

Modelling habitat distributions for multiple species using phylogenetics

Guillaume Guénard^{1*}, Gabriel Lanthier^{1†}, Simonne Harvey-Lavoie^{1‡},
Camille J. Macnaughton^{1§}, Caroline Senay^{1¶},
Michel Lapointe^{2||}, Pierre Legendre^{1**}, and Daniel Boisclair^{1††}

23rd August 2016

1. Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre-Ville
Montréal, QC, Canada H3C 3J7
2. Department of Geography
McGill University
Burnside Hall, 805 Sherbrooke W, Room 619
Montreal, QC, Canada H3A 2K6

Abstract

In this paper, we describe an empirical approach to model community structure using phylogenetic signals. That approach combines information about the species (i.e. traits and phylogeny) with information about the habitat (i.e. environmental conditions and spatial distribution of sampling sites) and their interactions to predict the species responses (e.g. the local densities). As an application, we use the approach to model fish densities in rivers. In the model, the different species and size classes were described using a functional trait, body length, and phylogenetic eigenvectors maps whereas the sites were described using water velocity, depth, substrate composition, macrophyte cover, degree-days, total phosphorus, and spatial eigenvector maps. The model (estimated using a regularised Poisson-family Generalised Linear Modelling approach) fitted the data well (likelihood-based $R^2_{adj} = 0.512$) and showed fair predictive power (likelihood-based cross-validation $R^2 = 0.283$) to predict the density of

*guillaume.guenard@gmail.com (corresponding author)

†gabriel.lanthier@umontreal.ca

‡simonne.harvey-lavoie@umontreal.ca

§camille.macnaughton@gmail.com

¶caroline.senay@umontreal.ca

||Michel.Lapointe@mcgill.ca

**pierre.legendre@umontreal.ca

††daniel.boisclair@umontreal.ca

fish pertaining to 48 species totalling 143 combinations of species and size classes in 15 unregulated Canadian rivers. Using the model as a baseline to estimate the effect of flow regulation on community composition, we found that, with few exceptions, the densities of most fish species were lower in regulated than in unregulated rivers. Phylogenetics have been proposed to study community structure, but this is, to our knowledge, the first time phylogenetic information is used explicitly for numerical habitat modelling. We expect that models of that type will be in increasing demand now that development projects are routinely assessed through impact studies.

Language: English (UK)

Keywords: Evolution, Species conservation, Water resource management, bilinear model, PEM, MEM, elastic net regression

Running title: Predicting species composition using phylogeny

Accepted for publication in Ecography

Introduction

Numerical Habitat Modelling (NHM; also known as species distribution models: SDM; Elith and Leathwick, 2009) seeks to explain and predict the distribution of organisms as a function of the environment (Guisan and Zimmermann, 2000; Boissclair, 2001; Guisan and Thuiller, 2005). Species distributions result from the interactions between species traits and environmental conditions (Dirnböck and Dullinger, 2004; McGill et al., 2006). A numerical habitat model can be build empirically after surveying the distribution of a species and that of the environmental (abiotic and biotic) conditions, exploring the resulting data for the presence of relationships. Building habitat models for multiple species classically involves repeating the exercise multiple times, each time producing a single model adapted to the requirements and particularities of a single species. Since life is so diverse (5 ± 3 million species, of which 1.5 million species are named; Costello et al., 2013), obtaining habitat models for any large number of species quickly becomes an impractical endeavour.

Species traits (e.g. physiological, behavioural) are structured with respect to phylogeny because they are the product of evolution (Felsenstein, 1985; Pillar and Duarte, 2010). Methods now exist that allow model builders to predict trait values using among-species phylogenetic patterns of trait variation, for instance, Phylogenetic Eigenvector Maps (PEM, Guénard et al., 2013, see also Diniz-Filho et al., 1998; Garland and Ives, 2000; Desdève et al., 2003; Ollier et al., 2006; Pavoine et al., 2008; Hardy and Pavoine, 2012; Swenson, 2014 for related approaches, and Diniz-Filho et al., 2015, for further evaluation of PEM’s properties in representing processes underlying trait evolution). Phylogenetic eigenvector methods produce sets of explanatory variables that have been used in regression models to predict species traits like, for instances, tolerance to toxic substances (Guénard et al., 2011) or metabolic costs (Guénard et al., 2015). Since any phylogenetic eigenvector can be regarded as a set of values of an underlying phylogenetic eigenfunction, and because we can calculate values for given nodes or other points on the phylogeny, PEM can be used to make predictions for related species that are outside of the model’s training data. Phylogenetic modelling provides modellers with a way to optimise information use by making a synthesis of the common features shared by multiple related species. That synthesis allows one to impute unknown species traits from that of phylogenetic relatives. It is therefore especially useful in cases when information is difficult to obtain or when many species are involved.

In addition to directly predicting trait values, PEM can be used to model phenomena that are the outcome of complex interactions between species traits and other factors. For instance, PEM have been used to model species tolerances of multiple species of a phylogeny to multiple organic pesticides (Guénard et al., 2014) and heavy metals (Malaj et al., 2015). In these studies, descriptors of among-compound variability in toxicity (e.g. toxic mode of action) were used alongside phylogenetic eigenvectors. Since habitat selection is a behavioural trait associated with the physiological requirements of each species and given that we can quantify habitat suitability as being a function of environmental variables, Numerical Habitat Modelling is another area that may benefit from the PEM analytic framework.

Phylogenetically-explicit habitat modelling brings new perspectives with respect to

producing multiple single species models. Rather than modelling single species and repeating the exercise for each species of interest, phylogenetic models target a group of species using estimated evolutionary interrelationships (i.e. the phylogeny). Rather than being the direct focus of modelling, columns of species data table are seen as a sample representing a larger population of potential species. Examples of such species may encompass species present at the study sites but not sampled for logistic or conservation reasons, potentially invasive species that have not yet reached the study site, locally extirpated species that have been targeted for reintroduction, extinct species, and colonising ancestral species. Phylogenetic habitat modelling can therefore handle tasks like, for instance, modelling the habitat of rare or endangered species, modelling the invasion of alien species, exploring habitat restoration scenarios, or testing biogeographic colonisation hypotheses.

In the present study we present a framework to develop multiple-species, phylogenetically-explicit habitat models. As an example, we employed that framework to model the habitat of river fish communities. Since fish is a diverse group ($\approx 32\,000$ species; Froese and Pauly, 2015) that has evolved over the last 530 million years, this application scenario will help demonstrate how habitat models can be improved by using phylogenetic patterns of habitat preference.

Methods

Modelling approach

The modelling approach used in the present study is computationally similar to that described in Guénard et al. (2014) for toxicity modelling; it is applied here to the context of habitat modelling. Instead of predicting the tolerance of multiple species to a set of toxic substances using their chemical descriptors, the model predicts the distribution of multiple species in a range of sites using environmental and spatial descriptors. At the core of that approach is a bilinear regression model, which is a multivariate model (i.e. a model with multiple response variables arranged in a matrix) using two tables of descriptors (Gabriel, 1998; Fig. 1). In the present study, the first table (\mathbf{X}) contains descriptors of the species (or other target units, for instance combinations of species and size classes), which correspond to the rows of the response matrix, and will hereafter be referred to as the row (or species) descriptors. The second table (\mathbf{Z}) contains descriptors of the variation among the sampling sites (e.g. rivers, forest, lakes, microcosms), which correspond to the columns of the response matrix, and will hereafter be referred to as the column (or habitat) descriptors. A bilinear model is represented as follows

$$\mathbf{Y} = \mathbf{XBZ}^T + \mathbf{E} \quad (1)$$

where $\mathbf{Y} = [y_{i,j}]$ is an $n \times m$ matrix (n : number of species, m : number of sampling sites) response variables whose elements $y_{i,j}$ are the response of species i to the conditions found in site j ; $\mathbf{X} = [x_{i,k}]$ is an $n \times p$ design matrix with a constant (all ones) column vector followed by $p - 1$ species descriptors; $\mathbf{Z} = [z_{j,l}]$ is an $m \times q$ design matrix with constant column

vector followed by the $q - 1$ site descriptors; $\mathbf{B} = [b_{k,l}]$ is a $p \times q$ matrix of bilinear regression coefficients; $\mathbf{E} = [\varepsilon_{i,j}]$ is a matrix of residuals; and T denotes matrix transposition. Since the first columns of both \mathbf{X} and \mathbf{Z} are constant vectors, the leftmost column vector of \mathbf{B} ($b_{k,1}$) contains the marginal (or main) effects of the species descriptors, the uppermost row vector of \mathbf{B} ($b_{1,l}$) contains the marginal effects of the site descriptors, while element $b_{1,1}$, which is the common element of the latter two vectors, is the intercept of the model. All elements of \mathbf{B} other than those of the first row and column are interaction terms between species and site descriptors. It is noteworthy that a multivariate regression is obtained by taking \mathbf{Z} as being an identity matrix. Such practice amounts to using a design matrix without a constant vector and including as many column descriptors as the number of columns in the response matrix \mathbf{Y} ($q = m$). When $q < m$, as is the case for most implementations of bilinear models, the latter is more parsimonious than a multivariate regression model.

Bilinear models can be easily estimated using readily-available software packages by first transforming it into ordinary (i.e. single-response) equation systems. That transformation is achieved using vectorisation (i.e. the operator $\langle \rangle$, which takes a matrix and transforms it into a single column vector by stacking its columns) and the Kronecker product (\otimes) as follows:

$$\langle \mathbf{Y} \rangle = (\mathbf{Z} \otimes \mathbf{X}) \langle \mathbf{B} \rangle + \langle \mathbf{E} \rangle. \quad (2)$$

The resulting equation system can be solved using linear models in the normally-distributed case, using Generalised Linear Models GLM in some non-normal cases, or even using Generalised Additive Models (GAM; Hastie and Tibshirani, 1990), Artificial Neural Network (ANN; Lek and Guégan, 1999) in the non-linear cases.

Model estimation

We estimated the bilinear regression model using elastic net regularisation (Zou and Hastie, 2005), to ensure that the resulting model was general, rendered dependable predictions, and avoided over-fitting. The method we implemented consisted in applying different values of penalty to different groups of row and column descriptors and their interactions using a factorial design matrix approach described in details in the appendix (see *Details on model estimation* in the appendix). It is implemented by estimating a global elastic net shrinkage between 0 and $+\infty$ for all the coefficients and factors between 0 and 1 for the different groups of descriptors.

