



Original research article

Multiple factors and thresholds explaining fish species distributions in lowland streams

Cristina Trigal^{a,*}, Erik Degerman^b^a Swedish Species Information Center, Swedish University of Agricultural Sciences, Uppsala, Sweden^b Department of Aquatic Resources, Swedish University of Agricultural Sciences, Drottningholm, Sweden

ARTICLE INFO

Article history:

Received 4 May 2015

Received in revised form 21 October 2015

Accepted 21 October 2015

Keywords:

Threshold values

Boosted regression trees

Random forest

Sweden

Headwater streams

ABSTRACT

Appropriate restoration and conservation measures require a good understanding of the factors limiting the distribution of species, the presence of steep changes in the distribution along environmental gradients and the effect of environmental interactions on species distribution. We used 12 environmental variables describing connectivity, hydrology, climate and stream morphology, to model the distributions of 17 fish species from 2005 Swedish stream sites that were sampled between 2000 and 2011. Modeling was performed using boosted regression trees and random forest, two machine learning techniques to assess the relationship between species distributions and their environment. Temperature, width and connectivity (minimum distance to lake or the sea and water discharge), were the most important variables explaining changes in species distribution at large spatial scales. Response curves of fitted occurrence probabilities along predictors often showed abrupt changes, however, clear threshold effects were difficult to detect. Our results show also differences across species and even in the outcomes of the two algorithms, implying that a simultaneous assessment of multiple species may provide a better signal of ecosystem change than the use of surrogate species.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Freshwater habitats, supporting ca 10% of all known species, are among the most threatened ecosystems in the world (Vorosmarty et al., 2010). Some of the most severe threats to freshwater biodiversity, such as habitat degradation, flow regulation, and species invasion, result in loss of taxa richness (Schinegger et al., 2012), declines in the distribution range and abundance of many species (Baxter et al., 2004; Byström et al., 2007), and eventually have negative effects on ecosystem functioning. Degradation of freshwater ecosystems will continue, as water demand and physical alterations will increase with human population density (Degerman et al., 2007; Schinegger et al., 2012), and as a result of anthropogenic induced climate change (Buisson et al., 2008; Griffiths et al., 2014). In an attempt to prevent European freshwater systems from further degradation, the Water Framework Directive (European Commission, 2000) and the EC Habitats Directive (Council of the European Communities, 1992) were developed, which aim at maintaining and restoring freshwater habitats to a favorable conservation status through the development of management and restoration strategies. This is, however, a difficult task because species are affected by multiple factors acting at different spatial scales (e.g. local and

* Correspondence to: Swedish Species Information Center, Swedish University of Agricultural Sciences, Box 7007, 750 07, Uppsala, Sweden. Tel.: +46 0 18672327.

E-mail address: cristina.trigal@slu.se (C. Trigal).

<http://dx.doi.org/10.1016/j.gecco.2015.10.009>

2351-9894/© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Table 1

Environmental predictors used in the models.

Predictor name	Type
Water temperature (°C)	Continuous (range, 5–27°C)
Substrate	Categorical (A, <0.02 cm; B, 0.02–0.2; C, 0.2–2; D, 2–10; E, 10–20; F, 20–30; G, 30–40; H, 40–200; I > 200 cm)
Annual discharge (m ³ s ⁻¹)	Continuous (range 0.002–966 m ³ s ⁻¹)
Width (m)	Continuous (range, 0.3–10 m)
CV discharge	Continuous (range, 3.8–242)
Minimum distance to lake or sea (km)	Continuous (range, 0.1–10 km)
Woody debris (number 100 m ⁻²)	Continuous (range, 0–167 pieces 100m ⁻²)
Mean depth (m)	Continuous (range, 0.02–1.3 m)
Shade (%)	Continuous (range, 0–100)
Barriers	Categorical (U, upstream; D, downstream; B, upstream and downstream the sampling point)
Sampling effort	Number of electrofishing passes (range, 1–3)
Flow velocity	Categorical (S, slow; F, fast)

catchment) (Degerman et al., 2007; Schinegger et al., 2012; Törnblom et al., 2011), and because they respond individually to environmental change (Olden et al., 2006; Parmesan and Yohe, 2003). To achieve a good conservation status and identify appropriate restoration and conservation measures, it is therefore necessary to first identify the factors limiting the distribution of individual species and to evaluate the effects of interactions among environmental drivers on species distribution (Guisan et al., 2013).

Conservation management also requires a good understanding of threshold effects along environmental gradients that may cause abrupt changes in species distribution (Roni et al., 2008), i.e. how much and what quality of habitat is required for different species in different environments? Thresholds are, however, difficult to predict, as they depend on a number of factors including landscape characteristics, species traits and non-linear relationships between species and the environment (Lindenmayer and Luck, 2005; Suding and Hobbs, 2008). In addition the interactions among environmental drivers may affect threshold values and produce complex responses in species distribution (Olden, 2007; Pittman and Brown, 2011), and complicate the outcomes of restoration. For example, many restoration programs in streams aim at increasing habitat heterogeneity through adding large woody debris or manipulating stream substrate; however, how fish species perceive environmental heterogeneity will depend on the interactions between the variable of interest and other local and regional variables (e.g. water level fluctuations, presence of barriers, etc.). Species distribution models are used to evaluate habitat suitability and the existence of thresholds in species occupancy over large spatial and temporal scales (Elith and Leathwick, 2009; Guisan et al., 2006, 2013). Those models often include non-linear relationships between species occurrence or abundance and habitat variability (Elith and Leathwick, 2009; Guisan et al., 2006). However, only a few studies have looked explicitly into the effect of interactions among drivers on threshold values and species occupancy.

In this study we use an extensive data set describing the distributions of 17 fish species across lowland streams in Sweden, sampled between 2000 and 2011. We use 12 environmental variables describing connectivity, hydrology, climate and stream morphology, which are important for fish (Degerman et al., 2004; Morin and Naiman, 1990; Riffart et al., 2009). The aims of the study are to: (a) identify the drivers that contribute most to the distribution of individual species and community turnover; (b) identify changes in environmental drivers that result in abrupt changes in species occurrence (threshold effects); and (c) evaluate the consistency of the species–environment relationships over time. We use boosted regression trees (De'ath, 2007) and random forest (Hothorn et al., 2006), two machine learning techniques to assess the relationship between species distributions and their environment.

2. Methods

2.1. Study area

Fish and environmental data were drawn from the Swedish Electrofishing Register (SERS), a database containing more than 56 500 records from 17 500 sites sampled across Sweden from 1951 onwards. For this study we selected a subset of 2005 lowland sites sampled at least once between 2000 and 2011. The study sites were located at altitudes lower than 200 m a.s.l (see Fig. 1). This boarder coincides roughly with the Swedish highest coastline, which acts as a natural barrier and plays a role in limiting the dispersal of lowland fish species into streams at higher altitudes (Ekman, 1922). We selected sampling sites with a wetted width less than 10 m, due to the reduced effectiveness of electrofishing by wading in wide streams (Kennedy and Strange, 1981). Water temperature at the time of sampling ranged from 5 °C to 27 °C. The surrounding landscape consisted of forest, with coniferous species dominating, and agricultural lands, particularly in southern Sweden. Other environmental variables are described in Table 1.

