled using the constrained null model may differ depending on the metric used to asses community level species association patterns. SES SD showed the greatest sensitivity, and is exemplified by the difference in the strength of the predicted relationship using the constrained (adjusted R^2 = 0.591) and unconstrained model (adjusted R^2 = 0.341) versus SES mean which had a difference in adjusted R^2 values of 0.002. The small differences in SES mean models but large differences in SES SD models between the constrained and unconstrained model suggests that the ability to predict the average effect of mechanisms on community structure does not depend on the species pool, but should be considered if the goal is to predict the variance of the effect.

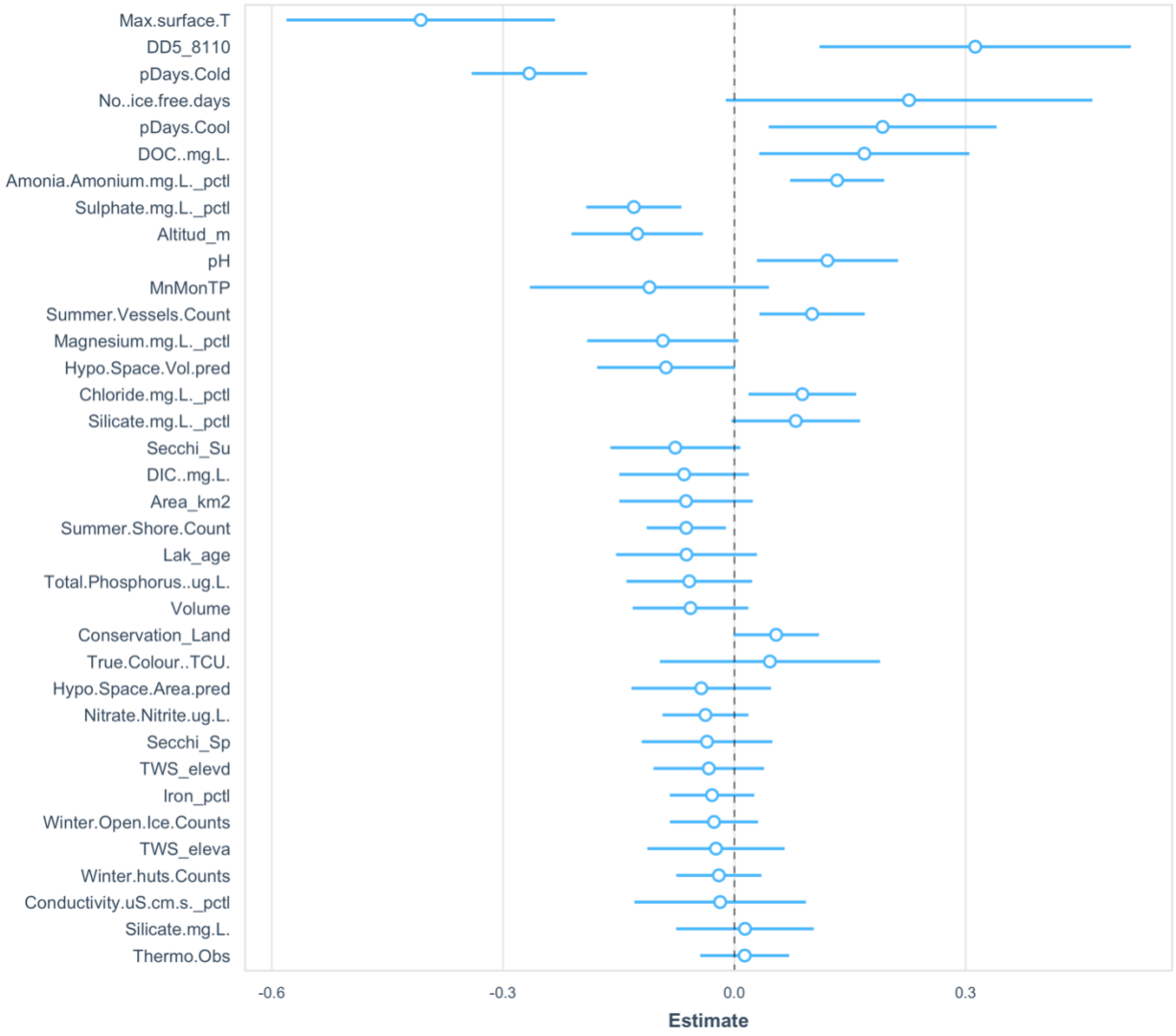

Figure S9. Estimated regression coefficients for SES mean estimated using the unconstrained null model and the environment. Coefficients are estimated using a linear model, with environmental variables in the model selected using LASSO.