We proposed that regularisation approach because the number of model coefficients of a bilinear model can be very large, especially when phylogenetic and spatial eigenfunctions are involved. Most regularisation methods enable model to be estimated with large numbers of coefficient, possibly outnumbering sample size. Analytically, models involving regularisation are also simpler to optimise than models involving variable selection (e.g. forward stepwise) as the former resort on a continuous process driven by continuous variables (i.e. the penalty parameters). Elastic net (and lasso) regularisation also enables to effectively withdraw variables from a model (i.e. by estimating its coefficient to be numerically 0) while conserving their influences on other collinear variables that may themselves not be effectively

deselected from the model. For instance, the marginal effect of a variable may be estimated to be 0 whereas its effect on the interaction terms in which it is involved will remain. Besides our choice of the elastic net regression, any other dependable method of model estimation can be used.

The bilinear habitat model was estimated using a Poisson GLM. We therefore proposed a generalised, likelihood-based, coefficient to assess the model predictive power. It was calculated as follows:

$$R_{like}^2 = 1 - \frac{\log L_{perfect} - \log L_{model}}{\log L_{perfect} - \log L_{null}} \quad (3)$$

where $L_{perfect}$ is the likelihood of a saturated model, i.e. one predicting the observations exactly, L_{null} is the likelihood of the null model, i.e. the one involving only the mean of the observed values, and L_{model} is the likelihood of the model being assessed. For an ordinary least-squares regression model (i.e. a Gaussian generalised linear model: GLM), the log-likelihood corresponds to the sum of squares and therefore $\log L_{perfect}$ always takes the value 0 (no residuals). In that particular case, the ratio $\log L_{model} / \log L_{null}$ is equal to the ration of the sum of squares around the model and around the mean, so that R_{like}^2 becomes numerically identical to a coefficient of prediction. For other families of GLM, $\log L_{perfect}$ will often take values other than 0, yet R_{like}^2 will nevertheless give an assessment of the predictive power that attributes the value 0 to models that are no better than the null model (i.e. which is the mean of the observed values) and the value 1 to models predicting the response perfectly, while producing negative values for very poor models (Fig. 2). The parameters of the penalty model (i.e. c_α , c_λ , and c_ξ) were estimated as those maximising the cross-validation R_{like}^2 (i.e. with predicted values compared to target values from observations that were not used to estimate the model) using gradient descent. We used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (see Nocedal and Wright, 1999, for a description) for that purpose. When L_{model} is calculated using fitted rather than predicted values, R_{like}^2 is the ordinary R^2 and it will always grow with the number of observations (i.e. $L_{model} > L_{null}$ for fitted values). To use R_{like}^2 as a determination rather than a prediction coefficient, we propose to adjust it using the approach proposed by Ezekiel (1930) as follows:

$$R_{like}^{2*} = 1 - \frac{(n m - 1)}{(n m - \nu)} (1 - R_{like}^2) \quad (4)$$

where ν is the number of degrees of freedom of the model (i.e. the number of non-zero regression coefficients, including the intercept). That formula, corresponds to that of the adjusted R^2 for the equivalent multiple regression model outlined in Eq. 2.

Types of predictions

Bilinear models can provide three types of predictions; we propose the following nomenclature to refer to them. We refer to the case where one makes predictions for unknown rows but known columns as “row predictions”. It would happen, for instance, when predicting the

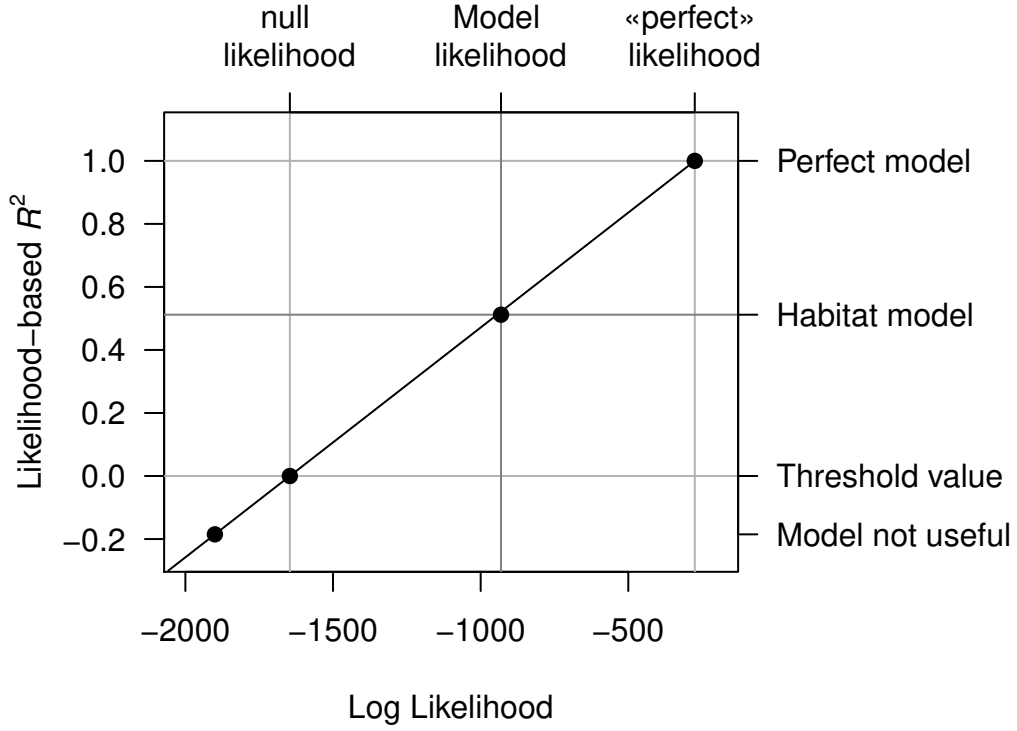


Figure 2: Metric of model performance used in the present study, which is 1 minus the ratio of the logarithms of two likelihood ratios. The first likelihood ratio is the likelihood of a model predicting the response perfectly compared to that of the model being assessed; it is divided by a second likelihood, which is the likelihood of a perfect model compared to that of the null model (i.e. with a single constant term). The value of the metric is smaller than or equal to 1 and will take a negative value when the cross-validated model likelihood is below the null likelihood.

density of some species that are not in the training data set for some rivers in the training data set. Such predictions would be useful when one is interested in predicting the density of an incoming invasive species in the eventuality that it would reach the rivers under study. On the other hand, we refer to the case where one makes predictions for unknown columns but known rows as “column predictions”. It would happen, for instance, when predicting the density of some species in the training data set for some rivers that are not in the training data set. For instance, one may be interested in predicting the densities of a set of widely-distributed reference species in potentially impacted rivers from densities observed in a set of reference rivers. Finally, we refer to the case where one makes predictions for unknown rows and columns as “full predictions”. In the present study, it would happen when predicting the density of some species that are not in the training data set for some rivers that are not in the training data set. While full predictions are potentially the most useful, they are also the most demanding for the bilinear model because they represent entirely new information. In the application that follows, full predictions were used when estimating the parameters of the regularisation model.

Application scenario

As a demonstration of the practical utility of the approach, we used phylogenetic habitat modelling to assess the effect of flow regulation by hydroelectricity production companies on fish species in rivers. That data set, had previously been used by Guénard et al. (2016), who analysed the effect of water regulation on total fish density and species richness. Here it has been analysed in greater details to describe fish community structure in river sampled over a broad geographic range and make specific predictions for regulated sites. For that demonstration, we built a spatially-explicit phylogenetic habitat model using data from the 15 unregulated rivers in the Hydronet river data set (see appendix), including the estimation of the regularisation model parameters by cross-validation. We used the resulting model to exemplify potential applications of phylogenetic habitat models, by estimating reference densities for fishes of different sizes in the 13 regulated rivers of the data set, in order to detect alterations of fish density and community structure that were potentially triggered by flow regulation.

Three flow regulation strategies used by hydroelectricity production companies were studied: run of the river (abbreviated RR; five rivers), storage (ST; five rivers), and flow spiking (FS, also known as «hydropeaking»; three rivers). A typical RR facility consists of a small reservoir (i.e. water storage sufficient for a few days of production) from which water flows constantly through either turbines or spillways, generally producing little effects (if any) on the downstream flow (Bratrich et al., 2004). A typical ST facility has a large reservoir (i.e. several months of water storage at mean flow), which induces a temporal shift of natural runoff, an attenuation of seasonal high flows, and an enhancement of low flows. A typical FS facility has a large reservoir like that of ST facilities, but implies frequent (sometimes once or twice daily), rapid (within minutes) and important (many folds) fluctuations of downstream flow caused by the operation of dam release structures, which are designed to produce power only at times of day when selling electricity is the most profitable (Cushman, 1985; Flodmark

et al., 2004).

Model data encompassed 143 response variables, each of which representing the density of a particular combination of fish species and size class sampled in the 28 rivers (15 unregulated +13 regulated; see section *Data collected*, in the appendix for further details about the Hydronet river data). The combinations of species and size classes involved a total of 48 species (Fig. 3) and each species was represented by one to nine size classes. The data set also included six environmental variables: water depth, z (cm); water velocity, v (cm s^{-1}); substrate median grain size, D_{50} (cm); proportion of macrophyte cover, MC (%); number of heating degree-days, DD ($^{\circ}\text{C d}$); and total phosphorus, TP ($\mu\text{g L}^{-1}$).

Two types of row descriptors were used (see *Appendix – Data processing* for further details). The first type was represented by a single trait, namely the median total length of the size class (TL), whereas the second type involved 47 phylogenetic eigenfunctions (number of species–1, referred to as PE1 through PE47). Also, the model involved 14 spatial eigenfunctions (number of rivers–1, referred to as SE1 through SE14) as a second type of column descriptors in addition to the environmental descriptors mentioned earlier. On top of these marginal effects, the bilinear model had 282 terms representing interactions between the phylogeny (47 eigenfunctions) and the environmental descriptors, six terms representing interactions between fish size and the environmental descriptors, 658 terms representing interactions between phylogeny and space, and 14 terms representing interactions between fish size and space. For the present scenario, interactions between phylogeny and the environment allowed the model to assess how evolution shaped the physical habitat requirements of the species whereas interactions between fish size and environment allowed the model to represent the influence of body size on habitat selection. On the other hand, interactions between space and phylogeny allow the model to represent potential phylogeographic patterns whereby fish evolved as they spread across the landscape after the last glacial age, whereas interactions between space and fish size allowed the model to represent potential geographic patterns in variation of each trait (e.g. areas in which fish are bigger, on average, than in other areas).