2.2. Fish sampling

Sampling was performed in August, according to national standards. At each site a 20–50 m long transect (total area 200–300 m²) was sampled by electric fishing, using a bank-based generator operated by a two-crew team wading and using

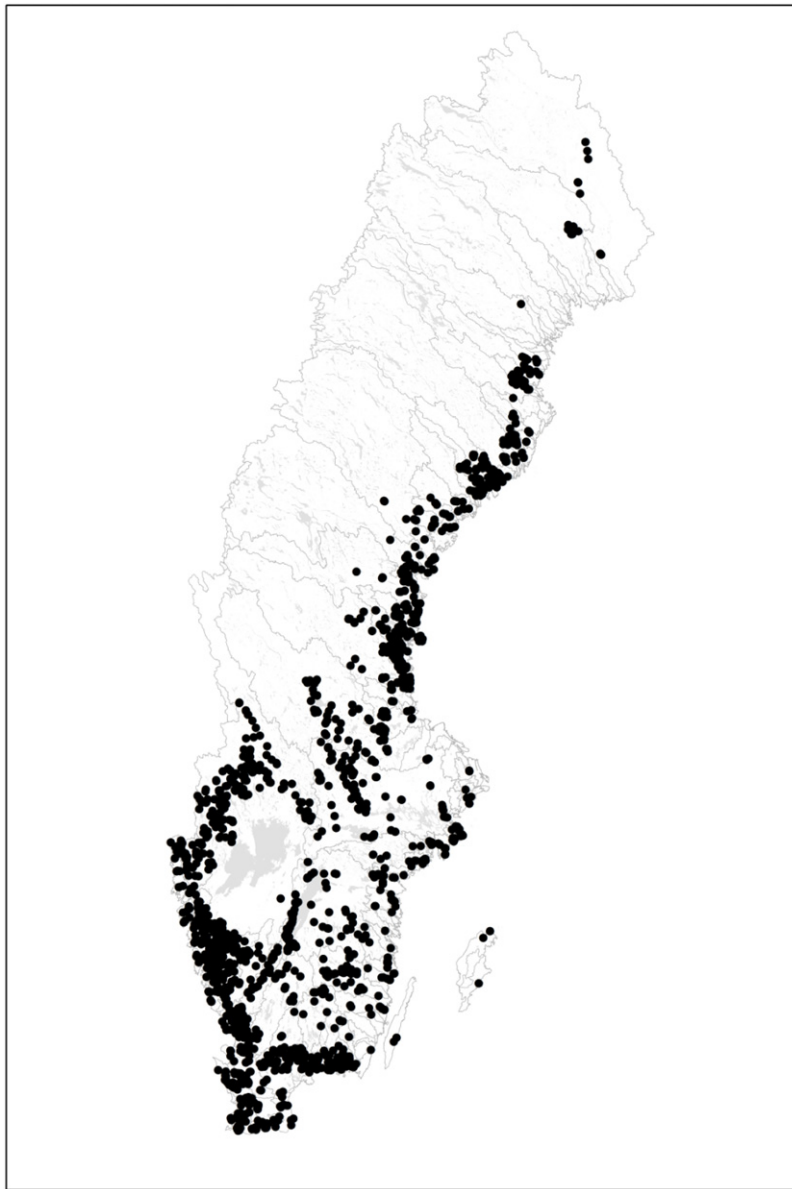


Fig. 1. Location of the sampling sites in small lowland (<200 m asl) streams in Sweden (black dots).

a single handheld anode. The electrofishing apparatus used were in 84% of the cases equipment from the national brand LUGAB using non-pulsed DC of 200–1000 V. In the remaining occasions other brands or pulsed-DC were used. The sampling effort, i.e. the number of successive removals at each site, varied between one and three. Fish were identified to species or genus (namely lampreys *Lampetra* sp.), counted and their total length measured in situ.

2.3. Environmental variables

Mean depth (m), wetted width (m), amount of large woody debris (at least 0.1 m in diameter and 0.5 m long) given as number of pieces per 100 m², water temperature (°C) at sampling, and percent of stream surface shaded from the sun at midday, were measured on each sampling occasion. The dominating bottom substrate was classified into seven groups based on grain size (Table 1). Water velocity was classified as slow (<0.2 m/s) or fast (≥0.2 m/s) at the time of sampling.

Data on the presence of dams blocking the way to nearest lake (or the sea), and the within stream distance from the sampling site to the nearest lake (or the sea) were measured in a GIS environment using ArcMap 10.2.

Catchment run-off was used as a proxy to estimate mean flow and water flow fluctuations at the sampling stations. For each site the average water discharge in the hydrological year was modeled using the S-HYPE (Arheimer et al., 2011;

Lindström et al., 2010). S-HYPE models flood regimes in each catchment from precipitation and land use data at large and fine spatial scales (www.smhi.se). We used October 1 as the beginning of the hydrological year and July 31, before fish sampling was conducted, as the end of the hydrological year. The intra-annual variation in water discharge was a proxy for water flow variation (i.e. flood events and droughts) and was expressed as the coefficient of variation of mean discharge over 10 months.

2.4. Statistical analysis

Boosted regression trees (hereafter, BRT De'ath (2007)) and random forest (hereafter RF Breiman (2001)) models were used to investigate the relationship between species and environmental variables.

BRT combine the results of a large number of classification (or regression) trees using a boosting algorithm to optimize model's predictive performance. Trees are added to the model sequentially to minimize prediction error in the model residuals, up to an optimal number of trees. BRT may accommodate missing values and different types of predictor variables, and is not affected by outliers, differing scales of measurement among predictors, or monotone transformations (De'ath, 2007; Elith et al., 2008). To identify the optimal number of trees BRT use a learning rate (lr), which is the contribution of each tree to the model, and tree complexity (tc), which is the number of splits in a tree, and allows for interactions among variables, 1 being main effects, 2 first order interactions, and so on (De'ath, 2007; Elith et al., 2008). We fitted models with varying values for tc (1, 3 and 5) and lr (0.05 and 0.001), and then used tenfold cross validation to identify the optimal number of trees, lr and tc values. We used lr = 0.001 and tc = 5 in the final models, as these values achieved the minimum predictive error and fitted more than 1000 trees for each species. The bag fraction, the random subset of data used to fit each new tree, was 0.5, which is adequate for presence–absence data (Elith et al., 2008).

The relative importance of each predictor was calculated based on the sum of squared improvements of all splits for a given predictor averaged over all trees and then scaled so that the sum of all variables is 100, the most important variable being the one with the highest score. Model performance was evaluated using the cv deviance and the area under the receiver operator characteristic curve. These two parameters were estimated with tenfold cross validation on a subset of calibration samples left out of the analysis during model training (cv deviance and cv AUC). Deviance is a measure of the goodness of fit of the model, whereas AUC is a measure of discrimination accuracy. An AUC value of 0.5 or less indicates predictions no better than chance alone (Hanley and McNeil, 1982).

In RF, each ensemble of classification tree models is created using bootstrap samples of the training data and a random selection of predictor variables. Predictor variables are split into two groups using the cut-off value that forms the most homogenous groups possible. The new groups can then be split again until a group contains only presences or absences or until a pre-set maximum number of splits have been made. We used extended forest which uses an unbiased algorithm to accommodate variables of different types (continuous and categorical) (Hothorn et al., 2006; Strobl et al., 2007) and may also handle missing values in a similar way to BRT. We run a total of 1000 tree models selecting 3 variables for each model ensemble. The latter approximates the square root of the total number of predictors, which is the standard option for classification in random forest. Other combinations of number of trees (1000 and 1500) and predictor variables (3 and 5), were also tested, but the results did not differ significantly.

Variable importance in RF was estimated by permuting one of the variables, and predicting the response with the permuted variable. If the variable in question is important for the response, the predictions from the permuted variable should be less accurate. Model performance in random forest was estimated using out-of-the bag observations. For each tree in the training model, about one-third of the calibration sites (out-of-bag observations, OOB) are set aside and the tree is built on the remaining sites. The OOB sites are a form of internal cross-validation and are used to estimate the accuracy of the calibration model (AUC OOB).

