Research

Ecography 39: 001–010, 2016
doi: 10.1111/ecog.02423

Subject Editor: Luis Mauricio Bini. Editor-in-Chief: Miguel Araújo. Accepted 2 August 2016

Modelling habitat distributions for multiple species using phylogenetics

Guillaume Guénard, Gabriel Lanthier, Simonne Harvey-Lavoie, Camille J. Macnaughton, Caroline Senay, Michel Lapointe, Pierre Legendre and Daniel Boisclair

G. Guénard (guillaume.guenard@gmail.com), G. Lanthier, S. Harvey-Lavoie, C. J. Macnaughton, C. Senay, P. Legendre and D. Boisclair, Dépt de sciences biologiques, Univ. de Montréal, Montréal, QC, Canada. – M. Lapointe, Dept of Geography, McGill Univ., Montreal, QC, Canada.

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 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.

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, Boisclair 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, Desdevise 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. 2016). 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 (≈ 32 000 species; Froese and Pauly 2015) that has evolved over the last 530 million yr, 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 (**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 (**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:

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

where $\mathbf{Y} = [y_{i,j}]$ is an n × 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 × p design matrix with a constant (all ones) column vector followed by p – 1 species descriptors; $\mathbf{Z} = [z_{j,l}]$ is an m × q design matrix with constant column vector followed by the q – 1 site descriptors; $\mathbf{B} = [b_{k,l}]$ is a p × 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 **X** and **Z** are constant vectors, the leftmost column vector of **B** ($b_{k,1}$) contains the marginal (or main) effects of the species descriptors, the uppermost row vector of **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 **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 **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 **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 ⟨⟩, which takes a matrix and transforms it into a single column vector by stacking its columns) and the Kronecker product (⊗) 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 Supplementry material Appendix 1, Details on model estimation. It is implemented by estimating a global elastic