The sample size and the number of descriptors in the model were large and, in the context of that particular data set, it was impractical to perform leave-one-out cross-validation to calculate predictive power and estimate the parameters of the regularisation model. Instead, we defined 12 cross-validation groups that were combinations of three groups of river data (each containing five unregulated rivers) and four groups of species data (each with 12 species). We obtained the river data groups by picking one river out of three in an order going from the westernmost to the easternmost rivers whereas the species data groups were obtained by picking one species out of four in the order that they appeared in the reference phylogenetic tree. We used that systematic data selection approach because it helped obtain more repeatable estimates of model predictive power compared to using, for instance, randomised sub-sampling. Given the relatively small number of rivers, this approach helped ensure that the sites used to make predictions had neighbouring sites representing their area and that the species used for prediction had relatives to represent their kind in the model. Also, a systematic cross-validation approach helps estimate regularisation parameters (Eq. 2

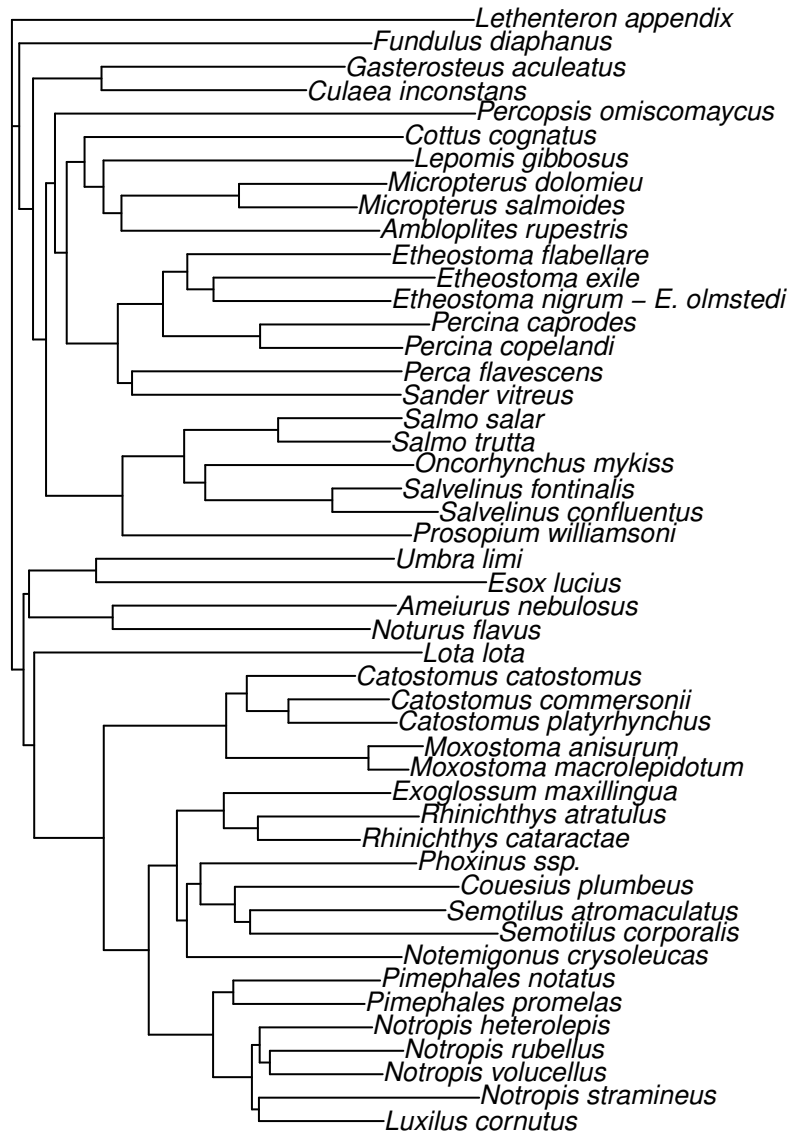


Figure 3: Phylogeny of the fish species observed in the 28 rivers (obtained from Hubert et al., 2008) after removing the species that were not observed during the surveys.

Table 1: Regularisation parameters used for the phylogenetic habitat model and estimated by cross-validation (α, λ , and $c_{k,l}$), their associated effective penalty ($\lambda \xi_{k,l}$) and the number of terms that were not discarded during elastic net regression (see *Details about Model estimation*, in the appendix, for details).

Type	$c_{k,l}$	$\lambda \xi_{k,l}$	Model terms		
			Total	Selected	% selected
<i>Intercept</i>	0	$1.722 \text{ e} - 1$	1	1	100%
<i>Trait</i>	-5.676	$1.176 \text{ e} - 3$	1	0	0%
<i>Phylogeny</i>	2.575	$3.200 \text{ e} - 1$	47	1	2.1%
<i>Environment</i>	-3.299	$1.226 \text{ e} - 2$	6	2	33%
<i>Space</i>	1.727	$2.924 \text{ e} - 1$	14	1	7.1%
<i>trait</i> \times <i>Env.</i>	2.885	$7.785 \text{ e} - 4$	6	5	83%
<i>Trait</i> \times <i>Space</i>	-0.265	$5.020 \text{ e} - 3$	14	10	71%
<i>Phylo.</i> \times <i>Env.</i>	-0.547	$7.548 \text{ e} - 2$	282	22	7.8%
<i>Phylo.</i> \times <i>Space</i>	2.718	$3.441 \text{ e} - 1$	658	1	0.2%
Model	$\alpha = 0.511$	$\lambda = 0.344$	1029	43	4.2%

in the appendix) by gradient descent as randomised sub-sampling would inevitably introduce noise in the objective function (R_{like}^2) and hamper convergence.

Calculations

All calculations other than sequence alignment and tree estimation were performed using the R language for statistical computing (R Development Core Team, 2015). Package *codep* (Guénard et al., 2010) was used to calculate spatial eigenfunctions, package *glmnet* (Friedman et al., 2010) was used to calculate elastic net regressions, and package *MPSEM* (Guénard et al., 2013) was used to calculate phylogenetic eigenfunctions.

Results

Phylogenetic habitat model (unregulated rivers)

Fish density ranged from 0 (1569 observations out of 2145, or 73.1%) to $23 \text{ fish} \cdot (100 \text{ m}^2)^{-1}$ for the unregulated rivers ($\log L_{perfect} = -274.81$, $\log L_{null} = -1646.18$; Fig. 2; $L_{perfect}$ and L_{null} are defined in Eq. 3; Fig. 4). The estimated values of c_α and c_λ were 0.0435 and -1.066, respectively, and the full prediction R_{like}^2 of the cross-validated model was 0.283 during parameter estimation (see Table 1, for the other regularisation parameters). The R_{like}^{2*} (Eq. 4), which is calculated using the fitted values of the model, was 0.512, and the fish densities found using the model ranged from $1.0 \cdot 10^{-5}$ to $6.4 \text{ fish} (100 \text{ m}^2)^{-1}$ (Fig. 4c).

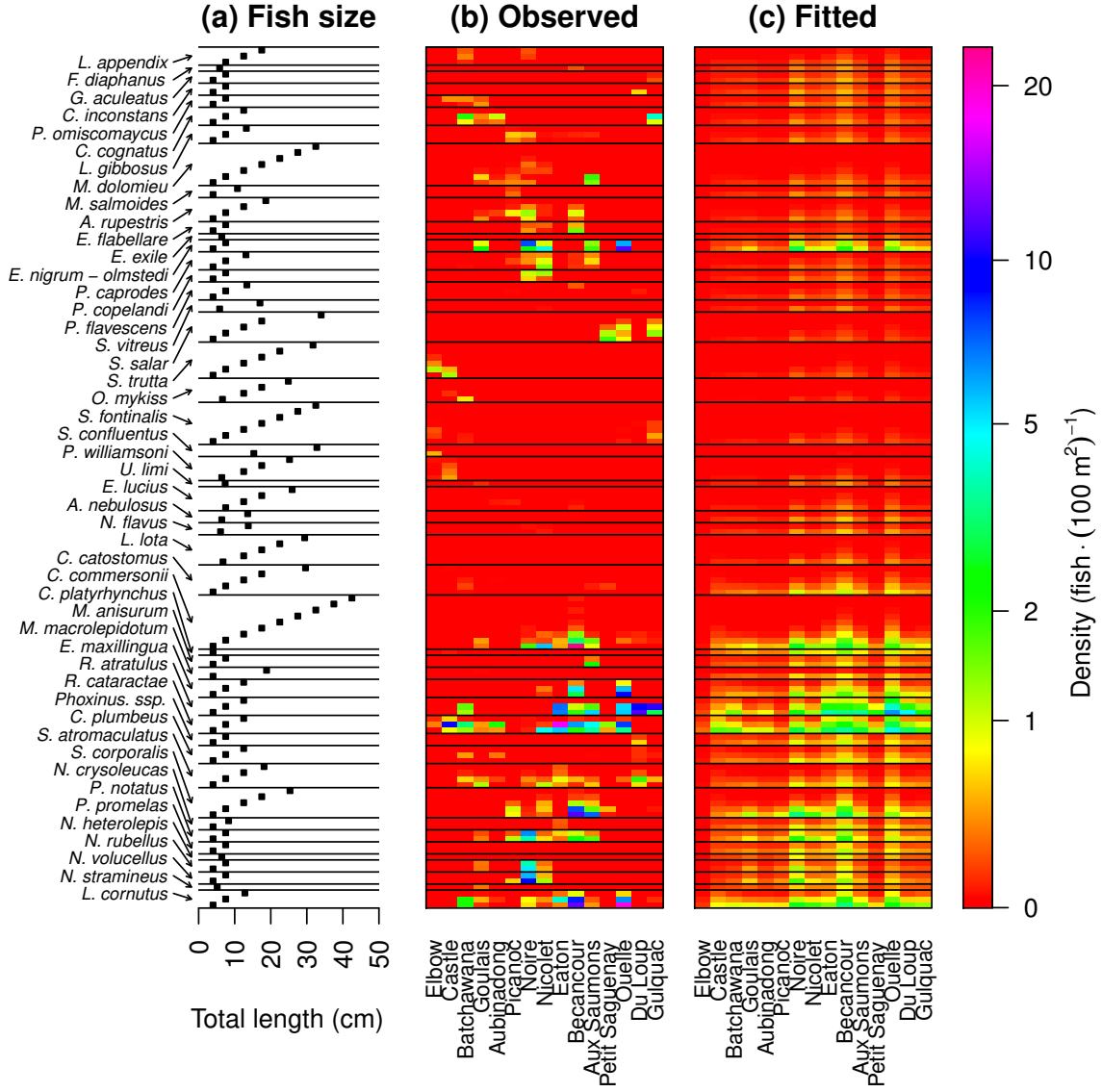


Figure 4: Different combinations of species and size classes found in 15 unregulated rivers, with (a) the median size of each class, (b) the observed fish densities, and (c) the densities fitted by the phylogenetic habitat model.