Models were fitted using presence–absence data for 17 species collected from the 2005 sites in August. The same analyses were even run on data for other 17 species with low occurrence frequency (<1%; i.e. less than 30 occurrences) but the models had low performance (AUC values <0.5) and those species were finally excluded from the study. A species was considered absent when not detected during electrofishing. Twelve environmental predictors describing habitat heterogeneity, connectivity and sampling effort were included in the models (Table 1). Variables were untransformed, except for the annual mean water discharge, which was square root transformed before analysis. All these environmental variables are known to be important in explaining the distribution and abundance of freshwater fish (Bohlin et al., 1989; Degerman et al., 2004; Shirvell and Dungey, 1983). Sampling effort, i.e. the number of electrofishing passes at a site, was included in the analysis to test if species detectability interferes with the response of the species to a given predictor. Prior to analysis, the correlation between pairs of predictors was tested and predictors with Spearman's rho >0.7 were excluded because estimates of variable importance may be biased to correlated variables (Strobl et al., 2008).

Samples for training the models were acquired by randomly selecting one year for each site (i.e. 2005 samples expanding 12 years, hereafter *calibration samples*). In this way, we intend to capture spatial and temporal variability and avoid potential problems introduced by temporal autocorrelation when two samples from the same site and different years are used in the analysis (Hallstan et al., 2013). We also tested for their ability to account for temporal variation using 1500 samples (i.e. same sites, different years, hereafter *validation samples*) not included in the calibration set. We applied the models to the validation sites and tested for significant differences in AUC values before and after the calibration years. Analyses were

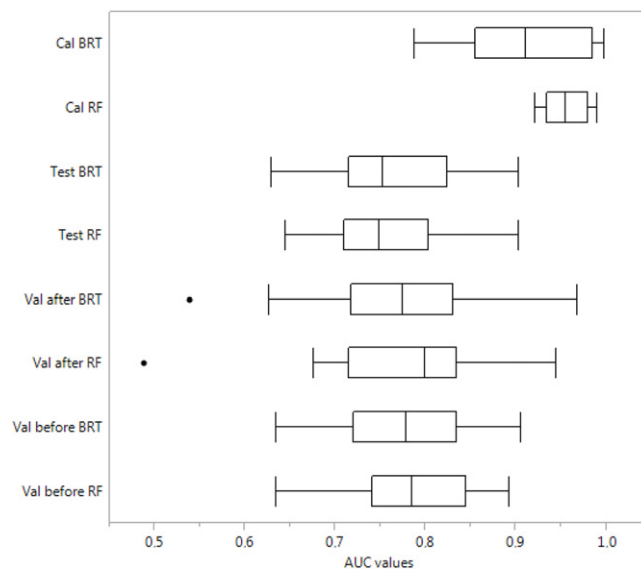


Fig. 2. Model statistics for the BRT and RF models averaged across 17 species. Median, quantiles and standard deviation of AUC values are shown for the calibration sets (Cal BRT and Cal RF), the cross validation (Test BRT) and out-of-the bag observations (Test RF), and the validation sets before (Val before BRT and Val before RF) and after (Val after BRT and Val after RF).

conducted in R 3.1.1 (R Core Team, 2010) with the Dismo (Hijmans et al., 2011) and party packages (Hothorn et al., 2006; Strobl et al., 2007).

3. Results

3.1. Model evaluation

There were no significant differences in the AUC values of the calibration, test and validations sets, i.e. before and after calibration year, between the BRT and RF models (Fig. 2, Table S1). The lowest model performance was for lamprey and pike, whereas the models for bullhead, nine-spine stickleback, brown trout, stone loach, common bleak and brook trout had AUC values greater than 0.8.

When the models were applied to validation time series, the AUC were consistent over time (Fig. 2, Table S1), i.e. no significant differences in AUC values between the years before and after the calibration year were found.

3.2. Relative importance of predictors

Width, water temperature and two connectivity variables (minimum distance to lake and water discharge) were among the 5 most important predictors according to BRT and RF models (Fig. 3, Table 2). Many species responded positively to temperature and annual discharge whereas the sign of the response to width and minimum distance to lake was more variable (Fig. 3). Substrate was the second most important variable in the BRT model, especially for eel, lamprey, pike, ruffe and grayling (Table 2). Sampling effort was the fifth most important variable in the RF model, affecting the occurrence probability of species such as trout and lamprey (Table 2). Water velocity (only two classes: slow and fast) had in general negligible effects. Other predictors with moderate to low contributions were woody debris, presence of barriers and shade. The relative importance of individual predictors and the response of species occurrence were highly variable among species.

3.3. Response of individual species

We analyzed the response of the 17 species along four of the most important predictors according to the BRT and RF models, respectively, water temperature, water discharge, distance to lake or sea and width. Many species, such as burbot, eel, minnow, perch, pike, roach (BRT and RF models), bleak and ruffe (BRT models) showed a monotonic response to temperature and their occurrence increased with temperatures above 15–20 °C (Fig. 4). Bullhead and lamprey were more common in cold waters (<15 °C; both models). Salmon, grayling and stone loach showed a non monotonic response and occurred more frequently at temperatures between 10 °C and 20 °C. For width, species such as bullhead, burbot, minnow, pike or salmon, preferred wider streams (>4 m), whereas others, such as sticklebacks, were more commonly found in small streams (Fig. 5). Most species avoided sites with large water discharges (Fig. 6). Many lacustrine species, such as perch, roach (both models) or burbot (RF models) and brook trout (BRT models) were more common within 4 km from the nearest lake

Table 2
The third most important predictors of species distribution for 17 species according to models: boosted regression trees (BRT) and random forest (RF).

	BRT		RF			
	V1	V2	V3	V1	V2	V3
Perch	Water temperature	Minimum distance to lake or sea	Mean depth	Minimum distance to lake or sea	Water temperature	Mean depth
Eel	Substrate	Water temperature	Flow Velocity	CV discharge	Minimum distance to lake or sea	Water temperature
Lamprey	Substrate	Water temperature	Mean depth	Sampling effort	Substrate	Water temperature
Brook trout	Annual discharge	Substrate	Water temperature	Width	Annual discharge	Water temperature
Bleak	Annual discharge	Substrate	Water temperature	Annual discharge	Minimum distance to lake or sea	Water temperature
Alpine bullhead	CV discharge	Water temperature	Annual discharge	Width	CV discharge	Annual discharge
Minnow	Annual discharge	Substrate	Width	Width	Annual discharge	CV discharge
Pike	Substrate	Barriers	Water temperature	Width	Barriers	Sampling effort
Ruffe	Substrate	Width	Woody debris	Minimum distance to lake or sea	Width	Substrate
Stone loach	Water temperature	Minimum distance to lake or sea	Barriers	Minimum distance to lake or sea	Water temperature	Sampling effort
Grayling	Substrate	Water temperature	CV discharge	CV discharge	Width	Shadowing
Burbot	Annual discharge	Water temperature	Minimum distance to lake or sea	Minimum distance to lake or sea	Width	Annual discharge
Salmon	Mean depth	Woody debris	Water temperature	Minimum distance to lake or sea	Sampling effort	Width
Roach	Water temperature	Substrate	Mean depth	Water temperature	Mean depth	Width
Trout	Sampling effort	Substrate	CV discharge	Sampling effort	Barriers	Substrate
Ninespine stickleback	Shadowing	Width	Substrate	Shadowing	Width	Woody debris
Bullhead	Width	CV discharge	Water temperature	Width	CV discharge	Water temperature