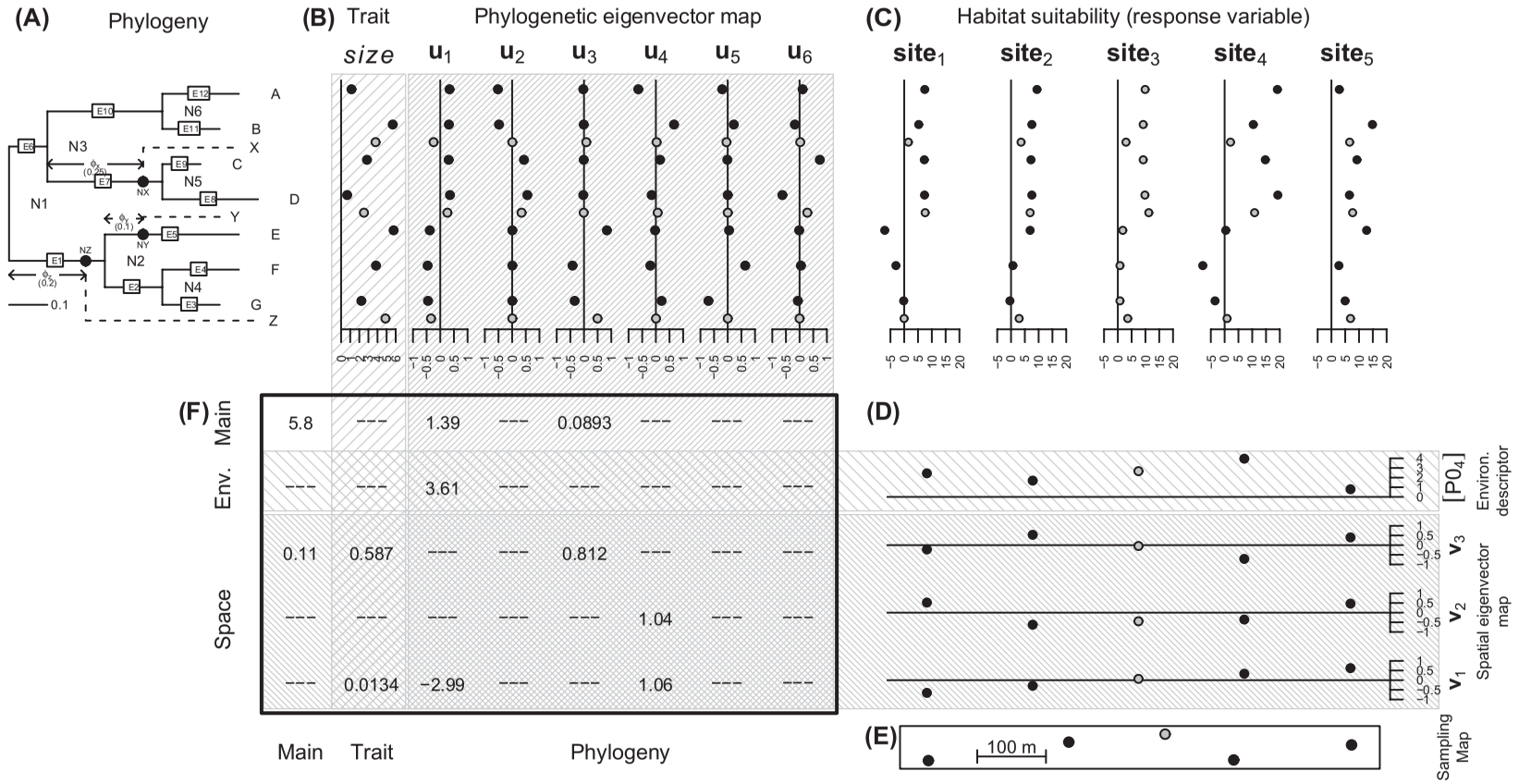

Figure 1. The phylogenetic habitat modelling approaches proposed in present study. A phylogeny (A) is used to calculate a phylogenetic eigenvector map (PEM; Guénard et al. 2013, for details on the computation of a PEM), which is used together with species traits to obtain (B) a matrix of row descriptors associated with the row of a response matrix (C) containing a metric of habitat suitability for every species. A second matrix (D) of column descriptors associated with the columns of the response matrix is made of environmental descriptors together with spatial eigenvectors, which are obtained from the geographic locations of the sampling sites ((E); see Legendre and Legendre (2012) for a description). The parameters of the model are stored in a matrix of bilinear regression coefficients ((F); the matrix is shown as its transpose in order for its content to align with (B) and (D)) The different shadings represent different types of descriptor and their interactions, only the non-zero coefficients are shown. The coefficients of that matrix are estimated from the observed values of the metric of habitat suitability (black markers on (C)) using the values of trait and PEM (black markers on (B) for species A–G), and environmental descriptor and spatial eigenvector (black markers on (C) for sites 1, 2, 4, and 5). Using these coefficients, one can use trait values and calculate the PEM values (grey markers on (B) for species X, Y, and Z) of new (target) species, and use environmental variables and calculate spatial eigenvector values (grey markers on (D) for site 3) of target sites to make predictions. There are three types of predictions (grey markers on (C)): row (or species) predictions (for species X, Y, and Z on sites 1, 2, 4, and 5), column (or site) predictions (for species A–G on site 3), and full predictions (for species X, Y, and Z on site 3).

net shrinkage between 0 and + ∞ 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:

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

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^2_{like} 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^2_{like} 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 $\mathbf{c}_\xi$) were estimated as those maximising the cross-validation R^2_{like} (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^2_{like} 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^2_{like} as a determination rather than a prediction coefficient, we propose to adjust it using the approach proposed by Ezekiel (1930) as follows:

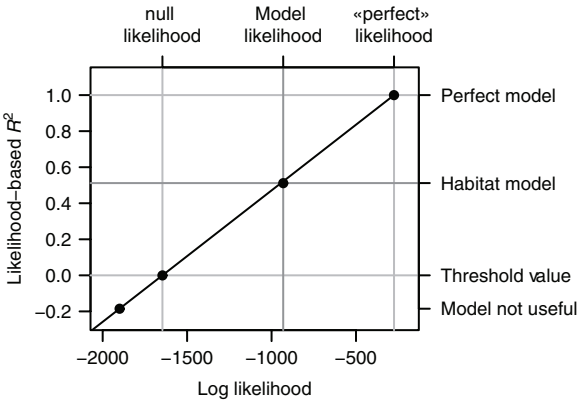

Figure 2. Metric of model performance used in the present study, which is 1 minus the ratio of the logarithms of two likelihood ratios. The first likelihood ratio is the likelihood of a model predicting the response perfectly compared to that of the model being assessed; it is divided by a second likelihood, 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.

$$R^{2*}_{like} = 1\frac{(n\,m-1)}{(n\,m-v)}\left(1-R^2_{like}\right) \quad (4)$$

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 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. (2016a), 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 (Supplementary material Appendix 1), 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 Supplementary material Appendix 1, Data collected, 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 represent 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 (°C d); and total phosphorus, TP (µg l^{-1}).