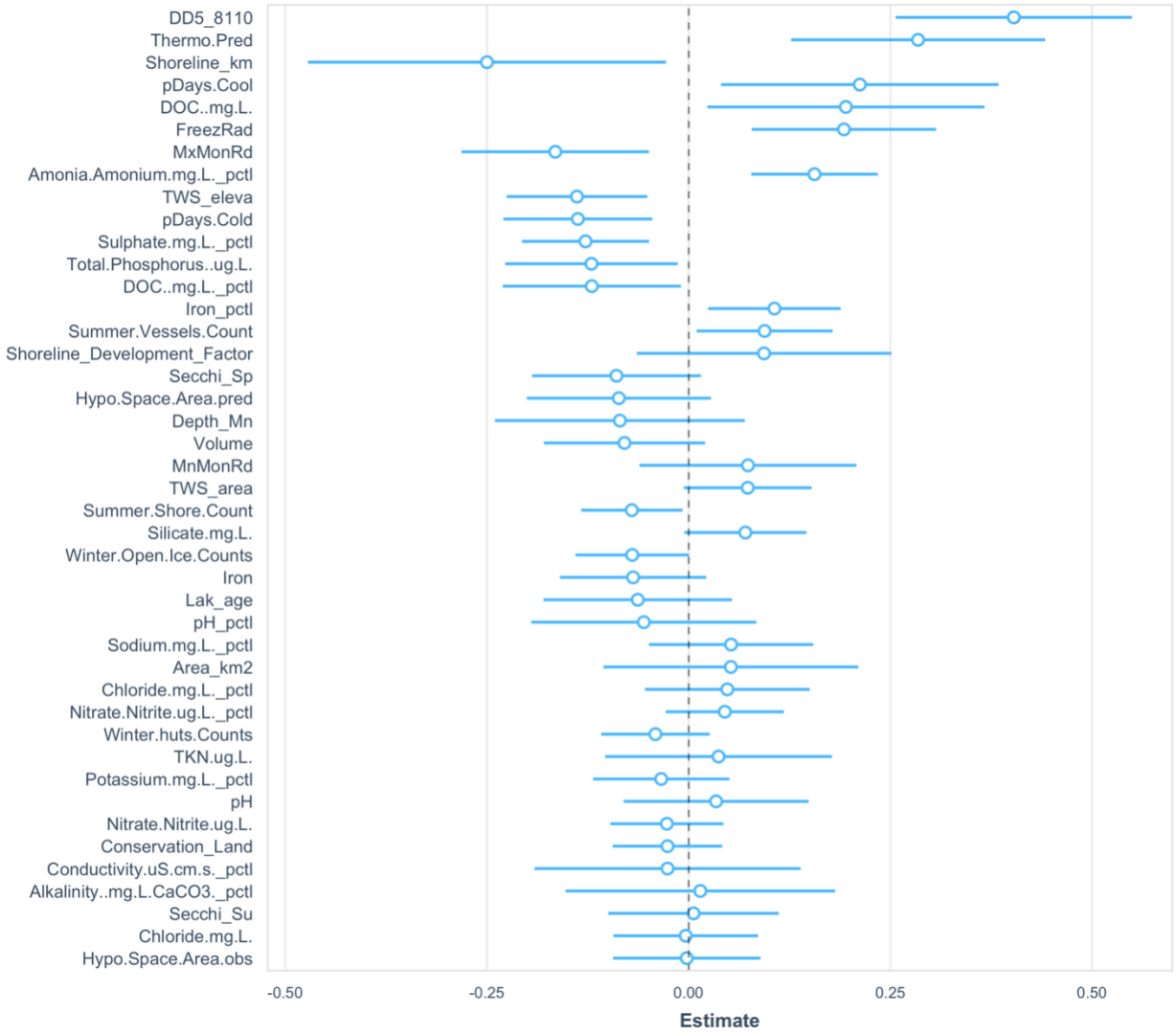

Figure S10. Estimated regression coefficients for SES SD estimated using the unconstrained null model and the environment. Coefficients are estimated using a linear model, with environmental variables in the model selected using LASSO.