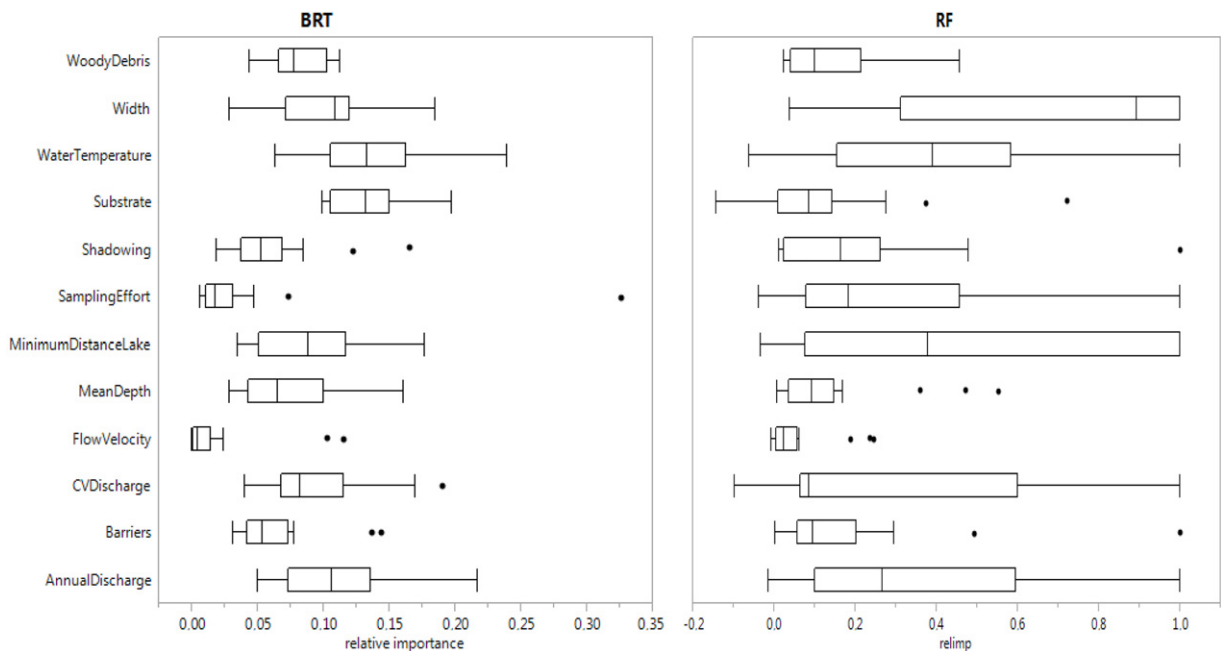


Fig. 3. The relative importance of 12 predictors for the occurrence of 17 species in the BRT and RF models. The median, quantiles and standard deviation are displayed in the plots.

(Fig. 7). In contrast, stone loach, preferred sites far from lakes. Overall, it was difficult to find clear thresholds that separate absences from occurrence along the environmental gradients.

To illustrate how interactions among predictor variables may be contributing to the large overlap between absent and present sites we evaluated the interactions among pairs of predictor variables in the BRT models (Fig. S1). These interactions tell us where the highest probability of occupancy for a given species is, but they also suggest that a change in a driver's threshold leads potentially to changes in the threshold of other drivers. For example, the distribution of bullhead was mostly affected by temperature, and the highest probability of occurrence was predicted in cold streams with high discharge variability. In contrast, for the same temperature thresholds its probability of occurrence was lower with low variations in discharge.

4. Discussion

Our results showed a low degree of congruence in the response and thresholds along environmental gradients for individual species, implying that a simultaneous assessment of multiple species may provide a better understanding of ecosystem change and restoration need than the use of surrogate species (Wiens et al., 2008). Species feeding, breeding strategies, life traits and biotic interactions, may explain the differences in the response and thresholds along local scale variables (e.g. width) for individual species. On the other hand, response and thresholds for variables acting at large spatial scales, such as connectivity or temperature, could be explained by historical and evolutionary processes, the presence of historical natural barriers and the ability of the species to disperse large distances. This is, however, difficult to ascertain from our data and further research is needed to clarify it.

Threshold effects that clearly separated occurrence from absence along the predictor variables were difficult to detect. A likely explanation is that interactions between predictors modulate threshold effects. In other words, changes in a given driver lead to changes in the response and threshold levels of another driver. Indeed, the response to temperature, stream morphometry (e.g. width) and connectivity for individual species was modulated by interactions among these variables according to the BRT models (Fig. S1). Further, the threshold effects may vary both with species, fish community and type of habitat. This is probable when thresholds are determined from competition for food or habitat, but not when thresholds are acting directly on physiological processes or dispersal. If so, a predictor as water temperature or barriers should have more distinct threshold effects than in-stream factors as width, substrate or large woody debris. A likely reason why we do not see a clear threshold along the temperature gradient for most species is that few data were collected from cold water streams.

Connectivity, described in our study by mean discharge, water flow variations, the distance to lake (or the sea) and the presence of natural and artificial barriers, is an important determinant of fish diversity (Fullerton et al., 2010). Small variations in natural flows and connectivity may result in drastic changes in community composition (Pringle and Pringle, 2006). Many species prefer high flows, which help maintain cool temperatures, high oxygen concentrations and low

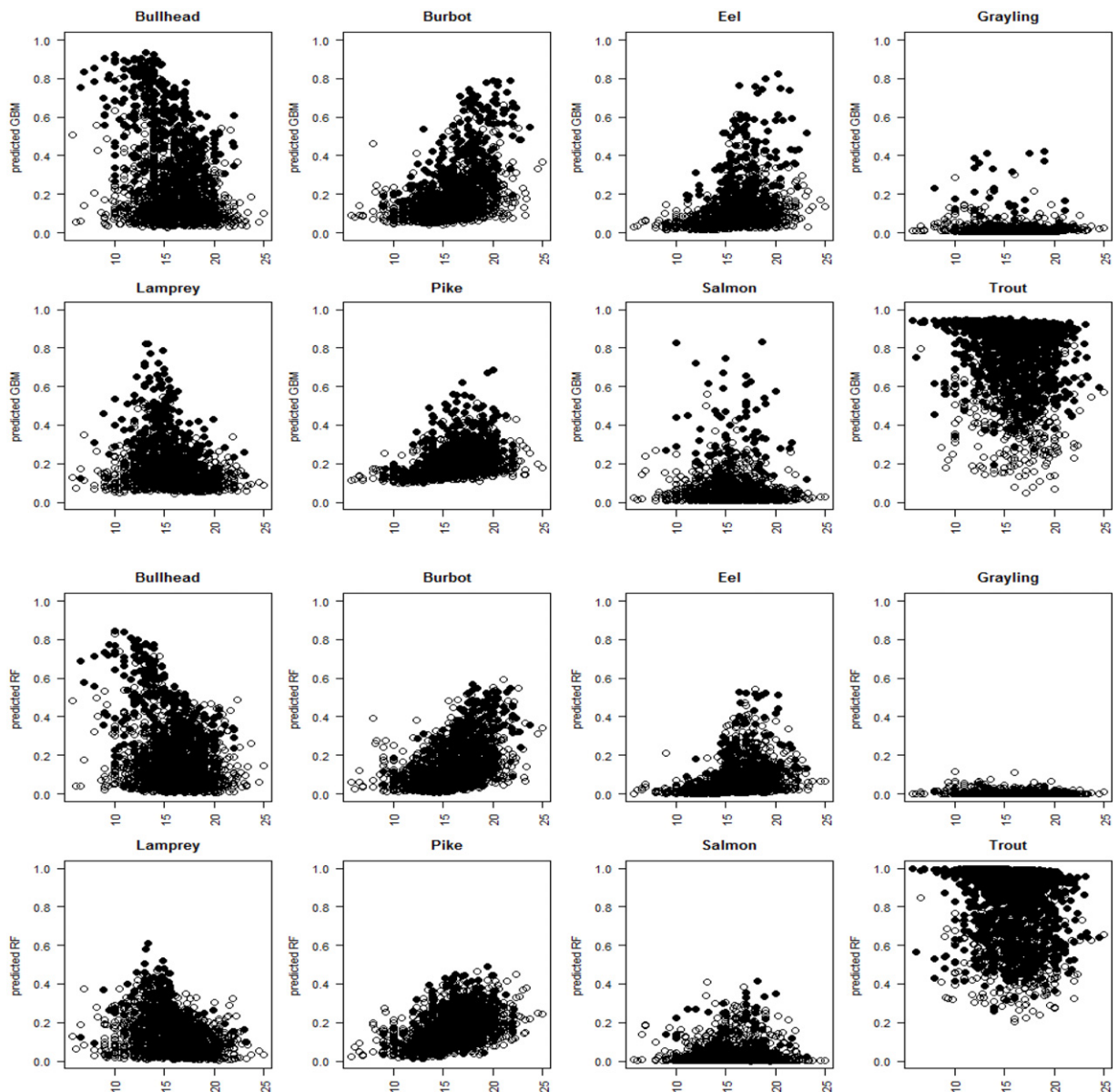


Fig. 4. The predicted probability of species occurrences in response to the water temperature (C) for 8 selected species in the BRT (predicted GBM) and RF (predicted RF) models. Black dots are the observed presences and white dots are the observed absences. A clear threshold in species occurrence along the predictor values would be signaled by a clear separation between present and absent sites. Bullhead, lamprey, burbot, grayling and salmon are species in the Habitats Directive. Eel is in the IUCN red list. Pike is a keystone predator and trout was the most abundant species in the study sites.