Predictions for regulated rivers

Observed fish densities ranged from 0 (1419 observations out of 1859, or 76.3%) to $18.5 \text{ fish} \cdot (100 \text{ m}^2)^{-1}$ for the regulated rivers and the densities predicted for the regulated rivers using information on unregulated rivers ranged from 0 to $15.8 \text{ fish} \cdot (100 \text{ m}^2)^{-1}$. All species observed in the unregulated rivers were also present in the regulated river and only site predictions were thus performed in the present application example. The difference between observed and predicted fish density ranged from -15.8 to $14.5 \text{ fish} \cdot (100 \text{ m}^2)^{-1}$ (Fig. 5; paired t -tests, see Appendix – Assessing the effect of flow regulation for details). We found statistically significant effects of flow regulation for 17 species out of 48 using paired t tests obtained by pooling the different size classes (Table 22). Flow regulation was found to have a negative effect on the density of most fish species for which an effect was detectable (RR: 16/17 species, ST: 11/17 species, FS: 4/17 species). A notable exception to that general observation was a positive effect of ST and FS dams on Brook trout (*S. fontinalis*) densities.

It is noteworthy that environmental variables such as flow velocity and water depth may be associated, to some degree, with flow regulation. Such an outcome is suitable when assessing the effect of flow regulation in order to isolate, as much as possible, that latter effect among that of other possible factors contributing to the observed difference in fish density.

Discussion

In the present study, we described a modelling approach whereby phylogeny, in the form of phylogenetic eigenfunctions, and space, in the form of spatial eigenfunctions, can be used in Numerical Habitat Modelling (NHM) to obtain models describing the distribution of multiple species (and size classes) in the landscape. We have also shown a regularisation method to address the issue of the numerous parameters that such a method involves; most of which described the interactions of row descriptors with column descriptors of the sites-by-species response table. Other regularisation methods exist besides elastic net regression (e.g. AIC-based variable selection; basis pursuit denoising; Dantzig selector, Candès and Tao, 2007) and it may be worthwhile to investigate their potential usefulness for phylogenetic NHM. The implementation of that approach was exemplified by modelling the distribution of a broad array (48 species) of freshwater fish species in a set of rivers covering a broad ($> 3400 \text{ km}$) geographic range across Canada. Different studies have proposed methods to use phylogeny to reach various goals such as predicting species traits (Bruggeman et al., 2009; Swenson, 2014), testing hypotheses about community structure (Ives and Helmus, 2011). However, this is the first time, to our knowledge, a phylogeny is used explicitly for NHM. We preferred to use phylogenetic eigenvectors instead of other approaches (e.g. Martins and Hansen, 1997; Garland and Ives, 2000) on the basis of its relative simplicity and ability to assess interaction between the species and site descriptors.

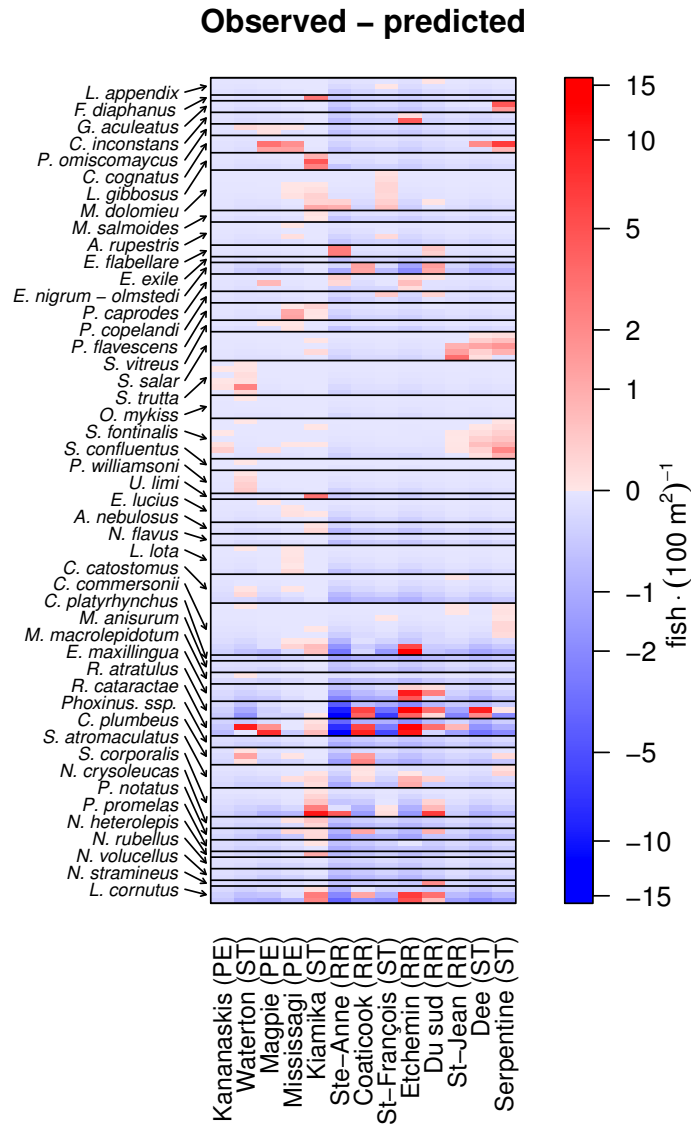


Figure 5: Difference between fish densities observed in the 13 regulated rivers for different species and size classes (see Fig. 4a for their corresponding median total lengths) and the baseline values predicted by the model built with information from the unregulated rivers. The differences are represented by colour codes in each river: red rectangles represent densities that are above the baseline, blue rectangles are densities that are below the baseline, and the intensity of the colour represents the absolute value of the difference. The dams are categorised with respect to three different types of flow regulation: run of the river (RR), storage (ST), and flow spiking (FS; also known as «hydropeaking»).

Table 2: Hypothesis tests of the differences between the fish densities observed and predicted in regulated rivers for the different types of flow regulation (RR: run of the river, ST: storage, and FS: flow spiking; also known as «hydropeaking»), for the species with significant differences between observed and predicted densities. Square parentheses: 95% confidence intervals.

Species	F-test	RR	ST	FS
Notropis rubellus	$F_{3,22} = 45.2^{(4)}$	-0.11 [-0.15, -0.069]	-0.060 [-0.10, -0.020]	-0.054 [-0.11, -0.0033]
Pimephales promelas	$F_{3,22} = 34.7^{(4)}$	-0.16 [-0.237, -0.0837]	-0.095 [-0.17, -0.025]	-0.061 [-0.16, 0.026]
Notemigonus crysoleucas	$F_{3,22} = 19.2^{(4)}$	-0.16 [-0.263, -0.0641]	-0.047 [-0.14, 0.041]	-0.037 [-0.16, 0.076]
Phoxinus ssp.	$F_{3,22} = 53.5^{(4)}$	-0.17 [-0.241, -0.0933]	-0.076 [-0.15, -0.0097]	-0.041 [-0.13, 0.040]
Moxostoma macrolepidotum	$F_{3,22} = 30.2^{(4)}$	-0.14 [-0.191, -0.0859]	-0.091 [-0.14, -0.041]	-0.075 [-0.14, -0.014]
Moxostoma anisurum	$F_{3,22} = 56.0^{(4)}$	-0.17 [-0.235, -0.112]	-0.10 [-0.16, -0.045]	-0.074 [-0.15, -0.0063]
Catostomus platyrhynchus	$F_{3,10} = 25.6^{(2)}$	-0.30 [-0.501, -0.119]	-0.16 [-0.34, -0.0017]	-0.11 [-0.34, 0.087]
Catostomus catostomus	$F_{3,61} = 35.4^{(4)}$	-0.14 [-0.187, -0.0968]	-0.080 [-0.12, -0.038]	-0.068 [-0.12, -0.016]
Lota lota	$F_{3,61} = 22.5^{(4)}$	-0.062 [-0.0851, -0.0398]	-0.038 [-0.060, -0.016]	-0.017 [-0.045, 0.0091]
Noturus flavus	$F_{3,22} = 27.2^{(4)}$	-0.094 [-0.138, -0.052]	-0.051 [-0.093, -0.010]	-0.029 [-0.082, 0.022]
Ameiurus nebulosus	$F_{3,22} = 17.0^{(3)}$	-0.12 [-0.179, -0.0578]	-0.052 [-0.11, 0.0036]	-0.039 [-0.11, 0.031]
Salvelinus fontinalis	$F_{3,87} = 6.14^{(1)}$	-0.013 [-0.0582, 0.0310]	0.056 [0.011, 0.10]	0.011 [-0.044, 0.067]
Sander vitreus	$F_{3,22} = 16.8^{(3)}$	-0.080 [-0.119, -0.0414]	-0.046 [-0.084, -0.0090]	-0.023 [-0.071, 0.023]
Ambloplites rupestris	$F_{3,48} = 25.6^{(4)}$	-0.064 [-0.0888, -0.0403]	-0.030 [-0.054, -0.0070]	-0.021 [-0.051, 0.0079]
Micropterus salmoides	$F_{3,22} = 25.5^{(4)}$	-0.065 [-0.102, -0.0298]	-0.027 [-0.062, 0.0072]	-0.014 [-0.058, 0.030]
Percopsis omiscomaycus	$F_{3,22} = 20.6^{(4)}$	-0.094 [-0.147, -0.0438]	-0.032 [-0.081, 0.016]	0.0044 [-0.058, 0.067]
Lethenteron appendix	$F_{3,35} = 26.8^{(4)}$	-0.048 [-0.0686, -0.0271]	-0.020 [-0.040, 2.20 e - 05]	-0.013 [-0.039, 0.013]

⁽¹⁾ $0.05 \leq P \leq 0.01$, ⁽²⁾ $0.01 < P \leq 0.001$, ⁽³⁾ $0.0001 < P \leq 0.0001$, ⁽⁴⁾ $P < 0.0001$

Fish habitat model

In addition to predicting the local densities of multiple fish species, we also showed how the predictions obtained from a phylogenetic model can be used to assess human impacts on ecosystems. Models of that type are in high demand now that development projects are routinely assessed through impact studies. We calibrated a phylogenetic NHM model using fish abundances in a set of 15 unregulated rivers to predict abundance in a separate set of 13 regulated rivers. With that approach, we showed that flow regulation had a deleterious effect on most fish species when analysed individually. The fish density model allowed us to highlight the many different ways in which fish and sites descriptors act and interact to describe fish habitat. More precisely, we were able to detect a general density deficit associated with flow regulation for many species. We regard that result as interesting since it has long been known that many ecological processes within rivers depend to a large extent on patterns of flow fluctuation and altering the flow regime is thus expected to affect organisms living in rivers (Bonner and Wilde, 2000; Nilsson et al., 2005). Here, we will use these results (Table 11) as an opportunity to illustrate the interpretation of model coefficients or groups of coefficients, as well as some requirements and assumptions of phylogenetic habitat modelling.