Appendix V – Predicting variation in community-level patterns of species association by geographic location and assessing spatial autocorrelation

We tested if latitude and longitude could predict variation in community-level patterns of species associations. As environmental variables like climate are frequently correlated with latitude, we were interested if a simply metric like latitude and longitude could also predict SES mean and SD. Using variation partitioning as implemented in the R package vegan (Oksanen et al. 2020), we partitioned the proportion of variation explained by the environmental variables selected by LASSO and geographic location together and exclusively. Here, we present results using community-level patterns calculated using the constrained null model. For SES mean, 17% of the variation explained by environmental variables was jointly explained by geographic location (Fig. S11). For SES SD, the proportion was higher, with 25% of the variation explained by environmental variables jointly explained by geographic location (Fig. S12). Our results show that a majority of the variation explained by the environment cannot be explained by just geographic location alone.

Lastly, to assess if we were missing any environmental variables that may explain variation in SES mean and SD, we calculated the spatial autocorrelation of our models. Spatial autocorrelation can increase Type 1 errors (e.g. detecting a relationship with the environment when there is none)(Diniz-Filho et al. 2003). Spatial autocorrelation in the residuals produced from the LASSO models can suggest that additional variables that can capture variation in SES mean and SD are missing. We calculated Moran's I for the residuals from the models for 15 distance classes as implemented in the R package pgirmess (Giraudoux 2021). Moran's I is a measure of spatial autocorrelation with values between 1 and -1, where a value of -1 represents dispersed residuals, 1 represents residuals that are aggregated, and a value of 0 represents residuals that are randomly dispersed. For both SES mean and SES SD, Moran's I was close to 0 across all distance classes (Fig. S13 and S14). These results show that residuals for both models are not spatially autocorrelated, suggesting that we are not missing any environmental variables that may explain variation in community-level patterns of species associations.