sedimentation rates, and enhance species movement and connectivity among habitat patches (Olden, 2007). Low water flows and the presence of barriers, in contrast, lead to a disruption in longitudinal connectivity and have negative effects on the population dynamics and distribution of many lotic species. The low importance of barriers and the negative effect of high discharges in our results were thus unexpected. Distance to barriers and the length of free flowing waters are likely more important than the presence of barriers upstream or downstream the sampling point in explaining species distribution, except for the long-migrating salmon and eel. Why these species were not singled out as impacted by connectivity was probably due to how this was measured. For these two species it is connectivity to the sea that is important, but in the data set was only accounted for barriers to the closest lake or sea. Even if migration routes to downstream lake were open, there could still be migration obstacles further downstream. This stresses that connectivity needs to be measured at different scales and that it is species dependent.

Hence, distance to lake or sea was important for most lacustrine species, namely pike, perch, roach and burbot, which are capable of colonizing some streams from nearby lakes during low flows (Degerman and Sers, 1994). Our results also signalize the existence of a zonation between 2 and 4 km from lake outlet (or inlet), with proximal areas to lakes characterized by

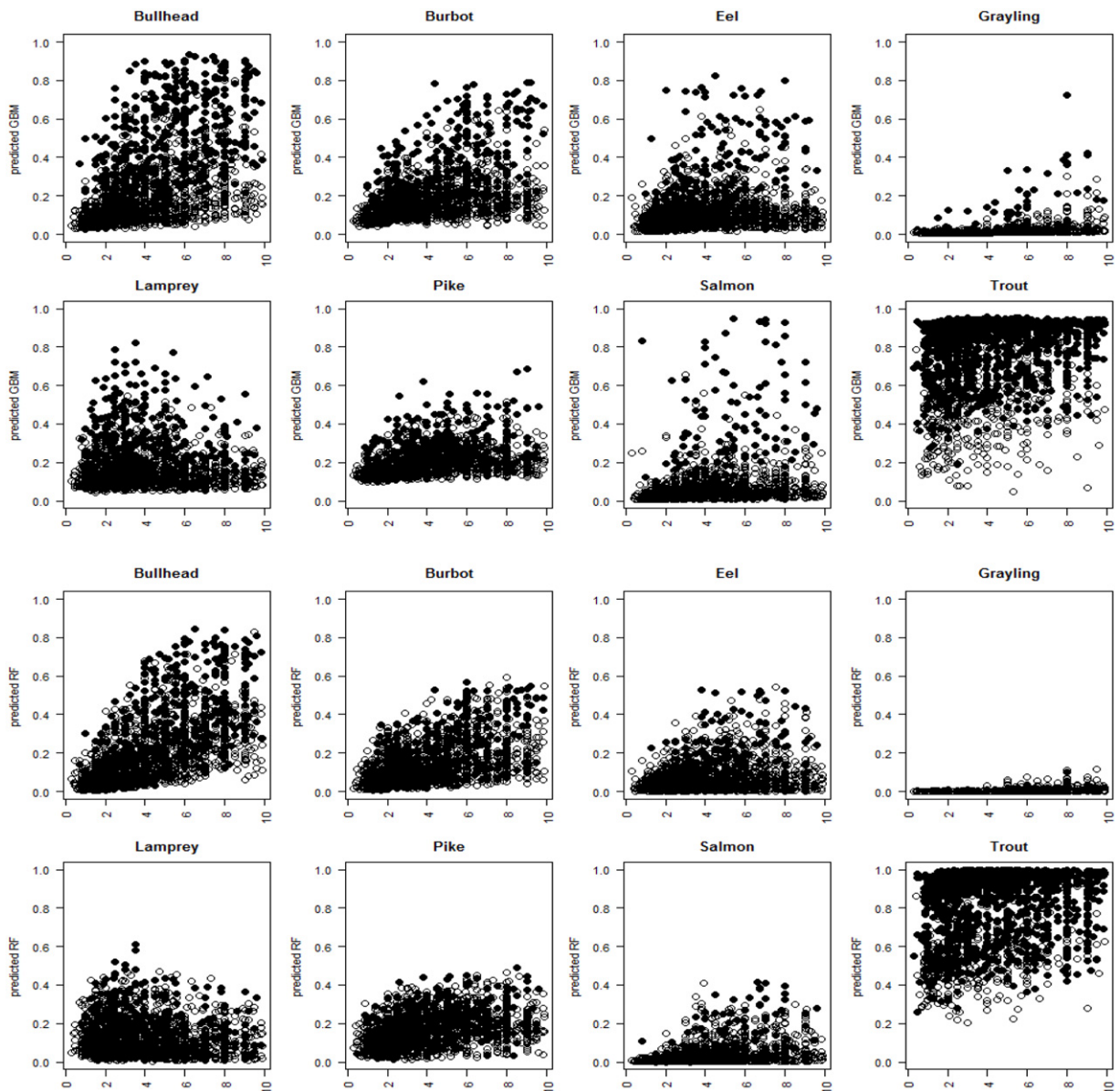


Fig. 5. The predicted probability of species occurrences in response to the width (m) for 8 species in the BRT (predicted GBM) and RF (predicted RF) models. Black dots are the observed presences and white dots are the observed absences. A clear threshold in species occurrence along the predictor values would be signaled by a clear separation between present and absent sites. Bullhead, lamprey, burbot, grayling and salmon are species in the Habitats Directive. Eel is in the IUCN red list. Pike is a keystone predator and trout was the most abundant species in the study sites.

high occurrence rates of limnetic species and low rates of some lotic fish. Proximity to lakes may alter sedimentation rates, water quality, substrate and flow velocity in streams, benefiting lentic species. In addition, lake closeness or accessibility may increase competition and predation by lacustrine fish on lotic species (op. cit., [Spens and Ball \(2008\)](#)).

Temperature affects the distribution of fish species directly through physiological effects on growth, hatching and reproduction ([Aigo et al., 2014](#); [McCormick et al., 1997](#); [Ohlberger et al., 2008](#); [Rodriguez-Munoz et al., 2001](#)) and indirectly, by modifying interspecific competition or favoring opportunistic predators ([Destaso and Rahel, 1994](#); [McMahon et al., 2007](#); [Mehner et al., 2011](#)). The results of our study suggest a community turnover at temperature values between 15 and 20 °C. These values signalize a zonation between cold water streams, characterized by species such as alpine bullhead and grayling ([Andreasson, 1972](#); [Wedekind and Kung, 2010](#)), in northern Sweden and warm water streams at lower latitudes, where other species, e.g. eel and most of the studied lacustrine fish, are likely to be found ([Degerman and Sers, 1992](#)). However, for species with restricted distributions in lowland streams their thermal requirements may insufficiently be quantified by a single measurement of water temperature as in the present study. For example, warm temperatures >15–20 °C were seemingly beneficial for burbot. This is in contrast to other studies, which have shown that burbot is a cold stenothermic species