Interpreting descriptor marginal effects

Bilinear models, like the one used for phylogenetic habitat modelling, have an intercept, which is the fitted value of the response variables when all row and all column descriptors have a value of 0. The intersection between a constant term in one of the descriptor matrices (either that of the rows or the columns) and a variable in the other descriptor matrix (either that of the columns or the rows) are modelling the marginal (i.e. main) effects of a given type of descriptor (i.e. the effect of a descriptor that is not conditional on that of another descriptor). In the present scenario, the marginal effect of fish size, which was the only trait used, was discarded by the elastic net regression (i.e., its effect was estimated to be non-existent). Consequently, fish size was not found, in itself, to be a useful variable to estimate fish density. On the other hand, we found a marginal effect of the phylogeny, involving PE1; this effect pointed out that, on average, species of order Cypriniformes (families Cyprinidae and Catostomidae) were ≈ 2.5 times more abundant than the other species, while Salmonidae were only half as abundant as the other species. A likely explanation for that contrast may be that Cypriniformes are generally smaller and have a greater propensity for schooling. For the site descriptors, the marginal effect of the number of heating degree-days DD and that of total phosphorus TP were retained by the elastic net procedure. Both variables have positive effects, indicating that warmer and more TP -rich rivers had greater fish densities. We also found a marginal effect of spatial eigenfunction SE1; fish density increasing (by 7.8%) from the south-easternmost end (Quebec, New Brunswick) to the north-westernmost end (Alberta) of the study area in Canada.

Interpreting interactions between descriptors

In addition to the marginal effects of the descriptors, bilinear models like the one used here for phylogenetic modelling have interaction terms that are the intersections between the variables in the matrices containing the row and column descriptors. In the present phylogenetic habitat model, we found significant interactions between the fish size trait and five environmental descriptors (depth z , water velocity v , substrate median grain size D_{50} , heating degree-days DD , and macrophyte cover MC). All five descriptors interacted negatively with fish size, indicating that greater densities of larger fish tended to be encountered in cold water, shallow and calm rivers having fine substrate sparsely covered with macrophyte. Comparatively, warmer, deeper, more agitated rivers having coarser substrate covered with more macrophyte harboured, on average, greater densities of smaller fish. The interaction term between total phosphorus TP and fish size was discarded from the model: fish of any size seemed to benefit from high TP in the rivers included in the model.

We also found fish size to interact with 10 of the spatial eigenfunctions, which together indicated that the larger fish were more abundant in specific areas, such as the Saint Lawrence lowlands and the western part of the study area (Alberta), whereas river communities in the north-eastern and central parts of the study area (central Quebec and Ontario), were dominated by smaller fish.

Many (22) interactions terms were found between phylogenetic eigenfunctions and five of the six environmental descriptors (v , D_{50} , DD , TP , and MC). These interactions are the consequences of the fact that the numerous features underlying the environmental requirements of the fish species are the outcome of evolution (see section “Details on the interactions between phylogeny and the environment” in the appendix for further details about these interactions).

Contrary to the many phylogeny×environment interactions that we found, a single interaction term was retained between phylogeny and space. That term involved the first spatial eigenfunction and PE35, and predicted higher densities for the Blacknose dace and the Cutlips minnow in the eastern portion of the study area, whereas the Longnose dace was more prominent in the western portion. Environmental differences among the rivers seem to be a more relevant driver of the density structure of river fish assemblages considered in the present study than geographic isolation. The model did not detect the expected phylogeographic interaction underlying differences between the native ranges of the Atlantic salmon, *Salmo salar*, and Brook trout, *Salvelinus fontinalis*, which are restricted to the eastern portion of the study area, and of the Rainbow trout, *Oncorhynchus mykiss*, and Bull trout, *Salvelinus confluentus*, which are restricted to the western portion. Had they been detected, these range differences would have most likely involved a set of phylogenetic eigenfunctions describing contrasts among Salmonidae (e.g. PE8, PE16, PE17, PE24, and PE41) with SE1, which describes a large-scale (> 3000 km) gradient spanning the whole study area. The relatively small number of study sites (15 over an area larger than 1.5 million km²) and the importance of the phylogeny×environment interactions, which were found to be better predictors, may explain that absence of significant effect. Hence, the model

estimation procedure imposed a smaller penalty to the phylogeny×environment interaction terms ($\lambda_{\xi_{phylo,envir.}} = 0.075$) than to phylogeny×space interactions terms ($\lambda_{\xi_{phylo,space}} = 0.34$; Table 11).

When using the model to predict the expected densities for the regulated rivers, we found that the fish density observed in the regulated rivers were, with very few exceptions, lower than those predicted from the unregulated river data (Table 22, Fig. 5). That result concurs partially with that of Guénard et al. (2016), who found a negative impact of flow regulation by FS dams on total fish density. In contrast with that study, however, we found here that RR and ST dams seem to affect even more species than FS facilities. These different conclusions come from the fact that the first study analysed total fish density in multiple 300 m² sampling sites within each river whereas the present study analysed the fish density a species at a time in whole rivers. Hence, the few species that were affected by FS dams were observed in large numbers, while many species affected by RR and ST dams were only sparsely observed and had little impact on total fish density. The analysis performed in the present study can therefore be regarded as more sensitive than that of Guénard et al. (2016) to detect the influence of flow regulation on fish densities, in addition to providing more detailed information about fish community responses to flow management.

Conclusion

We have described a powerful habitat modelling approach and hope that it will be put in due use when studying the impact of environmental changes on communities, which are composed of multiple species with sharing various levels of common ancestry. The framework we described is flexible and can easily be adapted to studies of other groups of organisms, e.g. bacteria, birds, evergreen trees, mammals, shrubs, yeast, weeds, zooplankton, or multiple taxonomic groups, and to any scenario regardless of the spatial scale of the study. Requirements are (1) information about the phylogeny of the organisms and traits that are relevant to the particular application; (2) that one has suitable descriptors of the environment at the study sites; (3) information about spatial variation (i.e. a model of spatial relationships among the study sites, based for example on the distances of the sampling sites to one another); and (4) sufficient sample size. Other types of linear model may also be adapted to other types of response variable by using different families of GLM. We are hoping that the present study will foster greater use of phylogenetic habitat modelling to take up further challenges in applied ecology.

Acknowledgements

We are thankful to the many people at Pierre Legendre’s lab and Daniel Boisclair’s lab (Université de Montréal) for their help and insightful hints during the elaboration of the present study. An early version of the manuscript was greatly improved from the comments

and suggestions of Pr. Cajo Ter Braak (Wageningen UR), as well as that of anonymous subject editor and reviewer. G. Guénard was supported by a Strategic Network Grant to D. Boisclair, P. Legendre, and 11 other scientists (NETGP/370899-2008), and an Individual Discovery grant from the Natural Sciences and Engineering Council of Canada (NSERC) to P. Legendre. This paper is an original contribution of NSERC Hydronet.

References

- Boisclair, D. (2001). Fish habitat modeling: from conceptual framework to functional tools. *Can. J. Fish. Aquat. Sci.*, 58:1–9.
- Bonner, T. H. and Wilde, G. R. (2000). Changes in the canadian river fish assemblage associated with reservoir construction. *J. Freshw. Ecol.*, 15(2):189–198.
- Bratrich, C., Truffer, B., Jorde, K., Markard, J., Meier, W., Peter, A., Scheneider, M., and Wehrli, B. (2004). Green hydropower: a new assessment procedure for river management. *River Res. Applic.*, 20:865–882.
- Bruggeman, J., Heringa, J., and Brandt, B. W. (2009). Phylopars: estimation of missing parameter values using phylogeny. *Nucleic Acid Res.*, 37(suppl. 2):W179–W184.
- Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much llarge than n . *Ann. Statist.*, 35(6):2313–2351.
- Costello, M. J., May, R. M., and Stork, N. E. (2013). Can we name earth’s species before they go extinct? *Science*, 339:413–416.
- Cushman, R. M. (1985). Review of ecological effects of rapidly varying flows downstream from hydroelectric facilities. *N. Am. J. Fish. Manag.*, 5:330–339.
- Desdevis, Y., Legendre, L., Azouzi, L., and Morand, S. (2003). Quantifying phylogenetically structured environmental variation. *Evolution*, 57:2647–2652.
- Diniz-Filho, J. A. F., de Sant’Ana, C. E. R., and Bini, L. M. (1998). An eigenvector method for estimating phylogenetic inertia. *Evolution*, 52(5):1247–1262.
- Diniz-Filho, J. A. F., Villalobos, F., and Bini, L. M. (2015). The best of both worlds: Phylogenetic eigenvector regression and mapping. *Genet. Mol. Biol.*, 38(3):396–400.
- Dirnböck, T. and Dullinger, S. (2004). Habitat distribution models, spatial autocorrelation, functional traits and dispersal capacity of alpine plant species. *J. Veg. Sci.*, 15:77–84.
- Elith, J. and Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Ann. Rev. Ecol. Evol. Syst.*, 40:677–697.
- Ezekiel, M. (1930). *Methods of correlation analysis*. John Wiley and Sons, New-York, USA.
- Felsenstein, J. (1985). Phylogenies and the comparative method. *Am. Nat.*, 125(1):1–15.