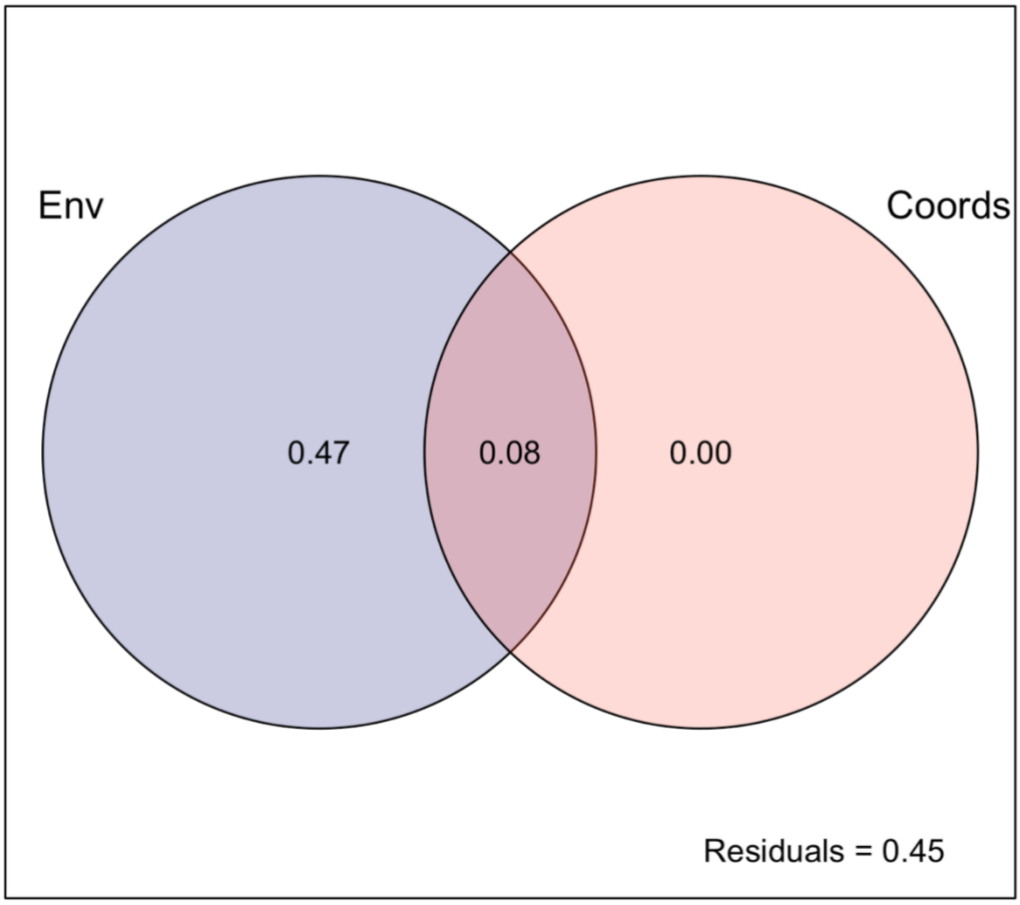

Figure S11. Variation partitioning of SES mean (n = 697) calculated using the constrained null model using environmental variables selected using LASSO and the latitude and longitude of each lake. The Venn diagram illustrates the proportion of variation that is explained by the environment (Env) and coordinates (Coords) respectively, and the proportion of variation that is jointly explained.

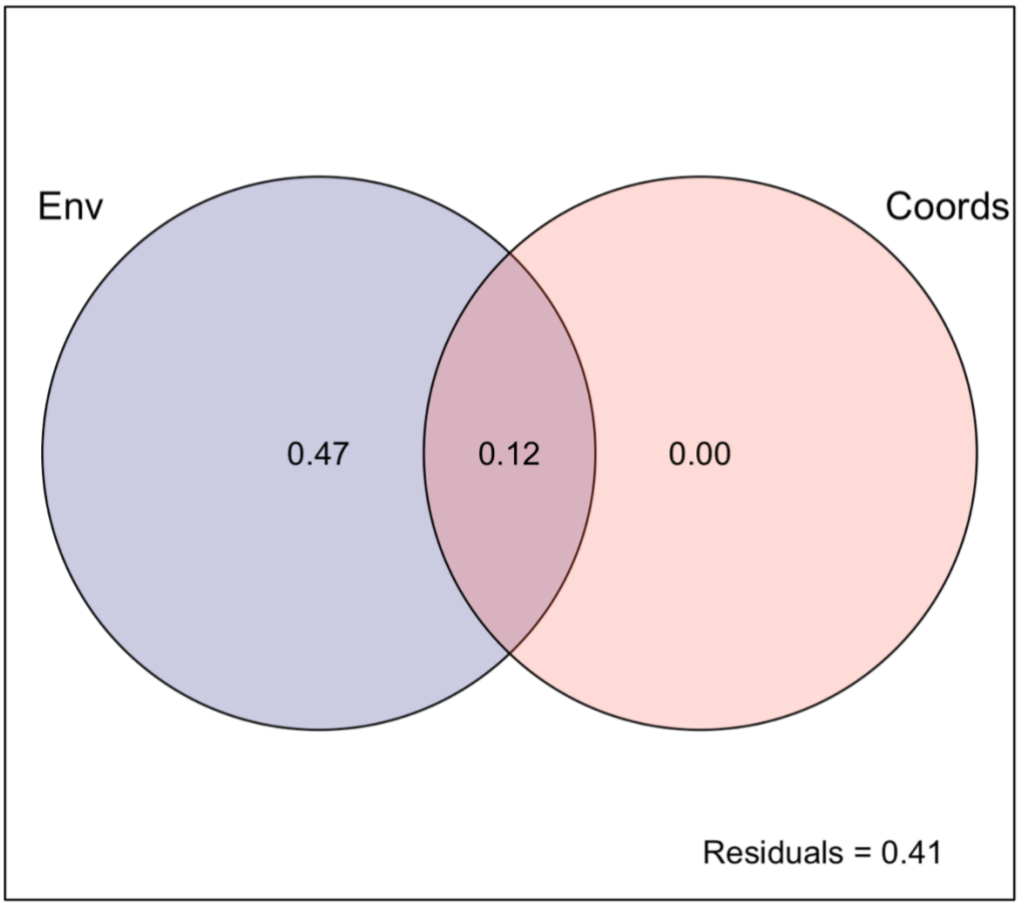

Figure S12. Variation partitioning of SES SD (n = 697) calculated using the constrained null model using environmental variables selected using LASSO and the latitude and longitude of each lake. The Venn diagram illustrates the proportion of variation that is explained by the environment and coordinates respectively, and the proportion of variation that is explained jointly.