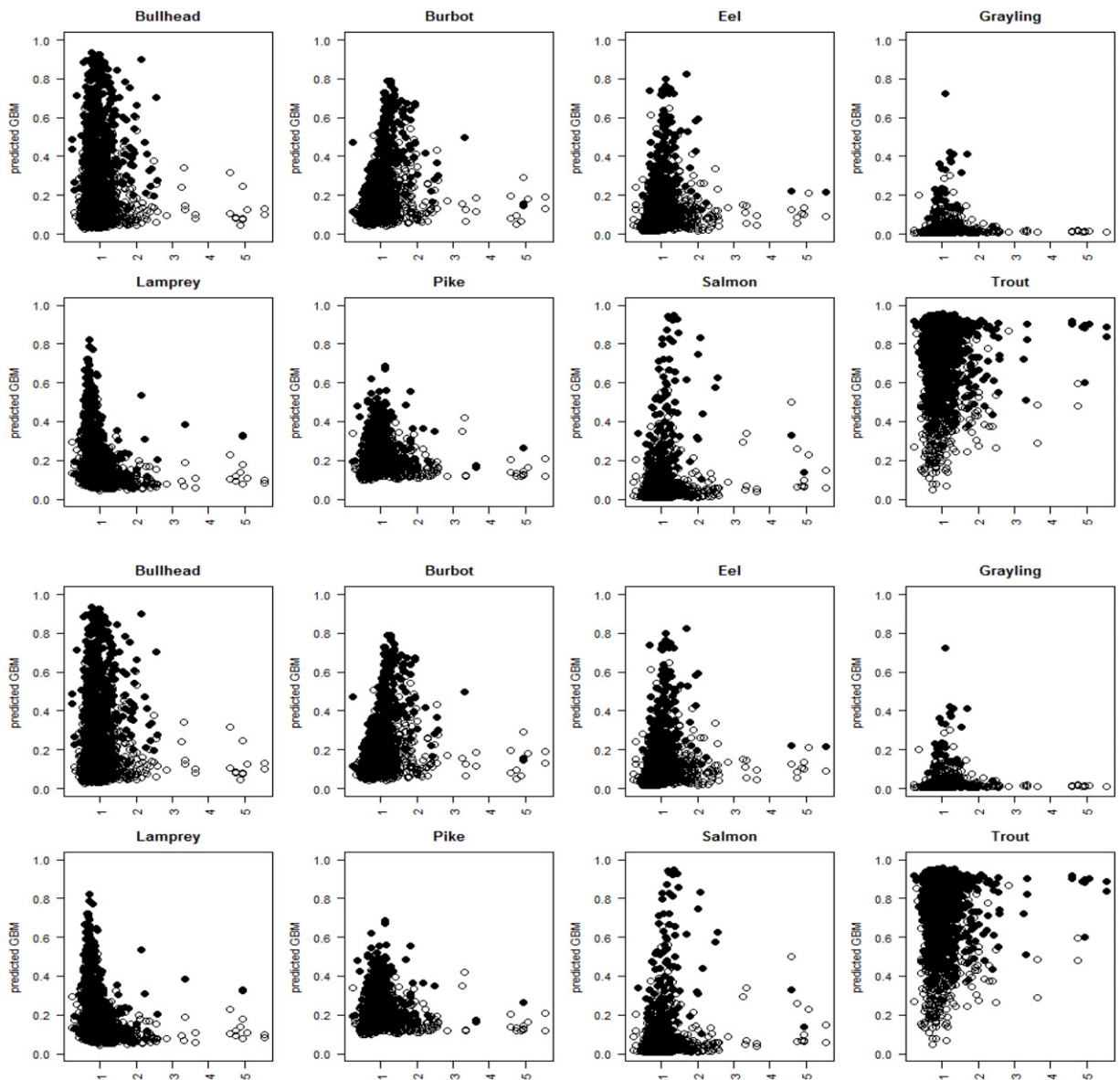


Fig. 6. The predicted probability of species occurrences in response to the water discharge ($\text{m}^3 \text{s}^{-1}$) for 8 species in the BRT (predicted GBM) and RF (predicted RF) models. Black dots are the observed presences and white dots are the observed absences. A clear threshold in species occurrence along the predictor values would be signaled by a clear separation between present and absent sites. Bullhead, lamprey, burbot, grayling and salmon are species in the Habitats Directive. Eel is in the IUCN red list. Pike is a keystone predator and trout was the most abundant species in the study sites.

(Hoffman and Fischer, 2002) and that high temperatures become stressful and negatively impact its distribution (Stapanian et al., 2010). However, the catch efficiency of electrofishing is generally improved with increased water temperature (Bodine and Shoup, 2010), as a consequence water temperature at sampling may be an ambiguous proxy for climate.

Morphometric in-stream variables (e.g. width), together with substrate and woody debris increase stream habitat heterogeneity. According to our results, wider streams ($>2\text{--}4\text{ m}$) supported a larger number of species, especially, lacustrine species, such as burbot or pike, as they thrive in wider streams with more vegetation and slow waters. However, these thresholds varied considerably across individual species. Other variables had moderate to low importance to species occurrence. In particular, current velocity was, in most cases, negligible. As water velocity is an important factor for species distribution in streams (Gosselin et al., 2012; Heggnes and Saltveit, 1990), the low contribution of current velocity in our study is surprising. Although water velocity as such did not stand out as an important predictor, correlated variables as discharge, substrate and width did. This is probably due to the fact that velocity was crudely measured (qualitative measure) and that the habitat sampled was mostly characterized by moderate to fast flows (90% of all sites).