- Flodmark, L. E. W., Vøllestad, L. A., and Forseth, T. (2004). Performance of juvenile brown trout exposed to fluctuating water level and temperature. *J. Fish Biol.*, 65:460–470.
- Friedman, J. H., Hastie, T. J., and Tibshirani, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Soft.*, 33(1):1–22.
- Froese, R. and Pauly, D. (2015). Fishbase.
- Gabriel, K. R. (1998). Generalized bilinear regression. *Biometrika*, 85:689–700.
- Garland, T. J. and Ives, A. R. (2000). Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *Am. Nat.*, 155:346–364.
- Guénard, G., Boisclair, D., and Legendre, L. (2015). Phylogenetics to help predict active metabolism. *Ecosphere*, 6(4):Article 62.
- Guénard, G., Lanthier, G., Harvey-Lavoie, S., Macnaughton, C. J., Senay, C., Lapointe, M., Legendre, P., and Boisclair, D. (2016). A spatially-explicit assessment of the fish population response to flow management in a heterogeneous landscape. *Ecosphere*, 7:e01252.
- Guénard, G., Legendre, P., Boisclair, D., and Bilodeau, M. (2010). Multiscale codependence analysis: an integrated approach to analyze relationships across scales. *Ecology*, 91:2952–2964.
- Guénard, G., Legendre, P., and Peres-Neto, P. (2013). Phylogenetic eigenvector maps (PEM): a framework to model and predict species traits. *Meth. Ecol. Evol.*, 4:1120–1131.
- Guénard, G., von der Ohe, P. C., de Zwart, D., Legendre, P., and Lek, S. (2011). Using phylogenetic information to predict species tolerances to toxic chemicals. *Ecol. Appl.*, 21:3178–3190.
- Guénard, G., von der Ohe, P. C., Walker, S. C., Lek, S., and Legendre, P. (2014). Using phylogenetic information and chemical properties to predict species tolerances to pesticides. *Proc. R. Soc. B*, 281(20133239):20133239.
- Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.*, 8:993–1009.
- Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecol. Model.*, 135:147–186.
- Hardy, O. J. and Pavoine, S. (2012). Assessing phylogenetic signal with measurement error: a comparison of Mantel tests, Blomberg et al.’s K, and phylogenetic distograms. *Evolution*, 66:2614–2621.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models.*, volume 43 of *Monographs on statistics and applied probability*. Chapman & Hall/CRC, New-York, NY, USA. ISBN 978-0-412-34390-2.

- Hubert, N., Hanner, R., Holm, E., Mandrak, N. E., Taylor, E., Burridge, M., Watkinson, D., Dumont, P., Curry, A., Bentzen, P., Zhang, J., April, J., and Bernatchez, L. (2008). Identifying canadian freshwater fishes through DNA barcodes. *PLOS ONE*, 3:e2490.
- Ives, A. R. and Helmus, M. R. (2011). Generalized linear mixed models for phylogenetic analyses of community structure. *Ecol. Monogr.*, 81(3):511–525.
- Legendre, P. and Legendre, L. (2012). *Numerical Ecology, 3rd English edition*. Elsevier Science B.V., Amsterdam, The Netherlands.
- Lek, S. and Guégan, J. F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecol. Model.*, 120:65–73.
- Malaj, E., Guénard, G., Schäfer, R. B., and von der Ohe, P. C. (2015). Evolutionary patterns and physicochemical properties explain macroinvertebrate sensitivity to heavy metals. *Ecol. Appl.*, Preprint.
- Martins, E. and Hansen, T. (1997). Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.*, 149:646–667.
- McGill, B. J., Enquist, B. J., Weiher, E., and Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends Ecol. Evol.*, 21(4):178–185.
- Nilsson, C., Reidy, C. A., Dynesius, M., and Revenga, C. (2005). Fragmentation and flow regulation of the world’s large river systems. *Science*, 308:405–408.
- Nocedal, J. and Wright, S. (1999). *Numerical Optimization*. Springer, New-York, NY, USA.
- Ollier, S., Couteron, P., and Chessel, D. (2006). Orthonormal transform to decompose the variance of a life-history trait across a phylogenetic tree. *Biometrics*, 62:417–477.
- Pavoine, S., Ollier, S., Pontier, D., and Chessel, D. (2008). Testing for phylogenetic signal in phenotypic traits: New matrices of phylogenetic proximities. *Theoretical Pop. Biol.*, 73:79–91.
- Pillar, V. D. and Duarte, L. d. S. (2010). A framework for metacommunity analysis of phylogenetic structure. *Ecology Letters*, 13:587–596.
- R Development Core Team, T. (2015). *R: A Language and Environment for Statistical Computing (version 3.2.2)*. R Foundation for Statistical Computing, Vienna, Austria.
- Swenson, N. G. (2014). Phylogenetic imputation of plant functional trait databases. *Ecography*, 37(2):105–110.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, 67:301–320.

Modelling habitat distributions for multiple species using phylogenetics (Appendix)

Guillaume Guénard^{1*}, Gabriel Lanthier^{1†}, Simonne Harvey-Lavoie^{1 ‡},
Camille J. Macnaughton^{1§}, Caroline Senay^{1¶},
Michel Lapointe^{2||}, Pierre Legendre^{1**}, and Daniel Boisclair^{1††}

23rd August 2016

1. Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre-Ville
Montréal, QC, Canada H3C 3J7
2. Department of Geography
McGill University
Burnside Hall, 805 Sherbrooke W, Room 619
Montreal, QC, Canada H3A 2K6

*guillaume.guenard@gmail.com (corresponding author)

†gabriel.lanthier@umontreal.ca

‡simonne.harvey-lavoie@umontreal.ca

§camille.macnaughton@gmail.com

¶caroline.senay@umontreal.ca

||Michel.Lapointe@mcgill.ca

**pierre.legendre@umontreal.ca

††daniel.boisclair@umontreal.ca

Field sampling

During the summer months (late June – early September) of the years 2011 – 2013, the Hydronet research group (hereafter simply referred to as “Hydronet”) surveyed 28 rivers (15 unregulated and 13 regulated) located in four Canadian provinces (New-Brunswick, Quebec, Ontario, and Alberta). Each river was sampled at multiple locations hereafter referred to as the sampling sites. Sampling sites had 300 m² surface area (5 m across by 60 m along the river) and were positioned in alternation (site width was often substantially narrower than river sections), near the left shore (facing downstream), in the middle, and near the right shore of the river, starting randomly, while ensuring that the habitat within them was fairly homogeneous. Locations of the beginning and end of each sampling site were recorded using a GPS unit (GPSMAP® 76sc: Garmin International Inc. 1200 E. 151st Street, Olathe, KS 66062, USA).

Fish sampling consisted in electrofishing (using an LR24® Backpack electrofisher: Smith-Root Inc., 14014 NE. Salmon Creek Ave., Vancouver, WA 98686, USA; electroshocking duration per site: 900 s; mean power: 200 W) and snorkelling surveys performed during daytime hours (between 0830 and 1800). These two sampling methods were used in tandem to minimise the bias associated with because each of these methods shows selectivity toward catching fish of particular species and size classes (Macnaughton et al., 2014). Sampling with both methods involved zigzagging in the upstream direction to cover the whole sampling area. Fish were identified to species or else to the nearest taxonomic level that electrofishers or snorkelers could discriminate visually. Electrofishing was performed by teams of three fishers: one operating the electrofisher and two catching the electroshocked fish with dip nets 38 mm long \times 33 mm wide \times 20 mm deep, mesh size: 6.35 mm (Smith-Root Inc.). Fish were identified, measured (Total length, $TL \pm 0.1$ cm), allowed to recover in cool aerated water, and released at the location where they had been captured. Visual observations were performed by teams composed of two trained snorkelers. Fish with total lengths ≥ 3 cm (fish with total lengths < 3 cm could not be identified reliably by snorkelers) were categorised into size classes, with the first class going from 3 to 5 cm (median size approximately 4 cm), and then in classes 5 cm apart (median sizes: 7.5 cm, 12.5 cm, 17.5 cm, etc.).

Descriptors of the local physical environment that are often regarded as key drivers shaping community structure were estimated in the sampling sites (Knouft et al., 2011; Michel and Knouft, 2014). We estimated water depth (z , in cm, measured using a graduated pole) and velocity (v , in cm s⁻¹; Flo-Mate 2000: Marsh-McBirney Inc., 4539 Metropolitan Court Frederick, MD 21704-9452, USA), substrate median grain size (D_{50} , in cm), and the proportion of macrophyte cover (MC , in %) in ten 50 cm \times 50 cm plots randomly dispatched within each sampling site (Wolman, 1954; Latulippe et al., 2001). Water temperatures was measured in the field using HOBO UA-002-62 logging thermometers (precision: $\pm 0.5^\circ\text{C}$; Onset Computer Corp., 470 MacArthur Blvd., Bourne, MA 02532, USA; 3-10 devices / river). Devices were deployed in riffle, run, or shallow pool habitats and set to record water temperature at 15 min intervals for at least 9 weeks. We used the number of heating degree-days (DD , in $^\circ\text{C d}$) as an environmental descriptor, which we estimated as the sum of mean daily water temperatures above 0°C for a period of nine weeks encompassing the four

weeks preceding the hottest week of the summer, and the four weeks following it (between late-May to late-September). During the low summer flows and on days without rain, we took between one and nine water samples in the main stream of the rivers. These samples were collected in 250 mL acid-washed, high-density polyethylene bottles (Nalgene®, Nalge Nunc International Corporation), kept at 4°C, and shipped to the University of Alberta’s Biogeochemical Laboratory for analysis. Total phosphorus (TP , in $\mu\text{g L}^{-1}$) was determined using persulfate oxidation (Ng, 2010). For each river, individual TP values that differed by 15 standard deviations or less from the river mean were retained; we took the average of the latter as the total phosphorus for the whole river.

Data processing

In the present study, we used fish density (fish $(100\text{ m}^2)^{-1}$) by species and size class as the response variable of the phylogenetic habitat model. To obtain a *species* \times *location* response matrix that was not overly sparse (i.e. did not contain too many zeros) as well as to facilitate computations, modelling was performed on the basis of whole rivers. For each river, values of z , v , D_{50} , and MC were averaged. Mean densities for whole rivers were estimated separately for each combination of species and size class using Generalised Linear Models (GLM: Hastie and Pregibon, 1991; Poisson distributed) fitted with the values of depth, velocity and substrate grain size for each sampling site as descriptors. We took the fitted values of each GLM (one per river for each combination of species and size-class), estimated for the average environmental conditions observed in the river, as the mean fish density. Species and size class combinations that were not observed in a given river were assumed to be absent from that river and assigned a density value of 0 fish $(100\text{ m}^2)^{-1}$.

For any given species, we discarded all size classes for which no fish was observed, and merged the size classes for which fish were present in two rivers or less. Size class merging was performed by coalescing size classes with insufficient numbers of observations with an adjacent size class. When two adjacent size classes were present on either side of the class with low number of observations, the one with the smallest number of observations was chosen. The mean size of the newly formed size class was taken as the mean of their median sizes weighted by their respective numbers of observations. Size class merging stopped when all those remaining had been observed in two rivers or more. The species represented by a single size class that were present in only one river were discarded.