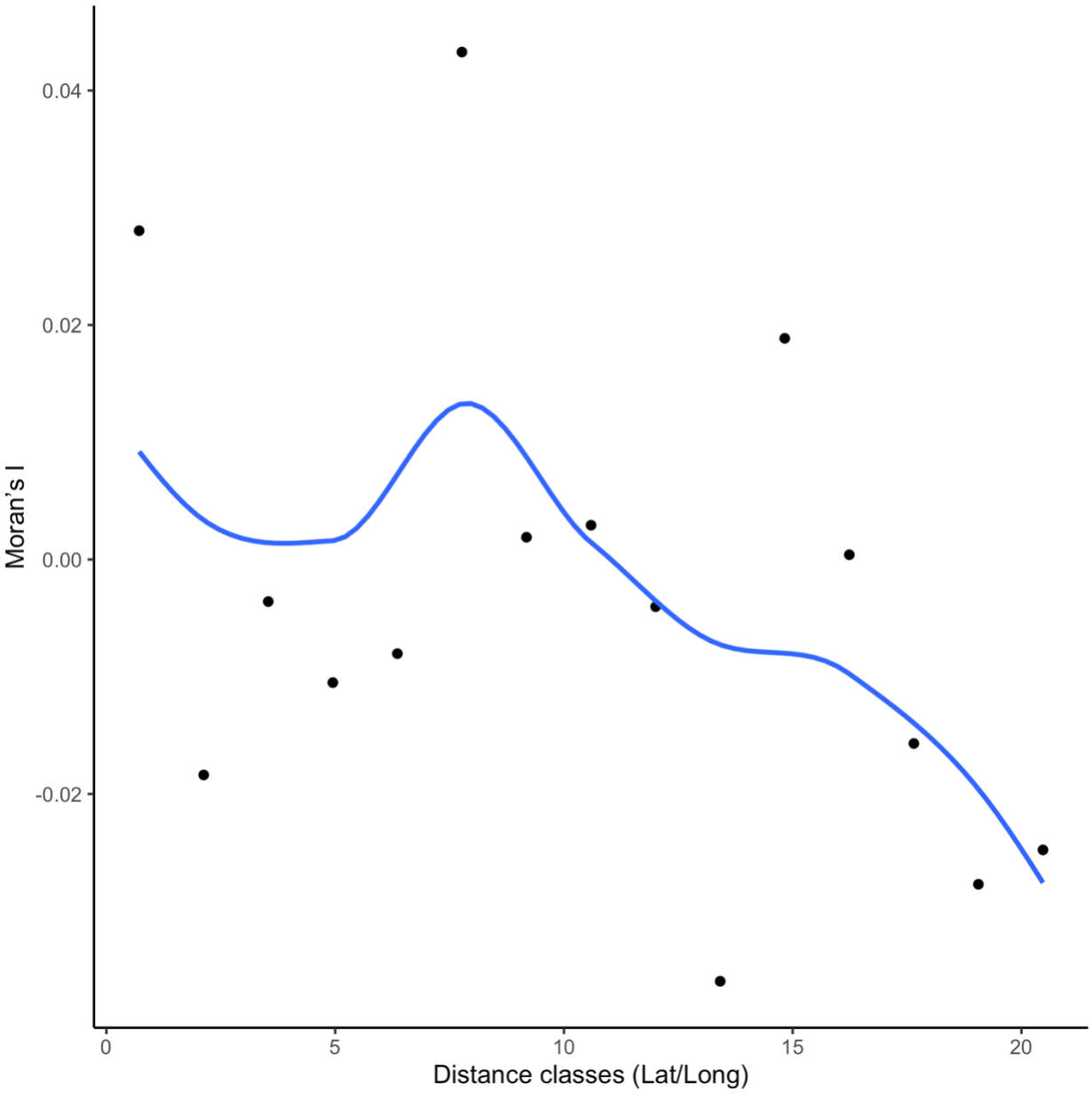

Figure S13. Assessing spatial autocorrelation of residuals from the LASSO model: SES mean and the environment. SES mean values were calculated using the constrained null model. Correlogram shows the change in Moran's I across increasing distance classes (distance 0 refers to closest neighbours, distance 1 for second-order neighbours, etc.).