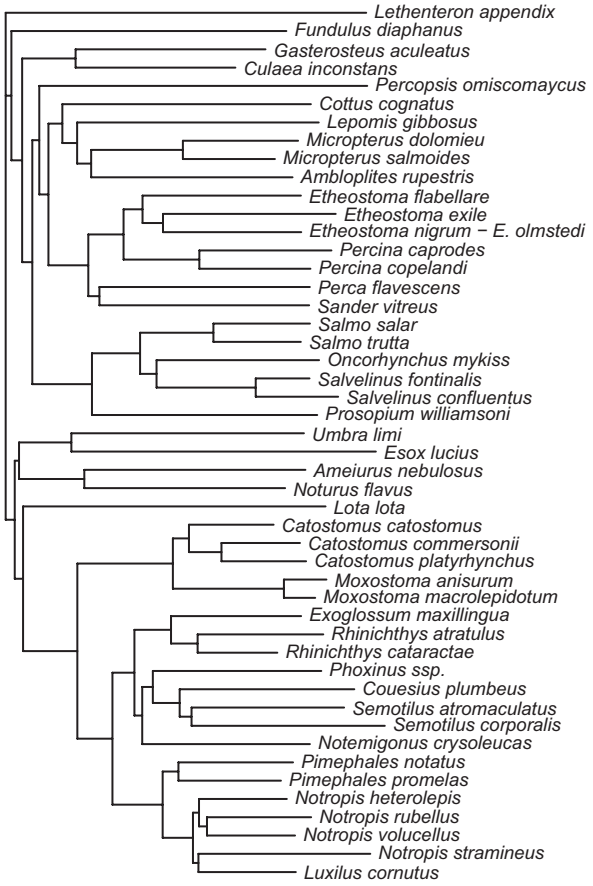

Figure 3. Phylogeny of the fish species observed in the 28 rivers (obtained from Hubert et al. 2008) after removing the species that were not observed during the surveys.

Two types of row descriptors were used (see Supplementary material Appendix 1, 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 (Supplementary material Appendix 1, Eq. 2) by gradient descent as randomised sub-sampling would inevitably introduce noise in the objective function (R^2_{like}) 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). 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.

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.60n52> (Guénard et al. 2016b).

Results

Phylogenetic habitat model (unregulated rivers)

Fish density ranged from 0 (1569 observations out of 2 145, or 73.1%) to 23 fish · (100 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^2_{like} of the cross-validated model was 0.283 during parameter estimation; see Table 1, for the other regularisation parameters. The R^2_{like}* (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 · 10–5 to 6.4 fish (100 m^2)$^{-1}$ (Fig. 4c).

Predictions for regulated rivers

Observed fish densities ranged from 0 (1419 observations out of 1859, or 76.3%) to 18.5 fish · (100 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 fish · (100 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 fish · (100 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 2). 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 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, Candes and Tao 2007) and it may be worthwhile to investigated 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 km) geographic range across Canada. Different study 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 (Martins and Hansen 1997, Garland and Ives 2000) on the basis of its relative simplicity and