We used the phylogeny published by Hubert et al. (2008), which was estimated using K2P distances (Kimura, 1980) calculated from a standard 652 base pairs «bar code» region of the Cytochrome Oxidase subunit I (COI) mitochondrial gene. That tree was used to calculate Phylogenetic eigenvector maps (PEM; Guénard et al., 2013). PEM required a tree tip for every combination of species and size class. Therefore, every single tree tip representing species with more than one size class was merged into a star tree having branch lengths of 0 and as many tips as the species had size classes. Since the merged star tree had branch lengths of 0, the number of phylogenetic eigenvectors with non-zero eigenvalues

obtained from the approach remain one minus the number of species, with each eigenvector defined for all the species and size classes; they are, however, invariant across observations involving the same species.

We used spatial eigenvector maps (SEM; Borcard and Legendre, 2002; Dray et al., 2006; Diniz-Filho et al., 2013; Legendre and Legendre 2012, Chapter 14) to model patterns of spatial variation among the rivers and account for spatial auto-correlation (Legendre, 1993; Dormann et al., 2007). These eigenfunctions were calculated from the geodesic distances among rivers.

In the present study, the design matrix \mathbf{X} contained the species descriptors, which included the median standard length of the fish in each size class as a species trait in addition to the PEM descriptors that represented the among-species phylogenetic patterns of trait variation. It is noteworthy that, in the model, variation among species and size classes was explained in part by the median fish size and the PEM, because fish of different species had different median sizes, on average, but within-species variation was only associated to variation in the median fish size, because PEM were constant across size classes for a given species. The design matrix \mathbf{Z} of site descriptors included environmental variables and the spatial eigenfunctions. Since two types of row descriptors were used with two types of column descriptors, there were eight types of bilinear model terms. Among them, four were modelling main effects acting alone in the model: 1) the traits, 2) the phylogeny (PEM), 3) the environment, 4) space (spatial eigenvectors). There were also four second-order interaction terms: 1) trait-environment, 2) trait-space, 3) phylogeny-environment, 4) phylogeny-space. Therefore the first two columns of matrix \mathbf{R} represent the different types of row descriptors, namely *trait* (fish total length; column 1) and *phylogeny* (PEM; column 2), the next two columns represent the column descriptors, namely *environment* (z , v , D_{50} , MC , DD , and TP ; column 3) and *space* (SEM; column 4), and the last four columns of \mathbf{R} represent their interactions (column 5: *trait* \wedge *environment*, column 6: *trait* \wedge *space*, column 7: *phylogeny* \wedge *environment*, and column 8: *phylogeny* \wedge *space*). We fitted the model using a Poisson-distributed Generalized Linear Model (GLM).

Details on model estimation

To ensure that the resulting model was general, rendered dependable predictions, and avoided over-fitting, we regularised the bilinear regression model using elastic net regularisation (Zou and Hastie, 2005), which is a combination of the Least Absolute Shrinkage and Selection Operator (LASSO; Tibshirani, 1996; the L_1 regularisation norm) and Tikhonov regularisation (Tikhonov and Arsenin, 1977, the latter being also known as ridge regression; the L_2 regularisation norm). That method involves estimating matrix \mathbf{B} as the solution minimising an objective function f involving size of the (standardised) regression coefficients as follows:

$$f_{\alpha,\lambda,\xi}(y; x, z|\beta) = \sum_{i=1}^n \sum_{j=1}^m \varepsilon_{i,j}^2 + \lambda \sum_{k=1}^p \sum_{l=1}^q \xi_{k,l} \left(\alpha |\beta_{k,l}| + \frac{(1-\alpha)}{2} |\beta_{k,l}|^2 \right) \quad (1)$$

where α is a parameter varying between 0 and 1, which controls the relative importance of the regularisation norms, λ is the average penalty of all the descriptors ($0 \leq \lambda < \infty$), and $\xi_{k,l}$ are penalty factors specific to particular terms of the model, $\beta_{k,l}$ are the standardised regression coefficients, and $\varepsilon_{i,j}$ are the residuals of the regression model. Therefore, any variable with $\lambda\xi_{k,l} > 0$ (the minimum value) experiences an effective shrinkage of its contribution to the model. In that framework, parameter λ controls the average amount of shrinkage for all the regression coefficients whereas parameters $\xi_{k,l}$ allows one to attribute greater shrinkage to certain groups of regression coefficients with respect to others. That feature can be exploited for the purpose of maximising the predictive power. We estimated α , λ , and $\xi_{k,l}$ using embedded models estimated by cross-validation as those maximising the predictive power of the models. Parameters α and λ were estimated as constants $c_\alpha = \text{logit } \alpha$ and $c_\lambda = \log \lambda$, respectively, to ease optimisation (since the domain of c_α and c_λ need not be bounded), whereas parameters $\xi_{k,l}$ were estimated through a logistic model with a factorial design that was defined as follows:

$$\mathbf{R}[c_\xi] = [\text{logit } \xi_{k,l}], \quad (2)$$

where \mathbf{R} is a binary design matrix with $(p \cdot q - 1)$ rows, each of them representing the columns of $\mathbf{Z} \otimes \mathbf{X}$, with the exception of the model intercept (the first column), which is assigned a constant penalty factor of 0.5, and $(r + 1)(s + 1) - 1$ columns, where r is the number of types of row descriptors and s is the number of types of column descriptors, while $[c_\xi]$ is a column vector of the regularisation model coefficients. It is noteworthy that following Eq.2, penalty values are equal for all the variables involved in the same model term. We chose a logistic rather than a linear model to represent the penalty factors because constraining the penalty factors between 0 and 1 avoids interference with the estimation of c_λ during the optimisation process. The columns of matrix \mathbf{R} represent the different types of row descriptors, column descriptors, as well as their interactions. Elements in each column of \mathbf{R} take the value 1 whenever the descriptor is involved in the type of descriptor it represents, or their interactions, and 0 otherwise.

For phylogenetic habitat modelling, there will be at least one type of each of the row and column descriptors (i.e. $r = 1$ and $s = 1$), namely a set of variables representing the phylogeny, and a set of variables representing the environment. In the simplest scenarios, matrix \mathbf{R} will therefore have three columns: one that represents the marginal effect of the phylogeny, one that represents the marginal effect of the environment, and a third representing the interaction between phylogeny and the environment. In such a scenario, rows of \mathbf{R} corresponding to the columns of $\mathbf{Z} \otimes \mathbf{X}$ representing the marginal effect of the phylogeny will take the value 1 in the first column of \mathbf{R} and 0 elsewhere. On the other hand, rows of \mathbf{R} corresponding to the columns of $\mathbf{Z} \otimes \mathbf{X}$ representing the phylogeny-environment interaction will also take the value 1 in the first column of \mathbf{R} , but also in the second, which represents the environment, and in the third column, which represents the phylogeny-environment interaction, thereby allowing the penalties applied to the marginal effects of phylogeny and environment to be non-additive with that applied to their interactions. In practice, it may be useful to use traits alongside the variables describing the phylogeny as row descriptors and variables describing spatial variation alongside environmental variables as column descriptors.

Assessing the effect of flow regulation

We assessed the effect of flow regulation on fish density by subtracting the predicted fish density values by species and size classes, obtained using the unregulated river model, from the fish densities observed at the sampling sites. The density differences thus obtained were negative when the model predicted higher densities than that observed in the field; positive values were obtained for observed densities that were higher than those predicted. Prior to further analysis, we $\log(x + 1)$ -transformed these density differences on the basis of their absolute values while conserving their signs as follows:

$$x_{transformed} = \text{sign}(x_{original}) \log(|x_{original}| + 1), \quad (3)$$

where $x_{original}$ and $x_{transformed}$ are the original and transformed differences in fish density, respectively, and function $\text{sign}(x)$ returns -1 when $x < 0$, 0 when $x = 0$, and 1 when $x > 0$. The necessity for that transformation was consequential to the fact that a Poisson distribution was assumed for modelling. It was used to mitigate over-dispersion and avoid conclusions driven only by a few extreme values.

Data collected

A total of 989 sites were sampled in the 28 rivers but because of malfunctioning equipment during field sampling, information about water velocity was not available for 48 sites located on four rivers (St-Jean: 17 sites, Elbow: 12, Kananaskis: 10, and Petit-Saguenay: 9); these sites were thus discarded, leaving us with 941 usable sites for modelling. Mean river environmental conditions and species densities were calculated using numbers of sites ranging from 16 (in Petit-Saguenay) to 50 (in Au saumon and Bécancour). Water depth was left-skewed and was $\log(x + 1)$ -transformed before further treatment as were water velocity and substrate grain size; both were $\log(x + 1)$ -transformed because values of 0 were present. Water depth (z) ranged 25.8–55.1 cm (mean= 36.82 cm), v ranged 0.07–0.59 m s⁻¹ (mean= 0.296 m s⁻¹), D_{50} ranged from 0.29 to 38.68 cm (7.057 cm), MC ranged from 0 to 21%, DD ranged from 367 to 1540 °C d, and TP ranged from 1.00 to 3.42 µg L⁻¹ among the rivers. A total of 244 combinations of species and size classes, involving 61 species, were observed in at least one of the 28 rivers. After merging the sparsely observed size classes, 143 combinations of species and size classes, involving 48 species, were retained (see Fig. 1 in the main text). On average, each species was represented by 2.6 size classes, with numbers ranging from one (six species) to nine (one species: the white sucker, *Catostomus commersoni*). The characteristic (median) sizes of these classes ranged from 4.0 to 42.5 cm (mean: 9.50 cm). The sample size (np) for the present application scenario was 2145 observations (143 combinations of species and size class sampled in the 15 unregulated rivers).