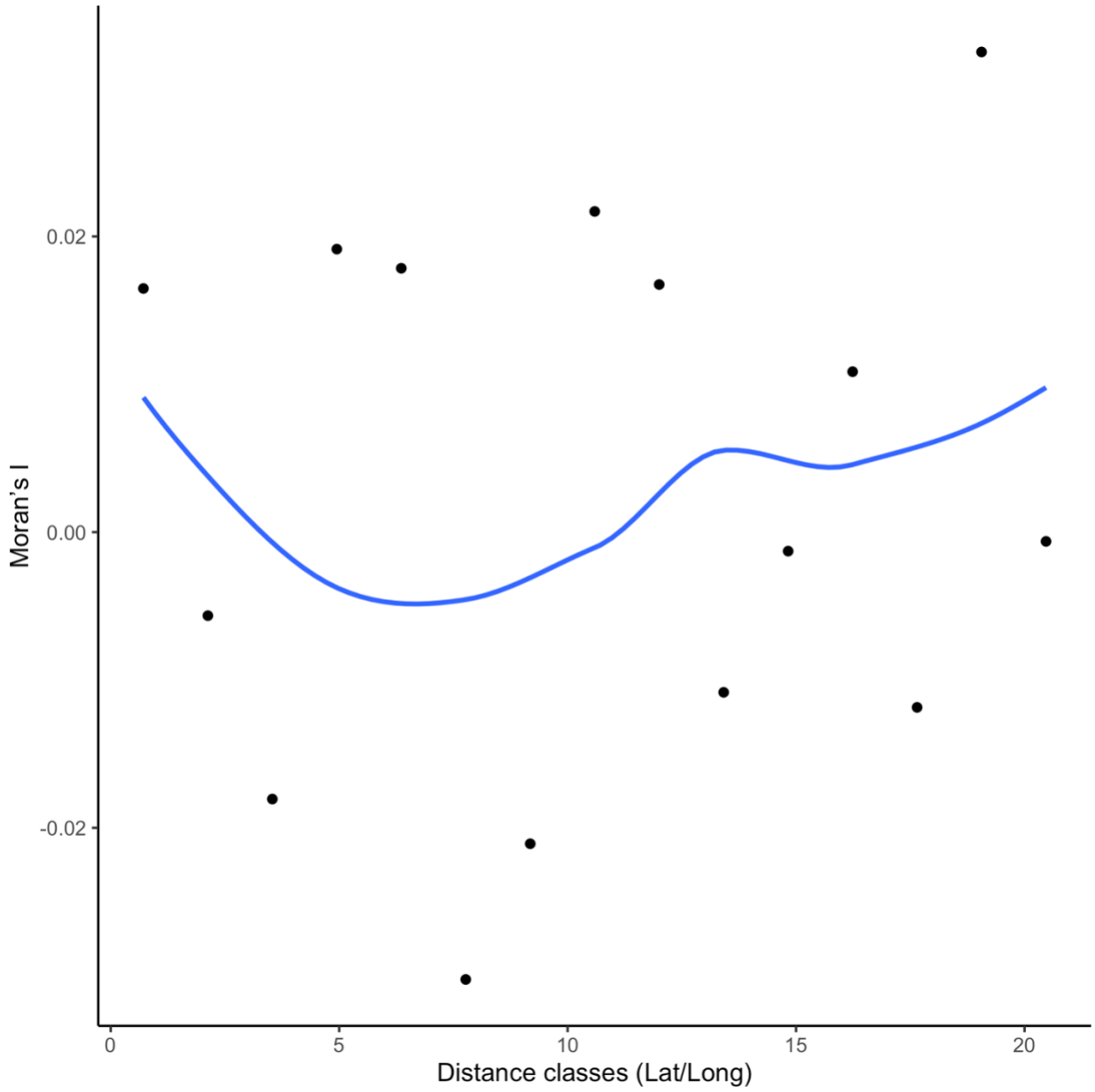

Figure S14. Assessing spatial autocorrelation of residuals from the LASSO model: SES SD and the environment. SES SD values were calculated using the constrained null model. Correlogram shows the change in Moran's I across increasing distance classes (distance 0 refers to closest neighbours, distance 1 for second-order neighbours, etc.).