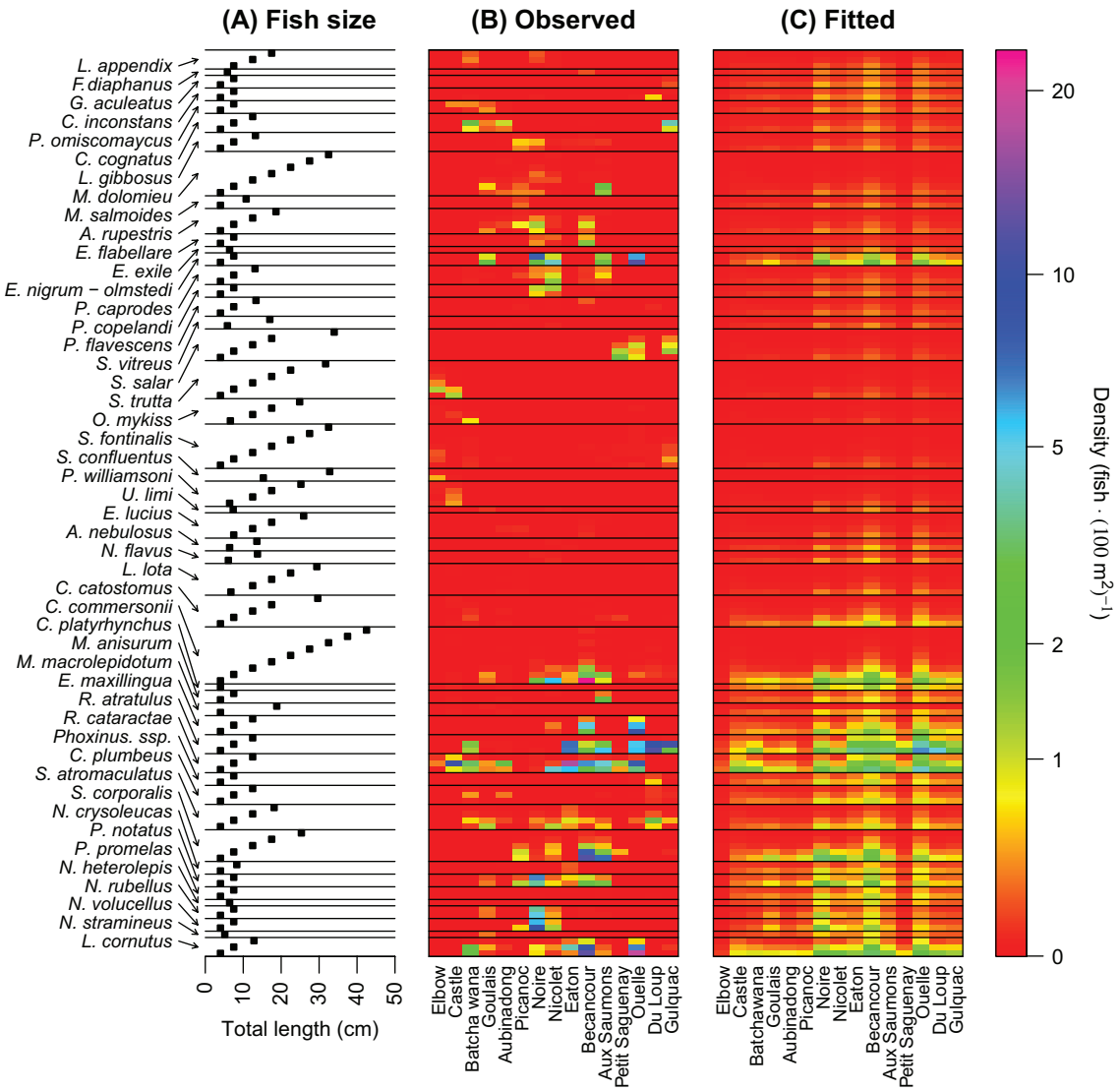

Figure 4. Different combinations of species and size classes found in 15 unregulated rivers, with (A) the median size of each class, (B) the observed fish densities, and (C) the densities fitted by the phylogenetic habitat model.

ability to assess interaction between the species and site descriptors.

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 Supplementary material Appendix 1, Details about model estimation, for details).

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

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

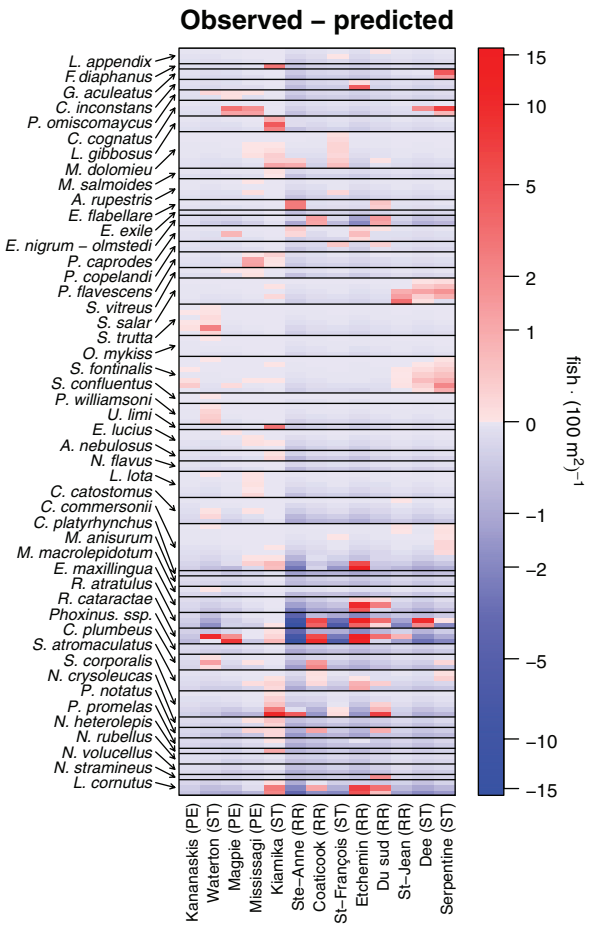

Figure 5. Difference between fish densities observed in the 13 regulated rivers for different species and size classes (see Fig. 4A for their corresponding median total lengths) and the baseline values predicted by the model built with information from the unregulated rivers. The differences are represented 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').

organisms living in rivers (Bonner and Wilde 2000, Nilsson et al. 2005). Here, we will use these results (Table 1) 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 Supplementary material Appendix 1, Details on the interactions between phylogeny and the environment, for further details about these interactions).