Details on the interactions between phylogeny and the environment

We found a total of 22 interactions terms between phylogenetic eigenfunctions and five of the six environmental descriptors (v , D_{50} , DD , TP , and MC). Mean current velocity interacted with one phylogenetic eigenfunction (PE36) and described the Common shiner (*Luxilus cornutus*), the Bluntnose (*Pimephales notatus*), and the Fathead (*P. promelas*) minnows as being more abundant in rivers with faster current than are the eastern shiners (i.e. species of genus *Notropis*). Substrate grain size interacted with five phylogenetic eigenfunctions (PE12, PE17, PE27, PE30, and PE36); their overall effect indicated that, in the family Cyprinidae, riffle daces (i.e. the Longnose dace, *Rhinichthys cataractae*, and the Blacknose dace, *R. atralutis*) tended to be more abundant in rivers with coarser substrate than the average, while the Fallfish (*Semotilus corporalis*), the Bluntnose and fathead (genus *Pimephales*) minnows, the eastern shiners, and the Golden shiner (*Notemigonus crysoleucas*) tended to be abundant in rivers with finer substrate. Temperature interacted with three phylogenetic eigenfunctions (PE17, PE19, and PE35), their overall contributions indicating that the White sucker (*Catostomus commersoni*), the redhorses (i.e. the Silver redhorse, *Moxostoma anisurum*, and the Shorthead redhorse, *M. macrolepidotum*), the Lake chub (*Couesius plumbeus*), and the Creek chub (*Semotilus atromaculatus*) were observed in greater densities in cold environments whereas the Longnose dace, the Fallfish, the White sucker, the Brown bullhead (*Ameiurus nebulosus*), and the Stonecat (*Noturus flavus*) were denser in warmer environments. Total phosphorus interacted with a total of 11 phylogenetic eigenfunctions and their combined effects were associated to a wide range of differential species responses to TP , which we enumerate as follows:

1. The complex formed of the Johnny darter and the Tessellated darter (*Etheostoma olmstedi nigrum* - *E. olmstedi*; these species could not be discriminated by phylogenetic analysis) is found in higher densities in high TP environments, while the Fantail darter (*E. flabellare*) is denser in lower TP environments. Two phylogenetic eigenfunctions were involved in that interaction: PE33 and PE42.
2. Among the Percidae, the Walleye (*Sander vitreus*) appears to be favoured in high TP conditions whereas the Yellow perch (*Perca flavescens*) and two roughbelly darters (Genus *Percina*: the Common logperch, *P. crapodes*, and the Channel darter, *P. copelandi*) were denser in lower TP environments. PE23 and PE45 were involved in that interaction.
3. Among the eastern shiners, the Sand shiner (*Notropis stramineus*) is more prevalent in high TP in comparison with the Mimic shiner (*N. volucellus*). PE44 was involved here.
4. In genus *Semotilus*, higher TP is concomitant with higher densities of the Fallfish compared to the Creek chub and the Lake chub. PE17 and PE22 were behind that interaction.

5. The riffle daces and the Cutlips minnow (*Exoglossum maxillingua*; the effect was described by PE29) as well as the Fathead minnow (the effect was, this time, associated with PE30 and PE40) seem to find high *TP* environments more favourable than the other Cyprinids.
6. Among the Catostomidae, the White sucker are found to be denser in high *TP* environments while the Longnose (*Catostomus catostomus*) and Mountain (*C. platyrhynchus*) suckers as well as the redhorses appear to be denser in rivers with lower *TP*. Here, PE19 and PE45 were involved.

Macrophyte cover interacted with three phylogenetic eigenfunctions; their combined effects described the riffle daces, the Cutlips minnow, the Golden shiner, and the Fathead minnow as being found in higher densities in rivers with low *MC*, compared to the eastern shiners, the Common shiner (*Luxilus cornutus*), the Fallfish and the Bluntnose minnow, which had higher densities in rivers with higher *MC*. These examples illustrate the capacity of the method described in this paper to represent, in a single model, many details about the habitat requirements of many species, using knowledge of their common evolutionary history, and how these factors drive community structure.

Analysis data and script

The raw data and analysis script used for the exemplary scenario are available online at <http://dx.doi.org/10.5061/dryad.60n52>.

Tables

Table 1: Bilinear model coefficients estimated by the elastic net procedure.

	<i>Int.</i>	<i>TL</i>	<i>PE</i> ₁	<i>PE</i> ₁₂	<i>PE</i> ₁₇	<i>PE</i> ₁₉	<i>PE</i> ₂₂	<i>PE</i> ₂₃	<i>PE</i> ₂₇
<i>Int.</i>	-3.58	—	-8.79	—	—	—	—	—	—
$\log z$	—	-0.0193	—	—	—	—	—	—	—
$\log(v + 1)$	—	-0.0837	—	—	—	—	—	—	—
$\log(D_{50} + 1)$	—	-0.0158	—	-1.11	0.107	—	—	—	0.72
<i>DD</i>	0.00131	-4.4e-05	—	—	0.000964	0.000785	—	—	—
$\log TP$	1.79	—	—	—	0.725	0.95	0.441	0.987	—
<i>MC</i>	—	-0.0242	—	0.395	—	—	—	—	—
<i>SE</i> ₁	0.313	-0.0316	—	—	—	—	—	—	—
<i>SE</i> ₂	—	0.103	—	—	—	—	—	—	—
<i>SE</i> ₃	—	-0.0777	—	—	—	—	—	—	—
<i>SE</i> ₄	—	0.0238	—	—	—	—	—	—	—
<i>SE</i> ₇	—	-0.00938	—	—	—	—	—	—	—
<i>SE</i> ₈	—	0.102	—	—	—	—	—	—	—
<i>SE</i> ₁₁	—	-0.00975	—	—	—	—	—	—	—
<i>SE</i> ₁₂	—	0.0272	—	—	—	—	—	—	—
<i>SE</i> ₁₃	—	0.0679	—	—	—	—	—	—	—
<i>SE</i> ₁₄	—	-0.0182	—	—	—	—	—	—	—

	<i>PE</i> ₂₉	<i>PE</i> ₃₀	<i>PE</i> ₃₃	<i>PE</i> ₃₅	<i>PE</i> ₃₆	<i>PE</i> ₄₀	<i>PE</i> ₄₂	<i>PE</i> ₄₄	<i>PE</i> ₄₅
<i>Int.</i>	—	—	—	—	—	—	—	—	—
$\log z$	—	—	—	—	—	—	—	—	—
$\log(v + 1)$	—	—	—	—	4.8	—	—	—	—
$\log(D_{50} + 1)$	—	0.127	—	—	0.318	—	—	—	—
<i>DD</i>	—	—	—	-0.000554	—	—	—	—	—
$\log TP$	-0.224	0.0298	-2.5	—	—	-0.295	0.256	0.624	-0.165
<i>MC</i>	—	—	—	—	—	-0.105	—	—	—
<i>SE</i> ₁	—	—	—	1.91	—	—	—	—	—
<i>SE</i> ₂	—	—	—	—	—	—	—	—	—
<i>SE</i> ₃	—	—	—	—	—	—	—	—	—
<i>SE</i> ₄	—	—	—	—	—	—	—	—	—
<i>SE</i> ₇	—	—	—	—	—	—	—	—	—
<i>SE</i> ₈	—	—	—	—	—	—	—	—	—
<i>SE</i> ₁₁	—	—	—	—	—	—	—	—	—
<i>SE</i> ₁₂	—	—	—	—	—	—	—	—	—
<i>SE</i> ₁₃	—	—	—	—	—	—	—	—	—
<i>SE</i> ₁₄	—	—	—	—	—	—	—	—	—

Int.: intercept, *TL*: total length (cm), *PE*_{*x*}: *x*th phylogenetic eigenvector, *z*: depth (cm), *v*: velocity (cm s⁻¹), *D*₅₀: median grain size (cm), *DD*: cumulative number of degree-day (°C d), *TP*: total phosphorus (μg L⁻¹), *MC*: percent macrophyte cover (%), *SE*_{*x*}: *x*th spatial eigenvector, —: placeholder for coefficients estimated to numerically 0; all logarithms are base *e*

References

- Borcard, D. and Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Model.*, 153:51–68.
- Diniz-Filho, J. A. F., Diniz, J. V. B. P. L., Rangel, T. F., Soares, T. F., de Campos Telles, M. P., Garcia Collevatti, R., and Bini, L. M. (2013). A new eigenfunction spatial analysis describing population genetic structure. *Genetica*, 141:479–489.
- Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., and Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30:609–628.
- Dray, S., Legendre, P., and Peres-Neto, P. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbor matrices (pcnm). *Ecol. Modelling*, 196:483–493.
- Guénard, G., Legendre, P., and Peres-Neto, P. (2013). Phylogenetic eigenvector maps (PEM): a framework to model and predict species traits. *Meth. Ecol. Evol.*, 4:1120–1131.
- Hastie, T. J. and Pregibon, D. (1991). *Generalized linear models*, volume Statistical models in S, chapter 6, pages 195–247. Wadsworth, Pacific Grove, CA.
- Hubert, N., Hanner, R., Holm, E., Mandrak, N. E., Taylor, E., BurrIDGE, M., Watkinson, D., Dumont, P., Curry, A., Bentzen, P., Zhang, J., April, J., and Bernatchez, L. (2008). Identifying canadian freshwater fishes through DNA barcodes. *PLOS ONE*, 3:e2490.
- Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120.
- Knouft, J. H., Caruso, N. M., Dupre, P. J., Anderson, K. R., Trumbo, D. R., and Puccinelli, J. (2011). Using fine-scale gis data to assess the relationship between intra-annual environmental niche variability and population density in a local stream fish assemblage. *Meth. Ecol. Evol.*, 2:303–311.
- Latulippe, C., Lapointe, M. F., and Talbot, T. (2001). Visual characterization technique for gravel-cobble river bed surface sediments; validation and environmental applications contribution to the programme of CIRSA (Centre Interuniversitaire de Recherche sur le Saumon Atlantique). *Earth Surf. Process. Landforms*, 26:307–318.
- Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology*, 74:1659–1673.
- Legendre, P. and Legendre, L. (2012). *Numerical Ecology*, 3rd English edition. Elsevier Science B.V., Amsterdam, The Netherlands.

- Macnaughton, C. J., Harvey-Lavoie, S., Senay, C., Lanthier, G., Bourque, G., Legendre, P., and Boisclair, D. (2014). A comparison of electrofishing and visual surveying methods for estimating fish community structure in temperate rivers. *River Res. Appl.*, 31:1040–1051.
- Michel, M. J. and Knouft, J. H. (2014). The effects of environmental change on the spatial and environmental determinants of community-level traits. *Landscape Ecol.*, 29:467–477.
- Ng, J. (2010). Subsampling water samples in the field. Technical report, Biogeochemical Analytical Laboratory, University of Alberta, Edmonton, AB, Canada.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, 58:267–288.
- Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solution of Ill-posed Problems*. Winston & Sons, Washington, DC, USA. ISBN 0-470-99124-0.
- Wolman, M. G. (1954). A method of sampling coarse river-bed material. *Trans. Am. Geophy. Union.*, 35:951–956.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, 67:301–320.