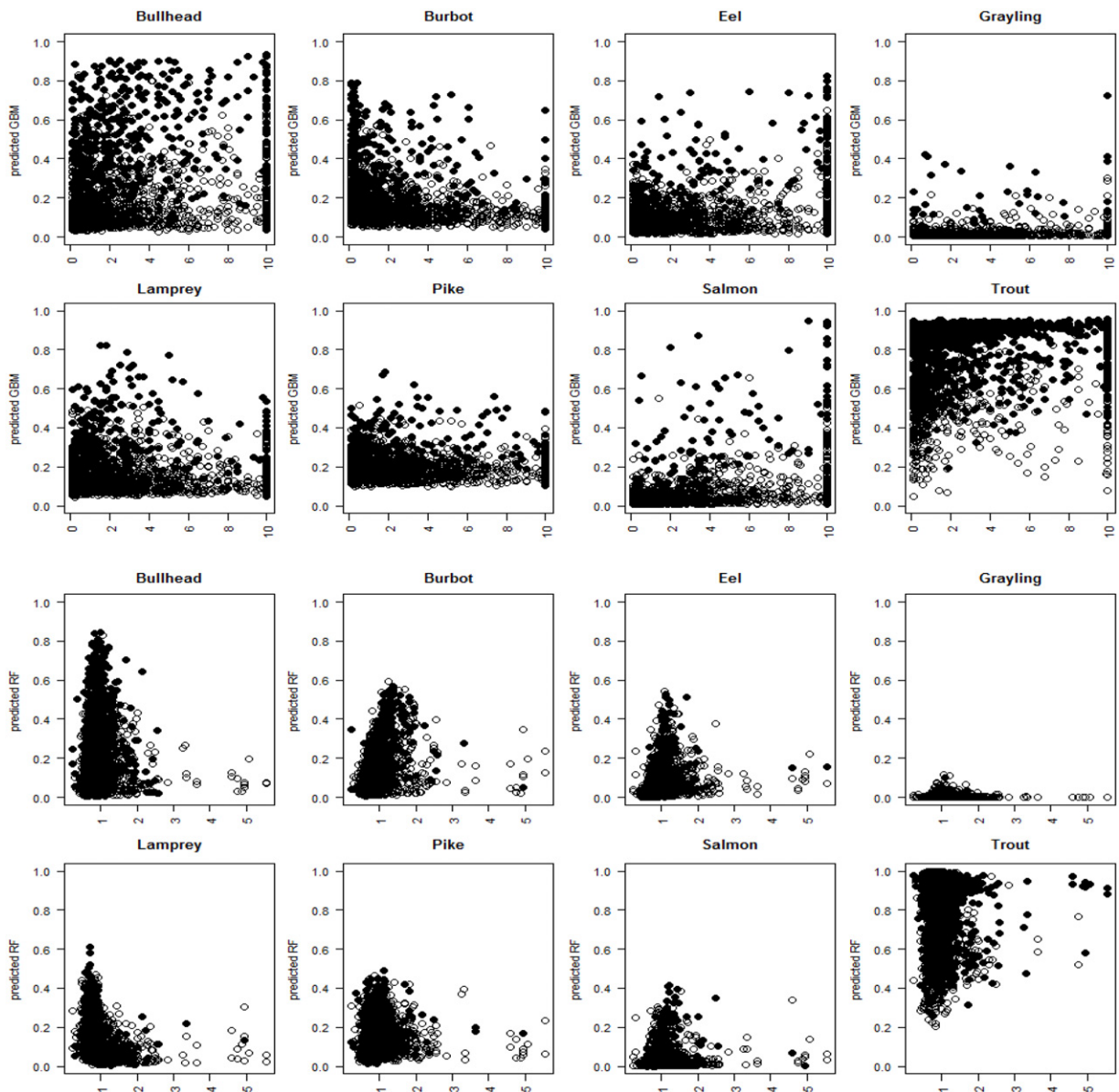


Fig. 7. The predicted probability of species occurrences in response to the minimum distance to lake or the sea (km) for 8 species in the BRT (predicted GBM) and RF (predicted RF) models. Black dots are the observed presences and white dots are the observed absences. A clear threshold in species occurrence along the predictor values would be signaled by a clear separation between present and absent sites. Bullhead, lamprey, burbot, grayling and salmon are species in the Habitats Directive. Eel is in the IUCN red list. Pike is a keystone predator and trout was the most abundant species in the study sites.

Sampling effort was especially important for some species. For example, trout was affected by sampling because the settings used for electrofishing are focused on salmonids, especially trout, and when trout are caught often consecutive runs are carried out, while this is not always the case if other species are caught. Some species are difficult to catch due to nocturnal habitat and being buried in hollows or beneath rocks during daytime. Stone loach is one such species where repeated runs improve catchability (unpubl. data SERS, cf. [Reyjol et al. \(2005\)](#)). Lampreys are difficult to detect with electrofishing, because they are hidden in the sediment ([Dunham et al., 2013](#)), and the probability of detection of this species increased with the number of electrofishing runs.

4.1. Model performance

AUC values for the BRT and RF models showed that the discriminatory power of the models was not random and that accuracy was high for most species. Nevertheless, for species that may be difficult to detect with electrofishing, such as pike or lamprey, data may be noisy and result in low model performance (i.e. low values of the AUC in the training and test sets)

and model overfitting (large differences in the AUC values between the calibration and test sets). When applied to time validation sites the accuracy of the models was higher than 0.5 and did not differ from the calibration sites, suggesting that the observed patterns are consistent over time.

Not surprisingly, the two algorithms produced different outputs including variable selection, response curves and threshold effects, corroborating that a combination of models should be evaluated when selecting variables that explain the distribution of species.

4.2. Conclusions

Our results also corroborate that a simultaneous assessment of multiple species may provide a better signal of ecosystem change than the use of surrogate species, as single species may not capture the attributes of the broader species pool (Wiens et al., 2008). Hence, multiple species rather than surrogate ones should be considered when the main conservation aim is to maintain or increase biological diversity and ecosystem function. Likewise, combining the results of different models on the important predictors determining species distribution seems more appropriate than relying on the result on a single model given the differences in the outcomes of different algorithms.

Temperature, connectivity and stream size (width) were the most important variables explaining the distribution of individual species, and implying a turn over in fish community composition along these drivers at large spatial scales. However, we must be cautious with the interpretation of these results since at least half of the species with low occurrences were not included in the final models. Another source of error in this study that may have contributed to lower model performance, e.g. low AUC or the low proportion of variation explained in the BRT models (Table S1), is the fact that we did not account for the detection probability of species (MacKenzie et al., 2002). Instead we considered a species to be absent if not captured during sampling. This may generate bias in the results of species that are difficult to detect during electrofishing, such as lamprey.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <http://dx.doi.org/10.1016/j.gecco.2015.10.009>.

References

- Aigo, J., Eugenia Lattuca, M., Cussac, V., 2014. Susceptibility of native perca (*Percichthys trucha*) and exotic rainbow trout (*Oncorhynchus mykiss*) to high temperature in Patagonia: different physiological traits and distinctive responses. *Hydrobiologia* 736, 73–82.
- Andreasson, S., 1972. Distribution of *Cottus poecilopus* Heckel and *Cottus gobio* L. (Pisces) in Scandinavia. *Zool. Scripta* 1, 69–78.
- Arheimer, B., Dahné, J., Lindström, G., Marklund, L., Strömquist, J., 2011. Multi-variable evaluation of an integrated model system covering Sweden (S-HYPE). *IAHS Publ.* 345, 145–150.
- Baxter, C.V., Fausch, K.D., Murakami, M., Chapman, P.L., 2004. Fish invasion restructures stream and forest food webs by interrupting reciprocal prey subsidies. *Ecology* 85, 2656–2663.
- Bodine, K.A., Shoup, D.E., 2010. Capture efficiency of blue catfish electrofishing and the effects of temperature, habitat, and reservoir location on electrofishing-derived length structure indices and relative abundance. *North. Am. J. Fish. Manag.* 30, 613–621.
- Bohlin, T., Hamrin, S., Heggberget, T.G., Rasmussen, G., Saltveit, S.J., 1989. Electrofishing—theory and practice with special emphasis on salmonids. *Hydrobiologia* 173, 9–43.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32.
- Buisson, L., Thuiller, W., Lek, S., Lim, P.U.Y., Grenouillet, G., 2008. Climate change hastens the turnover of stream fish assemblages. *Global Change Biol.* 14, 2232–2248.
- Byström, P., Karlsson, J.A.N., Nilsson, P.E.R., Van Kooten, T., Ask, J., Olofsson, F., 2007. Substitution of top predators: effects of pike invasion in a subarctic lake. *Freshw. Biol.* 52, 1271–1280.
- Council of the European Communities 1992. Council Directive 92/43/EEC of 21. May 1992 on the conservation of natural habitats and of wild fauna and flora. Official Journal of the European Communities, pp. 7–50.
- De'ath, G., 2007. Boosted trees for ecological modeling and prediction. *Ecology* 88, 243–251.
- Degerman, E., Beier, U., Breine, J., Melcher, A., Quataert, P., Rogers, C., Roset, N., Simoens, I., 2007. Classification and assessment of degradation in European running waters. *Fish. Manag. Ecol.* 14, 417–426.
- Degerman, E., Sers, B., 1992. Fish assemblages in Swedish streams. *Nord. J. Freshw. Res.* 67, 67–71.
- Degerman, E., Sers, B., 1994. The effect of lakes on the stream fish fauna. *Ecol. Freshw. Fish* 3.
- Degerman, E., Sers, B., Törnblom, J., Angelstam, P., 2004. Large woody debris and brown trout in small forest streams—towards targets for assessment and management of riparian landscapes. *Ecol. Bull.* 51, 233–239.
- Destaso, J., Rahel, F.J., 1994. Influence of water temperature on interactions between juvenile Colorado river cutthroat trout and brook trout in a laboratory stream. *Trans. Am. Fish. Soc.* 123, 289–297.
- Dunham, J.B., Chelgren, N.D., Heck, M.P., Clark, S.M., 2013. Comparison of electrofishing techniques to detect larval lampreys in Wadeable streams in the Pacific Northwest. *North. Am. J. Fish. Manag.* 33, 1149–1155.
- Ekman, S., 1922. Djurvärldens utbredningshistoria på Skandinaviska halvön (Fauna distribution on the Scandinavian peninsula).
- Elith, J., Leathwick, J.R., 2009. Species distribution models: Ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* 40, 677–697.
- Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. *J. Anim. Ecol.* 77, 802–813.
- European Commission 2000. Directive 2000/60/EC of the European Parliament and the Council of 23 October 2000. Establishing A Framework for Community Action in the Field of Water Policy. OJEC, pp. 1–73.
- Fullerton, A.H., Burnett, K.M., Steel, E.A., Flitcroft, R.L., Pess, G.R., Feist, B.E., Torgersen, C.E., Miller, D.J., Sanderson, B.L., 2010. Hydrological connectivity for riverine fish: measurement challenges and research opportunities. *Freshw. Biol.* 55, 2215–2237.
- Gosselin, M.P., Maddock, I., Petts, G., 2012. Mesohabitat use by brown trout (*Salmo trutta*) in a small groundwater-dominated stream. *River Res. Appl.* 28, 390–401.
- Griffiths, D., McGonigle, C., Quinn, R., 2014. Climate and species richness patterns of freshwater fish in North America and Europe. *J. Biogeography* 41, 452–463.

- Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J.M.C., Aspinall, R., Hastie, T., 2006. Making better biogeographical predictions of species' distributions. *J. Appl. Ecol.* 43.
- Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P., Buckley, I.M., 2013. Predicting species distributions for conservation decisions. *Ecol. Lett.* 1424–1435.
- Hallstan, S., Trigal, C., Johansson, K.L., Johnson, R., 2013. The impact of climate on the geographical distribution of phytoplankton species in boreal lakes. *Oecologia* 173, 1625–1638.
- Hanley, J., McNeil, B., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36.
- Heggenes, J., Saltveit, S.J., 1990. Seasonal and spatial microhabitat selection and segregation in young Atlantic salmon, *Salmo salar* L., and brown trout, *Salmo trutta* L., in a Norwegian river. *J. Fish Biol.* 36, 707–720.
- Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J., 2011. Package 'dismo'. Available online at: <http://cran.r-project.org/web/packages/dismo/index.html>.
- Hoffman, N., Fischer, P., 2002. Temperature preference and critical limits of burbot: implications for habitat selection and ontogenetic habitat shift. *Trans. Am. Fish. Soc.* 13, 1164–1172.
- Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., Van Der Laan, M., 2006. Survival ensembles. *Biostatistics* 7, 355–373.
- Kennedy, G.J.A., Strange, C.D., 1981. Efficiency of electric fishing for salmonids in relation to river width. *Fish. Manag.* 12, 55–60.
- Lindenmayer, D.B., Luck, G., 2005. Synthesis: Thresholds in conservation and management. *Biol. Cons.* 351–354.
- Lindström, G., Pers, C.P., Rosberg, R., Strömqvist, J., Arheimer, B., 2010. Development and test of the HYPE (Hydrological Predictions for the Environment) model—A water quality model for different spatial scales. *Hydrol. Res.* 41 (3–4), 295–319.
- MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J., Langtimm, C.A., 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83, 2248–2255.
- McCormick, S.D., Shrimpton, J.M., Zydlewski, J.D., 1997. Temperature effects on osmoregulatory physiology of juvenile anadromous fish. In: Wood, C.M., McDonald, D.G. (Eds.), *Global Warming: Implications for Freshwater and Marine Fish*. In: Society for Experimental Biology Seminar Series, pp. 279–301.
- McMahon, T.E., Zale, A.V., Barrows, F.T., Selong, J.H., Danahy, R.J., 2007. Temperature and competition between bull trout and brook trout: A test of the elevation refuge hypothesis. *Trans. Am. Fish. Soc.* 136, 1313–1326.
- Mehner, T., Emmrich, M., Kasprzak, P., 2011. Discrete thermal windows cause opposite response of sympatric cold-water fish species to annual temperature variability. *Ecosphere* 2, 16pp.
- Morin, R., Naiman, R.J., 1990. The relation of stream order to fish community dynamics in boreal forest watersheds. *Pol. Arch. Hydrobiol.* 37, 135–150.
- Ohlberger, J., Mehner, T., Staaks, G., Hoelker, F., 2008. Temperature-related physiological adaptations promote ecological divergence in a sympatric species pair of temperate freshwater fish, *Coregonus* spp. *Funct. Ecol.* 22, 501–508.
- Olden, J.D., 2007. Critical threshold effects of benthoscape structure on stream herbivore movement. *Phil. Trans. R. Soc. B* 362, 461–472.
- Olden, J.D., Joy, M.K., Death, R.G., 2006. Rediscovering the species in community-wide predictive modeling. *Ecol. Appl.* 16, 1449–1460.
- Parmesan, C., Yohe, G., 2003. A globally coherent fingerprint of climate change impacts across natural systems. *Nature* 421, 37–42.
- Pittman, S.J., Brown, K.A., 2011. Multi-scale approach for predicting fish species distributions across coral reef seascapes. *PLoS One* 6.
- Pringle, C., Pringle, C., 2006. *Hydrologic Connectivity: A Neglected Dimension of Conservation Biology*. Cambridge University Press.
- R Core Team 2010. *R: A Language and Environment for Statistical Computing*, Vienna, Austria.
- Reyjol, Y., Loo, G., Lek, S., 2005. Estimating sampling bias when using electrofishing to catch stone loach. *J. Fish Biol.* 66, 589–591.
- Riffart, R., Carrel, G., Le Coarer, Y., Fontez, B.N.T., 2009. Spatio-temporal patterns of fish assemblages in a large regulated alluvial river. *Freshw. Biol.* 54, 1544–1559.
- Rodriguez-Munoz, R., Nicieza, A.G., Brana, F., 2001. Effects of temperature on developmental performance, survival and growth of sea lamprey embryos. *J. Fish Biol.* 58, 475–486.
- Roni, P., Hanson, K., Beechie, T., 2008. Global review of the physical and biological effectiveness of stream habitat rehabilitation techniques. *North. Am. J. Fish. Manag.* 28, 856–890.
- Schinegger, R., Trautwein, C., Melcher, A., Schmutz, S., 2012. Multiple human pressures and their spatial patterns in European running waters. *Water Environ. J.* 26, 261–273.
- Shirvell, C.S., Dungey, R.G., 1983. Microhabitats chosen by brown trout for feeding and spawning in rivers. *Trans. Am. Fish. Soc.* 112, 355–367.
- Spens, J., Ball, J.P., 2008. Salmonid or nonsalmonid lakes: predicting the fate of northern boreal fish communities with hierarchical filters relating to a keystone piscivore. *Can. J. Fish. Aquat. Sci.* 65, 1945–1955.
- Stapanian, M.A., Paragamian, V.L., Madenjian, C.P., Jackson, J.R., Lappalainen, J., Evenson, M.J., Neufeld, M.D., 2010. Worldwide status of burbot and conservation measures. *Fish Fish.* 11, 34–56.
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forest. *BMC Bioinformatics* 9.
- Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bins in random forest variable importance measures: illustrations sources and a solution. *BMC Bioinformatics* 8.
- Suding, K.N., Hobbs, R.J., 2008. Threshold models in restoration and conservation: a developing framework. *Trends Ecol. Evolut.* 24, 271–275.
- Törnblom, J., Angelstam, P., Degerman, E., Henriksson, L., Edman, T., Temmerud, J., 2011. Catchment land cover as a proxy for macroinvertebrate assemblage structure in Carpathian Mountain streams. *Hydrobiologia* 673, 153–168.
- Vorosmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R., Davies, P.M., 2010. Global threats to human water security and river biodiversity. *Nature* 467, 555–561.
- Wedekind, C., Kung, C., 2010. Shift of spawning season and effects of climate warming on developmental stages of a grayling (*Salmonidae*). *Conserv. Biol.* 24, 1418–1423.
- Wiens, J.A., Hayward, G.D., Holthausen, R.S., Wisdom, M.J., 2008. Using surrogate species and groups for conservation planning and management. *BioScience* 58, 241–252.