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] $p < 0.0001$.

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 km^2) 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 1).

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 2, Fig. 4). That result concurs partially with that of Guénard et al. (2016a), 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 m^2 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. (2016a) 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 (Univ. 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 Cajo Ter Braak (Wageningen UR), as well as that of Subject Editor and reviewer. GG was supported by a Strategic Network Grant to DB, PL, and 11 other scientists (NETGP/370899-2008), and an Individual Discovery grant from the Natural Sciences and Engineering Council of Canada (NSERC) to PL. 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. Freshwater Ecol. 15: 189–198.

Bratrich, C. et al. 2004. Green hydropower: a new assessment procedure for river management. – River Res. Appl. 20: 865–882.

Bruggeman, J. et al. 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. Stat. 35: 2313–2351.

Costello, M. J. et al. 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. Manage. 5: 330–339.

Desdevise, Y. et al. 2003. Quantifying phylogenetically structured environmental variation. – Evolution 57: 2647–2652.

Diniz-Filho, J. A. F. et al. 1998. An eigenvector method for estimating phylogenetic inertia. – Evolution 52: 1247–1262.

Diniz-Filho, J. A. F. et al. 2015. The best of both worlds: phylogenetic eigenvector regression and mapping. – Genet. Mol. Biol. 38: 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. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.

Ezekiel, M. 1930. Methods of correlation analysis. – John Wiley and Sons.

Felsenstein, J. 1985. Phylogenies and the comparative method. – Am. Nat. 125: 1–15.

Flodmark, L. E. W. et al. 2004. Performance of juvenile brown trout exposed to fluctuating water level and temperature. – J. Fish Biol. 65: 460–470.

Friedman, J. H. et al. 2010. Regularization paths for generalized linear models via coordinate descent. – J. Stat. Softw. 33: 1–22.

Froese, R. and Pauly, D. 2015. Fishbase. – <www.fishbase.ca>.

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. et al. 2010. Multiscale codependence analysis: an integrated approach to analyze relationships across scales. – Ecology 91: 2952–2964.

Guénard, G. et al. 2011. Using phylogenetic information to predict species tolerances to toxic chemicals. – Ecol. Appl. 21: 3178–3190.

Guénard, G. et al. 2013. Phylogenetic eigenvector maps (PEM): a framework to model and predict species traits. – Methods Ecol. Evol. 4: 1120–1131.

Guénard, G. et al. 2014. Using phylogenetic information and chemical properties to predict species tolerances to pesticides. – Proc. R. Soc. B 281: 20133239.

Guénard, G. et al. 2015. Phylogenetics to help predict active metabolism. – Ecosphere 6: Article 62.

Guénard, G. et al. 2016a. A spatially-explicit assessment of the fish population response to flow management in a heterogeneous landscape. – Ecosphere 7: e01252.

Guénard, G. et al. 2016b. Data from: Modelling habitat distributions for multiple species using phylogenetics. – Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.60n52>.

Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.

Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.

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 and Hall/CRC.

Hubert, N. et al. 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: 511–525.

Legendre, P. and Legendre, L. 2012. Numerical ecology, 3rd English ed. – Elsevier Science.

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. et al. 2016. Evolutionary patterns and physicochemical properties explain macroinvertebrate sensitivity to heavy metals. – Ecol. Appl. 26: 1249–1259.

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. et al. 2006. Rebuilding community ecology from functional traits. – Trends Ecol. Evol. 21: 178–185.

Nilsson, C. et al. 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.

Ollier, S. et al. 2006. Orthonormal transform to decompose the variance of a life-history trait across a phylogenetic tree. – Biometrics 62: 417–477.

Pavoine, S. et al. 2008. Testing for phylogenetic signal in phenotypic traits: new matrices of phylogenetic proximities. – Theor. Popul. Biol. 73: 79–91.

Pillar, V. D. and Duarte, L. d. S. 2010. A framework for metacommunity analysis of phylogenetic structure. – Ecol. Lett. 13: 587–596.

Swenson, N. G. 2014. Phylogenetic imputation pf plant functional trait databases. – Ecography 37: 105–110.

Zou, H. and Hastie, T. 2005. Regularization and variable selection via the elastic net. – J. R. Stat. Soc. B 67: 301–320.

Supplementary material (Appendix ECOG-02423 at <www.ecography.org/appendix/ecog-02423>). Appendix 1.